THE BELL SYSTEM
TECHNICAL JOURNAL
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 52 September 1973 Number 7

A Coding Theorem for Multiple Access Channels With Correlated Sources

By DAVID SLEPIAN and JACK KEIL WOLF†

(Manuscript received February 22, 1973)

A communication system is studied in which two users communicate with one receiver over a common discrete memoryless channel. The information to be transmitted by the users may be correlated. Their information rates are described by a point in a suitably defined three-dimensional rate space.

A point in this rate space is called admissible *if there exist coders and decoders for the channel that permit the users to transmit information over it at the corresponding rates with arbitrarily small error probability. The closure of the set of all admissible rate points is called the* capacity region, $\mathcal{C}$, *and is the natural generalization of channel capacity to this situation.*

In this paper we show that $\mathcal{C}$, *which depends only on the channel, is convex and we give formulas to determine it exactly. Several simple channels are treated in detail and their capacity regions given explicitly.*

I. INTRODUCTION

The mathematical theory of communication has been concerned, for the most part, with the reliable transmission of information from a single information source to a single user. An extensive literature exists

† J. K. Wolf is Professor of Electrical Engineering at the University of Massachusetts, Amherst, Mass. Partial support for his research on this paper was furnished by the Air Force Office of Scientific Research under contract F-44620-72-C-0085.

on this problem: the basic concepts are contained in the classic papers of Shannon.[1]

In this work we consider the case in which messages from a *set* of information sources are communicated over a common channel to a single receiver. We impose constraints on the encoding techniques which can be employed.

A precise formulation of the problem is presented in a subsequent section. Here we describe in less mathematical terms the type of problem considered.

A particular multiple access communication channel with two inputs and one output is shown in Fig. 1. Here the two inputs, X_1 and X_2, and the output Y each take values from the set $\{0, 1\}$. The conditional probability of the output Y for each of the four possible input pairs (X_1, X_2) is also shown in the schema at the right in Fig. 1.

It is clear that if the transmitters can cooperate with each other they can transmit without error one bit per channel use by transmitting either the pair $(X_1 = 0, X_2 = 0)$ or the pair $(X_1 = 1, X_2 = 1)$. Such would be the case if a common binary source were connected to both inputs without any coding. If a message is to be transmitted by connecting it to only one input, say to input 1, and if the other input is unaware of the message, then even if no information is to be transmitted through input 2, the information rate for input 1 must be substantially less than one bit per channel use in order to achieve reliable communication. If two independent messages are to be connected separately to the inputs—message 1 to input 1, message 2 to input 2—the situation is even more difficult.

A general configuration that we consider is shown in Fig. 2. Three sources emitting statistically independent messages at rates R_0, R_1, and R_2 are connected to a multiple access channel via two encoders. The messages from source 1 and source 0 are inputs to encoder 1 and its output is connected to one of the input terminals of the channel.

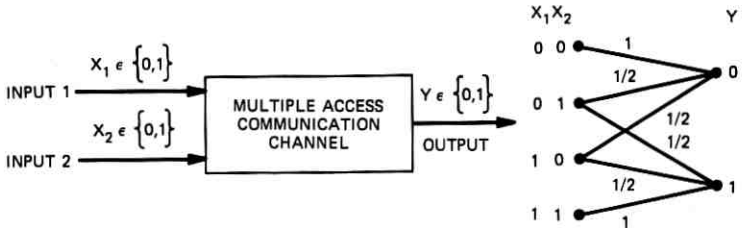

Fig. 1—A multiple access channel.

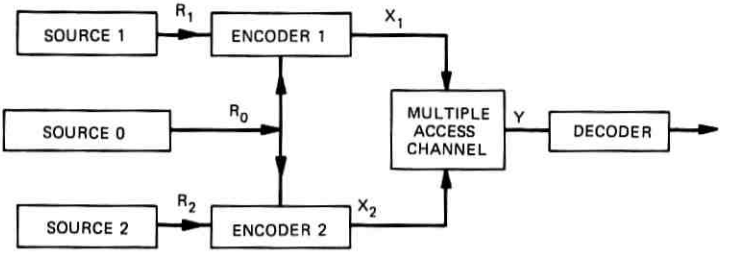

Fig. 2—Multiple access channel with correlated sources.

Encoder 2 has as inputs the messages from both source 2 and source 0 and its output is connected to the other input terminal of the channel. The channel output is connected to a decoder which estimates the three source messages. It is convenient to represent the rates of the three message sources by a point in a three-dimensional *rate space.*

For each given channel of the sort just described, there are certain rate triplets, R_0, R_1, R_2, for which it is possible to attain arbitrarily small probability of error in the system output by using sufficiently clever encoding and decoding schemes. For other points in the rate space this is not possible. We call the closure of the set of rate points for which the error probability can be made arbitrarily small the *admissible rate region* or the *capacity region* for this channel. It is a natural generalization to the multiple access channel of the channel capacity that is associated with the more commonly studied channel having a single input and a single output.

The main result of this paper is a complete determination of the capacity region $\mathcal{C}$. A typical case is shown in Fig. 3. The region always lies in the first octant and is bounded by the planes $R_0 = 0$, $R_1 = 0$,

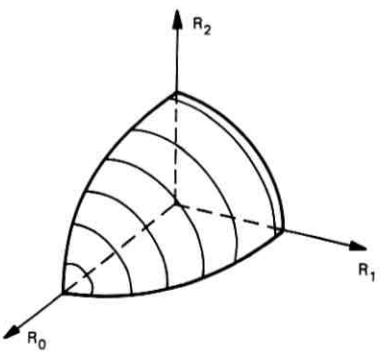

Fig. 3—An admissible rate region.

$R_2 = 0$ and a convex surface. The equations that describe $\mathcal{C}$ will be shown to involve various conditional and unconditional mutual informations. This is analogous to the single-user channel where the capacity is calculated from a mutual information.

Problems resembling ours have been treated by several authors. Shannon,[2] and then Van der Meulen,[3] consider a *two-way* channel with two inputs and two outputs. The configuration of the encoders and decoders is different than in our model, so that the problems are not the same. One similarity, however, is that the two sources are described by a pair of rates which are represented by a point in rate space. For certain points, encoders and decoders exist for which the probability of error can be made as small as desired.

The multiple access channel has been investigated by Liao,[4] Van der Meulen,[5] and Ahlswede.[6] Liao[4] and Ahlswede[6] both prove a coding theorem and a converse for the case of independent sources. Our results reduce to theirs for the case $R_0 = 0$. Correlation in the sources adds a totally new dimension to the problem (and literally to the region of admissible rates).

A problem which is the dual of the one considered here is the broadcast channel investigated by Cover[7] and Bergmans.[8] There, a channel with one input and two outputs is considered along with a single encoder and two decoders. Again the concept of an admissible rate region applies.

A brief outline of the paper follows. In Section II a detailed problem formulation is presented. Section III summarizes the main results of the paper and gives some examples. Sections IV and V and the associated appendixes give details of the derivation of a coding theorem and a converse. A more useful description of the admissible rate region is given in Section VI. We conclude in Section VII with some generalizations and comments.

II. PROBLEM FORMULATION

Consider the block diagram shown in Fig. 4. The three sources are described by a three-dimensional *rate vector* $\mathbf{R} = (R_0, R_1, R_2)$ with nonnegative components. For a fixed positive integer N, we define the components of the vector $\mathbf{M}$ by

$$\mathbf{M} = \mathbf{M}(\mathbf{R}, N) = (M_0, M_1, M_2), \tag{1a}$$

$$M_i = \lceil e^{R_i N} \rceil, \qquad i = 0, 1, 2, \tag{1b}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x. Every N time units the sources produce a triplet of numbers (i, j, k) that are the

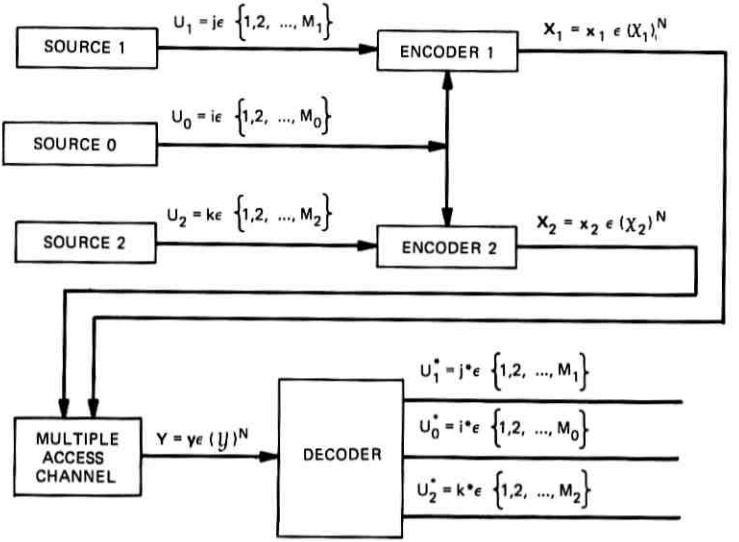

Fig. 4—Notation for multiple access channel.

corresponding values of the random variables (U_0, U_1, U_2). These random variables are assumed to be statistically independent and uniformly distributed over the rectangular lattice of dimensions $M_0 \times M_1 \times M_2$. That is, their joint probability distribution is

$$\begin{aligned} P_{U_0U_1U_2}(i, j, k) = \Pr[U_0 = i, U_1 = j, U_2 = k] = 1/M_0M_1M_2, \\ i \in (1, 2, \cdots, M_0) \equiv I_0, \\ j \in (1, 2, \cdots, M_1) \equiv I_1, \\ k \in (1, 2, \cdots, M_2) \equiv I_2. \end{aligned} \tag{2}$$

The channel is a probabilistic mapping which every unit of time maps a pair of real numbers (x_1, x_2) to the real number y. The real numbers x_1, x_2, and y belong to the finite alphabets $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{Y}$, respectively. The mapping is governed by the conditional probability distribution $P_{Y|X_1X_2}(y|x_1, x_2)$ for all x_1 in $\mathcal{X}_1$, x_2 in $\mathcal{X}_2$, and y in $\mathcal{Y}$. Here we describe the inputs by the pair of random variables (X_1, X_2) and the output by the random variable Y. Throughout this paper, it will be assumed that $P_{Y|X_1X_2}$ is specified *a priori* and cannot be altered.

To describe how the channel processes sequences of N input pairs, we define the N-vectors

$$\begin{aligned} \mathbf{x}_1 &= (x_{11}, x_{12}, \cdots, x_{1N}), & \mathbf{x}_1 \in (\mathcal{X}_1)^N, \\ \mathbf{x}_2 &= (x_{21}, x_{22}, \cdots, x_{2N}), & \mathbf{x}_2 \in (\mathcal{X}_2)^N, \\ \mathbf{y} &= (y_1, y_2, \cdots, y_N), & \mathbf{y} \in (\mathcal{Y})^N, \end{aligned} \tag{3}$$

and in a similar way the corresponding random vectors $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{Y}$. Here $(\mathfrak{X}_1)^N$ is the set of all N-vectors whose components are in $\mathfrak{X}_1$. The sets $(\mathfrak{X}_2)^N$ and $(\mathfrak{Y})^N$ are defined analogously. We assume the channel is stationary and memoryless; that is,

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2) = \prod_{t=1}^{N} P_{Y|X_1X_2}(y_t|x_{1t}, x_{2t}). \tag{4}$$

The superscript N on the joint probability distribution indicates the dimension of the vectors.

The encoders are deterministic mappings from the source outputs to channel input vectors. Encoder 1 is a mapping from the source pair (i, j) to an N-vector $\mathbf{x}_1 \in (\mathfrak{X}_1)^N$. The functional form for this mapping is written

$$\mathbf{x}_1 = \mathbf{f}_N(i, j), \quad i \in I_0, \quad j \in I_1, \quad \mathbf{x}_1 \in (\mathfrak{X}_1)^N. \tag{5}$$

Similarly, encoder 2 is a mapping from the source pair (i, k) to the N-vector $\mathbf{x}_2 \in (\mathfrak{X}_2)^N$. The functional form for this mapping is written

$$\mathbf{x}_2 = \mathbf{g}_N(i, k), \quad i \in I_0, \quad k \in I_2, \quad \mathbf{x}_2 \in (\mathfrak{X}_2)^N. \tag{6}$$

The collection of $(M_0 \times M_1 + M_0 \times M_2)$ N-vectors which result from these mappings is called a *code* of block length N. Usually, we will adopt the more suggestive notation $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ik}$ instead of $\mathbf{f}_N(i, j)$ and $\mathbf{g}_N(i, k)$.

To summarize the operation of the sources, encoders, and channel we note that:

(*i*) Every N time units, the three sources produce a triplet (i, j, k).

(*ii*) The two encoders act upon the source outputs to produce the two N-vectors $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ik}$.

(*iii*) The components of these vectors are impressed upon the channel, one pair of inputs each time unit. Corresponding to each pair of inputs the channel produces an output, so that in the N time units the channel produces an output N-vector, $\mathbf{y}$.

The *decoder* is a deterministic mapping from the vector $\mathbf{y}$ to the triplet (i^*, j^*, k^*) where $i^* \in I_0$, $j^* \in I_1$, $k^* \in I_2$. We describe this mapping by $(i^*, j^*, k^*) = \mathbf{h}_N(\mathbf{y})$. The triplet of decoder outputs is denoted by the vector random variable (U_0^*, U_1^*, U_2^*).

The deterministic mappings $(\mathbf{f}_N, \mathbf{g}_N, \mathbf{h}_N)$ will be called a *coding*. A coding with rate vector $\mathbf{R} = (R_0, R_1, R_2)$ and block length N will be denoted by $C_N(\mathbf{R})$. For a given coding, we can in principle calculate the probability of the error event $\mathcal{E}$, where $\mathcal{E}^c$ (the complement of $\mathcal{E}$)

is defined as

$$\mathcal{E}^c = \text{event } \{U_0^* = U_0 \quad \text{and} \quad U_1^* = U_1 \quad \text{and} \quad U_2^* = U_2\}. \tag{7}$$

For the coding $C_N(\mathbf{R})$, we denote the probability of the error event (hereafter called the probability of error) by $P_e(C_N(\mathbf{R}))$.

A rate vector $\mathbf{R}$ will be said to be *admissible* if, for every $\epsilon > 0$, there exists a positive integer N and a coding $C_N(\mathbf{R})$ such that $P_e(C_N(\mathbf{R})) \leqq \epsilon$. The closure of the set of admissible rate vectors is called the *admissible region* or *capacity region*, and is denoted by $\mathcal{C}$. Our purpose is to specify $\mathcal{C}$ for an arbitrary, discrete, memoryless, multiple access channel.

III. SUMMARY OF RESULTS AND EXAMPLES

The main results of this paper are two alternative descriptions of the admissible rate region $\mathcal{C}$ for any discrete memoryless channel. The proofs that these yield the correct region are contained in the remaining sections of the paper. Here we discuss only the simplest of the results.

We shall have much need of conditional mutual information expressions in the sequel. We remind the reader of the definition

$$I(\mathbf{A};\mathbf{B}|\mathbf{C}) = \sum_{\mathbf{i}}\sum_{\mathbf{j}}\sum_{\mathbf{k}} P_{\mathbf{ABC}}(\mathbf{i},\mathbf{j},\mathbf{k}) \log \frac{P_{\mathbf{AB}|\mathbf{C}}(\mathbf{i},\mathbf{j}|\mathbf{k})}{P_{\mathbf{A}|\mathbf{C}}(\mathbf{i}|\mathbf{k})P_{\mathbf{B}|\mathbf{C}}(\mathbf{j}|\mathbf{k})}. \tag{8}$$

Here

$$P_{\mathbf{ABC}}(\mathbf{i},\mathbf{j},\mathbf{k}) = \Pr\left[A_\alpha = i_\alpha, B_\beta = j_\beta, C_\gamma = k_\gamma, \quad \alpha = 1, 2, \cdots, L; \beta = 1, 2, \cdots, M; \gamma = 1, 2, \cdots, N\right]$$

is the joint distribution function of the discrete random variables $A_1, A_2, \cdots, A_L, B_1, B_2, \cdots, B_M, C_1, C_2, \cdots, C_N$. The conditional distributions $P_{\mathbf{AB}|\mathbf{C}}(\mathbf{i},\mathbf{j}|\mathbf{k})$, etc., are defined in the usual way.

Let us return now to consider a discrete memoryless channel with input alphabets $\mathfrak{X}_1$ and $\mathfrak{X}_2$, output alphabet $\mathcal{Y}$, and transition probabilities $P_{Y|X_1X_2}(y|x_1,x_2)$, $x_1 \in \mathfrak{X}_1$, $x_2 \in \mathfrak{X}_2$, $y \in \mathcal{Y}$. Let Z be a random variable which takes on values in the set

$$\mathfrak{z} = \{1, 2, \cdots, M\}. \tag{9}$$

From any set of three distributions $P_{X_1|Z}(x_1|z)$, $P_{X_2|Z}(x_2|z)$, and $P_Z(z)$, $x_1 \in \mathfrak{X}_1$, $x_2 \in \mathfrak{X}_2$, $z \in \mathfrak{z}$, form the joint distribution

$$P_{ZX_1X_2Y}(z, x_1, x_2, y) = P_Z(z)P_{X_1|Z}(x_1|z)P_{X_2|Z}(x_2|z)P_{Y|X_1X_2}(y|x_1,x_2). \tag{10}$$

Now denote by $\mathfrak{R}(P_{ZX_1X_2Y})$ the set of vectors $\mathbf{R} = (R_0, R_1, R_2)$ such that

$$0 \leqq R_1 \leqq I(X_1; Y|X_2, Z), \tag{11a}$$

$$0 \leqq R_2 \leqq I(X_2; Y|X_1, Z), \tag{11b}$$

$$0 \leqq R_1 + R_2 \leqq I(X_1, X_2; Y|Z), \tag{11c}$$

$$0 \leqq R_0 + R_1 + R_2 \leqq I(X_1, X_2; Y), \tag{11d}$$

where the mutual informations are computed according to (8) using the joint distribution (10). This region is a polyhedron such as is shown in Fig. 7, Appendix I. Then the admissible rate region $\mathcal{C}$ is given by

$$\mathcal{C} = \text{closure of the convex hull } \bigcup \mathcal{R}(P_{ZX_1X_2Y}), \tag{12}$$

where the union is taken over all possible choices of $P_{X_1|Z}$, $P_{X_2|Z}$, and P_Z, and all values of M, the size of the $\mathcal{Z}$ alphabet.†

To obtain the intersection of the admissible rate region $\mathcal{C}$ with the plane $R_0 = 0$, the size of the alphabet $\mathcal{Z}$ can be set equal to 1. The random variable Z no longer appears in the equations. For $R_0 = 0$, we then define $\mathcal{R}(P_{X_1}, P_{X_2})$ as the set of vector $\mathbf{R} = (0, R_1, R_2)$ such that

$$0 \leqq R_1 \leqq I(X_1; Y|X_2), \tag{13a}$$

$$0 \leqq R_2 \leqq I(X_2; Y|X_1), \tag{13b}$$

$$0 \leqq R_1 + R_2 \leqq I(X_1, X_2; Y). \tag{13c}$$

Then

$$\mathcal{C}|_{R_0=0} = \text{closure of the convex hull of } \bigcup \mathcal{R}(P_{X_1}, P_{X_2}), \tag{14}$$

where the union is taken over all possible choices for the unconditional distributions P_{X_1} and P_{X_2}. This is the solution found by Liao[4] for uncorrelated sources.

Other equations for specifying the region $\mathcal{C}$ are given in Section VI. They involve the calculation of mutual informations among long sequences of random variables and thus do not appear to be useful for computation.

Quite generally, $\mathcal{C}$ is convex. It is always bounded by portions of the three coordinate planes and a surface which encloses a finite volume in the first quadrant of rate space. If $\mathbf{R} = (R_0, R_1, R_2)$ is in $\mathcal{C}$, then for any $\boldsymbol{\delta} = (\delta_0, \delta_1, \delta_2)$ satisfying $0 \leqq \delta_i \leqq R_i$, $i = 0, 1, 2$, the rate vector $\boldsymbol{\delta}$ is also in $\mathcal{C}$.

In the remainder of this section, some simple examples are presented for which the admissible rate region has an explicit characterization.

† We suspect that it suffices to consider only values of $M \leqq \lceil e^{R_0} \rceil$, but have not been able to prove this conjecture.

Example 1 (Multiplier Channel)

Both the inputs, X_1 and X_2, and the output, Y, for this channel take values 0 and 1. The output is the product of the two inputs. Formally, $\mathfrak{X}_1 = \mathfrak{X}_2 = \mathfrak{Y} = \{0, 1\}$ and $P_{Y|X_1X_2}(0|0, 0) = P_{Y|X_1X_2}(0|0, 1) = P_{Y|X_1X_2}(0|1, 0) = P_{Y|X_1X_2}(1|1, 1) = 1$, and all other conditional probabilities are zero. Note that the channel is deterministic.

The pyramid described by the planes

$$\begin{gathered} R_0 = 0, \quad R_1 = 0, \quad R_2 = 0, \\ R_0 + R_1 + R_2 = \log 2 \end{gathered} \tag{15}$$

must contain the admissible rate region $\mathcal{C}$, as is seen from (11d) since $0 \leqq R_0 + R_1 + R_2 \leqq I(X_1, X_2; Y) \leqq \log 2$. But the rate vectors $\mathbf{R}_1 = (\log 2, 0, 0)$, $\mathbf{R}_2 = (0, \log 2, 0)$, and $\mathbf{R}_3 = (0, 0, \log 2)$ are all admissible by the following strategies:

$\mathbf{R}_1$: Choose $N = 1$, $M_0 = 2$, $M_1 = M_2 = 1$ and use code words $x_{111} = 0$, $x_{121} = 1$, $x_{211} = 0$, $x_{221} = 1$.

$\mathbf{R}_2$: Choose $N = 1$, $M_0 = 1$, $M_1 = 2$, $M_2 = 1$ and use code words $x_{111} = 0$, $x_{112} = 1$, $x_{211} = 1$.

$\mathbf{R}_3$: Choose $N = 1$, $M_0 = 1$, $M_1 = 1$, $M_2 = 2$ and use code words $x_{111} = 1$, $x_{211} = 0$, $x_{212} = 1$.

The probability of error for these codes is zero. Since the convex hull of these three rate points and the origin is the set bounded by the planes (15), the capacity region $\mathcal{C}$ is as shown in Fig. 5.

By similar arguments, we find that Fig. 5 gives the region of admissible rates for many other binary-input, binary-output deterministic

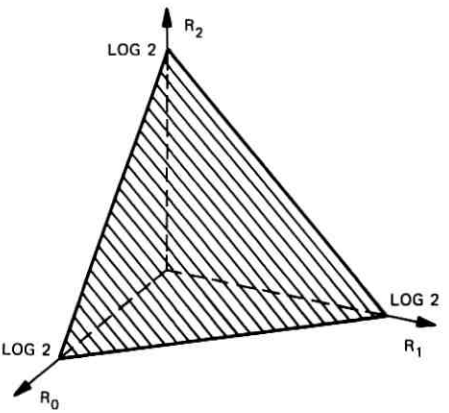

Fig. 5—Admissible rate region for the multiplier channel.

TABLE I—$P_{Y|X_1X_2}(y|x_1, x_2)$

x_1x_2 \ y	0	1	2	3
0 0	$1-p$	$p/3$	$p/3$	$p/3$
0 1	$p/3$	$1-p$	$p/3$	$p/3$
1 0	$p/3$	$p/3$	$1-p$	$p/3$
1 1	$p/3$	$p/3$	$p/3$	$1-p$

channels (ones with all transition probabilities equal to zero or one). Degenerate cases exist, however, in which the region $\mathcal{C}$ reduces to a portion of a plane. For example, if $P_{Y|X_1X_2}(0|0, 0) = P_{Y|X_1X_2}(0|0, 1) = P_{Y|X_1X_2}(1|1, 0) = P_{Y|X_1X_2}(1|1, 1) = 1$ and all other probabilities are zero, it is easy to verify that $\mathcal{C} = \{\mathbf{R} = (R_0, R_1, R_2): 0 \leqq R_0 + R_1 \leqq \log 2, R_2 = 0, R_1, R_0 \geqq 0\}$.

Example 2 (Symmetric Noisy Channel)

Let $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2, 3\}$ and let $P_{Y|X_1X_2}(y|x_1, x_2)$ be given as shown in Table I. Let M be as in (9). Define $P_Z(z_i) = \gamma_i$, $P_{X_1|Z}(0|z_i) = \alpha_i$, $P_{X_2|Z}(0|z_i) = \beta_i$, $i = 1, 2, \cdots, M$. Straightforward calculations then yield

$$I(X_1; Y|X_2, Z) = \sum_{i=1}^{M} \gamma_i(f_1(\alpha_i, p) - K(p)), \tag{16}$$

$$I(X_2; Y|X_1, Z) = \sum_{i=1}^{M} \gamma_i(f_1(\beta_i, p) - K(p)), \tag{17}$$

$$I(X_1, X_2; Y|Z) = \sum_{i=1}^{M} \gamma_i(f_2(\alpha_i, \beta_i, p) - K(p)), \tag{18}$$

where

$$f_1(\delta, p) = \frac{2}{3}p \log \frac{3}{p} + \left((1-p)\delta + \frac{p}{3}(1-\delta)\right) \times \log \frac{1}{(1-p)\delta + \frac{p}{3}(1-\delta)} + \left(\frac{p}{3}\delta + (1-p)(1-\delta)\right) \times \log \frac{1}{\frac{p}{3}\delta + (1-p)(1-\delta)}, \tag{19}$$

$$K(p) = (1-p) \log \frac{1}{(1-p)} + p \log \frac{3}{p}, \tag{20}$$

and

$$f_2(\alpha, \beta, p) = \left[(1-p)\alpha\beta + \frac{p}{3}(1-\alpha\beta)\right] \log \frac{1}{(1-p)\alpha\beta + \frac{p}{3}(1-\alpha\beta)}$$
$$+ \left((1-p)\alpha(1-\beta) + \frac{p}{3}(1-\alpha+\alpha\beta)\right)$$
$$\times \log \frac{1}{(1-p)\alpha(1-\beta) + \frac{p}{3}(1-\alpha+\alpha\beta)}$$
$$+ \left((1-p)\beta(1-\alpha) + \frac{p}{3}(1-\beta+\alpha\beta)\right)$$
$$\times \log \frac{1}{(1-p)\beta(1-\alpha) + \frac{p}{3}(1-\beta+\alpha\beta)}$$
$$+ \left((1-p)(1-\alpha)(1-\beta) + \frac{p}{3}(\alpha+\beta-\alpha\beta)\right)$$
$$\times \log \frac{1}{(1-p)(1-\alpha)(1-\beta) + \frac{p}{3}(\alpha+\beta-\alpha\beta)}. \quad (21)$$

It is easy to show that $f_1(\delta, p) \leqq f_1(\frac{1}{2}, p)$, $f_2(\alpha, \beta, p) \leqq f_2(\frac{1}{2}, \frac{1}{2}, p) = \log 4$. Therefore, the three mutual informations in (16), (17), and (18) are simultaneously maximized by setting $\alpha_i = \beta_i = \frac{1}{2}$, $i = 1, 2, \cdots, M$. Furthermore,

$$I(X_1, X_2; Y) = H(Y) - H(Y|X_1X_2) = H(Y) - K(p) \leqq \log 4 - K(p)$$

with equality when $\alpha_i = \beta_i = \frac{1}{2}$, $i = 1, 2, \cdots, M$. Thus all four mutual informations, $I(X_1; Y|X_2, Z)$, $I(X_2; Y|X_1, Z)$, $I(X_1, X_2; Y|Z)$, and $I(X_1, X_2; Y)$, are maximized for the same choice of the parameters α_i, β_i, and γ_i, and the maximum values are independent of M. The capacity region for this channel then is given by

$$0 \leqq R_1 \leqq f_1(\tfrac{1}{2}, p) - K(p), \quad (22)$$
$$0 \leqq R_2 \leqq f_1(\tfrac{1}{2}, p) - K(p), \quad (23)$$
$$0 \leqq R_0 + R_1 + R_2 \leqq \log 4 - K(p). \quad (24)$$

This region is shown in Fig. 6.

IV. EXISTENCE OF CODINGS WITH SMALL P_e

In this section we outline a proof of the existence of codings which have vanishingly small probability of error for certain values of the

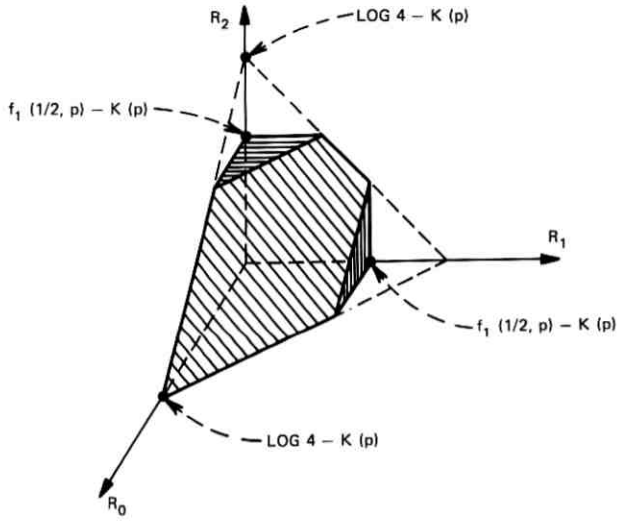

Fig. 6—Admissible rate region for the symmetric noisy binary channel.

rate vector $\mathbf{R}$ and sufficiently large block length N. Tedious details are relegated to the appendixes. A random coding argument is used. We calculate the average probability of error for an ensemble of codings, then argue that there must exist at least one member of the ensemble having error probability as small as this average. Actually, we can only compute an upper bound for this average error probability, but this bound is sufficiently small for our purposes.

For every coding, we shall use the same form of decoder mapping. Assume for the moment that the block length N, the rate vector $\mathbf{R}$, and the encoder functions $\mathbf{f}_N(i, j) = \mathbf{x}_{1ij}$ and $\mathbf{g}_N(i, k) = \mathbf{x}_{2ik}$ are fixed. For each $\mathbf{y} \in (\mathcal{Y})^N$, the decoder computes the $M_0 \times M_1 \times M_2$ numbers

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik}), \qquad i \in I_0, \qquad j \in I_1, \qquad k \in I_2.$$

Then $\mathbf{h}(\mathbf{y}) = (i_o, j_o, k_o)$ if and only if (i_o, j_o, k_o) is the smallest triplet (in lexicographic order) such that

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1i_oj_o}, \mathbf{x}_{2i_ok_o}) \geqq P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik}) \tag{25}$$

for all (i, j, k). Such a decoder mapping achieves a maximum likelihood decision among the possible source outputs.

We now describe the class of codings for which we obtain an upper bound to the average probability of error. It is specified by two positive integers, K and $N = KL$, where L is a positive integer, by a rate vector

$\mathbf{R}$, and by a particular probability distribution for the random variables $\mathbf{X}_1$, $\mathbf{X}_2$, and Z. The vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are K-dimensional and take on values in $(\mathfrak{X}_1)^K$ and $(\mathfrak{X}_2)^K$ respectively, the spaces of channel input K-vectors. The random variable Z takes values from an alphabet $\mathfrak{z}$ of size M as in (9). The joint distribution of these quantities is restricted to have the form

$$P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}(z, \mathbf{x}_1, \mathbf{x}_2) = P^{(K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1|z)P^{(K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2|z)P_Z(z) \tag{26}$$

and we denote this collection of distributions by $\mathcal{P}_K$. A class of codings is thus specified by K, $N = KL$, $\mathbf{R}$, and a $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2} \in \mathcal{P}_K$.

Now let K, $N = KL$, $\mathbf{R}$, and $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2} \in \mathcal{P}_K$ be given. A set of N-dimensional code vectors $\mathbf{x}_{111}, \cdots, \mathbf{x}_{11M_1}, \mathbf{x}_{211}, \cdots, \mathbf{x}_{21M_2}$ in the corresponding ensemble of codings is obtained as follows. Choose a sample, say z, from the distribution $P_Z(\cdot)$. Next independently choose M_1 K-vectors from $P^{(K)}_{\mathbf{X}_1|Z}(\cdot\,|z)$ and then M_2 K-vectors from $P^{(K)}_{\mathbf{X}_2|Z}(\cdot\,|z)$. These are respectively the first K components of the N-vectors $\mathbf{x}_{111}, \cdots, \mathbf{x}_{11M_1}$, $\mathbf{x}_{211}, \cdots, \mathbf{x}_{21M_2}$.

To obtain the next K components of the code words, independently choose a new sample z from $P_Z(\cdot)$ and repeat the process. After a total of L drawings from $P_Z(\cdot)$ the specification of the N-vectors $\mathbf{x}_{111}, \cdots, \mathbf{x}_{11M_1}, \mathbf{x}_{211}, \cdots, \mathbf{x}_{21M_2}$ is complete. The entire process is then repeated to obtain the remaining code words—those with second subscript equal to 2, 3, $\cdots$, M_0.

We now seek an upper bound to the average probability of error for the codings in this ensemble, an average in which the probability of error for each particular coding is weighted in accordance with its probability of occurrence in the ensemble. We denote this average probability of error by $P_e(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2})$ and we denote by $P_{e|i,j,k}(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2})$ the average probability of error given that the source triplet (i, j, k) was presented for transmission. A useful result is

Theorem 1: The average probability of error conditioned on the source triplet (i, j, k) has an upper bound

$$P_{e|i,j,k}(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) \leqq \sum_{\alpha=1}^{4} \exp\{-N[E_\alpha(\rho_\alpha, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) - \rho_\alpha \hat{R}_\alpha]\}, \tag{27}$$

where $0 \leqq \rho_\alpha \leqq 1$, $\alpha = 1, 2, 3, 4$,

$$\hat{R}_\alpha = \begin{cases} R_1, & \alpha = 1 \\ R_2, & \alpha = 2 \\ R_1 + R_2, & \alpha = 3 \\ R_0 + R_1 + R_2, & \alpha = 4, \end{cases} \tag{28}$$

and

$$E_1(\rho_1, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = -\frac{1}{K}\ln \sum_{\mathbf{y}}\sum_{\mathbf{x}_2}\sum_{z} P^{(K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2|z)P_Z(z) \times \Big(\sum_{\mathbf{x}_1} P^{(K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1|z)\big(P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1\mathbf{x}_2)\big)^{1/(1+\rho_1)}\Big)^{1+\rho_1}, \quad (29a)$$

$$E_2(\rho_2, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = -\frac{1}{K}\ln \sum_{\mathbf{y}}\sum_{\mathbf{x}_1}\sum_{z} P^{(K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1|z)P_Z(z) \times \Big(\sum_{\mathbf{x}_2} P^{(K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2|z)\big(P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)\big)^{1/(1+\rho_2)}\Big)^{1+\rho_2}, \quad (29b)$$

$$E_3(\rho_3, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = -\frac{1}{K}\ln \sum_{\mathbf{y}}\sum_{z} P_Z(z) \times \Big(\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1|z)P^{(K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2|z) \times \big(P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1\mathbf{x}_2)\big)^{1/(1+\rho_3)}\Big)^{1+\rho_3}, \quad (29c)$$

$$E_4(\rho_4, P^{(K)}_{\mathbf{X}_1\mathbf{X}_2}) = -\frac{1}{K}\ln \sum_{\mathbf{y}}\Big(\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(K)}_{\mathbf{X}_1\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) \times \big(P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)\big)^{1/(1+\rho_4)}\Big)^{1+\rho_4} - \frac{e^{-N(R_1+R_2)}\rho_4}{N}. \quad (29d)$$

A proof of this theorem is given in Appendix A. It follows closely the proof given in Gallager[9] for the single-input, single-output channel.

Since the bound proved in the theorem is independent of the triplet (i, j, k), we see that this same bound applies to the unconditioned average probability of error $P_e(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2})$. Finally, for fixed $N = KL$, and $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$, there must be at least one coding in the ensemble with probability of error no greater than the average probability of error. Thus we have

Theorem 2: For every positive integer K, for every positive integer N that is an integral multiple of K, for every joint distribution $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$ of form (26), and for every rate vector $\mathbf{R}$, there exists a coding $C_N(\mathbf{R})$ such that

$$P_e(C_N(\mathbf{R})) \leqq \sum_{\alpha=1}^{4} \exp\{-N[E_\alpha(\rho_\alpha, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) - \rho_\alpha \hat{R}_\alpha]\} \quad (30)$$

for all ρ_α, $0 \leqq \rho_\alpha \leqq 1$, $\alpha = 1, 2, 3, 4$. The E_α and $\hat{R}_\alpha$ are given by (28) and (29).

For a given $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$ and for certain values of the rate vector $\mathbf{R}$, the upper bound decreases exponentially in N. For these values of $\mathbf{R}$, by making N sufficiently large, we can insure a small probability of error. We now determine for what rate vectors this is the case.

Define $\mathcal{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ as the set of rate vectors $\mathbf{R}$ for which

$$0 \leqq R_1 < \frac{1}{K} I(\mathbf{X}_1; \mathbf{Y} | \mathbf{X}_2, Z) \tag{31a}$$

$$0 \leqq R_2 < \frac{1}{K} I(\mathbf{X}_2; \mathbf{Y} | \mathbf{X}_1, Z) \tag{31b}$$

$$0 \leqq R_1 + R_2 < \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y} | Z) \tag{31c}$$

$$0 \leqq R_0 + R_1 + R_2 < \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}), \tag{31d}$$

where the mutual informations are evaluated under the distribution

$$P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}(z, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}(z, \mathbf{x}_1, \mathbf{x}_2) P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y} | \mathbf{x}_1, \mathbf{x}_2), \tag{32}$$

where $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$ is given by (26) and $P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}$ by (4).

In Appendix B we prove

Theorem 3: For every $\epsilon > 0$ and every rate vector $\mathbf{R} \subset \mathcal{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$, there exists an L_0 and a sequence of codings, $C_N(\mathbf{R})$, such that

$$P_e(C_N(\mathbf{R})) \leqq \epsilon \qquad \text{for every} \qquad N = KL, L \geqq L_0. \tag{33}$$

This theorem holds for all $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$ of the form given in (26), that is, for all $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2} \in \mathcal{P}_K$. Now define

$$\mathcal{R}_K \equiv \bigcup_{P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2} \in \mathcal{P}_K} \mathcal{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}), \tag{34}$$

and finally define

$$\mathcal{R} \equiv \bigcup_K \mathcal{R}_K, \tag{35}$$

where $K = 1, 2, \cdots$. We then have the following main result:

Theorem 4: For every $\epsilon > 0$ and for every rate vector $\mathbf{R} \subset \mathcal{R}$, there exist values of K and L_0 and a sequence of codings $C_N(\mathbf{R})$ such that

$$P_e(\mathbf{B}_N(\mathbf{R})) \leqq \epsilon \qquad \text{for every} \qquad N = KL, L \geqq L_0. \tag{36}$$

Note that if we use the statement of Theorem 1 instead of Theorem 2, Theorem 4 becomes

Corollary 1: For every $\epsilon > 0$, for every message triplet (i, j, k), and for every rate vector $\mathbf{R} \subset \mathcal{R}$, there exist values of K and L_0 and a sequence of codings $C_N(\mathbf{R})$ such that

$$P_{e|ijk}(C_N(\mathbf{R})) \leqq \epsilon \qquad \text{for every} \qquad N = KL, L \geqq L_0. \tag{37}$$

V. CONVERSE THEOREM

In this section, we present a series of lemmas and theorems which yield a converse to the Coding Theorem 4. Let $\hat{\mathfrak{R}}$ denote the closure of the region $\mathfrak{R}$ given by (35) and let $\hat{\mathfrak{R}}^c$ be the complement of $\hat{\mathfrak{R}}$ with respect to the first octant. We shall ultimately show that every coding $C_K(\mathbf{R})$ with $\mathbf{R} \in \hat{\mathfrak{R}}^c$ transmits with a probability of error not less than a constant $\delta > 0$ which is independent of K.

Our notation is as before except that K, instead of N, will be used for the block length of a code. Let K, $\mathbf{R}$, and the channel be given. The associated vector $\mathbf{M}$ with components

$$M_\alpha = \lceil e^{KR_\alpha} \rceil, \qquad \alpha = 0, 1, 2, \tag{38}$$

is then determined. We shall no longer be concerned with ensembles of codes, but rather fix our attention on some given encoding functions $\mathbf{x}_{1ij} = \mathbf{f}(i, j)$, $\mathbf{x}_{2ik} = \mathbf{g}(i, k)$, where $i \in I_0$, $j \in I_1$, and $k \in I_2$. These vectors need not be distinct. Then with the source statistics given by (2), the given encoding defines a joint probability distribution

$$\begin{aligned} P^{(K)}_{U_0U_1U_2\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}(i, j, k, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) &= P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_1, \mathbf{x}_2) \\ &\times Q^{(K)}_{\mathbf{X}_1|U_0U_1}(\mathbf{x}_1\,|\,i, j)Q^{(K)}_{\mathbf{X}_2|U_0U_2}(\mathbf{x}_2\,|\,i, k)P_{U_0U_1U_2}(i, j, k) \end{aligned} \tag{39}$$

for the random variables in question. Here

$$Q^{(K)}_{\mathbf{X}_1|U_0U_1}(\mathbf{x}_1\,|\,i, j) = \delta_{\mathbf{x}_1\mathbf{x}_{1ij}}, \tag{40a}$$

$$Q^{(K)}_{\mathbf{X}_2|U_0U_2}(\mathbf{x}_2\,|\,i, k) = \delta_{\mathbf{x}_2\mathbf{x}_{2ik}}, \tag{40b}$$

where the right-hand terms are Kronecker deltas. Entropies and mutual informations can then be calculated from (39) by the usual formulas.

Several more definitions are needed. We shall make use of the *rate number vector* $\mathbf{R}' = (R_0', R_1', R_2')$ given by

$$R_\alpha' = \frac{1}{K} \log M_\alpha \geqq R_\alpha \qquad \alpha = 0, 1, 2 \tag{41}$$

and the elementary entropy function

$$h(x) \equiv -x \log x - (1 - x) \log (1 - x). \tag{42}$$

Finally, we define

$$P_{e1}(C_K(\mathbf{R})) = \Pr\,[U_1^* \neq U_1] \tag{43}$$

$$P_{e2}(C_K(\mathbf{R})) = \Pr\,[U_2^* \neq U_2] \tag{44}$$

$$P_{e3}(C_K(\mathbf{R})) = \Pr\,[U_1^* \neq U_1 \text{ or } U_2^* \neq U_2]. \tag{45}$$

Then, for the probability of error using the coding $C_K(\mathbf{R})$ we have

$$P_e(C_K(\mathbf{R})) = \Pr[U_0^* \neq U_0 \text{ or } U_1^* \neq U_1 \text{ or } U_2^* \neq U_2] \geqq \max[P_{e1}(C_K), P_{e2}(C_K), P_{e3}(C_K)]. \quad (46)$$

We now proceed to the first of the lemmas which is a generalization of Fano's inequality (Ref. 9, Theorem 4.3.1). The proof is given in Appendix C.

Lemma 1: For every K and $\mathbf{R}$ and for every $C_K(\mathbf{R})$:

$$H(U_1|\mathbf{Y}, U_0, U_2) \leqq P_{e1}(C_K) \log M_1 + h(P_{e1}(C_K)); \quad (47a)$$

$$H(U_2|\mathbf{Y}, U_0, U_1) \leqq P_{e2}(C_K) \log M_2 + h(P_{e2}(C_K)); \quad (47b)$$

$$H(U_1, U_2|\mathbf{Y}, U_0) \leqq P_{e3}(C_K) \log (M_1M_2) + h(P_{e3}(C_K)); \quad (47c)$$

$$H(U_0, U_1, U_2|\mathbf{Y}) \leqq P_e(C_K) \log (M_0M_1M_2) + h(P_e(C_K)). \quad (47d)$$

The next lemma, proved in Appendix D, is a generalization of the data processing theorem (Ref. 9, Theorem 4.3.3).

Lemma 2: For every K and $\mathbf{R}$ and every coding $C_K(\mathbf{R})$:

$$\text{(a)}, \quad I(U_1; \mathbf{Y}|U_2, U_0) \leqq I(\mathbf{X}_1; \mathbf{Y}|\mathbf{X}_2, U_0); \quad (48a)$$

$$\text{(b)}, \quad I(U_2; \mathbf{Y}|U_1, U_0) \leqq I(\mathbf{X}_2; \mathbf{Y}|\mathbf{X}_1, U_0); \quad (48b)$$

$$\text{(c)}, \quad I(U_1, U_2; \mathbf{Y}|U_0) \leqq I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0); \quad (48c)$$

$$\text{(d)}, \quad I(U_0, U_1, U_2; \mathbf{Y}) \leqq I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}). \quad (48d)$$

Lemma 3: For every K and $\mathbf{R}$ and every coding $C_K(\mathbf{R})$ with rate-number vector $\mathbf{R}'$:

$$\text{(a)}, \quad KR_1' - I(\mathbf{X}_1; \mathbf{Y}|\mathbf{X}_2, U_0) \leqq P_{e1}(C_K)KR_1' + h(P_{e1}(C_K)); \quad (49a)$$

$$\text{(b)}, \quad KR_2' - I(\mathbf{X}_2; \mathbf{Y}|\mathbf{X}_1, U_0) \leqq P_{e2}(C_K)KR_2' + h(P_{e2}(C_K)); \quad (49b)$$

$$\text{(c)}, \quad K(R_1' + R_2') - I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0) \leqq P_{e3}(C_K)K(R_1' + R_2') + h(P_{e3}(C_K)); \quad (49c)$$

$$\text{(d)}, \quad K(R_0' + R_1' + R_2') - I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) \leqq P_e(C_K)K(R_0' + R_1' + R_2') + h(P_e(C_K)). \quad (49d)$$

The proof is given in Appendix E.

With these lemmas in hand, we return to the matter of establishing a converse to Theorem 4. For a given K, $\mathbf{R}$, and encoding $C_K(\mathbf{R})$, there is established a joint probability distribution between the random variables $\mathbf{Y}$, $\mathbf{X}_1$, $\mathbf{X}_2$, and U_0 given by

$$Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}(i, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1\mathbf{x}_2)Q^{(K)}_{\mathbf{X}_1|U_0}(\mathbf{x}_1|i)Q^{(K)}_{\mathbf{X}_2|U_0}(\mathbf{x}_2|i)Q_{U_0}(i), \quad (50)$$

where $P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}$ is given by (4),

$$Q_{U_0}(i) = \frac{1}{M_0}, \tag{51a}$$

$$Q^{(K)}_{\mathbf{X}_1|U_0}(\mathbf{x}_1|i) = \frac{1}{M_1}\sum_{j=1}^{M_1}\delta_{\mathbf{x}_1\mathbf{x}_{1ij}}, \tag{51b}$$

$$Q^{(K)}_{\mathbf{X}_2|U_0}(\mathbf{x}_2|i) = \frac{1}{M_2}\sum_{k=1}^{M_2}\delta_{\mathbf{x}_2\mathbf{x}_{2ik}}, \tag{51c}$$

$$i \in I_0, \qquad \mathbf{x}_1 \in (\mathfrak{X}_1)^K, \qquad \mathbf{x}_2 \in (\mathfrak{X}_2)^K.$$

Here $C_K(\mathbf{R})$ is defined by the code words $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ik}$, $i \in I_0$, $j \in I_1$, $k \in I_2$. We denote by $\mathcal{Q}_K$ the set of all distributions $Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2}$ derived from code books, that is, all distributions of the form obtained by summing (50) over all $\mathbf{y} \in (\mathcal{Y})^K$.

With K, $\mathbf{R}$, and an encoding $C_K(\mathbf{R})$ now fixed, we define $\mathcal{S}^c(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ to be the set of all vectors $\mathbf{S} = (S_0, S_1, S_2)$ with non-negative components such that *at least one* of the following inequalities is satisfied,

$$S_1 > \frac{1}{K} I(\mathbf{X}_1; \mathbf{Y}|U_0, \mathbf{X}_2), \tag{52a}$$

$$S_2 > \frac{1}{K} I(\mathbf{X}_2; \mathbf{Y}|U_0, \mathbf{X}_1), \tag{52b}$$

$$S_1 + S_2 > \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0), \tag{52c}$$

$$S_0 + S_1 + S_2 > \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}). \tag{52d}$$

Next define

$$\mathcal{S}^c_K \equiv \bigcap_{Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2} \in \mathcal{Q}_K} \mathcal{S}^c(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}) \tag{53}$$

and finally

$$\mathcal{S}^c \equiv \bigcap_K \mathcal{S}^c_K. \tag{54}$$

Here c denotes complement with respect to the first octant $S_0 \geqq 0$, $S_1 \geqq 0$, $S_2 \geqq 0$. Thus, for example, $\mathcal{S}(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ is a closed convex polyhedron bounded by seven planes.

Note the similarity between (50) and (32) and between (31) which defines $\mathcal{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ and (52) which defines $\mathcal{S}^c(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$. For every distribution in $\mathcal{Q}_K$, there is a distribution in $\mathcal{P}_K$ that will give equality between the corresponding right-hand members of (31) and (52) when Z and U_0 are properly identified. We make this identification by

choosing $M = M_0$, and by taking Z to be uniformly distributed over its M possible values. With a member of $\mathcal{P}_K$ identified with each $Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2}$ in this way, we see that for this particular $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}$ one has

$$\mathcal{S}(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}) = \hat{\mathcal{R}}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}).$$

Here the caret, $\hat{}$, denotes closure. Comparison of (34) and (53) then shows that $\mathcal{S}_K \subset \hat{\mathcal{R}}_K$. Thus $\mathcal{S} \subset \hat{\mathcal{R}}$ or

$$\hat{\mathcal{R}}^c \subset \mathcal{S}^c. \tag{55}$$

In Appendix F we establish

Theorem 5: If $\mathbf{R}$ *is an interior point of* $\mathcal{S}^c$, *then for every* K *and every encoding* $C_K(\mathbf{R})$,

$$P_e(C_K(\mathbf{R})) \geqq \delta > 0,$$

where $\delta = \delta(\mathbf{R})$ *is independent of the encoding and of* K.

VI. SPECIFICATION OF THE CAPACITY REGION

At the end of Section II, the capacity region was defined as the closure of the set of admissible rate points. Theorems 4 and 5 along with (55) show that $\mathcal{C} = \hat{\mathcal{R}}$ where $\mathcal{R}$ is defined by (31), (34), and (35). This characterization of $\mathcal{C}$ is of little computational value. It entails the calculation of the mutual informations appearing on the right of (31) for all distributions of form (32). A further infinite union over all values of K is then required. In this section we shall show how a much simpler description of $\mathcal{C}$ can be obtained, one that is independent of K and hence much more suitable for numerical calculations.

Central to the development of this simpler characterization of $\mathcal{C}$ is

Theorem 6: The region $\mathcal{C}$ *of admissible rates is convex.*

This theorem is proved by a time-sharing argument in Appendix G. By deleting words from a code, one obtains an additional obvious feature of the region $\mathcal{C}$ which we state as

Theorem 7: Let $\mathbf{R} \in \mathcal{C}$. *Then if* $0 \leqq R''_\alpha \leqq R_\alpha$, $\alpha = 0, 1, 2$, *the rate vector* $\mathbf{R}''$ *is also contained in* $\mathcal{C}$.

We return to our simpler characterization of $\mathcal{C}$. Let $\mathcal{R}_1$ denote the region specified by (31), (32), and (34) for $K = 1$. Since $\mathcal{C}$ is the closure of $\mathcal{R}$ as given by (35), $\hat{\mathcal{R}}_1 \subseteq \mathcal{C}$. From Theorem 6 it follows that also

$$\hat{\mathcal{R}}' \equiv \text{convex hull } \mathcal{R}_1 \subseteq \mathcal{C}. \tag{56}$$

(The convex hull of a set $\mathcal{A}$ consists of all points in $\mathcal{A}$ and all points

on all line segments joining points of $\mathfrak{A}$.) We shall soon show that indeed $\mathfrak{R}' = \mathfrak{C}$.

As a step in this direction, in Appendix H we establish

Lemma 4: For every distribution $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}(z, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ *as in* (32),

$$I(\mathbf{X}_1; \mathbf{Y}|Z, \mathbf{X}_2) \leqq \sum_{t=1}^{K} I(X_{1t}; Y_t|Z, X_{2t}), \tag{57a}$$

$$I(\mathbf{X}_2; \mathbf{Y}|Z, \mathbf{X}_1) \leqq \sum_{t=1}^{K} I(X_{2t}; Y_t|Z, X_{1t}), \tag{57b}$$

$$I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|Z) \leqq \sum_{t=1}^{K} I(X_{1t}, X_{2t}; Y_t|Z), \tag{57c}$$

$$I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) \leqq \sum_{t=1}^{K} I(X_{1t}, X_{2t}; Y_t). \tag{57d}$$

Combined with (31) the lemma shows that

$$\mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}) \supseteq \mathfrak{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}), \tag{58}$$

where $\mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ is the set of rate vectors $\mathbf{R}$ for which

$$0 \leqq R_1 \leqq \frac{1}{K}\sum_{1}^{K} I(X_{1t}; Y_t|Z, X_{2t}), \tag{59a}$$

$$0 \leqq R_2 \leqq \frac{1}{K}\sum_{1}^{K} I(X_{2t}; Y_t|Z, X_{1t}), \tag{59b}$$

$$0 \leqq R_1 + R_2 \leqq \frac{1}{K}\sum_{1}^{K} I(X_{1t}, X_{2t}; Y_t|Z), \tag{59c}$$

$$0 \leqq R_0 + R_1 + R_2 \leqq \frac{1}{K}\sum_{1}^{K} I(X_{1t}, X_{2t}; Y_t), \tag{59d}$$

where the right sides of (59) are evaluated under $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}$. Note that $\mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$, unlike $\mathfrak{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$, is a closed set by definition.

Now, a typical term on the right of (59) depends only on the marginal distribution $P_{ZX_{1t}X_{2t}Y_t}(z, x_{1t}, x_{2t}, y_t)$. By summing (32) over the appropriate indexes and taking account of (26), it is seen that this marginal can be written

$$P_{ZX_{1t}X_{2t}Y_t}(z, x_{1t}, x_{2t}, y_t) = P_Z(z)P_{X_{1t}|Z}(x_{1t}|z)P_{X_{2t}|Z}(x_{2t}|z)P_{Y_t|X_{1t}X_{2t}}(y_t|x_{1t}, x_{2t})$$

which is a distribution of the form $P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}$ for $K = 1$. Thus the right-hand sides of (59), which are the parameters defining the box-like

region $\mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$, are averages of parameters that define the box-like region $\mathfrak{R}(P^{(1)}_{ZX_{1t}X_{2t}Y_t})$, $t = 1, 2, \cdots, K$. In Appendix I this fact and the convexity of the box-like regions are used to show that

$$\text{convex hull} \bigcup_{t=1}^{K} \hat{\mathfrak{R}}(P^{(1)}_{ZX_{1t}X_{2t}Y_t}) \supseteq \mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}) \tag{60}$$

from which it also follows that

$$\text{convex hull} \bigcup_{P^{(1)}_{ZX_1X_2}\in\mathcal{P}_1} \hat{\mathfrak{R}}(P^{(1)}_{ZX_1X_2Y}) \supseteq \bigcup_{P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}\in\mathcal{P}_K} \mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}_1}). \tag{61}$$

We now have

$$\begin{aligned}\mathfrak{R}' \equiv \text{convex hull } \hat{\mathfrak{R}}_1 &= \text{convex hull closure} \bigcup_{P^{(1)}_{ZX_1X_2}\in\mathcal{P}_1} \mathfrak{R}(P^{(1)}_{ZX_1X_2Y}) \\ &\supseteq \text{convex hull} \bigcup_{P^{(1)}_{ZX_1X_2}\in\mathcal{P}_1} \hat{\mathfrak{R}}(P^{(1)}_{ZX_1X_2Y}) \supseteq \bigcup_{P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}\in\mathcal{P}_K} \mathfrak{R}^*(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}) \\ &\supseteq \bigcup_{P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}\in\mathcal{P}_K} \mathfrak{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}}). \end{aligned} \tag{62}$$

Here the last inclusion follows from (58) and the next to last inclusion is (61). Using (34) and (62), we now see that $\mathfrak{R}' \supseteq \mathfrak{R}_K$ for every K. From (35) then $\mathfrak{R}' \supseteq \mathfrak{R}$, and since $\mathfrak{R}'$ is closed by definition, $\mathfrak{R}' \supseteq \hat{\mathfrak{R}} = \mathfrak{C}$. Combined with (56), this shows that $\mathfrak{R}' = \mathfrak{C}$, and the formulation (10)–(12) is thereby established.

It is to be noted that while this reduction permits calculation of $\mathfrak{C}$ by evaluating mutual informations involving no more than four random variables, the size of the Z alphabet is unrestricted. In this connection, see the footnote in Section III.

That we can indeed take the size of the Z alphabet to be 1 when computing the intersection of $\mathfrak{C}$ with the plane $R_0 = 0$, as claimed in Section III, is seen as follows. When $R_0 = 0$, (11d) is weaker than (11c), since always $I(X_1, X_2; Y) \geqq I(X_1,X_2; Y|Z)$. Thus we need only consider (11a), (11b), and (11c) in defining regions $\mathfrak{R}(P_{ZX_1X_2Y})$ in the $R_1 - R_2$ plane. But the right members of these equations are of the form

$$\begin{aligned} I(X_1; Y|X_2, Z) &= \sum_i P_Z(z_i) I(X_1; Y|X_2, Z = z_i) \\ I(X_2; Y|X_1, Z) &= \sum_i P_Z(z_i) I(X_2; Y|X_1, Z = z_i) \\ I(X_1, X_2; Y|Z) &= \sum_i P_Z(z_i) I(X_1, X_2; Y|Z = z_i). \end{aligned}$$

An argument just like that of Appendix I now shows that $\mathfrak{R} \subseteq$ convex hull $\bigcup_i \mathfrak{R}_i$ where $\mathfrak{R}_i$ is given by $0 \leqq R_1 \leqq I(X_1; Y|X_2, Z = z_i)$,

$0 \leqq R_2 \leqq I(X_2; Y|X_1, Z = z_i)$, $0 \leqq R_1 + R_2 \leqq I(X_1, X_2; Y|Z = z_i)$. Each box-like region $\mathfrak{R}_i$ can be thought of as obtained from a distribution in which Z takes a single value with probability one. The formulation of Section III follows at once.

VII. COMMENTARY

7.1 *Generalizations*

7.1.1 *N Input Users*

The foregoing can be generalized to the case of a memoryless channel with N input users and a single output. The channel then is specified by alphabets $\mathcal{Y}$, $\mathcal{X}_1$, $\mathcal{X}_2$, $\cdots$, $\mathcal{X}_N$ and transition probabilities $p(y|x_1, x_2, \cdots, x_N)$ for $y \in \mathcal{Y}$, $x_i \in \mathcal{X}_i$, $i = 1, 2, \cdots, N$. Again we allow the information supplied to the input users to be correlated in a special way.

We first write out the equations for $N = 3$ in full, and then indicate the general result. There are now seven independent sources, S_1, S_2, S_3, S_{12}, S_{13}, S_{23}, S_{123} producing information at rates R_1, R_2, R_3, R_{12}, R_{13}, R_{23}, R_{123} respectively. There are three encoders. Encoder 1 sees the outputs of only S_1, S_{12}, S_{13}, S_{123}; encoder 2 sees the outputs of only S_2, S_{12}, S_{23}, S_{123}; encoder 3 sees the outputs of only S_3, S_{13}, S_{23}, S_{123}. The decoder at the channel output attempts to reproduce separately the messages from the seven sources. Using block codes, for certain values of the rate vector $\mathbf{R} = (R_1, R_2, R_3, R_{12}, R_{13}, R_{23}, R_{123})$, the error probability of the system can be made arbitrarily small. The closure of the set of all such vector rates is called the *capacity region* $\mathfrak{C}$.

$\mathfrak{C}$ can be found as follows. Let

$$p_{123}(z_{123}),\ p_{12}(z_{12}),\ p_{13}(z_{13}),\ p_{23}(z_{23})$$
$$p_1(x_1|z_{123}, z_{12}, z_{13}),\ p_2(x_2|z_{123}, z_{12}, z_{23}),\ p_3(x_3|z_{123}, z_{13}, z_{23}) \tag{63}$$

be given probability distributions. Here $x_i \in \mathcal{X}_i$, $i = 1, 2, 3$. The Z_{12}, Z_{13}, etc., have finite alphabets of unspecified size. We denote by P the distribution

$$P = p_{123}(z_{123})p_{12}(z_{12})p_{13}(z_{13})p_{23}(z_{23})p_1(x_1|z_{123}, z_{12}, z_{13}) \times p_2(x_2|z_{123}, z_{12}, z_{23})p_3(x_3|z_{123}, z_{13}, z_{23})p(y|x_1, x_2, x_3). \tag{64}$$

Now let $\mathfrak{R}(P)$ be the set of $\mathbf{R}$ such that

$$0 \leqq R_1 \leqq I(X_1; Y|Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_2, X_3)$$
$$0 \leqq R_2 \leqq I(X_2; Y|Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_1, X_3)$$
$$0 \leqq R_3 \leqq I(X_3; Y|Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_1, X_2)$$

$$0 \leqq R_1 + R_2 \leqq I(X_1, X_2; Y | Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_3)$$
$$0 \leqq R_1 + R_3 \leqq I(X_1, X_3; Y | Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_2)$$
$$0 \leqq R_2 + R_3 \leqq I(X_2, X_3; Y | Z_{123}, Z_{12}, Z_{13}, Z_{23}, X_1)$$
$$0 \leqq R_1 + R_2 + R_3 \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{12}, Z_{13}, Z_{23})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{12} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{13}, Z_{23})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{13} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{12}, Z_{23})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{23} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{12}, Z_{13})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{12} + R_{13} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{23})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{12} + R_{23} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{13})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{13} + R_{23} \leqq I(X_1, X_2, X_3; Y | Z_{123}, Z_{12})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{12} + R_{13} + R_{23} \leqq I(X_1, X_2, X_3; Y | Z_{123})$$
$$0 \leqq R_1 + R_2 + R_3 + R_{12} + R_{13} + R_{23} + R_{123} \leqq I(X_1, X_2, X_3; Y), \quad (65)$$

where all the mutual informations here are computed with the distribution (64). Let $\mathfrak{R} = \bigcup \mathfrak{R}(P)$ where the union is over all distributions of form (64) as the factors listed in (63) are varied. Then $\mathfrak{C}$ is the closure of the convex hull of $\mathfrak{R}$.

The generalization to N users is simple in concept but awkward to describe. We do not dwell long on it here. There are now $2^N - 1$ sources and $\mathfrak{C}$ is a region in a $(2^N - 1)$-dimensional rate space. The list (63) is increased to contain $2^N - N - 1$ separate distributions for as many independent Z variables—$Z_{12}, Z_{13}, \cdots, Z_{23}, \cdots, Z_{123\cdots N}$—and N distributions of form $p_1(x_1 | z_{12}, z_{13}, \cdots, z_{12\cdots N})$, etc., where each z subscript contains the x subscript. Equations (64) and (65) are generalized in an obvious way. There are now $\sum_{j=1}^{N} \left(2^{\binom{N}{j}} - 1\right)$ equations (65). $\mathfrak{C}$ is given as the closure of the convex hull of the union of the regions defined by these equations.

These results for N users were obtained by cursory examination of the rigorous proofs given in this paper for two users. As we have not had the courage to write out all the details, however, the assertions made for the N-user case must still be regarded as conjectures, or educated guesses.

7.1.2 *Continuous Amplitudes*

It would appear that our results can be extended in a natural way to channels with more general alphabet structures. For example, the channel might be specified by a conditional probability *density* $P(y | x_1, x_2)$ where x_1, x_2, and y take all real values. Equation (11)

would remain the same, but the mutual informations are now given by integrals. Densities $P_Z(z)$, $P_{X_1|Z}(x_1|z)$, $P_{X_2|Z}(x_2|z)$ must be specified and the joint density of Z, X_1, X_2, and Y is the product (10) as before. Constraints, such as $EX_1^2 = \sigma_1^2$, $EX_2^2 = \sigma_2^2$ must be imposed on these densities in taking the union indicated in (12).

Again, we have not verified in detail the validity of the determination of $\mathfrak{C}$ just given for continuous amplitudes. *Caveat emptor.*

7.2 *Some Problems*

Many research problems related to the subject of this paper remain to be examined. A brief description of some of these follows:

(*i*) The footnote in Section III suggests that the size of the alphabet Z can be bounded in searching for the capacity of a particular channel. Is this conjecture true?

(*ii*) The explicit construction of good codes for use on specific multiple access channels is an untapped field that leads to new problems not found on single-input, single-output channels. For example, even for noiseless channels (all channel probabilities zero or one) a coding problem exists since users compete with each other for the use of the channel.

(*iii*) The region of rates for which error-free transmission with finite length codes is possible is not known. This region is analogous to the zero-error capacity of the single-input, single-output channel.

(*iv*) For a particular multiple access channel it has been found that the region of admissible rates can be enlarged by allowing the encoders to observe the output via a feedback channel. This is in contrast to the situation for the single-input, single-output channel where feedback does not alter the capacity. In the multi-user case, however, a feedback channel increases the cooperation possible between the users and in general increases the forward capacity. How to calculate the region of admissible rates for multiple access channels with feedback is not known.

(*v*) A special form has been assumed here for the correlation between the messages encoded by the two users. How does one handle more general correlations? Is the presently assumed form general in some asymptotic sense?

(*vi*) Can one calculate the capacity region for some class of multiple access channels with memory?

(*vii*) What is the rate distortion theory for these channels?

VIII. ACKNOWLEDGMENTS

We are deeply indebted to Professor N. T. Gaarder of the University of Hawaii for the many fruitful discussions we had with him on all aspects of this research. We also wish to acknowledge the work of Henry Liao whose investigation of the case of uncorrelated sources led us to the present study.

APPENDIX A

Proof of Theorem 1

Let $P_{e|i,j,k}(C)$ be the probability of error when the source triplet (i, j, k) is sent over the channel using coding C. Let Pr (C) be the probability of the particular coding C. Then

$$P_{e|i,j,k}(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = \sum \Pr(C) P_{e|i,j,k}(C), \tag{66}$$

where the sum is over all possible codings, that is, over all ways of choosing the code words $\mathbf{x}_{111}, \cdots, \mathbf{x}_{1M_0M_1}, \mathbf{x}_{211}, \cdots, \mathbf{x}_{2M_0M_2}$. But the right side of (66) can be interpreted as the probability of error in the joint experiment of drawing a code from the ensemble and transmitting (i, j, k) over the channel. With this interpretation in mind, we have

$$P_{e|i,j,k}(N, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = \sum_{i=1}^{4} P_i, \tag{67}$$

where

$$P_1 = \Pr[U_0^* = i, U_1^* \neq j, U_2^* = k | \mathfrak{B}] \tag{68a}$$

$$P_2 = \Pr[U_0^* = i, U_1^* = j, U_2^* \neq k | \mathfrak{B}] \tag{68b}$$

$$P_3 = \Pr[U_0^* = i, U_1^* \neq j, U_2^* \neq k | \mathfrak{B}] \tag{68c}$$

$$P_4 = \Pr[U_0^* \neq i | \mathfrak{B}], \tag{68d}$$

where $\mathfrak{B}$ is the event $\{U_0 = i, U_1 = j, U_2 = k\}$. We will find upper bounds for these four probabilities.

We first compute an upper bound for P_1. Fix values for the N-vectors $\mathbf{y}, \mathbf{x}_{1ij}, \mathbf{x}_{2ik}$. Let $\mathbf{z}_i$ denote the L-vector whose components $z_{i1}, z_{i2}, \cdots, z_{iL}$ were used in the choice of $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ik}$. Later we shall average over these quantities.

Define $\mathfrak{A}_{j'}$ as the event that

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y} | \mathbf{X}_{1ij'}, \mathbf{x}_{2ik}) \geq P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y} | \mathbf{x}_{1ij}, \mathbf{x}_{2ik}). \tag{69}$$

Note that the only random variable in this expression is $\mathbf{X}_{1ij'}$. Define

$$P^{(N,K)}_{\mathbf{X}|Z}(\mathbf{x} | \mathbf{z}) = \prod_{\alpha=1}^{L} P^{(K)}_{\mathbf{X}|Z}(\mathbf{x}_\alpha | z_\alpha), \tag{70}$$

where $\mathbf{x}$ is the N-vector obtained by concatenating the L K-dimensional vectors $\mathbf{x}_\alpha$, and $\mathbf{z}$ is an L-vector whose components are z_α, $\alpha = 1, 2, \cdots, L$. Then the probability of the event $\mathcal{A}_{j'}$ is

$$\Pr[\mathcal{A}_{j'}] = \sum_{\mathbf{x}_{1ij'}}{}' P^{(N,K)}_{\mathbf{X}_1|\mathbf{Z}}(\mathbf{x}_{1ij'}|\mathbf{z}_i),$$

where the sum is over all values of $\mathbf{x}_{1ij'}$ satisfying

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij'}, \mathbf{x}_{2ik}) \geqq P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij}, \mathbf{x}_{2ik}). \tag{71}$$

Following Gallager,[9] an upper bound to this expression is

$$\Pr[\mathcal{A}_{j'}] \leqq \sum_{\mathbf{x}_{1ij'}} P^{(N,K)}_{\mathbf{X}_1|\mathbf{Z}}(\mathbf{x}_{1ij'}|\mathbf{z}_i) \left(\frac{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij'}, \mathbf{x}_{2ik})}{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij}, \mathbf{x}_{2ik})} \right)^{s_1}, \tag{72}$$

for any $s_1 \geqq 0$. The summation is over all values of the N-vector $\mathbf{x}_{1ij'}$.

For the same fixed values of $\mathbf{y}$, $\mathbf{x}_{1ij}$, $\mathbf{x}_{2ik}$, and $\mathbf{z}_i$, let $\mathcal{A}$ be the event that (69) holds for *some* value of j' not equal to j. Then from Gallager[9] (page 136)

$$\Pr[\mathcal{A}] \leqq \left(\sum_{\substack{j'=1 \\ j' \neq j}}^{M_1} \Pr[\mathcal{A}_{j'}] \right)^{\rho_1}, \tag{73}$$

for any ρ_1 in the range $0 \leqq \rho_1 \leqq 1$. Combining (72) and (73) we have

$$\Pr[\mathcal{A}] \leqq (M_1 - 1)^{\rho_1} \Bigg[\sum_{\mathbf{x}_1} P^{(N,K)}_{\mathbf{X}_1|\mathbf{Z}}(\mathbf{x}_1|\mathbf{z}_i) \times \left(\frac{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_{2ik})}{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij}, \mathbf{x}_{2ik})} \right)^{s_1} \Bigg]^{\rho_1}, \tag{74}$$

where the summation is over all N-vectors in $(\mathfrak{X}_1)^N$.

The probability of interest, P_1, has an upper bound

$$P_1 \leqq \sum_{\mathbf{y}} \sum_{\mathbf{x}_{1ij}} \sum_{\mathbf{x}_{2ik}} \sum_{\mathbf{z}_i} P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_{1ij}, \mathbf{x}_{2ik}) P^{(N,K)}_{\mathbf{X}_1|\mathbf{Z}}(\mathbf{x}_{1ij}|\mathbf{z}_i) \times P^{(N,K)}_{\mathbf{X}_2|\mathbf{Z}}(\mathbf{x}_{2ik}|\mathbf{z}_i) P^{(L)}_{\mathbf{Z}}(\mathbf{z}_i) \Pr[\mathcal{A}], \tag{75}$$

where the inequality results from the fact that the occurrence of the event $\mathcal{A}$ does not necessarily imply the event $\{U_0^* = i, U_1^* \neq j, U_2^* = k\}$ but that the converse is true. Combining (74) and (75) and choosing $s_1 = 1/(1 + \rho_1)$, we obtain

$$P_1 \leqq (M_1 - 1)^{\rho_1} \sum_{\mathbf{y}} \sum_{\mathbf{x}_2} \sum_{\mathbf{z}} P^{(N,K)}_{\mathbf{X}_2|\mathbf{Z}}(\mathbf{x}_2|\mathbf{z}) P^{(L)}_{\mathbf{Z}}(\mathbf{z}) \times \Big[\sum_{\mathbf{x}_1} (P^{(N,K)}_{\mathbf{X}_1|\mathbf{Z}}(\mathbf{x}_1|\mathbf{z}) P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2))^{1/(1+\rho_1)} \Big]^{1+\rho_1}, \tag{76}$$

where the summations for $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{z}$ are taken over all elements in the spaces $(\mathcal{Y})^N$, $(\mathcal{X}_1)^N$, $(\mathcal{X}_2)^N$, and $(\mathcal{Z})^L$ respectively.

We note from (1b) that $(M_1 - 1) < e^{NR_1}$. Now use the product form of (70) and write the right-hand side of (76) as an exponential of a logarithm. We find the desired result

$$P_1 \leqq \exp\{-N[E_1(\rho_1, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) - \rho_1 R_1]\}, \tag{77}$$

where E_1 is given by (29a). The sums there are over all $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}$, and z contained respectively in $(\mathcal{X}_1)^K$, $(\mathcal{X}_2)^K$, $(\mathcal{Y})^K$, and $\mathcal{Z}$. Reversing the role of U_1 and U_2 one immediately obtains

$$P_2 \leqq \exp\{-N(E_2(\rho_2, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) - \rho_2 R_2)\}, \tag{78}$$

where E_2 is given by (29b).

The procedure for obtaining the upper bound for P_3 is very similar to that used for P_1. An outline of the proof follows. Fix $\mathbf{y}$, $\mathbf{x}_{1ij}$, $\mathbf{x}_{2ik}$, and $\mathbf{z}_i$. Define $\mathfrak{B}$ as the event

$$P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{X}_{1ij'}, \mathbf{X}_{2ik'}) \geqq P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik}) \quad \text{for some } j' \neq j \text{ and some } k' \neq k. \tag{79}$$

It can then be shown that for any $s_3 \geqq 0$ and $0 \leqq \rho_3 \leqq 1$

$$\Pr[\mathfrak{B}] \leqq (M_1 - 1)^{\rho_3}(M_2 - 1)^{\rho_3}\Bigg[\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(N,K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1\,|\,\mathbf{z}_i) \times P^{(N,K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2\,|\,\mathbf{z}_i)\left(\frac{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_1, \mathbf{x}_2)}{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik})}\right)^{s_3}\Bigg]^{\rho_3}. \tag{80}$$

Averaging over $\mathbf{y}$, $\mathbf{x}_{1ij}$, $\mathbf{x}_{2ik}$, and $\mathbf{z}_i$, and then setting $s_3 = 1/(1 + \rho_3)$, we obtain

$$P_3 \leqq (M_1 - 1)^{\rho_3}(M_2 - 1)^{\rho_3}\sum_{\mathbf{y}}\sum_{\mathbf{z}} P^{(L)}_Z(\mathbf{z}) \times \Big[\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(N,K)}_{\mathbf{X}_1|Z}(\mathbf{x}_1\,|\,\mathbf{z}) P^{(N,K)}_{\mathbf{X}_2|Z}(\mathbf{x}_2\,|\,\mathbf{z})(P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_1, \mathbf{x}_2))^{1/(1+\rho_3)}\Big]^{1+\rho_3}. \tag{81}$$

Replace $(M_1 - 1)^{\rho_3}(M_2 - 1)^{\rho_3}$ by the upper bound $e^{N\rho_3(R_1+R_2)}$, use (70) repeatedly, and write terms as exponentials of logarithms. One finds

$$P_3 \leqq \exp\{-N[E_3(\rho_3, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) - \rho_3(R_1 + R_2)]\}, \tag{82}$$

where E_3 is given by (29c).

One minor change is made in the procedure to compute the upper bound for P_4. We fix only the values of $\mathbf{y}$, $\mathbf{x}_{1ij}$, and $\mathbf{x}_{2ik}$ (but not of $\mathbf{z}_i$). Define $\mathfrak{D}$ as the event

$$P^{(N,K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{X}_{1i'j'}, \mathbf{X}_{2i'k'}) \geqq P^{(N,K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik}) \tag{83}$$

for some $i' \neq i$ and any j' and k'. Then for any $s_4 \geqq 0$, $0 \leqq \rho_4 \leqq 1$,

$$\Pr[\mathfrak{D}] \leqq (M_0 - 1)^{\rho_4}(M_1)^{\rho_4}(M_2)^{\rho_4} \times \left[\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(N,K)}_{\mathbf{X}_1\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)\left(\frac{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_1, \mathbf{x}_2)}{P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_{1ij}, \mathbf{x}_{2ik})}\right)^{s_4}\right]^{\rho_4}, \quad (84)$$

where

$$P^{(N,K)}_{\mathbf{X}_1\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{z}} P^{(N,K)}_{\mathbf{Z}\mathbf{X}_1\mathbf{X}_2}(\mathbf{z}, \mathbf{x}_1, \mathbf{x}_2). \quad (85)$$

Averaging over $\mathbf{y}$, $\mathbf{x}_{1ij}$, $\mathbf{x}_{2ik}$ and setting $s_4 = 1/(1 + \rho_4)$, we obtain

$$P_4 \leqq (M_0 - 1)^{\rho_4}(M_1)^{\rho_4}(M_2)^{\rho_4} \times \sum_{\mathbf{y}}\left[\sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} P^{(N,K)}_{\mathbf{X}_1\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)(P^{(N)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}\,|\,\mathbf{x}_1, \mathbf{x}_2)^{1/(1+\rho_4)}\right]^{1+\rho_4}. \quad (86)$$

From (1b) we see that $(M_0 - 1) < e^{NR_0}$. An upper bound for M_1 follows from (1b) as

$$M_1 < e^{NR_1} + 1 = e^{NR_1}(1 + e^{-NR_1}) = \exp\left\{N\left[R_1 + \frac{\log(1 + e^{-NR_1})}{N}\right]\right\} \leqq \exp\left\{N\left[R_1 + \frac{e^{-NR_1}}{N}\right]\right\}. \quad (87)$$

Using a similar upper bound for M_2, we have that

$$(M_0 - 1)(M_1)(M_2) \leqq \exp\left\{N\left[R_0 + R_1 + R_2 + \frac{e^{-N(R_1+R_2)}}{N}\right]\right\}. \quad (88)$$

From (88) and (70) we then obtain

$$P_4 \leqq \exp\{-N[E_4(\rho_4, P^{(K)}_{\mathbf{X}_1\mathbf{X}_2}) - \rho_4(R_0 + R_1 + R_2)]\}, \quad (89)$$

where E_4 is given by (29d). Summing (77), (78), (82), and (89) results in (27) which was to be proved.

APPENDIX B

Proof of Theorem 3

It can be easily verified that

$$E_\alpha(\rho_\alpha, P^{(K)}_{\mathbf{Z}\mathbf{X}_1\mathbf{X}_2})\,|_{\rho_\alpha=0} = 0 \qquad \text{for} \quad \alpha = 1, 2, 3, 4. \quad (90)$$

It can also be shown by a straightforward but tedious calculation that

$$\left.\frac{\partial E_\alpha}{\partial \rho_\alpha}\right|_{\rho_\alpha=0} = \begin{cases} \frac{1}{K} I(\mathbf{X}_1; \mathbf{Y}\,|\,\mathbf{X}_2, Z), & \alpha = 1 \\ \frac{1}{K} I(\mathbf{X}_2; \mathbf{Y}\,|\,\mathbf{X}_1, Z), & \alpha = 2 \\ \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}\,|\,Z), & \alpha = 3 \\ \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) - \frac{e^{-N(R_1+R_2)}}{N}, & \alpha = 4 \end{cases} \quad (91)$$

where the I's are mutual informations among K-vectors as computed under the joint distribution (32). Furthermore, from (29) it is seen that E_α is analytic in ρ_α in the neighborhood of $\rho_\alpha = 0$ and so can be expanded in a Taylor series about this point, $\alpha = 1, 2, 3, 4$:

$$E_\alpha(\rho_\alpha, P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2}) = \begin{cases} 0 + \frac{1}{K} I_\alpha \rho_\alpha + O_\alpha(\rho_\alpha^2), & \alpha = 1, 2, 3 \\ 0 + \left[\frac{1}{K} I_4 - \frac{e^{-N(R_1+R_2)}}{N} \right] \rho_4 + O_4(\rho_4^2), & \alpha = 4. \end{cases} \tag{92}$$

Here I_α is the appropriate expression from (91) and O is the usual Bachmann-Landau order-of-magnitude symbol. Furthermore, if $\mathbf{R} \subset \mathfrak{R}(P^{(K)}_{Z\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$, we have from (31) that

$$\frac{1}{K} I_\alpha - R_\alpha \equiv \delta_\alpha > 0, \qquad \alpha = 1, 2, 3\ 4. \tag{93}$$

Combining (92) and (93) with (30), we see that

$$P_e(C_N) \leqq \sum_{\alpha=1}^{3} \exp\{-N\rho_\alpha[\delta_\alpha + O_\alpha(\rho_\alpha^2)/\rho_\alpha]\} + \exp\left\{-N\rho_4\left[\delta_4 - \frac{e^{-N(R_1+R_2)}}{N} + O_4(\rho_4^2)/\rho_4\right]\right\}. \tag{94}$$

Now choose the integer $\hat{L}$ so large that

$$\bar{\delta}_4 \equiv \delta_4 - \frac{e^{-\hat{L}K(R_1+R_2)}}{\hat{L}K}$$

is positive. Next, choose sufficiently small positive values of $\rho_1, \rho_2, \rho_3, \rho_4$ so that $\delta_\alpha + O_\alpha(\rho_\alpha^2)/\rho_\alpha > 0$, $\alpha = 1, 2, 3$, and $\bar{\delta}_4 + O_4(\rho_4^2)/\rho_4 > 0$. The coefficient of N in each exponential of (94) is now negative, and we can increase N in multiples of K starting at $N = K\hat{L}$ until each term of (94) is less than $\epsilon/4$. Call this value of N, $N_0 = KL_0$. Then (33) follows. Q.E.D.

APPENDIX C

Proof of Lemma 1

By definition of $P_{e1}(C_K)$ and $H(U_1|\mathbf{Y}, U_0, U_2)$,

$$P_{e1}(C_K) = \sum_{\mathbf{y}} \sum_{i} \sum_{j \neq j^*(\mathbf{y})} \sum_{k} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}), \tag{95}$$

and

$$H(U_1|\mathbf{Y}, U_0, U_2) = \sum_{\mathbf{y}} \sum_{i} \sum_{j} \sum_{k} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}) \times \log \frac{1}{P^{(K)}_{U_1|U_0U_2\mathbf{Y}}(j|i, k, \mathbf{y})}. \quad (96)$$

By separating out the terms for which $j = j^*(\mathbf{y})$ in (96), one finds the identity

$$\begin{aligned} T &\equiv H(U_1|\mathbf{Y}, U_0, U_2) - P_{e1}(C_K) \log (M_1 - 1) - h(P_{e1}(C_K)) \\ &= \sum_{\mathbf{y}} \sum_{i} \sum_{j \neq j^*(\mathbf{y})} \sum_{k} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}) \\ &\quad \times \log \frac{P_{e1}(C_K)}{(M_1 - 1)P^{(K)}_{U_1|U_0U_2\mathbf{Y}}(j|i, k, \mathbf{y})} \\ &\quad + \sum_{\mathbf{y}} \sum_{i} \sum_{k} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j^*(\mathbf{y}), k, \mathbf{y}) \\ &\quad \times \log \frac{(1 - P_{e1}(C_K))}{P^{(K)}_{U_1|U_0U_2\mathbf{Y}}(j^*(\mathbf{y})|i, k, \mathbf{y})}. \end{aligned} \quad (97)$$

Now use the fact that $\log x \leqq x - 1$ to obtain

$$\begin{aligned} T &\leqq \sum_{\mathbf{y}} \sum_{i} \sum_{j \neq j^*(\mathbf{y})} \sum_{k} \left[\frac{P_{e1}P^{(K)}_{U_0U_1\mathbf{Y}}(i, k, \mathbf{y})}{(M_1 - 1)} - P_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}) \right] \\ &\quad + \sum_{\mathbf{y}} \sum_{i} \sum_{k} [(1 - P_{e1})P^{(K)}_{U_1|U_0U_2\mathbf{Y}}(j^*|i, k, \mathbf{y}) \\ &\quad - P_{U_0U_1U_2\mathbf{Y}}(i, j^*(\mathbf{y}), k, \mathbf{y})] \\ &= P_{e1} + (1 - P_{e1}) - 1 = 0. \end{aligned} \quad (98)$$

Replacing M_1 by $M_1 + 1$ yields (47a). Equations (47b), (47c), and (47d) are proved in a similar way starting from the definitions

$$P_{e2}(C_K) = \sum_{\mathbf{y}} \sum_{i} \sum_{j} \sum_{k \neq k^*(\mathbf{y})} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}), \quad (99a)$$

$$P_{e3}(C_K) = \sum_{\mathbf{y}} \sum_{i} \sum_{(j,k) \neq (j^*(\mathbf{y}), k^*(\mathbf{y}))} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}), \quad (99b)$$

and

$$P_e(C_K) = \sum_{\mathbf{y}} \sum_{(i,j,k) \neq (i^*, j^*, k^*)} P^{(K)}_{U_0U_1U_2\mathbf{Y}}(i, j, k, \mathbf{y}). \quad (99c)$$

APPENDIX D

Proof of Lemma 2

For part (a), we write a complicated conditional mutual information in two different ways:

$$\begin{aligned} &I(\mathbf{X}_1, U_1; \mathbf{Y}|U_2, \mathbf{X}_2, U_0) \\ &\quad = I(\mathbf{X}_1; \mathbf{Y}|U_2, \mathbf{X}_2, U_0) + I(U_1; \mathbf{Y}|\mathbf{X}_1, U_2, \mathbf{X}_2, U_0) \\ &\quad = I(U_1; \mathbf{Y}|U_2, \mathbf{X}_2, U_0) + I(\mathbf{X}_1; \mathbf{Y}|U_1, U_2, \mathbf{X}_2, U_0). \end{aligned} \quad (100)$$

Now

$$I(\mathbf{X}_1; \mathbf{Y} | U_0, U_2, \mathbf{X}_2)$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|U_0U_1U_2\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|U_0, U_1, U_2, \mathbf{X}_1, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|U_0U_2\mathbf{X}_2}(\mathbf{Y}|U_0, U_2, \mathbf{X}_2)}\right\}$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|U_0\mathbf{X}_2}(\mathbf{Y}|U_0, \mathbf{X}_2)}\right\} = I(\mathbf{X}_1; \mathbf{Y}|U_0, \mathbf{X}_2), \quad (101)$$

where the equalities result from the special form of the joint distributions as given by (39) and (40). For the next mutual information in (100), we have

$$I(U_1; \mathbf{Y}|U_0, U_2, \mathbf{X}_1, \mathbf{X}_2)$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|U_0U_1U_2\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|U_0, U_1, U_2, \mathbf{X}_1, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|U_0U_2\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|U_0, U_2, \mathbf{X}_1, \mathbf{X}_2)}\right\}$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2)}\right\} = 0. \quad (102)$$

The third mutual information in (100) can be written

$$I(U_1; \mathbf{Y}|U_0, U_2, \mathbf{X}_2)$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|U_0U_1U_2\mathbf{X}_2}(\mathbf{Y}|U_0, U_1, U_2, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|U_0U_2\mathbf{X}_2}(\mathbf{Y}|U_0, U_2, \mathbf{X}_2)}\right\}$$
$$= E\left\{\log \frac{P^{(K)}_{\mathbf{Y}|U_0U_1U_2}(\mathbf{Y}|U_0, U_1, U_2)}{P^{(K)}_{\mathbf{Y}|U_0U_2}(\mathbf{Y}|U_0, U_2)}\right\} = I(U_1; \mathbf{Y}|U_0, U_2). \quad (103)$$

Finally,

$$I(\mathbf{X}_1; \mathbf{Y}|U_0, U_1, U_2, \mathbf{X}_2) \geqq 0, \quad (104)$$

since all mutual informations are non-negative. Combining (100)–(104), we obtain (48a) which completes the proof of part (a).

The proofs for parts (b), (c), and (d) follow in a similar manner.

The equations corresponding to (100) are:
Part (b),

$$I(U_2, \mathbf{X}_2; \mathbf{Y}|U_0, U_1, \mathbf{X}_1)$$
$$= I(\mathbf{X}_2; \mathbf{Y}|U_0, U_1, \mathbf{X}_1) + I(U_2; \mathbf{Y}|U_0, U_1, \mathbf{X}_1, \mathbf{X}_2)$$
$$= I(U_2; \mathbf{Y}|U_0, U_1, \mathbf{X}_1) + I(\mathbf{X}_2; \mathbf{Y}|U_0, U_1, U_2, \mathbf{X}_1); \quad (105)$$

Part (c),

$$I(U_1, U_2, \mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0)$$
$$= I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0) + I(U_1, U_2; \mathbf{Y}|U_0, \mathbf{X}_1, \mathbf{X}_2)$$
$$= I(U_1, U_2; \mathbf{Y}|U_0) + I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0, U_1, U_2); \quad (106)$$

Part (d),

$$I(U_0, U_1, U_2, \mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) = I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) + I(U_0, U_1, U_2; \mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2) = I(U_0, U_1, U_2; \mathbf{Y}) + I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}|U_0, U_1, U_2). \quad (107)$$

Q.E.D.

APPENDIX E

Proof of Lemma 3

We use the identities:

$$\text{(a)}, \quad I(U_1; \mathbf{Y}|U_0, U_2) = H(U_1|U_0, U_2) - H(U_1|U_0, U_1, \mathbf{Y}); \quad (108a)$$

$$\text{(b)}, \quad I(U_2; \mathbf{Y}|U_0, U_1) = H(U_2|U_0, U_1) - H(U_2|U_0, U_1, \mathbf{Y}); \quad (108b)$$

$$\text{(c)}, \quad I(U_1, U_2; \mathbf{Y}|U_0) = H(U_1, U_2|U_0) - H(U_1, U_2|U_0, \mathbf{Y}); \quad (108c)$$

$$\text{(d)}, \quad I(U_0, U_1, U_2; \mathbf{Y}) = H(U_0, U_1, U_2) - H(U_0, U_1, U_2|\mathbf{Y}). \quad (108d)$$

From the joint distributions of the random variables U_0, U_1, and U_2 given in (2), we have:

$$(a), \quad H(U_1|U_0, U_1) = H(U_1) = KR_1'; \quad (109a)$$

$$\text{(b)}, \quad H(U_2|U_0, U_1) = H(U_2) = KR_2'; \quad (109b)$$

$$\text{(c)}, \quad H(U_1, U_2|U_0) = H(U_1, U_2) = H(U_1) + H(U_2) = K(R_1' + R_2'); \quad (109c)$$

$$\text{(d)}, \quad H(U_0, U_1, U_2) = H(U_0) + H(U_1) + H(U_2) = K(R_0' + R_1' + R_2'). \quad (109d)$$

Combining the appropriate equations in (108), (109), (41), (47), and (48), we have (49) which was to be proved.

APPENDIX F

Proof of Theorem 5

If $\mathbf{R}$ is an interior point of $\mathcal{S}^c$, then there is a sphere, σ, of radius $\eta(\mathbf{R}) > 0$, centered on $\mathbf{R}$ such that every point in σ is also in $\mathcal{S}^c$. Thus every point in $\mathcal{S}(Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2\mathbf{Y}})$ must be distant more than $\eta(\mathbf{R})$ away from $\mathbf{R}$, and this is true for every K, and every $Q^{(K)}_{U_0\mathbf{X}_1\mathbf{X}_2}$ in $\mathcal{Q}_K$. This in turn

implies that one of the inequalities

$$\begin{aligned} R_1 - \frac{1}{K} I(\mathbf{X}_1; \mathbf{Y} \mid U_0, \mathbf{X}_2) &> \eta(\mathbf{R}) \\ R_2 - \frac{1}{K} I(\mathbf{X}_2; \mathbf{Y} \mid U_0, \mathbf{X}_1) &> \eta(\mathbf{R}) \\ R_1 + R_2 - \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y} \mid U_0) &> \eta(\mathbf{R}) \\ R_0 + R_1 + R_2 - \frac{1}{K} I(\mathbf{X}_1, \mathbf{X}_2; \mathbf{Y}) &> \eta(\mathbf{R}) \end{aligned} \tag{110}$$

must hold for every encoding $C_K(\mathbf{R})$ of the sort under consideration, whenever $\mathbf{R}$ is interior to $\mathcal{S}^c$.

Now from (41), $R_\alpha' \geqq R_\alpha$, $\alpha = 0, 1, 2$, so that one of (110) holds also when the R's are replaced by R''s. From Lemma 3 we then find that

$$\hat{R}_\alpha' P_{e\alpha}(C_K(\mathbf{R})) + \frac{1}{K} h(P_{e\alpha}(C_K(\mathbf{R}))) > \eta(\mathbf{R}) \quad \text{for at least one } \alpha, \quad \alpha = 1, 2, 3, 4, \tag{111}$$

where we define

$$P_{e4}(C_K(\mathbf{R})) = P_e(C_K(\mathbf{R})) \tag{112}$$

and for any rate vector $\mathbf{R}$ we define an associated 4-vector $\hat{\mathbf{R}}$ by

$$(\hat{R}_1, \hat{R}_2, \hat{R}_3, \hat{R}_4) = (R_1, R_2, R_1 + R_2, R_0 + R_1 + R_2). \tag{113}$$

But from (87) and (88) we see that

$$\hat{R}_\alpha' \leqq \hat{R}_\alpha + \frac{e^{-K\hat{R}_\alpha}}{K} \leqq (\hat{R}_\alpha + e^{-\hat{R}_\alpha}), \tag{114}$$

so that

$$\hat{R}_\alpha' P_{e\alpha}(C_K(\mathbf{R})) + \frac{h(P_{e\alpha}(C_K(\mathbf{R}))}{K} \leqq (\hat{R}_\alpha + e^{-\hat{R}_\alpha}) P_{e\alpha}(C_K(\mathbf{R})) + h(P_{e\alpha}(C_K(\mathbf{R}))). \tag{115}$$

Combining (111) and (115) we find that

$$(\hat{R}_\alpha + e^{-\hat{R}_\alpha}) P_{e\alpha}(C_K(\mathbf{R})) + h(P_{e\alpha}(C_K(\mathbf{R})) \geqq \eta(\mathbf{R}) \quad \text{for at least one } \alpha, \quad \alpha = 1, 2, 3, 4. \tag{116}$$

Now

$$h(x) \leqq 2\sqrt{x}, \qquad 0 \leqq x \leqq 1, \tag{117}$$

as can be seen by the following simple argument. From

$$0 < \tfrac{1}{2}[1 + (1 - Z)^2],$$

it follows that $2Z < 1 + Z + Z^2/2 \leqq e^Z$, for $Z \geqq 0$. But for $Z < 0$ we also clearly have $2Z < e^Z$ so that $2Z < e^Z$ for all Z. Substitute $Z = \log \sqrt{(1-t)/t}$ to obtain

$$\log \frac{1-t}{t} < \frac{\sqrt{1-t}}{\sqrt{t}} < \frac{1}{\sqrt{t}}, \qquad 0 < t < 1$$

or

$$\int_\epsilon^x \log \frac{1-t}{t}\, dt < \int_\epsilon^x \frac{dt}{\sqrt{t}}, \qquad \epsilon < x < 1.$$

Perform the integration and take the limit as $\epsilon \to 0$. Equation (117) results.

Use (117) in (116) to find that $(\hat{R}_\alpha + e^{-\hat{R}_\alpha})P_{e\alpha} + 2\sqrt{P_{e\alpha}} \geqq \eta(\mathbf{R})$ for at least one α, $\alpha = 1, 2, 3, 4$. This implies that

$$P_{e\alpha}(C_K(\mathbf{R})) \geqq \left[\frac{\sqrt{1 + \eta(\mathbf{R})[\hat{R}_\alpha + e^{-\hat{R}_\alpha}]} - 1}{\hat{R}_\alpha + e^{-\hat{R}_\alpha}} \right]^2 \equiv \delta_\alpha(\mathbf{R}) > 0.$$

Since $P_e(C_K(\mathbf{R})) \geqq \max_\alpha [P_{e\alpha}(C_K(\mathbf{R}))]$, we find finally that

$$P_e(C_K(\mathbf{R})) \geqq \delta(\mathbf{R}) > 0, \tag{118}$$

where $\delta(\mathbf{R}) \equiv \min_\alpha \delta_\alpha(\mathbf{R})$ is independent of K and the encoding $C_K(\mathbf{R})$. Q.E.D.

APPENDIX G

Proof of Theorem 6

We first show that for every positive integer n and every integer r such that $0 \leqq r \leqq n$,

$$\mathbf{R}_1 \in \mathfrak{R}, \quad \mathbf{R}_2 \in \mathfrak{R} \Rightarrow \mathbf{R}_3 \equiv \frac{r}{n}\mathbf{R}_1 + \frac{n-r}{n}\mathbf{R}_2 \in \mathfrak{R}. \tag{119}$$

Since $\mathfrak{C}$ is the closure of $\mathfrak{R}$, and since the rationals are dense in the reals, (119) implies that if $\mathbf{R}_1 \in \mathfrak{C}$ and $\mathbf{R}_2 \in \mathfrak{C}$, then for every λ, $0 \leqq \lambda \leqq 1$, $\mathbf{R}_3 \equiv \lambda\mathbf{R}_1 + (1-\lambda)\mathbf{R}_2 \in \mathfrak{C}$, which shows $\mathfrak{C}$ to be convex.

To establish (119), we use the notion of time sharing to generate new codings from old ones. Suppose we have two codings $C_N(\mathbf{R}_1)$ and $C_N(\mathbf{R}_2)$ both of block length N and with numbers of words $\mathbf{M}_1$ and $\mathbf{M}_2$

respectively, where as usual

$$\mathbf{M}_\alpha = (M_{0\alpha}, M_{1\alpha}, M_{2\alpha}) = (\lceil e^{NR_{0\alpha}} \rceil, \lceil e^{NR_{1\alpha}} \rceil, \lceil e^{NR_{2\alpha}} \rceil) \qquad \alpha = 1, 2. \tag{120}$$

Denote by P_{e1} and P_{e2} the respective error probabilities achievable with $C_N(\mathbf{R}_1)$ and $C_N(\mathbf{R}_2)$. Now consider the possible channel input vectors that can be obtained by using $C_N(\mathbf{R}_1)$ r times followed by $(n - r)$ uses of $C_N(\mathbf{R}_2)$. The totality of these input vectors, each of nN components, can be thought of as the words of a new code of block length nN. Denoting its word size parameter by $\mathbf{M}$, we have

$$M_i = (M_{i1})^r (M_{i2})^{n-r}, \qquad i = 0, 1, 2. \tag{121}$$

If we use the decoders for $C_N(\mathbf{R}_1)$ and $C_N(\mathbf{R}_2)$ to decode the appropriate blocks of length N in this new larger code, the error probability for the new code, P_e, will satisfy

$$1 - P_e = (1 - P_{e1})^r (1 - P_{e2})^{n-r} \geqq (1 - rP_{e1})[1 - (n - r)P_{e2}] \geqq 1 - rP_{e1} - (n - r)P_{e2}$$

so that

$$P_e \leqq rP_{e1} + (n - r)P_{e2}. \tag{122}$$

Here we have used the fact that the channel is memoryless.

We now use this time-sharing notion to establish (119). Suppose that integers n and r are given with $n > 0$, $0 \leqq r \leqq n$ and that $\mathbf{R}_1$ and $\mathbf{R}_2$ are rate points in $\mathfrak{R}$. Suppose further that $\epsilon > 0$ is given. Then, from Theorem 4, there exist positive integers K_1 and L_1 and a sequence of codings $C_{N_1}(\mathbf{R}_1)$, $N_1 = K_1L_1$, $K_1(L_1 + 1)$, $K_1(L_1 + 2)$, $\cdots$ such that for each coding of the sequence $P_e(C_{N_1}(\mathbf{R}_1)) > \epsilon/n$. Similarly there exist integers K_2 and L_2 and a second sequence of codings $C_{N_2}(\mathbf{R}_2)$, $N_2 = K_2L_2$, $K_2(L_2 + 1)$, $K_2(L_2 + 2)$, $\cdots$ such that for each coding in the sequence $P_e(C_{N_2}(\mathbf{R}_2)) < \epsilon/n$. We now choose one coding out of each of these sequences of codings in such a way that they are of the same block length N. A suitable choice for N is the least common multiple of K_1L_1 and K_2L_2. Call the two codings $C_N(\mathbf{R}_1)$ and $C_N(\mathbf{R}_2)$. Their error probabilities are $P_{e1} < \epsilon/n$ and $P_{e2} < \epsilon/n$. Time sharing them as discussed earlier yields a new coding C, of block length $N_3 = nN$, code book size $\mathbf{M}$ given by (121) and (120), and error probability

$$P_e \leqq rP_{e1} + (n - r)P_{e2} = r\frac{\epsilon}{n} + (n - r)\frac{\epsilon}{n} = \epsilon$$

from (122). Now from the fact that $\lceil x \rceil \lceil y \rceil \geqq \lceil xy \rceil$, (121) and (120) give

$$M_i = \lceil e^{NR_{i1}} \rceil^r \lceil e^{NR_{i2}} \rceil^{n-r} \geqq \lceil e^{N[rR_{i1}+(n-r)R_{i2}]} \rceil \geqq \lceil e^{NR_3 R_{i3}} \rceil, \qquad i = 0, 1, 2,$$

where $\mathbf{R}_3$ is as in (119). Thus by deleting some words from the code C we can obtain a coding with rate $\mathbf{R}_3$, and block length N_3, that has error probability $P_e < \epsilon$. Q.E.D.

APPENDIX H

Proof of Lemma 4

Consider (57a). We write

$$\begin{aligned} I(\mathbf{X}_1; \mathbf{Y} | Z, \mathbf{X}_2) &= E \log \frac{P^{(K)}_{\mathbf{Y}|Z\mathbf{X}_1\mathbf{X}_2}(\mathbf{Y} | Z, \mathbf{X}_1, \mathbf{X}_2)}{P^{(K)}_{\mathbf{Y}|Z\mathbf{X}_2}(\mathbf{Y} | Z, \mathbf{X}_2)} \\ &= E \log \frac{\prod_{t=1}^{K} P_{Y|X_1X_2}(Y_t | X_{1t}, X_{2t})}{\prod_{t=1}^{K} P^{(K)}_{Y_t|Z\mathbf{X}_2Y_1\cdots Y_{t-1}}(Y_t | Z, \mathbf{X}_2, Y_1, \cdots, Y_{t-1})} \\ &= \sum_{t=1}^{K} [H(Y_t | Z, \mathbf{X}_2, Y_2, \cdots, Y_{t-1}) - H(Y_t | X_{1t}, X_{2t})]. \end{aligned} \tag{123}$$

Here, for $t = 1$, the conditioning on $Y_1, \cdots, Y_{t-1}$ is to be omitted. But

$$H(Y_t | Z, \mathbf{X}_2, Y_1, \cdots Y_{t-1}) \leqq H(Y_t | Z, X_{2t}), \tag{124}$$

since removing conditioning random variables cannot decrease an entropy. Combining (123) and (124) we have

$$I(\mathbf{X}_1; \mathbf{Y} | Z, \mathbf{X}_2) \leqq \sum_{t=1}^{K} [H(Y_t | Z, X_{2t}) - H(Y_t | X_{1t}, X_{2t})], \tag{125}$$

or

$$I(\mathbf{X}_1; \mathbf{Y} | Z, \mathbf{X}_2) \leqq \sum_{t=1}^{K} I(X_{1t}; Y_t | Z, X_{2t}). \tag{126}$$

The proofs for (57b), (57c), and (57d) are similar.

APPENDIX I

Let numbers A_t, B_t, C_t, and D_t be given that satisfy the inequalities

$$0 \leqq A_t \leqq C_t, \tag{127a}$$

$$0 \leqq B_t \leqq C_t, \tag{127b}$$

$$0 \leqq C_t \leqq A_t + B_t, \tag{127c}$$

$$0 \leqq C_t \leqq D_t, \qquad t = 1, 2, \cdots, K. \tag{127d}$$

Let $\mathcal{R}_t$ denote the set of points (x, y, z) in three-space such that

$$0 \leqq x \leqq A_t, \tag{128a}$$

$$0 \leqq y \leqq B_t, \tag{128b}$$

$$0 \leqq x + y \leqq C_t, \tag{128c}$$

$$0 \leqq x + y + z \leqq D_t, \tag{128d}$$

for $t = 1, 2, \cdots, K$. A sketch of $\mathcal{R}_t$ is shown in Fig. 7, corresponding to the case in which all the inequalities in (127) are strict. We further define

$$\mathcal{R} \equiv \bigcup_{t=1}^{K} \mathcal{R}_t. \tag{129}$$

Now consider the region $\mathcal{R}_0$ consisting of all points (x, y, z) such that

$$0 \leqq x \leqq A_0 \equiv \frac{1}{K} \sum_{t=1}^{K} A_t, \tag{130a}$$

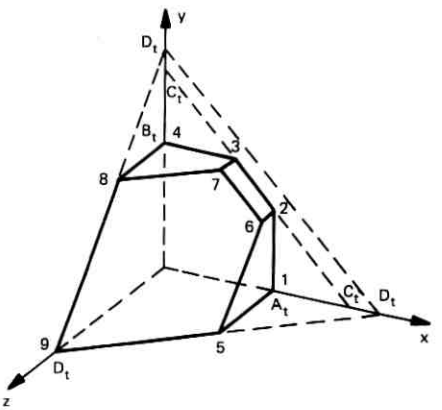

Fig. 7—The convex region $\mathcal{R}_t$.

$$0 \leqq y \leqq B_0 \equiv \frac{1}{K} \sum_{t=1}^{K} B_t, \tag{130b}$$

$$0 \leqq x + y \leqq C_0 \equiv \frac{1}{K} \sum_{t=1}^{K} C_t, \tag{130c}$$

$$0 \leqq x + y + z \leqq D_0 \equiv \frac{1}{K} \sum_{t=1}^{K} D_t. \tag{130d}$$

Our first goal in this appendix is to show that

$$\mathfrak{R}_0 \subseteq \text{convex hull } \mathfrak{R}. \tag{131}$$

By summing the inequalities (127) and using the definitions of A_0, B_0, C_0, and D_0 given in (130), we see that (127) also holds for $t = 0$. $\mathfrak{R}_0$, too, then has the form shown in Fig. 7. As is seen, each region $\mathfrak{R}_t$, $t = 0, 1, \cdots, K$, is convex and has ten extreme points, the components of which are listed below:

$$\begin{aligned}
\mathbf{r}_{0t} &= (0, 0, 0) \\
\mathbf{r}_{1t} &= (A_t, 0, 0) \\
\mathbf{r}_{2t} &= (A_t, C_t - A_t, 0) \\
\mathbf{r}_{3t} &= (C_t - B_t, B_t, 0) \\
\mathbf{r}_{4t} &= (0, B_t, 0) \\
\mathbf{r}_{5t} &= (A_t, 0, D_t - A_t) \\
\mathbf{r}_{6t} &= (A_t, C_t - A_t, D_t - C_t) \\
\mathbf{r}_{7t} &= (C_t - B_t, B_t, D_t - C_t) \\
\mathbf{r}_{8t} &= (0, B_t, D_t - B_t) \\
\mathbf{r}_{9t} &= (0, 0, D_t).
\end{aligned} \tag{132}$$

[Some of these points may coincide if there are equalities in (127) instead of strict inequalities.] For the extreme point of $\mathfrak{R}_0$ we also have

$$\mathbf{r}_{i0} = \frac{1}{K} \sum_{t=1}^{K} \mathbf{r}_{it}, \qquad i = 0, 1, \cdots, 9 \tag{133}$$

which follows directly from (132) and the definitions on the right of (130). We recall that a convex body is characterized by its extreme

points: $\mathbf{r} \in \mathfrak{R}_t$ if and only if

$$\mathbf{r} = \sum_{i=0}^{9} \lambda_i \mathbf{r}_{it}, \tag{134}$$

where

$$\lambda_i \geqq 0, \quad i = 0, 1, \cdots, 9 \quad \text{and} \quad \sum_{0}^{9} \lambda_i = 1, \quad t = 0, 1, \cdots, 9. \tag{135}$$

Equation (131) is now easy to establish. It is clear that the convex hull of $\mathfrak{R}$ is the set of all points that can be written in the form

$$\mathbf{r}' = \sum_{i=0}^{9} \sum_{t=1}^{K} u_{it} \mathbf{r}_{it}, \tag{136}$$

where

$$u_{it} \geqq 0, \quad i = 0, 1, \cdots, 9, \quad t = 1, \cdots, K, \quad \sum_{i=0}^{9} \sum_{t=1}^{K} u_{it} = 1. \tag{137}$$

Now let $\mathbf{r}$ be any element in $\mathfrak{R}_0$. Then $\mathbf{r}$ can be written in the form (134)–(135) with $t = 0$. Substituting from (133) yields

$$\mathbf{r} = \sum_{i=0}^{9} \sum_{t=1}^{K} \frac{\lambda_i}{K} \mathbf{r}_{it}. \tag{138}$$

But defining

$$u'_{it} = \frac{\lambda_i}{K}, \qquad i = 0, \cdots, 9, \qquad t = 1, 2, \cdots, K, \tag{139}$$

we see that $u'_{it} \geqq 0$ and

$$\sum_{i=0}^{9} \sum_{t=0}^{K} u'_{it} = 1. \tag{140}$$

Comparison with (136) now shows that $\mathbf{r}$ is in the convex hull of $\mathfrak{R}$. Equation (131) then follows.

The application of the foregoing to (60) is immediate. Let

$$\begin{aligned} A_t &= I(X_{1t}; Y_t | X_{2t}, Z) \\ B_t &= I(X_{2t}; Y_t | X_{1t}, Z) \\ C_t &= I(X_{1t}, X_{2t}; Y_t | Z) \\ D_t &= I(X_{1t}, X_{2t}; Y_t) \end{aligned}$$

$t = 1, \cdots, 9$. Equations (127) are satisfied. We then identify $\mathfrak{R}_t$ of this appendix with $\mathfrak{R}(P_{ZX_{1t}X_{2t}Y_t})$ of (60), $t = 1, 2, \cdots, K$, and $\mathfrak{R}_0$ with

$\mathcal{R}^*(P^{(K)}_{ZX_1X_2Y})$ of (59) which is consistent with (130). Then (129) and (131) yield (60). Q.E.D.

REFERENCES

1. Shannon, C. E., "A Mathematical Theory of Communications," B.S.T.J., *27*, No. 3 (July 1948), pp. 379–423, No. 4 (October 1948), pp. 623–656.
2. Shannon, C. E., "Two Way Communication Channels," *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1961, pp. 611–644.
3. Van der Meulen, E. C., "A Note and Counterexample on the Two Way Channel," Center for System Science, 70–15, University of Rochester, 1970.
4. Liao, H., "A Coding Theorem for Multiple Access Communications," *1972 International Symposium on Information Theory*, Asilomar, California, 1972. Also Ph.D. dissertation, "Multiple Access Channels," Dept. of Electrical Engineering, University of Hawaii, 1972.
5. Van der Meulen, E. C., "The Discrete Memoryless Channel with Two Senders and One Receiver," *2nd International Symposium on Information Transmission*, USSR, 1971.
6. Ahlswede, R., "Multi-way Communication Channels," *2nd International Symposium on Information Transmission*, USSR, 1971. To appear in Problems of Control and Information Theory.
7. Cover, T., "Broadcast Channels," IEEE Trans. Information Theory, *IT-18*, 1972, pp. 2–13.
8. Bergmans, P. P., "Random Coding Theorem for Broadcast Channels with Degraded Components," unpublished work.
9. Gallager, R. G., *Information Theory and Reliable Communication*, New York: John Wiley and Sons, Inc., 1968.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

A Compromise Equalizer Design Incorporating Performance Invariance

By F. J. BROPHY, G. J. FOSCHINI, and R. D. GITLIN

(Manuscript received February 20, 1973)

We give a solution to the problem of designing a fixed compromise equalizer for use in transmission systems involving an ensemble of random channels. The signal and noise spectra, along with the second-order statistics of the channel ensemble, are used to find the equalizer characteristic that minimizes the mean-square distortion between the equalizer output and a scaled version of the transmitter output. The key departure from previous work is that the criterion better captures practical performance invariance; specifically, the cost function incorporates the insensitivity of a well-designed demodulator to any amplitude scaling or time delay introduced by a particular channel. After demonstrating that the optimum equalizer shape is related to the principal eigenfunction of a normalized channel correlation function, we consider several special cases that give further insight into the properties of the solution. We find that the equalizer amplitude is attenuated over those frequencies where the signal-to-noise or signal-to-channel-variance ratios are small. The analysis confirms the standard engineering practice of inverting the average channel in the absence of noise and when the variance of the channel characteristics is small.

I. INTRODUCTION

A fixed compromise equalizer is frequently employed in data transmission systems to compensate for linear distortion introduced by a channel drawn from a random ensemble.[1] Typically, compromise equalizers find application in systems which, because of economic or other considerations, do not use an adaptive equalizer. It is possible that, even when adaptive equalization is used to compensate for a particular channel characteristic, one might use a compromise equalizer to provide a good initial channel and thereby reduce the receiver

adaptation time. As its name suggests, the equalizer is a *fixed* linear filter that effects a compromise by compensating for an "average" channel. We propose a procedure that uses the statistics of the channel ensemble, the modulated signal, and the additive noise at the demodulator to design a filter that minimizes a performance measure appropriate for most transmission systems. The performance measure is an adaptively scaled mean-square error. For example, it is particularly well suited for use in a data transmission system and results in a filter that is significantly different from that obtained by directly minimizing the mean-square error.[2]

In order to avoid being restricted by nonlinear demodulation techniques (e.g., those used in FSK or DPSK data systems) and to be able to accommodate asynchronous signalling, the equalizer operates directly on the received passband signal and is thus a channel, rather than a synchronous, equalizer. Our performance criterion is the continuous time mean-square error between the equalizer output and an adaptively scaled version of the transmitter output. This scaling needs to be done only once in the design of the filter and not each time a new channel is dialed up. The scaling is such that the filter is invariant to any amplitude scaling, sign inversion, or time delay encountered in transmission over a particular channel; this type of invariance is appropriate for a compromise equalizer used in most transmission systems, since amplitude scaling and a fixed time delay will not be "seen" by a well-designed demodulator.

Since the criterion is quadratic in nature, the optimum filter response is determined by the second-order statistics of the signal, noise, and channel ensemble. The best filter shape is shown to be the principal eigenfunction of an integral operator whose kernel is a weighted channel correlation function. An explicit design procedure is described in the text, and several interesting questions are discussed as well. Some of these questions are:

(*i*) How different is the compromise equalizer from the inverse of the average channel characteristic?

(*ii*) Suppose there is no amplitude distortion but only delay distortion. What is the nature of the compensation? Is there amplitude as well as delay compensation?

(*iii*) How sensitive is the filter design to different signal spectra?

In Section II the compromise equalization problem is formulated, and the distortion measure is discussed in considerable detail. The determination of the optimum filter is described in Section III, and an example is provided in Section IV illustrating the design technique.

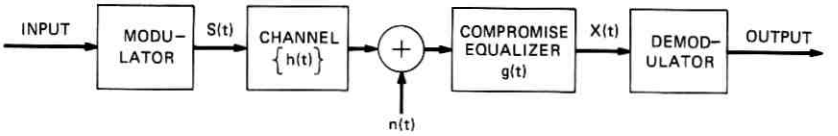

Fig. 1—Transmission system with a compromise equalizer.

II. PROBLEM FORMULATION

2.1 *System Configuration*

We begin by specifying the framework of our discussion. To accommodate a system employing nonlinear demodulation, the compromise equalizer is assumed to be placed in the passband. Since we also want to provide a setting general enough to include asynchronous transmission, no restriction is placed on the form of the filter other than requirement of a finite energy impulse response. As shown in Fig. 1, $s(t)$ denotes the modulated signal, $n(t)$ the received additive noise, $x(t)$ the compromise equalizer output, and $g(t)$ the equalizer impulse response (which we ultimately seek to specify). The ensemble of channel impulse responses, $\{h(t)\}$, as well as $n(t)$ and $s(t)$, are all assumed to be continuous in quadratic mean and statistically independent of each other. Both $s(t)$ and $n(t)$ will be taken to be zero mean, to be wide-sense stationary with finite power and to possess power spectral densities $S(\omega)$ and $N(\omega)$, respectively. To make the subsequent analysis precise, we assume $h(t)$ has bounded absolutely integrable sample paths (with probability one); thus, the ensemble of channel frequency characteristics $\{H(\omega)\}$ is well defined.†

Our problem is first to select an appropriate optimization criterion and then to choose the compromise equalizer that minimizes this measure. To avoid the specifics of particular demodulators, we will be concerned with preserving the fidelity between $s(t)$ and $x(t)$. A natural first choice for a distortion measure is the continuous-time mean-square error

$$\mathcal{E}_0 = \langle [x(t) - s(t - \tau)]^2 \rangle, \tag{1}$$

where $\langle\ \rangle$ denotes the average with respect to the signal, noise, and channel ensemble, and τ is an arbitrary delay. Using this measure, Maurer and Franks[2] found that the optimum filter is given by

$$G_o(\omega) = \frac{S(\omega)\langle H(\omega)\rangle^* e^{-j\omega\tau}}{S(\omega)\langle |H(\omega)|^2 \rangle + N(\omega)}, \tag{2}$$

† The process $\{H(\omega)\}$ is assumed to have a square integrable covariance, and the variance is assumed to be nonzero for the frequency range of interest.

where the asterisk denotes the complex conjugate and $H(\omega)$ is a member of the channel ensemble. It is easy to see that $G_o(\omega)$, given by (2), is not just the Wiener filter for the average channel, since the mean-square channel dispersion $\langle |H(\omega)|^2 \rangle$, rather than $|\langle H(\omega) \rangle|^2$, modifies the filter characteristic. We note that, even as $N(\omega) \to 0$, the filter does not generally invert the average channel. This will be the case when the channel variance is extremely small so that $\langle |H(\omega)|^2 \rangle \approx |\langle H(\omega) \rangle|^2$; this would occur, for instance, when the ensemble consists of only one characteristic.

Care must be exercised in determining how to use the design technique that results in the filter given by (2). For example, suppose the individual transfer characteristics are greatly varied in the degree of attenuation they impart across the band of interest. It can follow (depending on how probabilities are attached) that the low-loss channels determine the character of the averages in (2)—the "whomper" effect. Hence, linear distortion in the lossy channels may not be suitably equalized. However, if linear distortion is uniformly the dominant impairment, the noise level can be set to zero and the "whomper" effect easily eliminated by preparing the data by normalizing the channels prior to averaging. For example, the normalization can be accomplished by individually scaling the characteristics so they have the same energy in response to some pulse or so they have the same gain at some central frequency. Even if linear distortion is not the uniformly dominant impairment, there may be applications where the noise is set to zero and a normalization of the channels made prior to computing the optimum equalizer. In such a procedure, one is trading noise immunity for immunity to linear distortion. Another case of importance in applications is when the channels have the same amplitude characteristic but different phase characteristics; here, of course, the "whomper" effect is nonexistent. The considerations of this paragraph will also apply to the design technique we shall develop in the following sections.

While the $\mathcal{E}_0$ criterion provides an interesting and tractable formulation, it has the shortcoming that it understates the capability of most demodulators. A striking example of this occurs if the ensemble $\{h(t)\}$ has zero mean. The zero mean is reasonable for systems subject to occasional "phase hits," the mathematical implication being that, if $h(t)$ is a possible channel, then $-h(t)$ is just as probable. Then $G_o(\omega) \equiv 0$. Yet, in practice, one would expect to do better than "pull out the plug and go home."

At the other extreme we could use an information theoretic type

criterion. For example, we could choose $G(\omega)$ to maximize the average mutual information between $x(t)$ and $s(t)$. This criterion is mathematically tractable; in fact, the solution is trivial—$G(\omega)$ can be any characteristic which is nonzero over the same frequency range as $S(\omega)$.† The shortcoming of this criterion is (as the solution suggests) that it overestimates the demodulation capability.

As the state of the art in demodulation advances, we would anticipate an evolution of criteria away from the mean-square error toward the information theoretic. Clearly, a good criterion should give the demodulator credit for what it can realistically accomplish and at the same time pose a tractable optimization problem for determining $G(\omega)$.

With this motivation, we propose a mean-square-error criterion which reflects the fact that the performance of a well-designed demodulator is insensitive to scaling of the input (the automatic gain control feature), an input sign change (the information is generally differentially encoded), and, of course, a time delay. Under such a criterion, the compromise equalizer will no longer be implicitly constrained by attempting to faithfully reproduce the modulated signal.

2.2 *An Adaptively Scaled Mean-Square Error*

Based upon the above discussion, we consider the following adaptively scaled mean-square error

$$\mathcal{E} = \min_{A,B} \langle [x(t) - As(t - B)]^2 \rangle_{s,n}, \tag{3}$$

where A and B are real numbers and the averaging is over the signal and noise statistics with both the channel and equalizer held fixed. The criterion is meaningful under the assumption that the signal-to-noise ratio at the receiver does not change appreciably from channel to channel. We stress that the considerations mentioned in the third paragraph of Section 2.1 apply here, as well. The optimum equalizer is obtained by averaging $\mathcal{E}$ over the channel ensemble and then minimizing the result with respect to the equalizer transfer function $G(\omega)$, subject to a power constraint on the demodulator input. The quantities A and B, which are channel-dependent, provide an adaptively scaled reference signal $As(t - B)$. The reference is adaptive in that, for each realization of the channel, A and B are chosen to minimize $\mathcal{E}$. Notice, for example, that, if a particular channel introduces a sign inversion, then $A = -1$ will remove this effect. When a channel

† This comes about because no information is lost when the signal is subject to a reversible operation (such as a channel).

has a more complicated phase and/or gain characteristic, it is no longer apparent what the optimal value of A should be; however, we shall shortly see that the minimizing value of A can be determined analytically. We shall also consider the determination of B. Thus, the filter will not expend any of its degrees of freedom by attempting to compensate for a sign inversion, amplitude scaling, or time delay introduced during transmission. It should be clear that, since A and B depend on the channel characteristics, they are random variables. Simply put, the criterion given by (3) forces the equalizer to minimize only the portion of the output signal that does not look like a scaled or delayed version of the transmitted signal.

We now consider the properties of the adaptively scaled mean-square error. We begin by letting

$$I = \langle [x(t) - As(t - B)]^2 \rangle_{s,n}, \tag{4}$$

i.e.,

$$\mathcal{E} = \min_{A,B} I.$$

Carrying out the indicated average gives

$$I = \int_{-\infty}^{\infty} \{S(\omega)[|F(\omega)|^2 - 2Ae^{j\omega B}F(\omega) + A^2] + |G(\omega)|^2 N(\omega)\} \frac{d\omega}{2\pi}, \tag{5}$$

where $S(\omega)$ is the power spectral density of $s(t)$ and $F(\omega)$ is the product of $G(\omega)$ and $H(\omega)$. (Notice I does not change when $F(\omega)$ is replaced by $Re\{F(\omega)\}$.) To find the minimum of I with respect to A and B, we set to zero the partial derivatives of I with respect to these variables and find that

$$\frac{\partial I}{\partial A} = -2\int_{-\infty}^{\infty} S(\omega)e^{j\omega B}F(\omega) \frac{d\omega}{2\pi} + 2A\int_{\infty}^{\infty} S(\omega) \frac{d\omega}{2\pi} = 0 \tag{6a}$$

$$\frac{\partial I}{\partial B} = -2\int_{-\infty}^{\infty} j\omega S(\omega)e^{j\omega B}F(\omega) \frac{d\omega}{2\pi} = 0. \tag{6b}$$

From (6a) we have A_{opt} given by the correlation ratio

$$A_{\text{opt}} = \frac{\int_{-\infty}^{\infty} S(\omega)e^{j\omega B}F(\omega) \frac{d\omega}{2\pi}}{\int_{-\infty}^{\infty} S(\omega) \frac{d\omega}{2\pi}} = \frac{\langle s(t - B)x(t) \rangle}{\langle s^2(t - B) \rangle}. \tag{7}$$

The interpretation of B_{opt} is facilitated by letting

$$y(t) \equiv \int_{-\infty}^{\infty} j\omega F(\omega)S(\omega)e^{j\omega t} \frac{d\omega}{2\pi}, \tag{8}$$

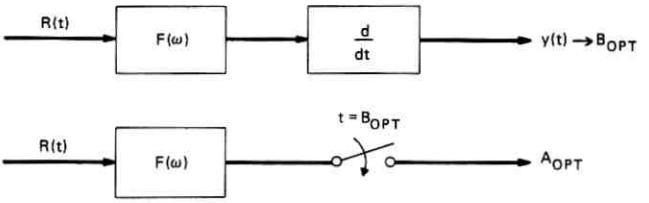

Fig. 2—Interpretation of optimum delay (B_{opt}) and amplitude (A_{opt}) scaling. B_{opt} is one of the instants when $y(t)$ is zero.

which is recognized as the response of a linear filter with transfer function $j\omega F(\omega)$ [i.e., $F(\omega)$ followed by a differentiator] to the input $R(t)$, where $R(t)$ is the (inverse) Fourier transform of $S(\omega)$ and is the signal correlation function. Comparing (6b) and (8), we see that B_{opt} is one of the instants when $y(t)$ is zero. We illustrate this interpretation of B_{opt} in Fig. 2, where we also indicate how A_{opt} may be obtained in a similar manner. Determining B_{opt} is a very difficult problem, since it is tantamount to asking for the zero crossing of a signal from knowledge of its Fourier transform. In order to proceed further, we will approximate B_{opt} by the delay at midband, which we conveniently take to be zero for each channel.

Using the value of A_{opt} given by (7) and setting $B = 0$, we have

$$\mathcal{E} = \int_{-\infty}^{\infty} |G(\omega)|^2 [S(\omega)|H(\omega)|^2 + N(\omega)] \frac{d\omega}{2\pi} - \frac{1}{\alpha} \left| \int_{-\infty}^{\infty} S(\omega) G(\omega) H(\omega) \frac{d\omega}{2\pi} \right|^2, \quad (9)$$

where

$$\alpha = \int_{-\infty}^{\infty} S(\omega) \frac{d\omega}{2\pi}.$$

The optimum filter is obtained by first averaging (9) with respect to the ensemble of channels and then minimizing this average with respect to $G(\omega)$.† Before doing this, we give geometric interpretations to both the criterion and the optimum filter.

2.3 *A Geometric Interpretation of the Problem*

Finding a geometric framework in which to view the optimization problem will be quite helpful in understanding the nature of the solution. To put the problem at hand in such a setting, we introduce

† The minimization is done subject to a constraint on the average output power.

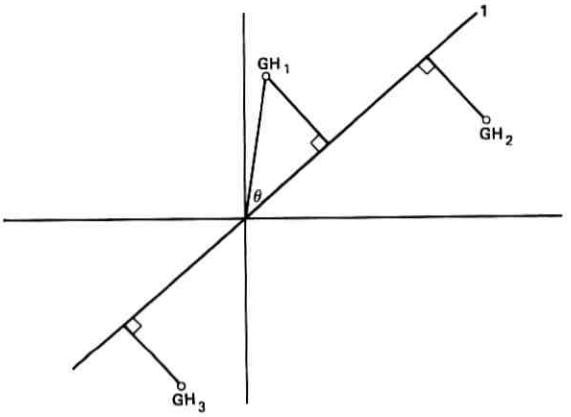

Fig. 3—Geometric interpretation of criterion. The adaptively scaled mean-square error may be interpreted as the average squared distance from the point **GH** (which represents an equalized channel) to the ray **1** (which represents a distortionless channel).

the signal-weighted inner product

$$(\mathbf{U}, \mathbf{V}) \equiv \int_{-\infty}^{\infty} U(\omega) V^*(\omega) S(\omega) \frac{d\omega}{2\pi}, \tag{10}$$

where the vectors **U** and **V** represent the functions $U(\omega)$ and $V(\omega)$ respectively, and the vector **1** will correspond to the real function of unit amplitude. Suppose that the noise is set to zero; then, in terms of the above notation, we can write†

$$\mathcal{E} = \|\mathbf{GH}\|^2 - \frac{1}{\alpha}(\mathbf{GH}, \mathbf{1})^2, \tag{11}$$

where the vector **GH** corresponds to the function $G(\omega)H(\omega)$. For convenience we set the signal power equal to unity, i.e., $\alpha = 1$, and apply the Schwarz inequality, which gives

$$\mathcal{E} = \|\mathbf{GH}\|^2 - (\mathbf{GH}, \mathbf{1})^2 \geqq \|\mathbf{GH}\|^2 - \|\mathbf{GH}\|^2 \cdot \|\mathbf{1}\|^2 = 0, \tag{12}$$

where the lower bound is achieved when **GH** is proportional to **1**.‡ From (11) we have

$$\begin{aligned} \mathcal{E} &= (\mathbf{GH}, \mathbf{GH}) - (\mathbf{GH}, \mathbf{1})^2 = (\mathbf{GH}, \mathbf{GH})\left[1 - \frac{(\mathbf{GH}, \mathbf{1})^2}{(\mathbf{GH}, \mathbf{GH})}\right] \\ &= \|\mathbf{GH}\|^2[1 - \cos^2\theta] = [\|\mathbf{GH}\| \sin\theta]^2, \end{aligned} \tag{13}$$

† The norm of the vector **U**, denoted by $\|\mathbf{U}\|$, is given by $(\mathbf{U}, \mathbf{U})^{\frac{1}{2}}$.

‡ Thus, if we have only one channel characteristic and no noise, the equalizer will invert the channel.

where θ is the angle between **GH** and **1**. Equation (13) provides a very useful interpretation of the error $\mathcal{E}$. As shown in Fig. 3, $\|\mathbf{GH}\| \sin\theta$ is the distance from the vector **GH** to the ray colinear with the vector **1**. Hence, the average of $\mathcal{E}$ with respect to the channel ensemble, which we denote by $\langle\mathcal{E}\rangle_{\mathbf{H}}$, is the average squared distance from **GH** to the ray **1**. Thus, for a given channel-equalizer constellation, $\{\mathbf{GH}_i\}$, the equalizer is chosen to minimize the dispersion about the ray **1**.

In order to get a feeling for the capability of the equalizer to modify the channel constellation $\{\mathbf{H}_i\}$, we associate with $G(\omega)$ a linear operator $\mathcal{G}$ that maps a particular channel $\mathbf{H}_i$ into $\mathbf{GH}_i$; we write this operation symbolically as

$$\mathcal{G}\colon\ \mathbf{H}_i \rightarrow \mathbf{GH}_i. \tag{14}$$

The equalizer operator, $\mathcal{G}$, is called diagonal since it modifies $H(\omega)$ in a pointwise fashion to produce $G(\omega)H(\omega)$. Suppose, for the purpose of illustration, we relax our hypothesis and assume that the channel ensemble has energy only at two values of ω, ω_1 and ω_2. In this case, the operator $\mathcal{G}$ is particularly simple since the point $\mathbf{H} = (h_1, h_2)$ is mapped into the point $\mathbf{GH} = (g_1h_1, g_2h_2)$, where $h_i \equiv H(\omega_i)$ and $g_i \equiv G(\omega_i)$. The locus of points (g_1h_1, g_2h_2), subject to the average power constraint on the demodulator input $x(t)$

$$g_1^2k_1 + g_2^2k_2 = 1 \qquad (k_i \triangleq E\{|H(\omega_i)|^2S(\omega_i)\}), \tag{15}$$

describes the manner in which the equalizer redistributes the channel ensemble so as to minimize the average squared distance to the ray **1**.

By noting that the point (g_1h_1, g_2h_2) satisfies the relation

$$\frac{(g_1h_1)^2k_1}{(h_1)^2} + \frac{(g_2h_2)^2k_2}{(h_2)^2} = 1, \tag{16}$$

we see that the locus of the points (g_1h_1, g_2h_2) subject to (15) is an ellipse centered around the origin whose major and minor axes are parallel to the x–y axes and are of length $|h_1/k_1^{\frac{1}{2}}|$ and $|h_2/k_2^{\frac{1}{2}}|$. Thus, the effect of the equalizer on the channel array, as shown in Fig. 4, is to allow each equalized channel, $\mathbf{GH}_i = \{gh_1^{(i)}, gh_2^{(i)}\}$, to move on an elliptical surface. The nature of the compromise by which the optimum filter shape is chosen should now be clear. Each channel wants to move as close to the unit ray as possible—this will usually lead to conflicting requirements, i.e., as one channel is brought closer to the unit ray, other channels will move further away from this ray. For a multiplicity of channels there will be an "elliptical flow" along the respective ellipses which terminates when the best filter shape has been found.

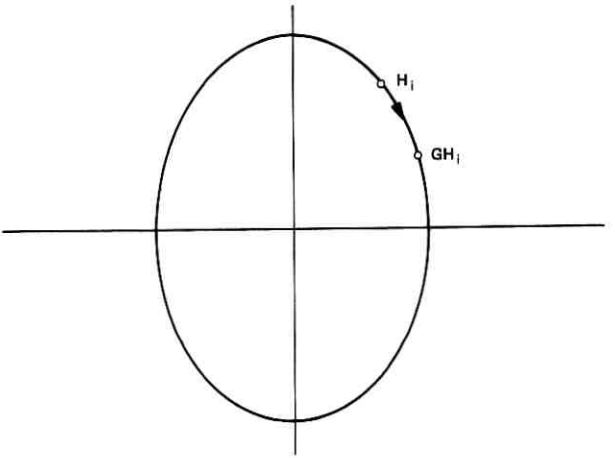

Fig. 4—Geometric interpretation of the effect of the compromise equalizer on the channel ensemble. A channel, represented by the point $\mathbf{H}_j$, is modified by the equalizer to give the equalized channel $\mathbf{GH}_j$.

The above interpretation would of course be valid in a Hilbert space of dimension large enough to accurately represent the function $H(\omega)$.

III. DETERMINING THE OPTIMUM EQUALIZER CHARACTERISTIC

3.1 *Analytic Solution*

Having developed a geometric representation as an aid to understanding the equalizer design problem, we are now in a position to explicitly determine the best filter shape. Returning to (9) and averaging with respect to the channel ensemble, we have

$$\langle \mathcal{E} \rangle_H = \int_{-\infty}^{\infty} |G(\omega)|^2 [S(\omega) H_{\text{rms}}^2(\omega) + N(\omega)] \frac{d\omega}{2\pi} - \frac{1}{\alpha} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(\omega) S(\nu) G^*(\omega) G(\nu) \langle H^*(\omega) H(\nu) \rangle \frac{d\omega}{2\pi} \frac{d\nu}{2\pi}, \quad (17)$$

where

$$H_{\text{rms}}(\omega) \equiv \sqrt{\langle |H(\omega)|^2 \rangle}. \quad (18)$$

If we let

$$Q(\omega) = \sqrt{S(\omega) H_{\text{rms}}^2(\omega) + N(\omega)}\, G(\omega) \quad (19)^\dagger$$

and introduce the weighted channel covariance kernel

$$K(\omega, \nu) = \frac{1}{\alpha} \frac{S(\omega) S(\nu) \langle H^*(\omega) H(\nu) \rangle}{\sqrt{[S(\omega) H_{\text{rms}}^2(\omega) + N(\omega)][S(\nu) H_{\text{rms}}^2(\nu) + N(\nu)]}}, \quad (20)$$

† Note that $|Q(\omega)|^2$ is the power spectral density of the equalizer output for the "rms channel."

then our cost function can be rewritten in the convenient form

$$\langle \mathcal{E} \rangle_H = \int_{-\infty}^{\infty} Q(\omega) Q^*(\omega) \frac{d\omega}{2\pi} - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(\omega, \nu) Q^*(\omega) Q(\nu) \frac{d\omega}{2\pi} \frac{d\nu}{2\pi}. \quad (21)$$

Introducing the Hermitian integral operator $\mathcal{K}$, whose kernel is $K(\omega, \nu)$, permits us to write the criterion as the quadratic form

$$\langle \mathcal{E} \rangle_H = (\mathbf{Q}, \mathbf{Q}) - (\mathcal{K}\mathbf{Q}, \mathbf{Q}), \quad (22)$$

where the inner product does not include the signal-spectrum weighting introduced in (10).† We now wish to minimize (22) with respect to Q, subject to an appropriate constraint. Referring to Fig. 1, we see that the average power present at the demodulator input is

$$P = \int_{-\infty}^{\infty} [S(\omega) H_{\text{rms}}^2(\omega) + N(\omega)] |G(\omega)|^2 \frac{d\omega}{2\pi}.$$

Thus, a natural constraint is to require that the power be constant. In terms of $Q(\omega)$, this constraint takes the form

$$(\mathbf{Q}, \mathbf{Q}) = P, \quad (23)$$

and the optimization problem consists of minimizing the positive definite form (22) subject to (23). The solution, which we denote by Q_{opt}, is easily obtained[3] by using a Lagrange multiplier and is recognized as the *principal eigenfunction*‡ of the operator $\mathcal{K}$. The best filter shape, $G_{\text{opt}}(\omega)$, is obtained from $Q_{\text{opt}}(\omega)$ by using (19), and can be regarded as the first term in a Karhunen-Loève representation[4] of the random process

$$\frac{S(\omega)}{\sqrt{S(\omega) H_{\text{rms}}^2(\omega) + N(\omega)}} H(\omega). \quad (24)$$

The residual value of the criterion, evaluated when $\mathbf{Q} = \mathbf{Q}_{\text{opt}}$, is given by

$$\langle \mathcal{E} \rangle_{\text{opt}} = (\mathbf{Q}_{\text{opt}}, \mathbf{Q}_{\text{opt}})[1 - \lambda] = (1 - \lambda) P, \quad (25)$$

where λ is the maximum eigenvalue and is a measure of how effectively the compromise equalizer performs for various channel ensembles and various signal and noise spectra. The degree to which the random process given by (24) is described by the first term in a Karhunen-Loève expansion will, of course, determine the equalizer performance.

† In the sequel, the inner product between the vector $\mathbf{U}$ and $\mathbf{V}$ will be taken to be $(\mathbf{U}, \mathbf{V}) = \int_{-\infty}^{\infty} U(\omega) V^*(\omega) (d\omega/2\pi)$.

‡ That is, the eigenfunction corresponding to the maximum eigenvalue (it is well known that the eigenvalues of a Hermitian operator are real).

In order to get more insight into the nature of $G_{opt}(\omega)$ we will, in the next paragraph, consider the form of the optimum filter under some special circumstances. This discussion, along with the examples treated in the next section, will reveal some properties of the optimum filter shape.

3.2 *Special Cases*

3.2.1 *One Channel*

Suppose the channel ensemble consists of only one member,† $H(\omega)$. The principal eigenfunction of $\mathcal{K}$ is easily determined to be

$$Q_{opt}(\omega) = \frac{S(\omega)}{\sqrt{S(\omega)\,|H(\omega)|^2 + N(\omega)}}\,H^*(\omega). \tag{26}$$

Thus, the best filter shape is

$$G_{opt}(\omega) = \frac{S(\omega)H^*(\omega)}{S(\omega)\,|H(\omega)|^2 + N(\omega)}. \tag{27}$$

The filter given by (27) is just the well-known Wiener filter for estimating a random signal that has been passed through a nonrandom channel and then corrupted by additive noise. The amplitude characteristic of the filter, which is given by

$$\frac{S(\omega)\,|H(\omega)|}{S(\omega)\,|H(\omega)|^2 + N(\omega)}, \tag{28}$$

provides *noise rejection* at those frequencies where $S(\omega)$ is small relative to $N(\omega)$. Thus, as the spectrum of the modulating signal is changed, the noise-rejecting regions of the filter will be altered. It is worth noting that, if the channel has only phase distortion, amplitude as well as phase compensation is required.‡ It should also be noted that the amplitude response of the filter will be greatly attenuated at those frequencies where $|H(\omega)|^2$ dominates $N(\omega)/S(\omega)$; this phenomenon, which we call *channel-variance rejection*, is observed in practice for an arbitrary channel ensemble at the frequencies where the variance of $H(\omega)$, defined by

$$\mathrm{Var}[H(\omega)] \equiv \langle |H(\omega)|^2 \rangle - |\langle H(\omega) \rangle|^2, \tag{29}$$

becomes large. Since the phase of $G_{opt}(\omega)$, for the simple case of one channel, is just the negative of the channel phase, we have, under the further specialization of vanishingly small noise power, the not-

† This assumption gives good insight when there is small dispersion about the average channel, i.e., the channels all look pretty much alike.

‡ This, along with the preceding sentence, provides a partial answer to the second and third questions posed in the introduction.

surprising result that the equalizer should invert the channel. On the other hand, as the noise becomes dominant, the solution is observed to approach a filter matched to the corresponding average channel characteristic and signal spectrum.

3.2.2 *Deterministic Amplitude Distortion, Random Delay Distortion, and No Noise*

Suppose the members of the channel ensemble can be written as

$$H(\omega) = a(\omega)e^{j\theta(\omega)}, \tag{30}$$

where the amplitude response $a(\omega)$ does not vary from channel to channel, and the phase response $\theta(\omega)$ is randomly selected.

We consider first the kernel $K(\ ,\ )$ when there is only delay distortion (i.e., $a(\omega) = 1$) and no noise. Since $H_{\text{rms}}(\omega)$ is unity, we have

$$K(\omega, \nu) = \frac{\sqrt{S(\omega)S(\nu)}}{\alpha} \langle e^{j\theta(\omega)-j\theta(\nu)} \rangle, \tag{31}$$

and the filter is given by

$$G_{\text{DELAY}}(\omega) = \frac{Q_{\text{opt}}(\omega)}{\sqrt{S(\omega)}}. \tag{32}$$

If we no longer restrict $a(\omega)$ to be unity, the kernel is still given by (31) while the filter is seen to be

$$G(\omega) = \frac{1}{a(\omega)} G_{\text{DELAY}}(\omega). \tag{33}$$

The above indicates that, in the presence of deterministic amplitude distortion and random phase distortion, the compensation is decoupled in the sense that the filter shape is a cascade of the compensation for the component distortions. While this sort of decoupling is generally not the case, we have found in the example described in the next section that the *phase* characteristic of the equalizer is rather insensitive to changes in amplitude distortion and noise level.

3.2.3 *Small Variation in the Amplitude and Phase Characteristics*

We now wish to give a suggestive description of the optimum filter when the channel characteristics have small variation about the average characteristic. To this end, let us denote a typical characteristic by

$$\begin{aligned} H(\omega) &= a(\omega)e^{j\theta(\omega)} \\ &= \bar{a}(\omega)b(\omega)e^{j[\bar{\theta}(\omega)+\phi(\omega)]}, \end{aligned} \tag{34}$$

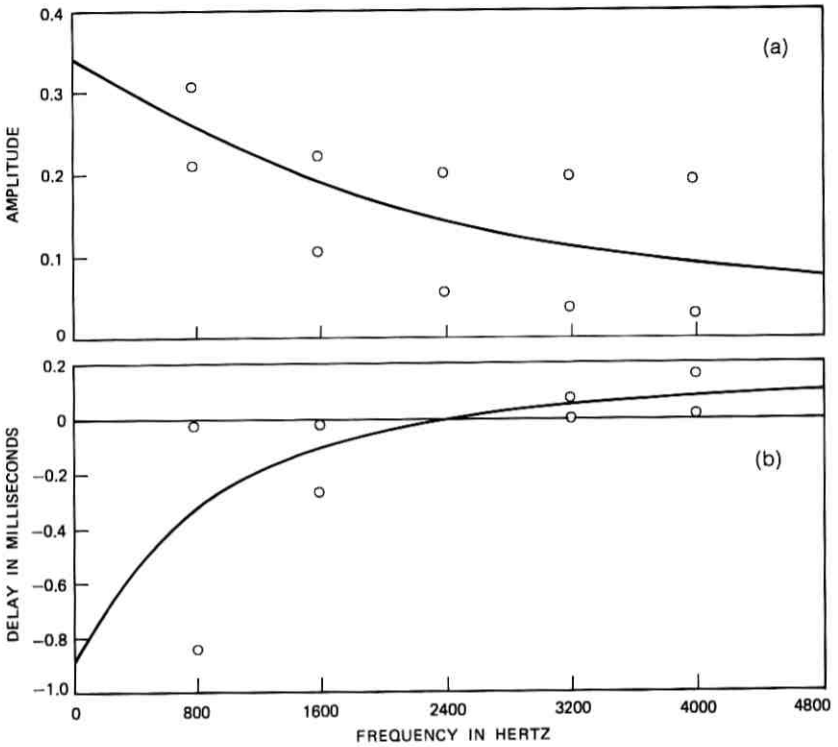

Fig. 5—(a) amplitude and (b) delay characteristics for average channel. The worst case amplitude and phase variation are indicated by the isolated circles.

where $\bar{a}(\omega)$ and $\bar{\theta}(\omega)$ are the average channel amplitude and phase, respectively, and $b(\omega)$ and $\phi(\omega)$ represent small (random) perturbations about these quantities. We first write $H(\omega)$ as

$$H(\omega) = \bar{a}(\omega)e^{j\bar{\theta}(\omega)}e^{\ell n\, b(\omega)+j\phi(\omega)} \tag{34a}$$

$$= \bar{H}(\omega)e^{q(\omega)}, \tag{34b}$$

where we have let

$$\bar{H}(\omega) = \bar{a}(\omega)e^{j\bar{\theta}(\omega)} \tag{35}$$

$$q(\omega) = \ell nb(\omega) + j\phi(\omega), \tag{36}$$

and we note that $\bar{H}(\omega)$ is composed of the average amplitude and phase of the ensemble. Since $|b(\omega)| \approx 1$ and $|\phi(\omega)| \approx 0$, we see that $|q(\omega)| \approx 0$; thus, retaining only the leading term in the series expansion of $e^{q(\omega)}$ gives

$$e^{q(\omega)} \approx 1 + q(\omega). \tag{37}$$

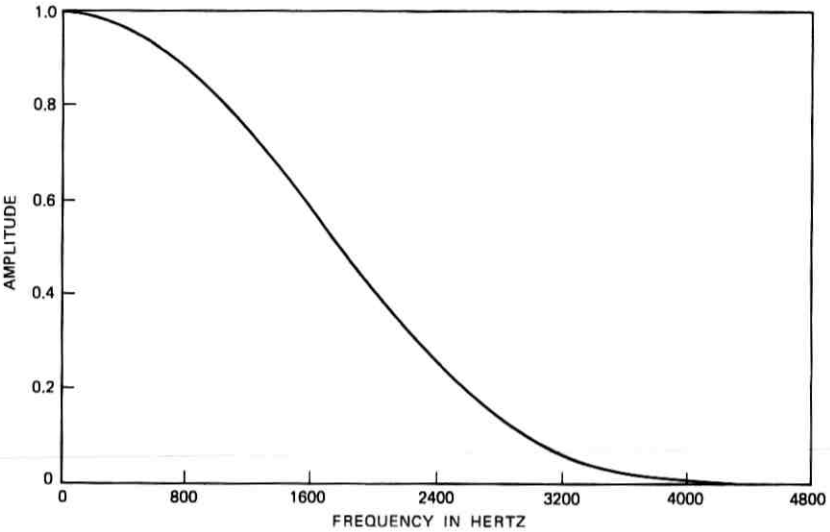

Fig. 6—Raised cosine power spectral density (rolloff = 1).

We are now in a position to evaluate the kernel $K(\omega, \nu)$ under the above assumptions. Using (37) we have

$$\begin{aligned}\langle H^*(\omega)H(\nu)\rangle &= \bar{H}^*(\omega)\bar{H}(\nu)\langle e^{q^*(\omega)}e^{q(\nu)}\rangle \\ &\approx \bar{H}^*(\omega)\bar{H}(\nu)\langle 1 + q^*(\omega) + q(\nu)\rangle, \end{aligned} \tag{38}$$

where we have kept only first-order terms in the expansion of $e^{q^*(\omega)}e^{q(\nu)}$. If we assume that the perturbations $\ell nb(\omega)$ and $\phi(\omega)$ have zero average value, then (38) separates, reducing to

$$\langle H^*(\omega)H(\nu)\rangle = \bar{H}^*(\omega)\bar{H}(\nu). \tag{39}$$

By a similar argument, the denominator of the kernel is found to be

$$\sqrt{S(\omega)\,|\,\bar{H}(\omega)\,|^2 + N(\omega)} \times \sqrt{S(\nu)\,|\,\bar{H}(\nu)\,|^2 + N(\nu)}, \tag{40}$$

and combining (39) and (40) gives the separable kernel

$$K(\omega, \nu) = \frac{S(\omega)S(\nu)\bar{H}^*(\omega)\bar{H}(\nu)}{\sqrt{S(\omega)\,|\,\bar{H}(\omega)\,|^2 + N(\omega)}\sqrt{S(\nu)\,|\,\bar{H}(\nu)\,|^2 + N(\nu)}}. \tag{41}$$

The equalizer shape is then given by

$$G_{\text{opt}}(\omega) = \frac{S(\omega)\bar{H}^*(\omega)}{S(\omega)\,|\,\bar{H}(\omega)\,|^2 + N(\omega)}, \tag{42}$$

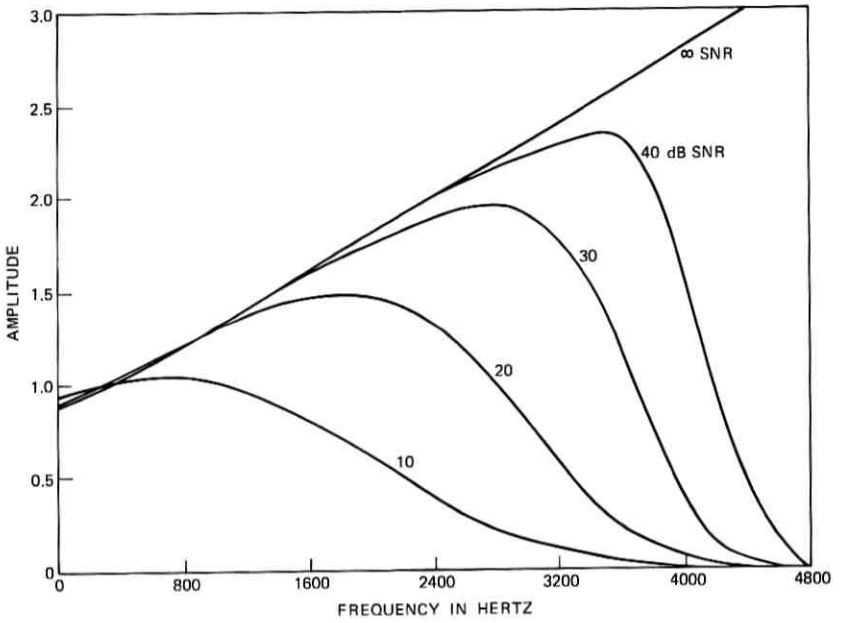

Fig. 7—Equalizer amplitude characteristics for various noise levels.

which is of the same form as the one-channel result [see (27)], as well as the filter obtained by Maurer and Franks (see Ref. 2). Clearly, for those frequencies where the noise is negligible (compared to the signal) the equalizer characteristic is $[1/\bar{a}(\omega)]e^{-j\bar{\theta}(\omega)}$, i.e., the best filter is one which *inverts* the average channel. Recalling the first question posed in the introduction, we can see that, apart from noise rejection (which occurs in the frequency range of small signal spectrum), the filter will invert the average channel when the variance of the ensemble is not appreciable. This is a useful rule-of-thumb for rapid (and approximate) compromise equalizer design.

IV. EXAMPLES USING THE 1964 CUSTOMER LOOP SURVEY

In this section we consider the design of a compromise equalizer for use over an ensemble of data customer loops. Some knowledge of the characteristics of this ensemble can be obtained from the 1964 Loop Survey.[5] This survey collected information about transmission parameters and channel makeup (e.g., gauges and lengths of sections which compromise the loop, locations of load coils and bridged taps, etc.) for

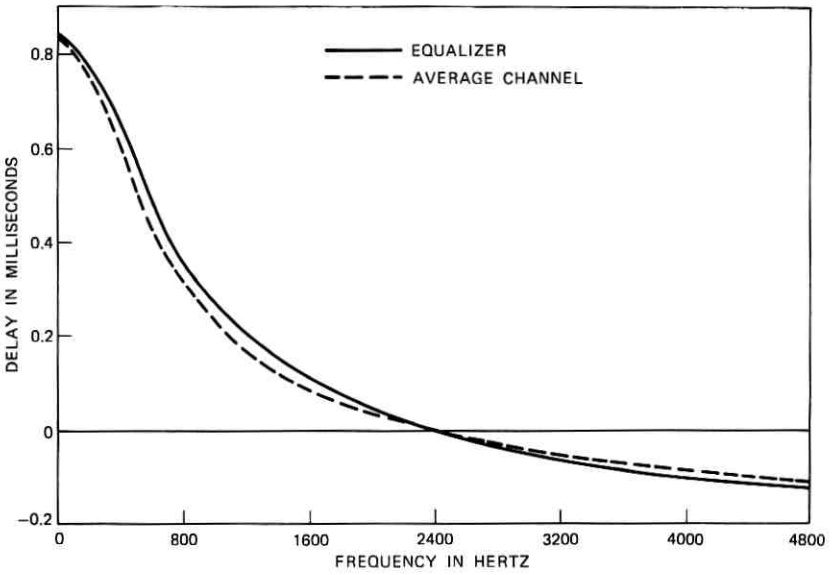

Fig. 8—Average channel delay (inverted) and equalizer delay (SNR = 30 dB).

a random selection of loops. For our purposes we extracted an ensemble of business loops (as opposed to residential loops) described in the survey. An existing computer program was used to obtain the frequency responses of the channels (with load coils and bridged taps removed) from the descriptions of the physical makeup of the channels. A recent study[6] shows that, at voiceband frequencies, the functions $\{e^{-\sqrt{i\omega\alpha}}\}_{\alpha>0}$ provide a good approximation of the loop transfer characteristics. The parameter α is a random variable (whose distribution can be approximated by the survey information), specifically $\alpha = RC\ell^2$, where R is the average series resistance per mile, C is the capacitance per mile, and ℓ is the loop length in miles. The average channel is depicted in Figs. 5a and 5b in terms of gain and delay.† These loop characteristics display appreciable amplitude variation (as well as delay variation) and the extent of these variations is indicated by the isolated circles.

To illustrate the design technique on these loop networks, we determined a compromise equalizer characteristic to be employed in a maxentropic 4.8-kb/s stream of randomly signed pulses. The basic

† The channel delays each include an estimate of B_{opt}.

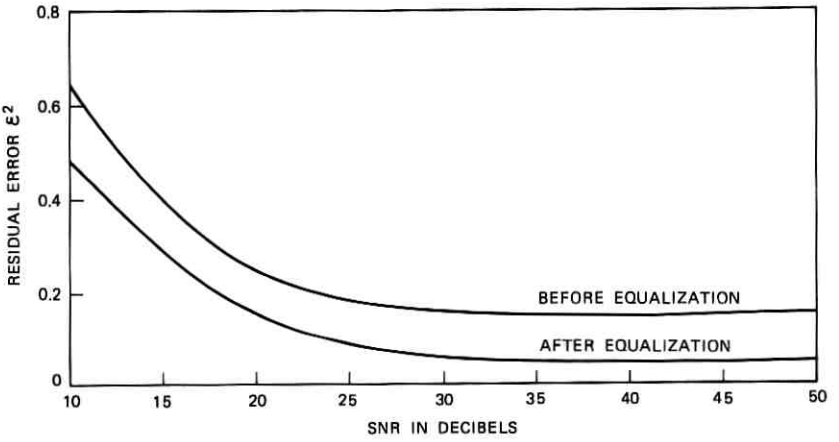

Fig. 9—Reduction of residual error as a result of equalization.

pulse was assumed to have a raised cosine spectrum, and the resulting power spectral density is shown in Fig. 6. It was assumed that only the "worst" 70 of the 143 customer loops required equalization. Figure 7 shows the resulting equalizer amplitudes for various noise levels. In this example, the channels were normalized to avoid the "whomper" effect mentioned in Section II. So when the signal-to-noise ratio is not infinite, the noise levels should be interpreted as if the lowest signal-to-noise ratio for the entire ensemble was assumed for all channels. Observe that, in the absence of noise, the equalizer tends to invert the average channel amplitude characteristic, and as the noise level is increased the equalizer tends to attenuate the high end of the spectrum where the signal power is lowest. Also note that, as the signal-to-noise ratio decreases, the solution tends to a filter matched to the raised cosine characteristic times a characteristic approximating the average channel. We observed that, for each of the noise levels, the compromise equalizer delay is close to the inverse delay of the average channel. The specific delay curve for the case of 30-dB SNR is provided in Fig. 8, which also shows the average channel delay (inverted). Finally, a direct computation shows that, for typical signal-to-noise ratios, the equalizer reduces the residual error by 4.8 dB, and Fig. 9 displays this improvement in mean-square error as a function of signal-to-noise ratio. The improvement is greatest at high SNR since the equalizer is combating only linear distortion rather than a combination of linear distortion and noise.

VI. CONCLUSIONS

Using the second-order statistics of the channel ensemble, as well as the signal and noise spectra, the equalizer characteristics were easily computed by solving a matrix eigenvalue problem. Several interesting conclusions can be drawn concerning the properties of the equalizer:

(*i*) The equalizer amplitude is attenuated in those frequency regions where the signal-to-noise ratio or signal-to-channel-variance ratio is small.

(*ii*) When the channel ensemble has only delay distortion, amplitude as well as phase compensation is required.

(*iii*) The delay characteristics of the equalizer are rather insensitive to changes in noise level or (non-random) amplitude distortion.

(*iv*) When the channel ensemble has small variation about the average characteristic, a situation that commonly arises in practice, then the equalizer will invert the average channel.

REFERENCES

1. Bennett, W. R., and Davey, J. R., *Data Transmission*, New York: McGraw-Hill, 1965.
2. Maurer, R. E., and Franks, L. E., "Optimal Linear Processing of Randomly Distorted Signals," IEEE Trans. Circuit Theory, *17*, No. 1 (February 1970).
3. Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, New York: Interscience Publishers, 1966, Chapter 3.
4. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
5. Gresh, P. A., "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3385.
6. Brophy, F., Foschini, G. J., and Smith, J. W., "Some Distortion Versus Speed Results for High-Speed Cable Transmission Systems," unpublished work.

The Bell System Technical Journal
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Distribution of $\sum a_n/n$, a_n Randomly Equal to ± 1

By S. O. RICE

(Manuscript received February 6, 1973)

When $a_1, a_2, \cdots$ are independent random variables, each equal to ± 1 with probability $\frac{1}{2}$, the sum $\sum_1^\infty a_n/n$ is a random variable whose distribution is difficult to determine theoretically. This sum is of interest in the study of intersymbol interference in digital communication systems. Here the distribution of the sum is computed by numerical integration and the results tabulated. Asymptotic expressions are given for the tails of the distribution.

I. INTRODUCTION

The distribution of the random variable

$$x = \sum_{n=1}^{\infty} a_n/n, \tag{1}$$

where $a_1, a_2, \cdots$ are independent random variables equal to $+1$ or -1 with probability $\frac{1}{2}$, is of some interest in the study of intersymbol interference in a digital communication system. For example, the sum of two independent expressions of the form (1) occurs when the pulse train $\sum_{-\infty}^{\infty} a_n \sin(t - n\pi)/(t - n\pi)$, a_n randomly equal to ± 1, is sampled at regularly spaced instants which are slightly out of step with the zeros of $\sin t$. The theory of random variables of type (1) (in particular with $a_n\beta^n$ in place of a_n/n) has been studied by a number of investigators. A survey of the field has been made recently by Hill and Blanco.[1] Here we evaluate the distribution of x numerically and give expressions for its behavior when x is large. Questions of continuity and convergence are put aside.

Since the distribution is even about $x = 0$, only values for $x \geqq 0$ need be considered. From the characteristic function

$$f(u) = \text{avg}\,[\exp(ixu)] = \prod_{n=1}^{\infty} [\cos(u/n)] \tag{2}$$

we get an expression for the probability density $p(x)$ of x:

$$p(x) = \frac{1}{\pi} \int_0^\infty \cos(xu) f(u) du, \tag{3}$$

$$\text{Prob}\ (x > x_1) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin(x_1 u)}{u} f(u) du. \tag{4}$$

The values of $p(x)$ and Prob $(x > x_1)$ shown in Table I were obtained by evaluating these integrals by the trapezoidal rule[2,3] which works well for (3) and (4).

The asymptotic expressions (8) and (18) for $p(x)$ follow from a saddle point analysis of (3). Both $p(x)$ and Prob $(x > x_1)$ decrease rapidly when x (or x_1) > 3, the decrease being dominated by the factor $\exp[-\exp(x - A)]$ where $A = 1.39 \cdots$.

The rapid decrease of $p(x)$ is interesting because the divergence of $\sum 1/n$ might lead one to expect that $p(x)$ would decrease slowly as $x \to \infty$. Instead, $p(x)$ actually decreases much faster than a Gaussian probability density. I am indebted to a referee for the observation that the second, fourth, and sixth moments of $p(x)$ are, respectively, $\pi^2/6$, $11\pi^4/180$, and $233\pi^6/7560$.

The reader may wonder why as many as six decimal places are given in Table I. There are several reasons. One is that the cost was low. About 3 seconds were required by a Honeywell 6000 Processor to compute the values shown in Table I, and about 40 terms were required in each trapezoidal sum. This illustrates the fact (apparently not well known) that when integrals like (3) and (4) are to be evaluated numerically, the trapezoidal rule often performs better than most of the other conventional quadrature methods (better than Simpson's rule, for instance).[2] The six-figure accuracy is also used to gain an idea of the values of x for which the asymptotic expansion (18) for $p(x)$ begins to be valid. This degree of accuracy also shows that $p(0)$ is equal to $0.249\ 994 \cdots$ and not to $\frac{1}{4}$, as might be inferred from a four-figure tabulation.

II. TRAPEZOIDAL RULE CALCULATION

Preliminary computations showed that $|f(u)| < 10^{-10}$ when $u > 15$. Furthermore, it was found that (3) and (4) could be evaluated to within the desired accuracy by using a trapezoidal-rule spacing of $\Delta u = h = 0.4$. In line with these values the trapezoidal sum was truncated at the 40th term ($15/0.4 \approx 40$).

TABLE I—VALUES OF $p(x)$ AND Prob $(x > x_1)$

x or x_1	$p(x)$	Prob $(x > x_1)$	x or x_1	$p(x)$	Prob $(x > x_1)$
0.0	0.249 994	0.500 000	2.0	0.125 000	0.056 599
0.2	0.249 970	0.450 003	2.2	0.091 768	0.034 949
0.4	0.249 802	0.400 021	2.4	0.061 647	0.019 683
0.6	0.249 073	0.350 118	2.6	0.037 148	0.009 912
0.8	0.246 778	0.300 494	2.8	0.019 592	0.004 357
1.0	0.241 222	0.251 623	3.0	0.008 777	0.001 623
1.2	0.230 408	0.204 357	3.2	0.003 222	0.000 494
1.4	0.212 852	0.159 912	3.4	0.000 927	0.000 118
1.6	0.188 353	0.119 683	3.6	0.000 198	0.000 021
1.8	0.158 232	0.084 949	3.8	0.000 029	0.000 003
2.0	0.125 000	0.056 599	4.0	0.000 003	0.000 000

The infinite product (2) for $f(u)$ was computed by using

$$f(u) = \left(\prod_{n=1}^{N-1} [\cos (u/n)]\right) \exp\left[\sum_{n=N}^{\infty} \ln \cos (u/n)\right], \tag{5}$$

where N is a large number such that $u/N \ll 1$ for all values of u used in the computation ($0 \leqq u \leqq 16$). The product $\prod_1^{N-1}$ in (5) was computed by straightforward multiplication. The sum in (5) was computed by setting $g(n) = \ln [\cos (u/n)]$ in the Euler-Maclaurin sum formula:

$$\sum_{n=N}^{\infty} g(n) = \int_N^{\infty} g(t)dt + \frac{1}{2} g(N) - \frac{B_2}{2!} g^{(1)}(N) - \frac{B_4}{4!} g^{(3)}(N) - \cdots - \frac{B_{2k}}{(2k)!} g^{(2k-1)}(N) + R_k. \tag{6}$$

Here the B's denote Bernoulli's numbers, $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = 1/42$ [we stopped at B_6 in our use of (6)], $g^{(k)}(N)$ denotes the value of $(d/dt)^k g(t)$ at $t = N$, and the remainder R_k is the integral of $g^{(2k+1)}(t)$ times the Bernoulli "polynomial" of degree $2k + 1$ and period 1 (see pages 520–540 of Ref. 4).

The integral in (6) can be evaluated to within the desired accuracy by setting $u/t = y$, $dt = -udy/y^2$, expanding ln (cos y) in powers of y with the help of

$$-\ln (\cos y) = \int_0^v \tan v dv = \int_0^v \left[v + \frac{v^3}{3} + \frac{2v^5}{15} + \frac{17v^7}{315} + \cdots\right] dv,$$

and integrating termwise:

$$\int_N^\infty g(t)dt = u\int_0^{u/N} y^{-2}\ln(\cos y)dy$$
$$= \frac{u^2}{2N} + \frac{u^4}{36N^3} + \frac{u^6}{225N^5} + \frac{17u^8}{17640N^7} + \cdots. \quad (7)$$

Expressions for the higher derivatives of $g(N)$ in (6) can be obtained by differentiating the series in (7) with respect to N.

In using (6) we stopped at $k = 3$ and neglected R_3.

Three separate trapezoidal-rule evaluations of $p(x)$ and Prob $(x > x_1)$ were made using eqs. (3) to (7) with $h = 0.4$, $N = 201$; $h = 0.38$, $N = 201$; and $h = 0.36$, $N = 301$, respectively. Here h is the spacing used in the trapezoidal-rule evaluations of (3) and (4). The three sets of computed values differed only in the 7th or 8th decimal places, i.e., all agreed with the values shown in the table. The values 201 and 301 of N are so large that terms beyond $k = 3$ in (6) and those shown in (7) are not needed. To check the computations, the integrals of $p(x)$ amd $x^2p(x)$ from $x = 0$ to $x = \infty$ (the upper limit used was actually $x = 5$) were computed by the trapezoidal rule with a spacing of $\Delta x = 0.1$. The trapezoidal values agreed with the known values, respectively $\frac{1}{2}$ and $\pi^2/12$, to within 6 significant figures or better.

III. DISCUSSION OF TABLE I

Table I shows that $p(x)$ remains nearly equal to $p(0) = 0.249994$ for $0 \leqq x \leqq 1$, passes through $p(2) = 0.125000$ (is it exactly $\frac{1}{8}$?), and then decreases rapidly. The question as to whether $p(2)$ is exactly $\frac{1}{8}$ remains unanswered, but $p(0) = 0.249994$ does not seem to be an erroneous calculation of $\frac{1}{4}$. For if $p_2(u)$ is the probability density of $u = \sum_2^\infty a_n/n$, then

$$p(x) = \tfrac{1}{2}p_2(x-1) + \tfrac{1}{2}p_2(x+1).$$

Setting x equal to 0 and 2 and combining the results give a result I owe to J. E. Mazo,

$$p(2) = \tfrac{1}{2}p(0) + \tfrac{1}{2}p_2(3) > \tfrac{1}{2}p(0).$$

Furthermore, replacing $\frac{1}{2}p_2(3)$ by $p(4) - \frac{1}{2}p_2(5)$ gives

$$p(2) = \tfrac{1}{2}p(0) + p(4) - \tfrac{1}{2}p_2(5)$$
$$0.125000 = 0.124997 + 0.000003 - \tfrac{1}{2}p_2(5)$$

which is satisfied by the tabulated values when $p_2(5)$ is assumed to be negligibly small.

IV. ASYMPTOTIC EXPRESSIONS FOR LARGE x

We shall show that the rapid decrease of $p(x)$ for $x > 3$ is described by

$$\begin{aligned} p(x) &\sim (y_0/\pi)^{\frac{1}{2}} e^{-y_0}, \\ y_0 &= \exp [x - 2\gamma + \ln (\pi/4)] = \exp [x - 1.39599 \cdots], \end{aligned} \tag{8}$$

where γ denotes Euler's constant, 0.577215 $\cdots$. Integrating (8) gives

$$\text{Prob } (x > x_1) \sim \text{erfc } (y_0^{\frac{1}{2}}) \sim p(x_1)/y_0, \tag{9}$$

where y_0 is computed from the second of equations (8) with x_1 in place of x. Bounds for the distribution involving exponential functions of e^x have been obtained by L. A. Shepp in unpublished work.

To obtain (8) we rewrite the integral (3) for $p(x)$ as

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ixu} f(u) du. \tag{10}$$

As is often the case for such integrals, the asymptotic value of $p(x)$ is given by the contribution of a saddle point, $u_0 = iy_0$, lying far out on the positive imaginary u-axis (the path of integration being deformed so as to pass through the saddle point). For $u = iy$ the integrand in (10) becomes

$$\exp [-xy + \varphi(y)], \qquad \varphi(y) = \sum_{n=1}^{\infty} \ln [\cosh (y/n)]. \tag{11}$$

When y is large we can show that

$$\begin{aligned} \varphi(y) &= y \ln y - y + Ay + \tfrac{1}{2} \ln 2 + r(y), \\ A &= 2\gamma - \ln (\pi/4) = 1.39599 \cdots, \end{aligned} \tag{12}$$

where $r(y)$ has roughly the same magnitude as $\exp (-2y)$.

We first outline the derivation of the expression (12) for $\varphi(y)$, and then apply (12) to obtain the asymptotic expression (8) for $p(x)$.

The derivation of (12) is based upon the Euler-Maclaurin sum formula (6) with $N = 1$ and $g(t) = \ln [\cosh (y/t)]$. The integral in the sum formula is

$$\begin{aligned} &\int_1^{\infty} \ln [\cosh (y/t)] dt \\ &= y \int_0^{y} v^{-2} \ln (\cosh v) dv \\ &= -\ln (\cosh y) + y(\ln y) \tanh y - y \int_0^{y} (\ln v) \operatorname{sech}^2 v dv \\ &= \ln 2 - y + y \ln y - y \int_0^{\infty} (\ln v) \operatorname{sech}^2 v dv + 0[ye^{-2y} \ln y], \end{aligned} \tag{13}$$

where we have integrated by parts twice. The last integral in (13) has the value

$$\int_0^\infty (\ln v)\,\operatorname{sech}^2 v\,dv = -\gamma + \ln(\pi/4) \tag{14}$$

which can be obtained by (*i*) replacing $\ln v$ in (14) by $v^{\mu-1}$, (*ii*) differentiating the known value (formula 3.527-3, page 352 of Ref. 5) of the resulting integral with respect to μ, and (*iii*) setting $\mu = 1$. The derivative of $g(t)$ with respect to t, $g^{(1)}(t)$, in the sum formula (6) is

$$\begin{aligned} g^{(1)}(t) &= -yt^{-2}\tanh(y/t) \\ &= -yt^{-2}(1 - 2e^{-2y/t} + 2e^{-4y/t} - \cdots), \end{aligned}$$

where the exponential terms become negligible when y becomes large and $t = 1$. In general, for $l = 0, 1, 2, \cdots$, $g^{(2l+1)}(1)$ is equal to $-(2l+1)!\,y$ plus negligible terms. Therefore, the right side of the sum formula (6) is the integral plus

$$\tfrac{1}{2}(y - \ln 2) + \frac{B_2}{2}y + \frac{B_4}{4}y + \cdots + yR_k' \tag{15}$$

plus terms which are negligible when y is large. The sum of the coefficients of y in (15) is known to be equal to γ (page 529 of Ref. 4). Hence (15) is equal to $\gamma y - \frac{1}{2}\ln 2$. Addition of (13) and (15) and use of (14) gives the expression (12) for $\varphi(y)$.

Next we use the expression for $\varphi(y)$, with the small term $r(y)$ neglected, to obtain the asymptotic form of $p(x)$. The saddle point of interest occurs at $u_0 = iy_0$ where y_0 is the zero of the derivative of the exponent in (11). The exponent is $-xy + \varphi(y)$, and y_0 is the zero of

$$-x + \varphi'(y) = -x + \ln y + A.$$

Thus $y_0 = \exp(x - A)$. This y_0 is the same as the y_0 appearing in the asymptotic expression for $p(x)$ stated in (8). The exponent itself has the value $-xy_0 + \varphi(y_0) = -y_0 + \frac{1}{2}\ln 2$ at y_0. By making use of $\varphi''(y) = 1/y$ and the higher derivatives of $\varphi(y)$, the exponent can be expanded in a Taylor series about y_0. From this expansion it follows that near u_0 the integrand in the integral (10) for $p(x)$ can be written as

$$\exp\left[-y_0 + \frac{1}{2}\ln 2 + y_0\sum_{k=2}^{\infty}\frac{(iz/y_0)^k}{(k-1)k}\right], \tag{16}$$

where $z = u - u_0$. Setting (16) in (10), changing the variable of integration from u to $\tau = z/y_0 = (u - u_0)/y_0$, and assuming that $p(x)$ is given asymptotically (as x and y_0 tend to ∞) by the contribution of

the saddle point at u_0 give

$$p(x) \sim (2\pi)^{-1} 2^{\frac{1}{2}} y_0 e^{-y_0} \int \exp\left[y_0 \sum_{k=2}^{\infty} \frac{(i\tau)^k}{(k-1)k} \right] d\tau. \tag{17}$$

Here the nominal path of integration is the real τ-axis. The classical saddle point asymptotic expansion obtained from (17) is

$$p(x) \sim (y_0/\pi)^{\frac{1}{2}} e^{-y_0} \left[1 + \frac{1}{24y_0} - \frac{23}{1152y_0^2} + \cdots \right]. \tag{18}$$

The coefficients of the powers of $1/y_0$ in the series can be determined by a general procedure described in Appendix D, page 1999, of Ref. 6.

The asymptotic expression stated in (8) is the leading term in (18). An idea of the accuracy of the asymptotic expressions can be obtained by considering the case $x = 3$. For $x = 3$, y_0 is 4.973 and Table I gives the "exact" value $p(3) = 0.008777$. The asymptotic value of $p(3)$ obtained from (8) is 0.008710, the first two terms in (18) give 0.008783, and the first three terms give 0.008776. Table I gives the "exact" value Prob $(x > 3) = 0.001623$ and eq. (9), namely Prob $(x > 3) \sim \text{erfc}\,(y_0^{\frac{1}{2}}) = \text{erfc}\,(2.230)$, gives 0.001612.

REFERENCES

1. Hill, F. S., Jr., and Blanco, M. A., "Random Geometric Series and Intersymbol Interference," accepted for publication in IEEE Trans. Information Theory.
2. Rice, S. O., "Efficient Evaluation of Integrals of Analytic Functions by the Trapezoidal Rule," B.S.T.J., *52*, No. 5 (May–June 1973), pp. 707–722.
3. Kendall, D. G., "A Summation Formula Associated With Finite Trigonometric Integrals," Quart. J. Math. (Oxford Ser.), *13* (1942), pp. 172–184.
4. Knopp, K., *Theory and Application of Infinite Series*, London: Blackie and Son, 1928.
5. Gradshteyn, I. S., and Ryzhik, I. W., *Tables of Integrals, Series, and Products*, New York and London: Academic Press, 1965.
6. Rice, S. O., "Uniform Asymptotic Expansions for Saddle Point Integrals—Application to a Probability Distribution Occurring in Noise Theory," B.S.T.J., *47*, No. 9 (November 1968), pp. 1971–2013.

The Bell System Technical Journal
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Adaptive Quantization in Differential PCM Coding of Speech

By P. CUMMISKEY, N. S. JAYANT, and J. L. FLANAGAN

(Manuscript received March 12, 1973)

We describe an adaptive differential PCM (ADPCM) coder which makes instantaneous exponential changes of quantizer step-size. The coder includes a simple first-order predictor and a time-invariant, minimally complex adaptation strategy. Step-size multipliers depend only on the most recent quantizer output, and input signals of unknown variance can be accommodated. We derive appropriate multiplier values from computer simulations with speech signals and with Gauss-Markov inputs. We compare performance of the ADPCM coder with conventional log-PCM, using both objective and subjective criteria. Finally, we describe an economical integrated hardware implementation of the ADPCM coder. We believe that at bit rates of 24 to 32 kb/s, ADPCM provides a robust and efficient technique for speech communication and for digital storage of speech.

I. INTRODUCTION

The advantages of coding speech digitally are well known.[1] Expected benefits include low costs per line, ease of maintenance, and high-quality signal regeneration at repeaters. Furthermore, digital coding is well matched to current technology in terms of readily available integrated circuit hardware. Results from speech-coding research are now beginning to specify techniques that are nearly optimal for a given bit rate, a given channel quality, and a given degree of coder complexity. Finally, the subject of direct digital conversion between alternative code formats is being widely studied, and simple techniques have already been proposed for some specific conversions.

The coder discussed in this paper is believed to be efficient and robust for speech coding at bit rates of 24 to 32 kilobits/second (kb/s). Other refinements of differential PCM (DPCM)[2] are based, at least in part, on adaptive prediction.[3–6] These techniques offer considerable potential for bandwidth compression,[3] but are typically hard to

implement. Therefore, for the type of bit rates mentioned earlier, it seems much more reasonable to tap the advantages of a more simply implemented adaptive quantizer.

Our adaptive DPCM (ADPCM) coder, therefore, operates on the basis of a fixed, first-order predictor in the DPCM loop, and a time-invariant, adaptation strategy for instantaneous changes of quantizer step-size. The technique has obvious advantages over conventional PCM (due to redundancy removal) and over conventional DPCM (due to increased dynamic range). Further, the quality of speech reproduction in the 24- to 32-kb/s range is believed to be perceptually better than that provided by adaptive delta modulation (ADM) which, however, has the advantage of even greater simplicity.[7,8]

Besides digital telephone applications, appropriate utilizations of ADPCM coding are seen in computer storage of digital speech (for voice answer-back, "voice-wiring," and similar functions), in mobile radio telephony, and in special applications such as deep-space communication and digital encryption.

II. DEFINITION OF THE ADPCM CODER: ADAPTIVE QUANTIZATION WITH A ONE-WORD MEMORY

A schematic block diagram of the coder appears in Fig. 1. It follows the conventional differential PCM structure with a first-order, fixed

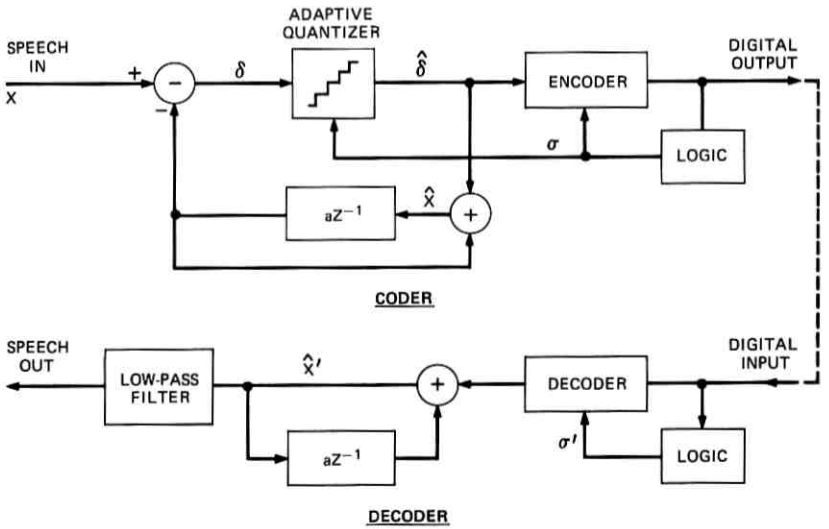

Fig. 1—Block diagram of ADPCM coder.

predictor in the feedback loop.[2] It has, however, the additional box labeled LOGIC, which provides adaptation of quantizer step-size on the basis of the most recent quantizer output. In the absence of channel errors, the step-size controls σ and σ' are identical, and so are the signal estimates $\hat{x}$ and $\hat{x}'$.

Step-size adaptations are motivated by the assumption that the variance of the quantizer input δ is unknown. The empirical adaptation rule is that, for every new input sample, the step-size is changed by a factor depending only on the knowledge of which quantizer slot was occupied by the previous signal sample.

Formally, if the outputs of a uniform B-bit quantizer are of the form

$$Y_u = P_u \frac{\Delta_u}{2}; \qquad \pm P_u = 1, 3, \cdots 2^B - 1; \qquad \Delta_u > 0,$$

the step-size Δ_r is given by the previous step-size multiplied by a time-invariant function of the code-word magnitude $|P_{r-1}|$;

$$\Delta_r = \Delta_{r-1} \cdot M(|P_{r-1}|),$$

subject, of course, to maximum and minimum limits on Δ_r, as specified in specific implementations. Step-size multipliers for a 3-bit uniform quantizer are illustrated in Fig. 2. Note that there are only four distinct

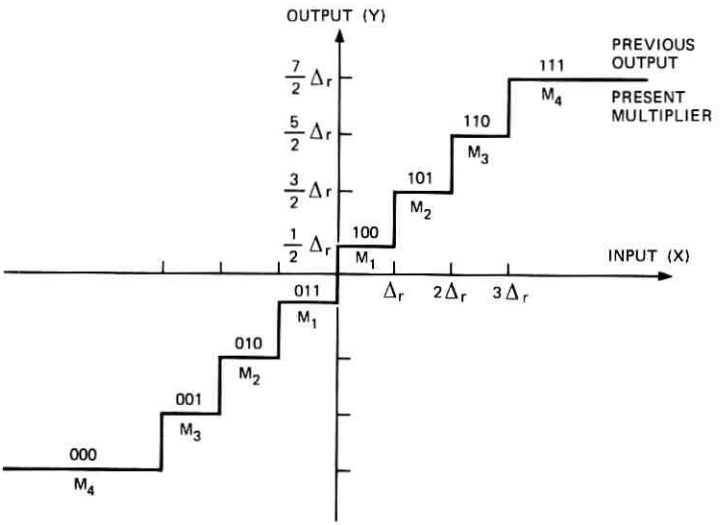

Fig. 2—Adaptation multipliers associated with quantizer levels for 3-bit ADPCM coder.

TABLE I—ADAPTATION MULTIPLIERS FOR SPEECH AND FOR GAUSS-MARKOV INPUTS (3-BIT CODER)

Previous Output Word	Multiplier	Speech Input	Gauss-Markov Input (Correlation Between Adjacent Samples = 0.5)
111 or 000	M_4	2	1.75
110 or 001	M_3	5/4	1.25
101 or 010	M_2	7/8	0.90
100 or 011	M_1	7/8	0.90

multipliers because the polarity of quantizer output is not utilized in the adaptation logic. Furthermore, meaningful adaptation requires that the step-size be increased on the detection of quantizer overload ($M_4 > 1$) and decreased during underload ($M_1 < 1$), and that $M_1 \leqq M_2 \leqq M_3 \leqq M_4$. Derivation of specific multiplier values is outlined in the next section.

III. DESIGN OF STEP-SIZE MULTIPLIERS

Two conflicting requirements are encountered in designing step-size multipliers. The first is the need to respond quickly to abrupt changes of input variance (suggesting the use of $M_4 \gg 1$, $M_1 \ll 1$ for the 3-bit example). The second requirement is the prevention of excessive step-size alterations in a stationary or steady-state situation (suggesting the use of $M_4 = 1 + \epsilon_4$; $M_1 = 1 - \epsilon_1$; $\epsilon_4, \epsilon_1 \to 0$). Compromise values of multipliers are therefore suggested for an input signal, or for a class of input signals.

Extensive computer simulations were carried out to determine the most desirable multiplier values for an illustrative speech sample. The sample was a male utterance of "This circuit operates on the same principle as N. S. Jayant's simulation." The speech was bandpass-filtered (200–3200 Hz), and was sampled at 8 kHz. Multiplier values were sought that maximized the signal-to-quantization-error (power) ratio (SNR), as averaged over the entire duration of the above utterance. Rounded values of these multipliers are shown in Table I for a 3-bit quantizer. Also shown are the values found to be desirable for the quantization of a Gauss-Markov input with an input signal correlation similar to that expected for Nyquist-sampled speech.[2] The similarity is interesting, particularly because the speech quantizer had a Max nonuniformity[9] (to take into account the observed Gaussian tendencies of the quantizer input), while the Gauss-Markov simulation utilized a uniform quantizer. The latter simulation also showed that desirable multiplier values are only slightly dependent on signal correlation,

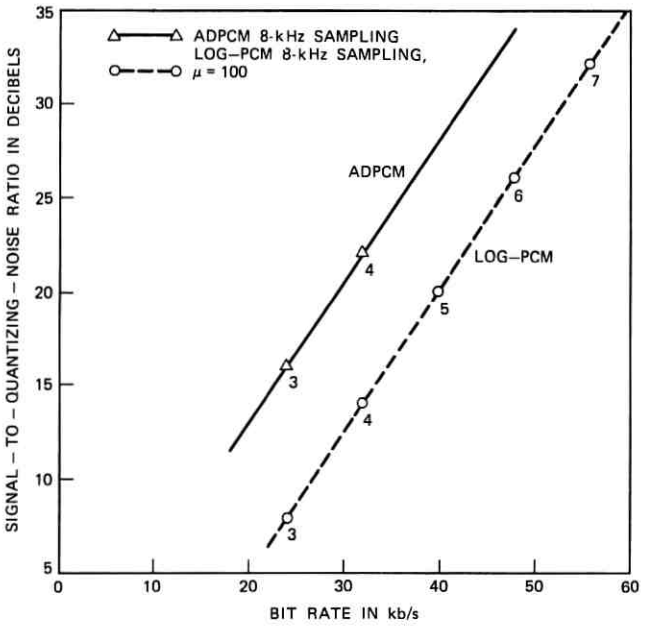

Fig. 3—Signal-to-quantizing-noise ratios for speech signals.

suggesting a robust adaptation strategy. Finally, the similarity of the multiplier values found for speech and for Gauss-Markov inputs suggests that the coder has a versatility that might extend to facsimile and video signals.[10]

The general problem of determining most desirable multiplier values is discussed at length in a companion paper.[10] That paper also points out the possibility of near-optimal adaptation strategies that have nontrivial ($\neq 1$) values only for the end multipliers (M_1 and M_4 in the 3-bit example), and compares our adaptation logic with that of Stroh.[4]

IV. PERFORMANCE COMPARISONS OF THE ADPCM CODER WITH CONVENTIONAL PCM

4.1 *SNR Data*

Computer simulations using speech input showed the signal-to-error-power ratio to be 16 dB for a 3-bit (24 kb/s) ADPCM coder which has the multipliers of Table I and a maximum step-size which was $D = 128$ times the minimum. The SNR of 16 dB represents an 8-dB gain over 3-bit logarithmic PCM with $\mu = 100$.[11]* It turns out that

* This value was chosen on the basis of past experience with speech coders. Present trends are toward a higher value for μ.

TABLE II—COMPARISON OF OBJECTIVE AND SUBJECTIVE PERFORMANCE OF ADPCM AND LOG-PCM

Objective Rating (SNR)	Subjective Rating (Preference)	
7-bit PCM	7-bit PCM	(High)
6-bit PCM	4-bit ADPCM	
4-bit ADPCM	6-bit PCM	
5-bit PCM	3-bit ADPCM	
3-bit ADPCM	5-bit PCM	
4-bit PCM	4-bit PCM	(Low)

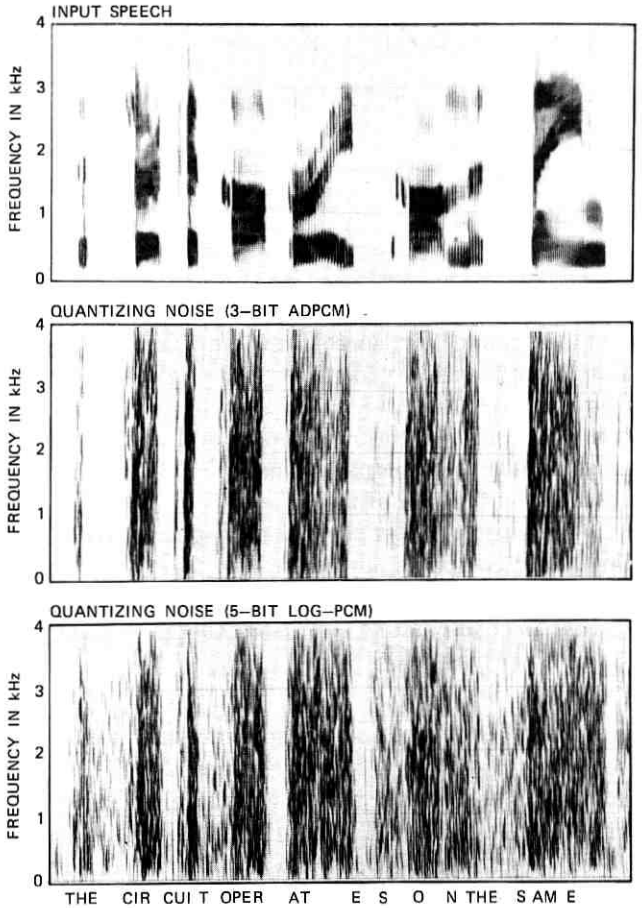

Fig. 4—Spectograms of speech and quantization error.

this improvement includes a 4-dB gain due to differential encoding and a 4-dB gain due to the quantizer adaptation. Figure 3 gives a more complete comparison of speech signal SNR's measured for the ADPCM and for log-PCM.

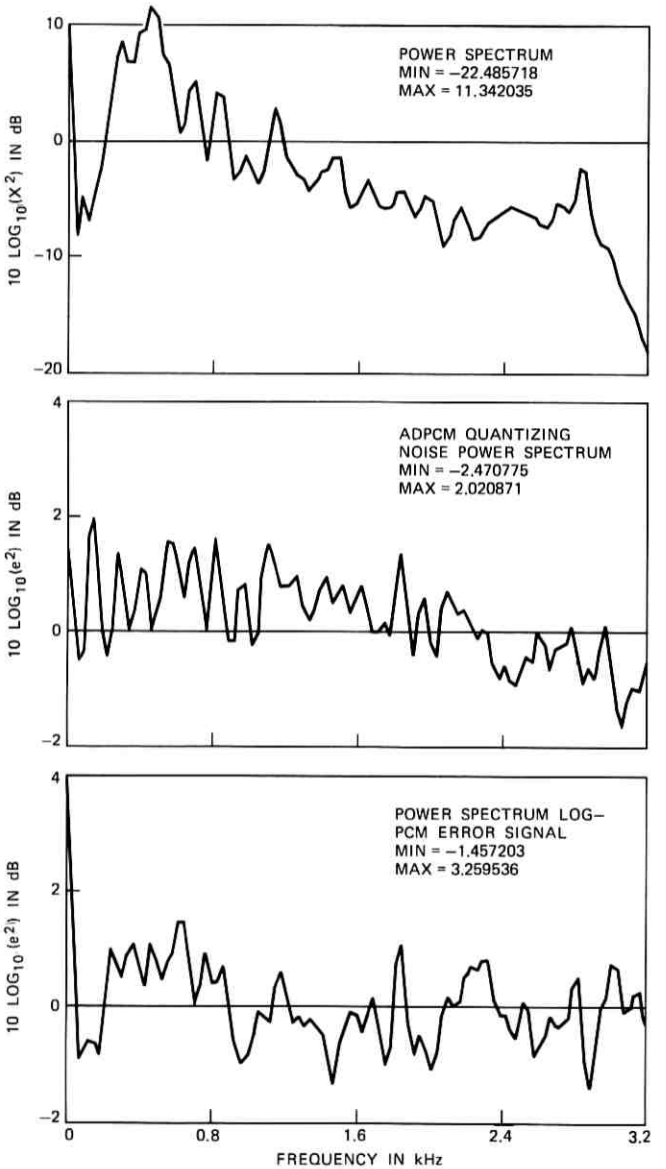

Fig. 5—Long-term error spectra.

4.2 *Error Spectrograms and Long-Term Spectra*

Figure 4 displays spectrograms of a section of the input speech and the associated quantizing noise spectrograms with 3-bit ADPCM and 5-bit log-PCM. Note that ADPCM provides considerably less noise during the silent intervals in speech, although the total noise power is greater in this coder (Fig. 3). This suggests that adaptive quantization in ADPCM provides greater dynamic range than the logarithmic compandor used in PCM. (The observation results, no doubt, from the specific numerical values $\mu = 100$ and $D = 128$ in Section 4.1. However, these values are believed to be very representative.)

Figure 5 illustrates another interesting difference between the quantizing noise in ADPCM and that in PCM. Note the high-frequency rolloff in ADPCM noise and the relative whiteness of the noise spectrum in PCM.

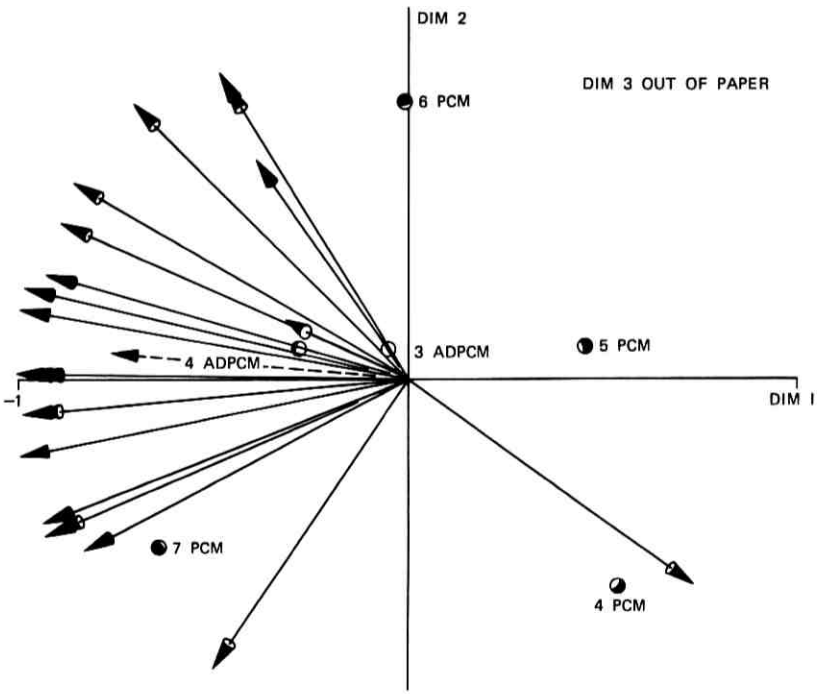

Fig. 6—Subjective preference judgments of various ADPCM and log-PCM codings. Dimensions 1 and 2 account for most of intersubject variance. Increasing preference is in the $-x$ direction. Individual subject vectors are plotted, and projection of coding conditions onto a subject vector indicates how that individual rank-ordered the coding systems.

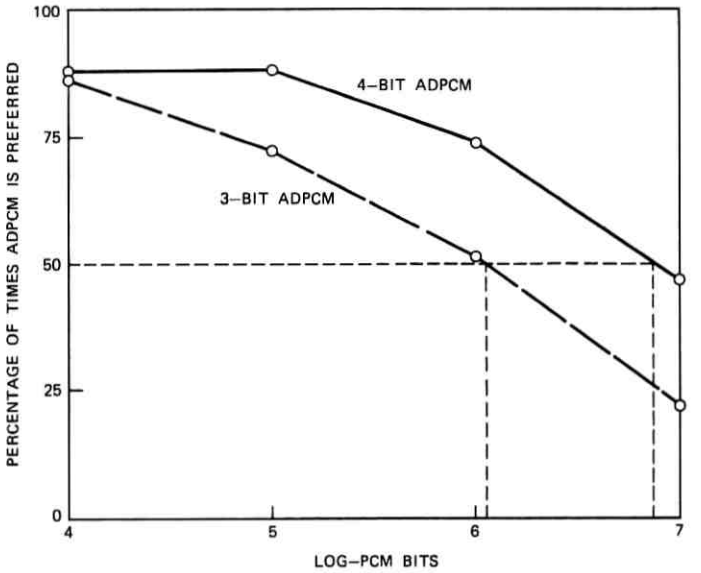

Fig. 7—Alternative interpretation of (ADPCM-versus-PCM) preference scores.

4.3 *Subjective Tests*

Apart from the measured differences mentioned in Section 4.2, informal listening tests indicated that ADPCM noise had a perceptually more palatable character* than PCM noise of equal variance; in other words, the quality of speech reproduction from the ADPCM (relative to PCM) was much better than was suggested by the SNR comparisons in Fig. 3. This observation was borne out in the following perceptual experiment.

The experiment involved 3- and 4-bit ADPCM stimuli and 4-, 5-, 6-, and 7-bit log-PCM stimuli. The total number of cross comparisons possible was 16 (2 stimuli × 4 stimuli × 2 orders of presentation). Twenty-two listeners participated in the tests and made preference judgments of signal quality for each of the 16 A-B comparisons. The preference judgments were submitted to a multidimensional scaling program,[12] and the results were plotted in terms of two subjective dimensions which accounted for most of the perceived differences. Dimension 1, in particular, accounted for 75 percent of the variance in the preference data.

The results are shown in Fig. 6. Individual subject vectors are

* Related, perhaps, to a lesser proportion of the noise getting into the idle circuit; and, also, due to some correlations of ADPCM noise with pitch information.

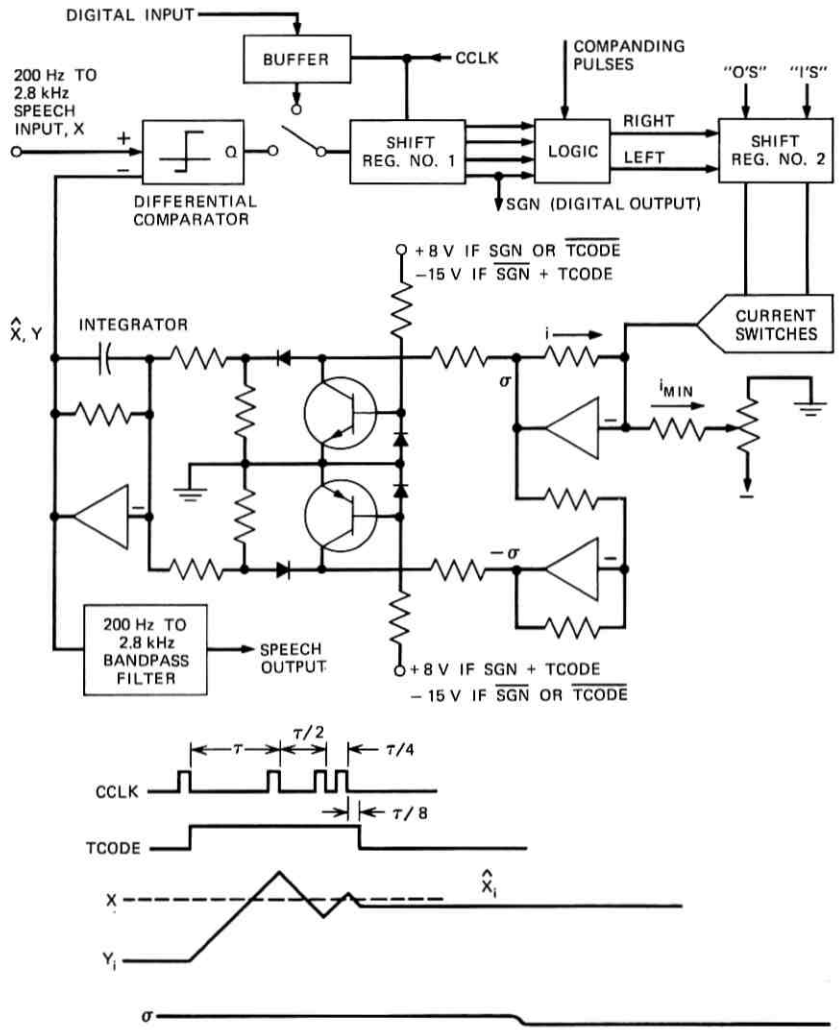

Fig. 8—Circuit block diagram for the hardware ADPCM coder.

displayed (solid lines), and projection of the coding conditions into a subject's vector reveals how that individual rank-ordered the signal qualities.* A resultant of the subject vectors is also shown (dashed), and projections onto this resultant indicate subject consensus in rank-

* One subject (vector in quadrant IV) apparently misunderstood the test instructions and gave essentially complementary preference judgments.

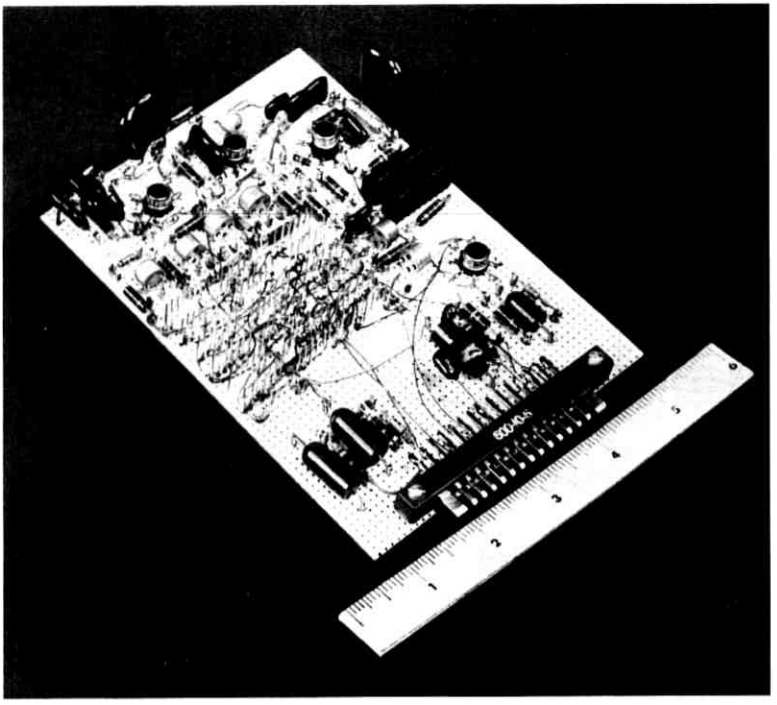

Fig. 9—ADPCM coder.

ing the qualities. This averaged subjective ranking can be compared with the objective SNR performance for the same codings (using the data of Fig. 3).

A comparison of the objective (SNR) performance and the subjective (preference) ranking of the codings tested is shown in Table II. It is clear from these data that subjectively the ADPCM does an even better job than the objective SNR's indicate. Arrows indicate two subjective "promotions" of the ADPCM. Table II (and, of course, Fig. 6) show, too, that 4-bit ADPCM is perceptually better than 6-bit log-PCM.

Figure 7 provides yet another means of comparing ADPCM and PCM using the original preference scores from the perceptual test. Ordinates in Fig. 7 are overall percentages (including all listeners) of A-B judgments where an ADPCM stimulus was preferred to a certain PCM stimulus as shown on the x-axis of Fig. 7. The 50-percent probability-of-preference line intersects the curves at points whose ab-

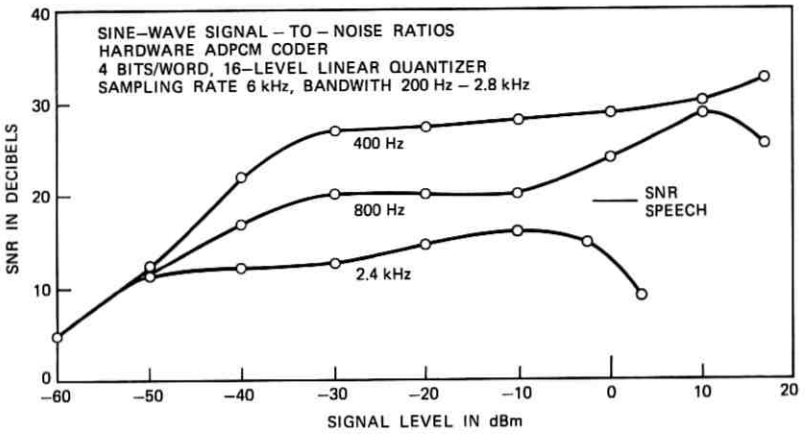

Fig. 10—Signal-to-quantizing-noise ratios for sine-wave input to the hardware ADPCM coder.

scissas represent quantitative log-PCM equivalences for 3- and 4-bit ADPCM.

V. HARDWARE DESIGN OF ADPCM CODER

The computer simulations described above established the design criteria for the ADPCM. To assess hardward viability of the technique, we constructed a 4-bit ADPCM coder in integrated circuit hardware. A circuit block diagram of the hardware coder is given in Fig. 8, and a photograph of the circuit card is shown in Fig. 9. State-of-the-art circuit technology is used and all circuit components are "off the shelf."

The circuit incorporates a uniform quantizer realized by using a serial logic (shown at the bottom of Fig. 8) to code the difference between the input X and the signal stored on the integrator Y. The logic provides four consecutive increments of integrator voltage within a duration much smaller than the sampling period. Each of these increments follows the latest sign of $(X - Y)$ and has a magnitude that is one-half that of the previous increment in the cycle.

Step-size adaptations are controlled by the current switches which provide a dictionary of 21 step-sizes. These are spaced with a ratio of $2^{\frac{1}{4}}$ between adjacent steps, and therefore provide an overall step-size range of 128:1. The step-size multipliers for speech (Table I) are approximated in the circuit as positive and negative exponents of $2^{\frac{1}{4}}$.

Measurements on the hardware realization confirm the computer observations on SNR for speech signals. Also, SNR's for sine-wave

inputs are conventionally used to assess digital coders (although sine-wave performance can be very deceptive in terms of perceptual acceptability). Figure 10 shows SNR measured on the hardware coder for sine-wave input. The 800-Hz behavior is reasonably consistent with the SNR measured for speech input.

VI. DIRECT DIGITAL CONVERSION BETWEEN ADPCM AND OTHER SIGNAL FORMATS

An important issue in the compatibility of digital systems is the provision of graceful and virtually transparent conversions among different code formats.[13] Digital techniques for directly converting between ADPCM and the conventional formats of DPCM and PCM have been proposed and are being studied.[14] One of the indications of these studies is that direct conversion between ADPCM and DPCM is quite feasible, especially when the conversion incorporates an intermediate stage of PCM.

VII. CONCLUSION

Results of this study indicate that the ADPCM technique leads to an economic, efficient digital coding of speech for the bit-rate range 24 to 32 kb/s. This range constitutes a channel capacity saving of over 2:1 compared to conventional PCM and produces a signal coding of comparable quality. Hardware implementation is relatively straightforward and noncritical.

Studies presently in progress are examining ADPCM coding for operation at 18 kb/s. Preliminary indications are that signal quality attractive for mobile radio application can be achieved at this low bit rate. This low rate also makes digital encryption for privacy attractive in mobile telephone.

Although not specifically discussed in this exposition, ADPCM proves reasonably robust in the presence of errors in the transmit channel. The computer simulations described above incorporated preliminary studies of error vulnerability which show the coder to perform well for channels with error probability $\leqq 10^{-4}$. Typical error rates in "clean" PCM channels are routinely maintained lower than this.

Further study of ADPCM is anticipated in objective analysis of its quantizing characteristics. This should be coupled with more complete perceptual tests to better understand the "perceptual palatability" of ADPCM coding. Further, a close competitor is adaptive delta modulation,[7,8] and subjective comparisons are planned that will include this coding technique.

One present utilization of the hardware ADPCM coder is in a computer voice response system for generating computer-spoken wiring instructions.[15] Speech coding at 24 kb/s provides economy of digital storage and simple A/D–D/A communication with the computer.

REFERENCES

1. Flanagan, J. L., "Focal Points in Speech Communication Research," IEEE Trans. Commun. Technology, December 1971 (Part I), pp. 1006–1015.
2. McDonald, R. A., "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems—Particular Applications to Voice Signals," B.S.T.J., *45*, No. 7 (September 1966), pp. 1123–1151.
3. Atal, B. S., and Schroeder, M. R., "Adaptive Predictive Coding of Speech Signals," B.S.T.J., *49*, No. 8 (October 1970), pp. 1973–1986.
4. Stroh, R. W., "Optimum and Adaptive Differential Pulse Code Modulation," Ph.D. Thesis, Polytechnic Institute of Brooklyn, 1970.
5. Noll, P. W., "Non-Adaptive and Adaptive DPCM of Speech Signals," Overdruk int Polytechnisch Tijdschrift, editie elektrotechnick-Elektronica, Nr. 19, 1972.
6. Cummiskey, P., "Adaptive Differential Pulse-Code Modulation for Speech Processing," Ph.D. Thesis, Newark College of Engineering, 1973.
7. Jayant, N. S., "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
8. Jayant, N. S., Cummiskey, P., and Flanagan, J. L., "Design and Implementation of an Adaptive Delta Modulator," Proc. IEEE Int. Conf. Speech Commun. and Processing, Boston, Massachusetts, April 1972.
9. Max, J., "Quantization for Minimum Distortion," IRE Trans. Information Theory, *IT-6*, March 1960, pp. 7–12.
10. Jayant, N. S., "Adaptive Quantization With a One-Word Memory," B.S.T.J., this issue, pp. 1119–1144.
11. Smith, B., "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 653–709.
12. Carroll, J. D., "Individual Differences and Multidimensional Scaling," in R. N. Shepard, A. K. Romney, and S. Nerlove (eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, vol. I, New York: Seminar Press, 1972, pp. 105–155.
13. Goodman, D. J., and Flanagan, J. L., "Direct Digital Conversion between Linear and Adaptive Delta Modulation Formats," Proc. Int. Conf. Commun., Montreal, June 1971, pp. 1–33 to 1–36.
14. Flanagan, J. L., U. S. patent application.
15. Shipley, K., Rabiner, L. R., Schafer, R. W., and Flanagan, J. L., "Computer-Spoken Instructions for ESS-1 Central Office Wiring," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Adaptive Quantization With a One-Word Memory

By N. S. JAYANT

(Manuscript received March 12, 1973)

We discuss a quantizer which, for every new input sample, adapts its step-size by a factor depending only on the knowledge of which quantizer slot was occupied by the previous signal sample.[1] *Specifically, if the outputs of a uniform B-bit quantizer ($B > 1$) are of the form*

$$Y_u = P_u \frac{\Delta_u}{2}\ ; \quad \pm P_u = 1, 3, \cdots, 2^B - 1; \quad \Delta_u > 0,$$

the step-size Δ_r is given by the previous step-size multiplied by a time-invariant function of the code-word magnitude $|P_{r-1}|$:

$$\Delta_r = \Delta_{r-1} \cdot M(|P_{r-1}|).$$

The adaptations are motivated by the assumption that the input signal variance is unknown, so that the quantizer is started off, in general, with a suboptimal step-size Δ_{START}. Multiplier functions that maximize the signal-to-quantization-error ratio (SNR) depend, in general, on Δ_{START} and the input sequence length N. For example, if the signal is stationary and $N \to \infty$, best multipliers, irrespective of Δ_{START}, have values arbitrarily close to unity. On the other hand, small values of N and suboptimal values of Δ_{START} necessitate M values further away from unity. By including an adequate range of values for N and Δ_{START} in a generalized SNR definition, we show how one can determine stable multiplier functions M_{OPT} that are optimal for a given signal.

In computer simulations of 2- and 3-bit quantizers with first-order Gauss-Markovian inputs, we note that, except when the magnitude of the correlation C between adjacent samples is very high, M_{OPT} has the property of calling for fast increases and slow decreases of step-size. We derive optimum multipliers theoretically for two simple cases:

$$M_r^{\text{OPT}} = \left[\frac{1}{2} + \frac{K^2}{8} P_{r-1}^2\right]^{\frac{1}{2}} + \delta^2(|P_{r-1}|); \quad C = 0$$

$$M_r^{\text{OPT}} = \frac{|P_{r-1}|}{2^{B-1}} + \delta^2(|P_{r-1}|); \quad C \to 1.$$

K is a constant depending only on B, and δ^2 is a positive correction that is significant only for the last slot: $|P_{r-1}| = 2^B - 1$. Using the example of $C = 0$, we also show how the approach of specifying P_{r-1}, explicitly, in the determination of Δ_r, is more effective than an earlier procedure[2] where Δ_r is determined by past output values Y_{r-1} (rather than by a function of their components, P_{r-1} and Δ_{r-1}).

Computer simulations with speech and picture signals have shown, once again, that SNR-maximizing multiplier functions demand step-size increases that are relatively faster than step-size decreases. Values of M_{OPT} depend, interestingly, on whether the quantizer is used in a PCM or a DPCM-type coder. In the case of speech signals, we propose corresponding tables of M_{OPT} values for $B = 2$, 3, 4, and 5. DPCM coding of speech with 3- and 4-bit adaptive quantizers is the subject of a companion paper.[1]

I. INTRODUCTION

Quantization error, in general, can take one of two distinct forms, overload distortion or granular noise, reflecting, respectively, situations where the quantizer step-size is too small or too large relative to the signal being quantized. This distinction has been widely noted for 1-bit quantizers (delta modulators), and variable step-size quantization has therefore been widely discussed in this context.[3–6] The general idea is to increase the step-size during overload and decrease it during granularity, and to detect those conditions on the basis of observations of the delta modulator bit stream. The step-size adaptations can be either instantaneous[3,5,6] or "syllabic,"[4] and the advantages of adaptation have been shown, among other means, by demonstrations of dynamic range and of SNR gains over nonadaptive quantizers.[6]

The problem of step-size adaptations, as applied to quantizers with more that two output levels, has been less widely studied. It is conventional in such quantizers to take signal nonstationarity into account by means of a suitably designed, time-invariant, nonuniform quantizer.[7] Recently, however, two proposals have incorporated time-variant step-size logics in multibit quantization. The first of these techniques is a syllabically adapting PCM which Wilkinson empirically designed for speech encoding at 10 kb/s.[8] The second proposal is an instantaneously adapting quantizer discussed by Stroh, in the context of differential encoding of Gaussian signals.[2] Syllabic adaption has the advantages that it can be better tailored to a given signal such as

speech and that it can also be designed to provide better resistance to bit errors[4] than instantaneous adaptation. The latter, on the other hand, has the advantages of minimal structure and applicability to different types of signals, and, in relatively noise-protected environments, it constitutes an efficient and simple encoding procedure for signal storage or transmission.

The adaptation that we discuss is instantaneous, and we indicate, at the end of this paper, how it can perform better than Stroh's compandor[2] when working with one word of quantizer (output) memory. We must emphasize here that in each case what is being gained by the adaptation is increased dynamic range rather than an inherent signal-to-noise ratio advantage over a nonadaptive technique. The adaptive techniques presuppose that the input signal variance is unknown. The quantizer step-size cannot therefore be meaningfully preset to an optimized constant value, but must be allowed to adapt itself to signal statistics in a fashion determined by a (time-invariant) adaptation strategy.

The specific quantizer configuration that we consider is characterized by a uniform spacing of nonzero output levels, and Fig. 1 shows a snapshot of the quantizer at sampling instant r for the example of

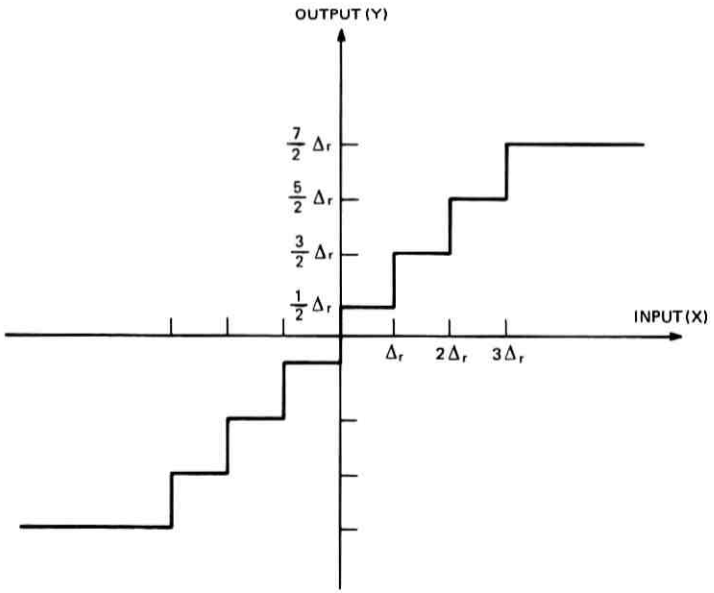

Fig. 1—Uniform quantizer with 8 levels ($B = 3$).

$B = 3$. The step-size Δ_r is adapted, for every new input sample, by a factor depending only on the knowledge of which quantizer slot was occupied by the previous signal sample. More precisely, if the outputs of a B-bit quantizer ($B > 1$) are of the form

$$Y_u = P_u \frac{\Delta_u}{2}; \qquad P_u = \pm 1, 3, \cdots, 2^B - 1; \qquad \Delta_u > 0, \tag{1}$$

the step-size Δ_r is given by the previous step-size multiplied by a time-invariant function of the previous code-word magnitude $|P_{r-1}|$:

$$\Delta_r = \Delta_{r-1} \cdot M(|P_{r-1}|). \tag{2}$$

Note that, according to (2), the entire quantizer is "accordioned in" when $M < 1$ and stretched out when $M > 1$. The resulting quantizer is also uniform, with a step-size or "slot width" equal to Δ_r. Practical implementations will also include upper and lower limits Δ_{MAX} and Δ_{MIN} for Δ_r. This is discussed later in the paper.

The above logic has been recently employed[1] for efficient differential encoding of speech signals at bit rates of 20 to 30 kb/s. The adaptation strategy (2) is indeed arbitrary.* But it represents, in the manner of the adaptive delta modulator discussed earlier by the author,[6] a very simple, yet nontrivial, type of exponential adaptation, and sets a lower bound on the performance of possible sophistications that may include nonexponential adaptations and the use of longer word memories, i.e., the use of P_{r-2}, P_{r-3}, etc.

An interesting result of this paper is that, for many interesting input signals, the step-size multiplier function $M(|P|)$ which minimizes the mean-squared quantization error has the interesting property that it demands step-size decreases significantly slower than step-size increases. This is shown to be true for illustrative speech and picture signals and for first-order Gauss-Markovian inputs where the magnitude of the correlation between adjacent signals is not too high (say, less than 0.9).

In Section II, we discuss computer simulations with a first-order Gauss-Markov input. We discuss the simple case of a white signal ($C = 0$) at length. Results show the dependence of signal-to-quantization-error ratio (SNR) on the function $M(|P|)$ for different values of B (number of quantizer bits), N (number of samples in input sequence), and Δ_{START} (initial step-size). We then specify adequate ranges of

* Schlink has recently described another useful, but perhaps less general, empirical system.[9] Here, the adaptation consists in switching between only two quantizing characteristics.

variation for N and Δ_{START}, and thence determine a stable multiplier function that is optimal for a white Gaussian signal. Further results include the cases of $C = 0.5$ and 0.99, and show, for $B = 2$ and 3, values of M_{OPT} and SNR gain over a nonadaptive quantizer. We also provide illustrative histograms of slot occupancies and observed step-sizes and a family of companding curves for a 4-bit quantizer.

In Section III, we derive optimum multipliers theoretically for the examples of $C = 0$ and $C \to 1$. Results substantiate the values of M_{OPT} from the computer simulation. We also compare our technique with that of Stroh[2] and discuss the greater efficacy of our adaptation strategy using the example of $C = 0$. Finally, in Section III, we discuss quantizer simulations with speech and picture inputs. We present multiplier functions basically similar to those for Gauss-Markov inputs. Optimal multipliers are found to be slightly different for PCM and DPCM coders. In the case of speech, we provide separate tables of M_{OPT} for $B = 2$, 3, 4, and 5.

II. GAUSS-MARKOV INPUTS

Our simulations have employed, as quantizer input, a first-order Gauss-Markovian sequence $\{X_r\}$ of 10,000 samples generated by the recursive rule

$$X_u = C \cdot X_{u-1} + \sqrt{1 - C^2} \cdot N_u; \qquad X_0 = 0, \tag{3}$$

where the samples N_u are drawn from a zero-mean, unit variance, white Gaussian sequence that is independent of past values of $\{X_r\}$. The input sequence generated in (3) is itself Gaussian with a mean of zero, a variance of unity, and a correlation between adjacent samples equal to the preset constant C.

The quantizer output, by definition, is the output level nearest to the input X_r. It is formally written as

$$\begin{aligned} Y_r &= \left\{\left(2\left[\frac{X_r}{\Delta}\right] + 1\right)\frac{\Delta}{2}\right\} \operatorname{sgn} X_r; \qquad \frac{X_r}{\Delta} < 2^{B-1} \\ &= \left\{(2^B - 1)\frac{\Delta}{2}\right\} \operatorname{sgn} X_r; \qquad \frac{X_r}{\Delta} \geqq 2^{B-1}, \end{aligned} \tag{4}$$

where $[\cdot]$ stands for "greatest integer in."

The quantization error

$$E_r = Y_r - X_r \tag{5}$$

has a magnitude that is bounded by $\Delta/2$ except during overload which is expressed by the second line in eq. (4).

A conventional performance measure is the signal-to-quantization-error ratio

$$\text{SNR} = \frac{\sum X_r^2}{\sum E_r^2}, \tag{6}$$

where summations are assumed to be over the duration of a statistically adequate input sequence.

We also refer in this paper to nonadaptive quantizers for which

$$\begin{aligned} M(|P_{r-1}|) &= 1; \quad \text{all } P_{r-1} \\ \Delta_r &= \Delta; \quad \text{all } r, \end{aligned} \tag{7}$$

and the variation of signal-to-quantization-error ratio SNR_{NA} is a function of the constant step-size Δ for this case. The step-size which maximizes SNR_{NA} for a nonadaptive quantizer will be referred to as the optimum step-size Δ_{OPT}. Values of Δ_{OPT} and the corresponding values of SNR_{NA}, for different values of B, have in fact been tabulated by Max[10] for the case of $C = 0$. Max's results also specify (via the Gaussian probability density function) the probability P_s that the sth slot is occupied in an optimized nonadaptive quantizer:

$$\begin{aligned} P_s &= \text{Prob}\,(P_u = 2s - 1) + \text{Prob}\,(-P_u = 2s - 1); \\ s_u &= 1, 2, \cdots, 2^{B-1}, \end{aligned} \tag{8}$$

where P_u is defined by (1). We will see presently that the probability P_s is also very relevant in the study of an adaptive quantizer when $C = 0$.

2.1 *A General Performance Criterion*

Adaptive quantizers are needed, as mentioned earlier, when nonstationary input signals are expected. Our simulations with Gaussian signals utilized a stationary input (3). To make the study of adaptation strategies meaningful in this stationary environment, we shall introduce some unconventional performance measures. For example, consider the ratio

$$\text{SNR}(N, \Delta_{\text{START}}) = \sum_1^N X_r^2 \Big/ \sum_1^N E_r^2, \tag{9}$$

where summations are over the first N samples of the input sequence. The dependence of SNR on Δ_{START} is significant only for small values of N. For large N, (9) tends to an asymptotic value that is independent of Δ_{START}:

$$\text{SNR}(\infty) \triangleq \lim_{N \to \infty} \text{SNR}(N, \Delta_{\text{START}}). \tag{10}$$

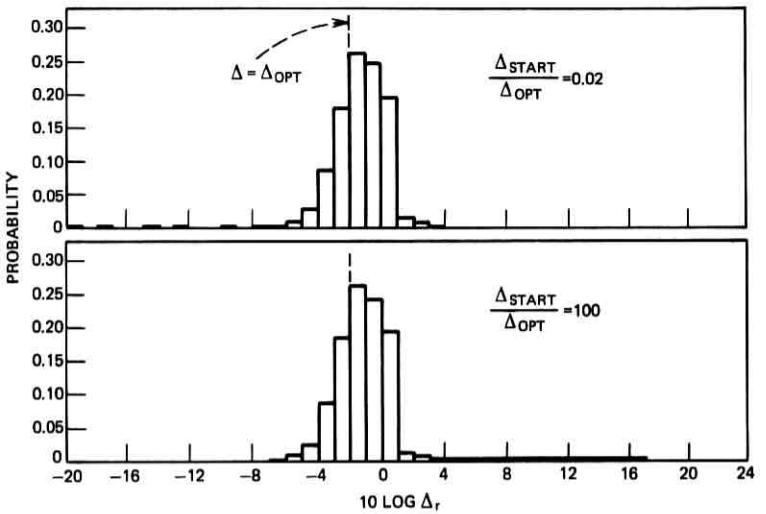

Fig. 2—Step-size histograms ($B = 3$, $C = 0.5$, $N = 10{,}000$).

In fact, if N is sufficiently large, the value of Δ_{START} is entirely academic in the study of adaptive quantizers. See the step-size histograms in Fig. 2, for example. Notice how they are independent of Δ_{START}, except for the flat tails representing transient values of Δ.

In adaptive quantization, a suitable multiplier function for a given signal should provide a compromise between quickness of response [as measured by the magnitude of (9) for small values of N and bad values of Δ_{START}] and satisfactory steady-state performance [as measured by the magnitude of (9) for large values of N and values of Δ_{START} close to Δ_{OPT}]. With these opposing factors in mind, we define an average performance index

$$\mathrm{SNR}_{AVE} \triangleq \frac{1}{20} \sum_{N} \sum_{\Delta_{START}} \mathrm{SNR}(N, \Delta_{START}) \tag{11}$$

for values of $N = 10$, 100, 1000, and 10,000, and

$$\Delta_{START} = \left[\frac{1}{10}, \frac{1}{\sqrt{10}}, 1, \sqrt{10}, 10 \right] \Delta_{OPT}.$$

The target values of N and Δ_{START} above have been chosen with the following factors in mind:

(*i*) First, as mentioned earlier, infinitesimally small ranges of values (for example, $\Delta_{START} \cong \Delta_{OPT}$; any N) are uninteresting

because they can result in M_{OPT} values arbitrarily close to the trivial value of unity.

(*ii*) On the other hand, overly wide ranges of parameters which include combinations like ($N = 1$, $\Delta_{START} = 10^4 \Delta_{OPT}$) reflect pathological situations and lead to multiplier specifications that tend to be quite uncorrelated with the statistical nature of the signal being quantized.

(*iii*) As long as the extreme situations in (*i*) and (*ii*) are avoided, it has been found that M_{OPT} values are not overly sensitive to the actual N and Δ_{START} values employed in the performance criterion (11), but depend mainly on the statistics of the signal being encoded. In fact, optimal multipliers in this case are merely the best multipliers in a variance-estimating problem (see the theory for $C = 0$ in Section III) that includes neither N nor Δ_{START} as a significant parameter.

(*iv*) With the aforementioned factors in mind, the specific values of N and Δ_{START} in (11) were selected to have the following significance for a typical application such as speech quantization. First, the 40-dB range for Δ_{START} reflects an extent of uncertainty (about signal power) which is reasonably characteristic of telephone conversation.[7] Second, when one considers Nyquist-sampled speech for applications like adaptive PCM or adaptive DPCM,[1] the values of N in (11) correspond at the lower end to about 1 millisecond of speech, and at the higher end to about 1 second of speech. This range clearly includes the range of durations that one may associate with "steady-state" or "stationary" segments in the acoustic waveform. In fact, if one considers phoneme durations, values of N in the range 100 to 5000 seem to provide an adequate model. It is our contention that by using N values of this type in an index of performance such as (11), we can very usefully assess M-functions for quantizing locally stationary signals such as speech, even when simulating the quantizer with a (standard and easily duplicated) stationary Gaussian input. Actually, however, we have carried out completely independent simulations with real speceh signals as well (Section IV), and the results of this section are directed toward the quantization of Gaussian inputs as such.

2.2 *Multiplier Functions for $B = 2$, $C = 0$*

Table I illustrates the nature of the SNR function (9) for two multiplier functions in a 2-bit quantizer. The first multiplier function

TABLE I—EXAMPLE OF SNR FUNCTIONS FOR $B = 2$, $C = 0$ (ENTRIES IN dB)

$20 \log \left(\frac{\Delta_{START}}{\Delta_{OPT}} \right)$	Values of N			
	10	100	1000	10,000
	$M_1 = 0.8$		$M_2 = 1.6$	
−20	6.4	7.2	7.4	7.3
−10	10.5	8.9	7.9	7.3
0	9.7	8.3	7.6	7.3
10	5.8	7.2	7.4	7.2
20	−5.9	4.2	7.1	7.3
	$M_1 = 0.98$		$M_2 = 1.04$	
−20	1.6	3.8	8.4	9.1
−10	5.2	5.8	8.9	9.2
0	10.7	8.0	9.4	9.2
10	0.0	5.9	9.0	9.2
20	−13.2	−5.0	3.5	8.1

shows quicker response (better SNR values for $N = 10$ and 100), while the second function achieves a better asymptotic value of SNR (at $N = 10{,}000$). Obviously, the poor asymptotic performance of the first M-function is due to overly abrupt step-size oscillations in the "steady-state," while the inferior performance of the second M-function for small N is due to sluggish adaptations of Δ when Δ_{START} is suboptimal.

Table II compares several M-functions* for a 2-bit quantizer on the basis of (11). The functions included represent a subset of many more functions which were simulated and compared on the basis of SNR_{AVE}. The best value of 6.8 dB has been noted for $M_1 = 0.80$, $M_2 = 1.60$, although this function provides a clearly nonmaximal asymptotic performance (Table I). The first five functions in Table II also satisfy

TABLE II—COMPARISON OF MULTIPLIER FUNCTIONS ($B = 2$, $C = 0$)

M_1	M_2	SNR_{AVE} (dB)
0.71	2.00	5.9
0.80	1.60	6.8
0.90	1.20	6.5
0.95	1.10	6.1
0.98	1.04	5.3
0.95	1.20	5.9
0.50	2.00	5.8
0.90	1.10	5.2

* Whenever there is no scope for confusion, we shall use the symbols M_1, M_2, M_3, and M_4 instead of $M_r(1)$, $M_r(3)$, $M_r(5)$, and $M_r(7)$.

the interesting constraint suggested by Goodman:[11]

$$M_1^2 \cdot M_2 \cong M_1^{0.67} M_2^{0.33} \cong M_1^{P_1} \cdot M_2^{P_2} \cong 1, \tag{12}$$

where $P_1 \cong 0.67$ and $P_2 \cong 0.33$ are the probabilities of inner- and outer-slot occupancy in a nonadaptive quantizer with an optimal Δ_{OPT} for the Gaussian input. Goodman conjectures that the probabilities of using M_1 and M_2 in a well-designed adaptive quantizer should indeed be equal to the parameters P_1 and P_2 of the nonadaptive quantizer. A constraint of the form (12) then represents a stability

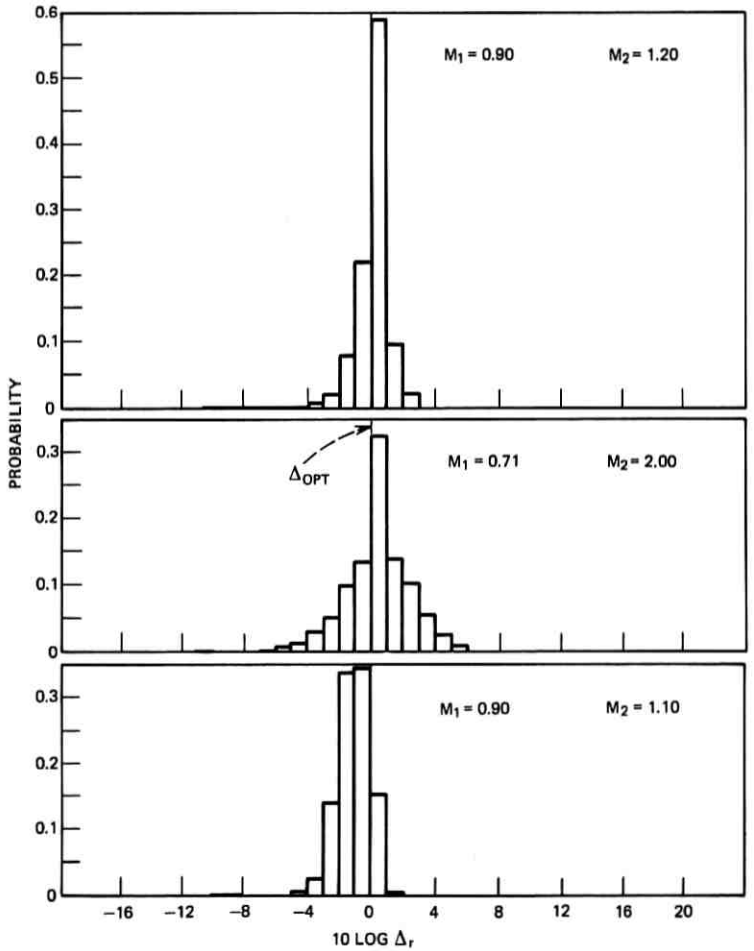

Fig. 3—Step-size histograms ($B = 2$, $C = 0$, $N = 10{,}000$).

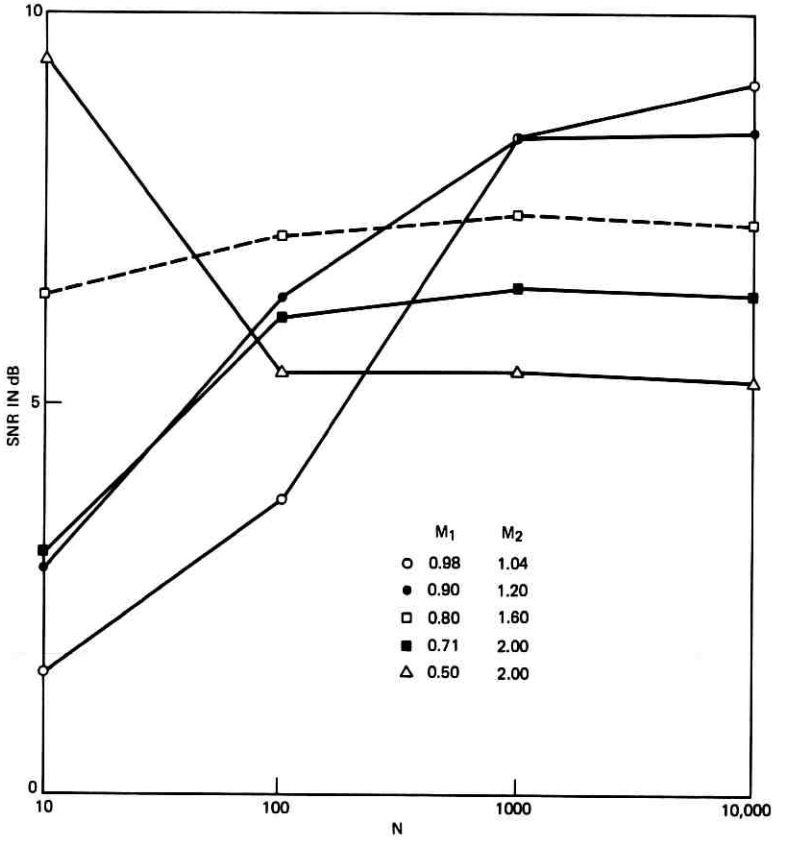

Fig. 4—Comparison of multiplier functions ($B = 2$, $C = 0$, $\Delta_{START} = 0.1\ \Delta_{OPT}$).

criterion which specifies that the random step-size Δ_r neither grows out of bounds, independently of the input, nor decays to infinitesimal values. This criterion has been discussed earlier in the context of adaptive delta modulation with a 1-bit memory.[6]

The desirability of constraint (12) on step-size multipliers is also demonstrated by the step-size histograms in Fig. 3. The multiplier pairs (0.9, 1.2) and (0.71, 2.0) satisfy constraint (12), and the corresponding histograms have the desirable property that they are centered on Δ_{OPT} although they have different dispersions (suggesting differences in quickness of response and steady-state performance). The function (0.9, 1.10), on the other hand, produces a histogram whose mode is clearly displaced from Δ_{OPT}. This suggests that (0.9, 1.10) falls in a

TABLE III—SNR FUNCTION FOR $M_1 = 0.90$, $M_2 = 0.90$, $M_3 = 1.25$, $M_4 = 1.75$ ($B = 3$, $C = 0$, ENTRIES IN dB)

$20 \log \left(\frac{\Delta_{START}}{\Delta_{OPT}}\right)$	Values of N			
	10	100	1000	10,000
−20	11.9	11.2	12.7	12.7
−10	14.5	11.8	12.6	12.7
0	15.7	11.2	12.9	12.7
+10	11.8	11.6	12.5	12.7
+20	−1.6	8.7	12.3	12.7

class of inefficient multiplier functions; this is attributed to the fact that the function (0.9, 1.10) clearly violates requirement (12) above.[11]

Finally, in Fig. 4, we show SNR (6) as a function of N for a fixed value of Δ_{START}, and for different M-functions. It is once again apparent that the adaptation function (0.8, 1.6) provides an attractive combination of responsiveness and asymptotic performance for $B = 2$.

2.3 *Multiplier Functions for $B = 3$, $C = 0$*

Table III demonstrates the nature of the SNR function (9) for $B = 3$ and a specific multiplier function. Table IV uses the performance criterion (11) to show the efficiency of this multiplier function (0.9, 0.9, 1.25, 1.75). As in the 2-bit example, the M-functions in Table IV are only a subset of a much larger set of M-functions which were simulated and compared on the basis of SNR_{AVE}. We have only included the most interesting functions from our search for maximum SNR_{AVE}. The first three M-functions in Table IV satisfy a stability constraint analogous to (11):

$$M_1^{0.46} M_2^{0.31} M_3^{0.16} M_4^{0.07} = M_1^{P_1} \cdot M_2^{P_2} \cdot M_3^{P_3} \cdot M_4^{P_4} = 1. \tag{13}$$

It is interesting that the best function in Table IV belongs to the class of functions obeying (13). Notice also that the reduction of the number of distinct step-size multipliers (second row in Table IV) leads to a

TABLE IV—COMPARISON OF MULTIPLIER FUNCTIONS ($B = 3$, $C = 0$)

M_1	M_2	M_3	M_4	SNR_{AVE} (dB)
0.90	0.90	1.25	1.75	11.7
0.90	1.00	1.00	1.75	11.4
0.5	1.0	1.0	2.0	9.6
0.3	0.9	1.5	2.1	8.9

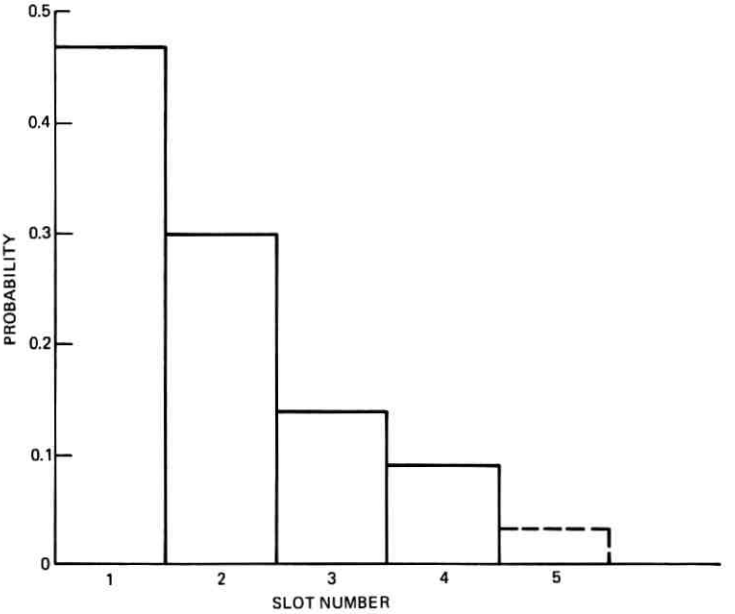

Fig. 5—Histogram of slot occupancies ($B = 3$, $C = 0$, $N = 10{,}000$).

marginal decrease of $\mathrm{SNR}_{\mathrm{AVE}}$. This tolerance to a reduction of the number of distinct multipliers does seem to extend, although in lesser measure, to larger values of B and to speech and picture signals.

Finally, Fig. 5 shows a histogram of slot occupancies for the best M-function in Table IV. The number of quantizer slots or output levels is equal to four (neglecting signs), and the dotted fifth slot refers to the overload probability that has been accumulated into the fourth bar of the histogram. It is interesting that, despite step-size adaptations, the Gaussian nature of the input density function shows up in the histogram. The heights of the bars in Fig. 5 represent experimental slot probabilities of 0.47, 0.30, 0.14, and 0.09. Notice again that, in the manner of (13):

$$0.9^{0.47} \cdot 0.9^{0.30} \cdot 1.25^{0.14} \cdot 1.75^{0.09} = 0.994 \cong 1. \tag{14}$$

2.4 *Comparison of Adaptive and Nonadaptive Quantizers*

Table V summarizes the nature of optimal multiplier functions for $B = 2$ and 3. These functions are obtained on the basis of criterion (11). Values of M are generally rounded, representing broad optima,

TABLE V—QUANTIZATION OF GAUSS-MARKOV INPUTS [ENTRIES ARE SNR (10,000, Δ_{OPT}) VALUES IN dB]

B	C	0.00	0.50	0.99
2	SNR_{NA}	9	9	9
	SNR_A	7	8	11
	$M(1)$	0.8	0.8	0.5
	$M(2)$	1.6	1.6	2.0
3	SNR_{NA}	14	14	14
	SNR_A	13	13	16
	$M(1)$	0.90	0.90	0.30
	$M(2)$	0.90	0.90	0.90
	$M(3)$	1.25	1.25	1.50
	$M(4)$	1.75	1.75	2.10

and the precision in the specification of M values may be as bad as ± 5 percent in some cases.

To provide a fair comparison with optimal nonadaptive quantizers, the performance figure used in Table V is the asymptotic value (10). Formally, the notation used in the table is as follows:

$$\text{SNR} = \text{SNR}(10{,}000, \Delta_{OPT}). \tag{15}$$

The subscript A refers to the adaptive quantizer with step-size multipliers optimized using (11), while the subscript NA refers to a non-

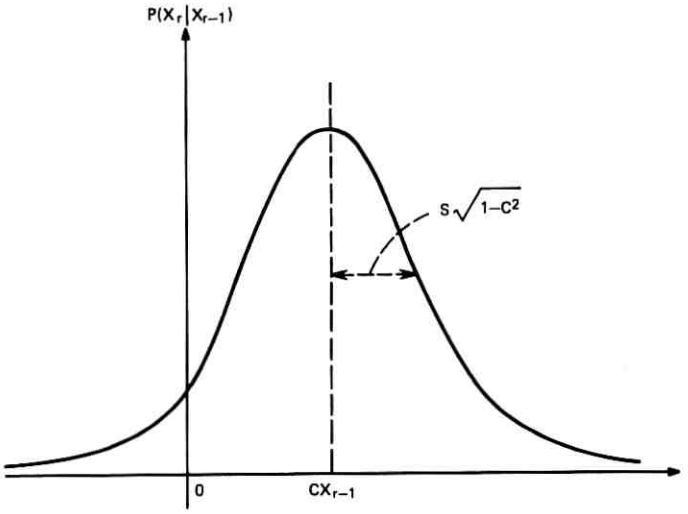

Fig. 6—Conditional density function of quantizer input.

adaptive quantizer with constant step-size Δ_{OPT}. The SNR values are in dB, and are rounded to the nearest integer.

Note that negative values of C are not included in the table. The assumption of a symmetrical quantizer (Fig. 1) renders the quantizer design independent of the sign of C. Specifically, the quantizer input X_r has a probability density function (conditioned to X_{r-1}) that is sketched in Fig. 6; the optimum step-size is that which fits the quantizer to this density function in a way that minimizes the sum of overload error variance and granular error power. This optimum depends only on the disposition of the PDF in Fig. 6 and the magnitude of the nonzero mean, not the sign of it.

Finally, Table V assumes that no constraints exist on the minimum and maximum values of step-size. Practical implementations will, of course, involve such constraints (see Fig. 8), as well as constraints on actual multiplier values. Significant conclusions from Table V are the following:

(*i*) Except for $C = 0.99$, optimal multipliers are such that step-size decreases are always slower than step-size increases. The observation has been found to extend for $B = 4$ also and, as seen later (Section IV), to the quantization of speech and picture signals as well.

The need for fast increases of step-size and slow decreases thereof may be physically explained as follows. Quantization errors during overload tend to be more harmful than those during granularity, in that the magnitude of granular error is restricted, by definition, to a half step-size, while no such simple constraint exists for an overload error. It is therefore reasonable to decrease step-sizes (relatively) slowly to avoid unduly small step-sizes leading to the harmful overload errors. The observation is obviously less significant for a coarser quantizer than for a finer quantizer because granular errors in the former are more comparable in magnitude to overload errors and hence more equally harmful. This is indeed reflected in Table V. Note that, for a given value of C, the disparity in rates of step-size increases and step-size decreases is least for the coarser quantizer ($B = 2$).

There is an alternative explanation for (*i*) above, which also clarifies why the disparity between the speeds of step-size increase and step-size decrease is less apparent for large values of C. Refer to the stability constraints (12) and (13), as discussed for the case of $C = 0$. It turns out that in the uniform (nonadaptive) quantization of a Gaussian signal, the probability P_s (8) is a monotonically decreasing function of s. It follows then, as seen in (12) and (13), that multipliers for step-size decreases have greater probabilities of being employed, and hence

must lead to slower step-size changes each time they are actually used. Explicitly, for $B = 2$, (12) can be rewritten:

$$\frac{\ell n M_2}{\ell n(1/M_1)} = \frac{P_1}{P_2}. \tag{16}$$

Obviously, then, if $P_1 > P_2$, the step-size increase (as given by M_2) is faster than the step-size decrease (as given by M_1).

The argument for nonzero values of C is very similar, except that the probabilities P_s peculiar to a Gaussian probability density function should now be replaced by probabilities $P_s(C)$ that refer to the uniform quantization of the asymmetrical conditional PDF in Fig. 6. Apparently, the probabilities $P_s(C)$ are not monotonically decreasing for $C = 0.99$. This is why the requirement of relatively more rapid step-

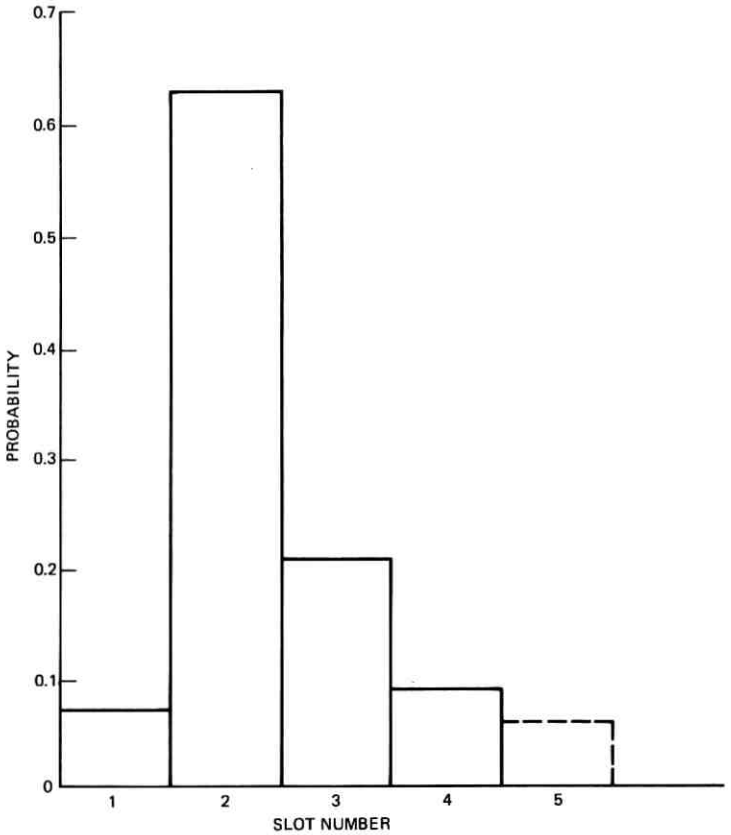

Fig. 7—Histogram of slot occupancies ($B = 3$, $C = 0.99$, $N = 10{,}000$).

size increases is waived for the example of $C = 0.99$ (while being still true for $C = 0.5$).

Figure 7 shows a histogram of slot-occupancies for $B = 3$ and $C = 0.99$. Compare this non-monotonic PDF with the Gaussian histogram for $C = 0$ (Fig. 5). In analogy with (13), the stability criterion associated with Fig. 7 is

$$0.3^{0.07}\,0.9^{0.63}\,1.5^{0.21}\,2.1^{0.09} = 0.993 \cong 1. \tag{17}$$

Finally, it should be mentioned that the (relatively) slow step-size decreases in Table V are fast enough, in an absolute sense, for typical quantizer applications. For example, if $M = 0.95$, and a step-size decrease of 20 dB is needed for adaptation to an idle-channel situation in speech quantization, the time needed for such adaptation will be 45 samples. For Nyquist-sampled speech, this is only about 5 ms.

(*ii*) Although the quantization problem for $C = 0.5$ is qualitatively similar to that for $C = 0.99$ (Fig. 6), we note that results for $C = 0.5$ (Table V) are nearly identical with those for $C = 0$. The differences in M_{OPT} values that are caused by a nonzero $C = 0.5$ were apparently too small to be detected in our finite search for best multipliers.

(*iii*) Referring again to Table V, the best adaptive quantizers seem to have an SNR advantage over the nonadaptive scheme (working with an optimal step-size) only for very highly correlated inputs. In fact, in many instances, the SNR gain resulting from adaptation is seen to be negative (due, evidently, to overly abrupt manipulations of step-size).

The reason for using an adaptive quantizer in these situations is only to facilitate quantizations with much less knowledge of the input—equivalently, with much less knowledge of Δ_{OPT} than is necessary for an equivalent performance in the nonadaptive case. In other words, step-size adaptions increase the dynamic range of the quantizer and enable it to handle inputs with large amplitude variations, such as nonstationary signals.

The above idea has already been demonstrated by the asymptotic SNR values in Tables I and III. To provide a more application-oriented illustration, we undertook two extensions of our computer simulation. These experiments employed $B = 4$, $C = 0.5$, and the following multiplier function:

$$(0.90, 0.90, 0.95, 1.0, 1.2, 1.5, 1.8, 2.1). \tag{18}$$

Finite step-size dictionaries were used, determined by maximum and minimum step-sizes Δ_{MAX} and Δ_{MIN}. The starting step-size was set

equal to Δ_{OPT}, subject, however, to modification because of the constraints Δ_{MAX} and Δ_{MIN}.

In the first of these extensions, the step-size dictionary had the characterization

$$\Delta_{MAX} \cdot \Delta_{MIN} = \Delta_{OPT}^2 \tag{19}$$

$$\Delta_{MAX}/\Delta_{MIN} = R \tag{20}$$

and the quantizer performance was studied in terms of SNR (10,000, Δ_{OPT}) as a function of R. It was reassuring to note that the SNR was constant to within 1 dB for sample values of R in the range 1 to ∞—due, no doubt, to the safe design feature (19). In fact, a maximum SNR was noted for a noninfinite value of R.

In a more revealing second experiment, the quantizer was "centered" at a value Δ_{MID} not necessarily equal to Δ_{OPT}:

$$\Delta_{MAX}\Delta_{MIN} = \Delta_{MID}^2 \tag{21}$$

and the performance was measured as a function of Δ_{MID} for values of $R(20)$ equal to 1, 10, and 100. Note that $R = 1$ refers to the nonadaptive case.

Figure 8 plots these results. The monotonic improvement of dynamic

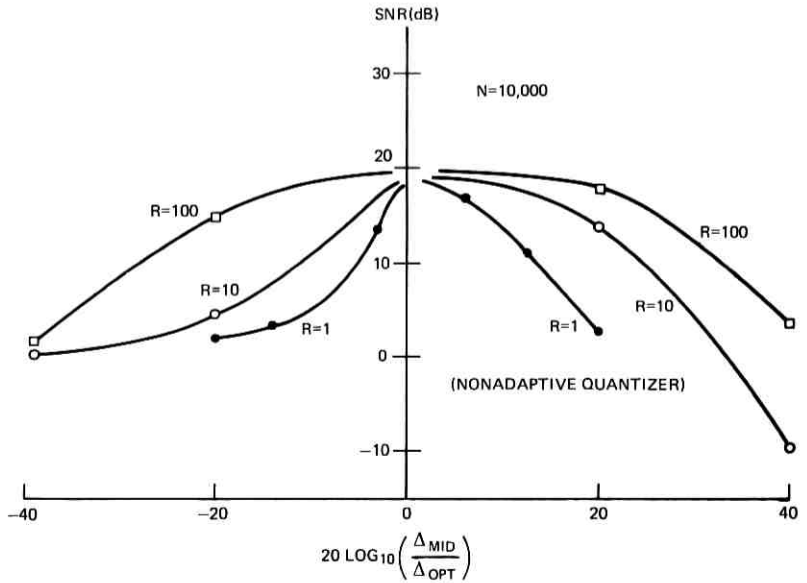

Fig. 8—Companding characteristics ($B = 4$, $C = 0.5$).

range with increasing R is apparent. It is expected[1] that practical quantizers can be designed with values of R equal to 100 or more.

III. THEORETICAL DERIVATION OF OPTIMAL MULTIPLIERS

In this section, we shall regard the adaptive quantization problem as one of learning signal variance. In other words, the problem of determining an optimum instantaneous step-size Δ_r is regarded as being tantamount to that of finding the best estimate at time r of the conditional standard deviation S_r of the quantizer input; and of setting Δ_r proportional to this estimate:

$$\Delta_r^{\text{OPT}} = K(B) \cdot \hat{S}_r(\Delta_{r-1}, P_{r-1}). \tag{22}$$

3.1 *Case of $C = 0$*

The constant K is an obvious function of the number of quantizer levels and, hence, of B. For the problem of uniform quantization of a zero-mean Gaussian signal, Max's Table II[10] specifies the following values for $K(B)$:*

$$\begin{aligned} K(1) &= 1.596, \qquad K(2) = 0.996, \\ K(3) &= 0.586, \qquad K(4) = 0.335. \end{aligned} \tag{23}$$

The dependence of $\hat{S}_r$ on Δ_{r-1} and P_{r-1} (22) is, of course, characteristic of an adaptation strategy which uses a 1-word memory.

We now propose that the variance of X_r be estimated as the average of the squares of (*i*) X_{r-1}, the most recent quantizer input, and (*ii*) $\hat{S}_{r-1}$, the most recent estimate of S. In other words, let

$$\hat{S}_r^2 = \tfrac{1}{2}(X_{r-1}^2 + \hat{S}_{r-1}^2).^{\dagger} \tag{24}$$

We next recall the identity

$$X_{r-1} = Y_{r-1} - E_{r-1} = \frac{P_{r-1}\Delta_{r-1}}{2} - E_{r-1}, \tag{25}$$

where E_{r-1} is the quantization error. Furthermore, by virtue of the basic algorithm (22), we suggest that

$$\hat{S}_{r-1}^2 = \Delta_{r-1}^2 / K^2 \tag{26}$$

Let us use (25) and (26) in (24) and set the resulting value of $\hat{S}_r$ in (22). We obtain, after some algebra:

$$(\Delta_r^{\text{OPT}}/\Delta_{r-1})^2 = \frac{K^2}{2}\left[\frac{P_{r-1}^2}{4} + \frac{1}{K^2} + \frac{1}{\Delta_{r-1}^2}(E_{r-1}^2 - E_{r-1}\Delta_{r-1}P_{r-1})\right]. \tag{27}$$

* These K values are relevant for $C = 0$ because, in this case, the conditional density function (Fig. 6) is indeed zero-mean Gaussian.

† In general, one may consider a weighted average of the type $ux^2 + vs^2$. The case of $u = 0$ will be appropriate for "steady-state" operation, and the use of $v = 0$ will be appropriate for a "transient" situation. The need for time-invariant step-size multipliers suggests a compromise design characterized by a weighting of the type $u = v = 0.5$.

E_{r-1} is an unknown random variable, but the following can be said about its role in (27):

First, the E_{r-1}^2 term is significant only for the last quantizer slot in which, due to possible overload, E_{r-1}^2 can be arbitrarily large. Furthermore, for this end slot the $-E_{r-1}\Delta_{r-1}P_{r-1}$ term tends to be positive. Notice, from definition (25), E is negative in overload when P is positive and vice versa.

For the remaining quantizer slots, E_{r-1}^2 is again positive but no longer significant, and $-E_{r-1}\Delta_{r-1}P_{r-1}$ is expected to be negligible as well, on the average. This is by virtue of the uniform PDF approximation for granular errors

$$P(E) = 1/\Delta; \qquad -\frac{\Delta}{2} < E_{\text{gran}} < \frac{\Delta}{2} \tag{28}$$

and a consequent decorrelation of output $P\Delta$ and error E.

The optimum multiplier function [square root of (27)] can therefore be expressed in the form

$$M_r^{\text{OPT}} = \left[\frac{1}{2} + \frac{K^2}{8} P_{r-1}^2\right]^{\frac{1}{2}} + \delta^2(|P_{r-1}|); \qquad C = 0; \tag{29}$$

where δ^2 is a positive correction term that is significant only for the end slot:

$$\delta^2(|P_{r-1}|) \cong 0 \quad \text{if} \quad |P_{r-1}| \neq 2^B - 1. \tag{30}$$

Table VI compares the M values from (29) with those from the simulation in Section II.

3.2 *Comparison With Stroh's Adaptation Logic* ($C = 0$)

Consider, in place of (24), a simpler variance estimation of the type considered by Stroh:[2]

$$\hat{S}_r^2 = X_{r-1}^2. \tag{31}$$

This results in, by virtue of (25), (22), and arguments similar to those at the end of the previous paragraph, a multiplier function of the form

$$\begin{aligned} M_r &= \frac{K}{2}\,|P_{r-1}| + \delta^2(|P_{r-1}|); \\ \delta^2(|P_{r-1}|) &\cong 0 \quad \text{if} \quad |P_{r-1}| \neq 2^B - 1. \end{aligned} \tag{32}$$

Table VI lists values of M_r^{OPT} (29), M_r (32), and the experimental optima M_{EXP} from Table V. Values of K have been taken from (23).

Notice how M_r^{OPT} provides a better specification of optimal multi-

TABLE VI—COMPARISON OF MULTIPLIER FUNCTIONS

$B = 2$				$B = 3$		
M_r	M_r^{OPT}	M_{EXP}	$\lvert P_{r-1}\rvert$	M_r	M_r^{OPT}	M_{EXP}
0.50	0.79	0.80	1	0.29	0.75	0.90
$1.50 + \delta^2$	$1.27 + \delta^2$	1.60	2	0.87	0.94	0.90
—	—	—	3	1.45	1.26	1.25
—	—	—	4	$2.00 + \delta^2$	$1.61 + \delta^2$	1.75

pliers than does M_r. Furthermore, as B increases, the constant K approaches zero and the theoretical multiplier functions for the innermost slot ($P_{r-1} = \pm 1$) have the following limiting behavior:

$$\lim_{B \to \infty} M_r(1) = 0 \tag{33}$$

$$\lim_{B \to \infty} M_r^{\text{OPT}}(1) = \sqrt{1/2} = 0.71. \tag{34}$$

Simulations with $B = 4$ and 5 have verified that the trend in (34) is indeed more realistic than that in (33).

It should be mentioned that the adaptation strategy (31) is only the simplest case of Stroh's[2] method which has a general variance estimator of the form

$$\hat{S}_{r,n}^2 = \frac{1}{n} \sum_{u=1}^{n} X_{r-u}^2. \tag{35}$$

It is interesting, nevertheless, that for the same length ($n = 1$ or one-word) of quantizer memory, our adaptation rule specifies better step-size multipliers, as seen in Table V. In fact, the use of M_r^{OPT} yields for ($B = 3$, $C = 0$), an SNR (for Gaussian signals) which is better than what Stroh reports for $n = 2$ (10 dB; $N = 2500$) in his Fig. 3.3. With the experimentally optimized M-function (Table V), we indeed do significantly better and the SNR value of 12.7 dB for this case is equivalent to $n = 6$ in Stroh's logic and falls short of the optimum ($n = \infty$ in Stroh) by not much more than 1 dB.

The efficiency of our logic is clearly attributable to the way we exploit quantizer memory, namely, in terms of P and Δ, rather than in terms of the product of the two quantities (the quantizer output Y used by Stroh). Physically, the use of $P\Delta$ for adaptation seems to wipe out some of the "overload" and "underload" cues that an individual knowledge of P and Δ preserves.

TABLE VII—COMPARISON OF THEORETICAL(M_r^{OPT}) AND EXPERIMENTAL (M_r^{EXP}, IN PARENTHESES) MULTIPLIERS

B	C	0	0.99
2	$M(1)$	0.79(0.8)	0.5(0.5)
	$M(2)$	$[1.27 + \delta^2](1.6)$	$[1.5 + \delta^2](2.0)$
3	$M(1)$	0.75(0.9)	0.25(0.3)
	$M(2)$	0.94(0.9)	0.75(0.9)
	$M(3)$	1.26(1.25)	1.25(1.5)
	$M(4)$	$[1.61 + \delta^2](1.75)$	$[1.75 + \delta^2](2.1)$

3.3 *Case of* $C \to 1$

When the adjacent signal correlation C approaches unity, the conditional PDF (probability density function) of X_r approaches a Gaussian spike centered at CX_{r-1} (Fig. 6). The width of the spike is proportional to the square root of $(1 - C^2)$, and therefore approaches zero irrespective of the value of signal variance S. The adaptive quantization problem is no longer one of variance estimation. It will consist, instead, in a "fool-proof" strategy of the following type: Select a step-size Δ_r such that the PDF spike at CX_{r-1}^* falls right in the middle of the positive (or negative) half of the quantizer range, assuming that CX_{r-1} is positive (or negative). If we recall that a B-bit quantizer has a half-range width equal to $2^{B-1}\Delta$, we see the requirement (assuming positive quantities throughout) is:

$$CX_{r-1} = \frac{2^{B-1}\Delta_r}{2}\,; \qquad C \to 1\,. \tag{36}$$

The logic clearly provides simultaneous protection against both overload and underload. Utilizing the estimate of X_{r-1} (25) in (36), we obtain the condition

$$C\left[\frac{P_{r-1}\Delta_{r-1}}{2} - E_{r-1}\right] = 2^{B-2}\Delta_r; \qquad C \to 1. \tag{37}$$

Equivalently, with usual assumptions on the quantization error E_{r-1},

$$\lim_{C \to 1} M_r^{OPT} = \frac{\Delta_r}{\Delta_{r-1}} = \frac{|P_{r-1}|}{2^{B-1}} + \delta^2(|P_{r-1}|) \tag{38}$$
$$\delta^2(|P_{r-1}|) \cong 0 \qquad \text{if} \quad P_{r-1} \neq 2^B - 1.$$

* See the spike in the histogram of Fig. 7.

3.4 *Comparison with Simulation Results*

Results for a general value of $C(0 < C < 1)$ can in principle be attempted on the basis of a general PDF such as Fig. 6. However, tractable derivations seem to require too many simplifying assumptions to make the theory worthwhile, especially in view of the observation (Table V) that the correlation becomes significant only if $C \to 1$. We therefore conclude this section by merely listing, in Table VII, theoretical step-size multipliers for $C = 0$ and 0.99 [from (29) and (38)] together with the experimentally optimized multipliers from Section II.

IV. QUANTIZER SIMULATIONS WITH SPEECH AND PICTURE SIGNALS

In this section, we present results from computer simulations of the adaptive quantizer with speech and picture inputs.

The results in Table VIII refer to a low-pass-filtered speech signal (about a second long), and a single frame of picture input (the face of Karen in *Picturephone*® format[6]). Listed are step-size multipliers found, by search procedure, to maximize an asymptotic SNR (10) as measured over the entire length ($N \gg 10{,}000$) of the input sequences.
The following observations are of interest:

(*i*) The signal PDF seems to have a significant effect (presumably through overload statistics and the end-slot correction $\delta^2(|P_{r-1}|)$ of Section III) on the largest step-size multiplier. Note the value of M_4 for picture input.

TABLE VIII—STEP-SIZE MULTIPLIERS FOR ILLUSTRATIVE SPEECH AND PICTURE SIGNALS (ENTRIES IN PARENTHESES REFER TO PICTURES)

B \ Coder Type	PCM	DPCM
2	0.6, 2.2	0.8, 1.6
3	0.85, 1, 1, 1.5 (0.9, 0.95, 1.5, 2.5)	0.9, 0.9, 1.25, 1.75 (0.9, 0.95, 1.5, 2.75)
4	0.8, 0.8, 0.8, 0.8, 1.2, 1.6, 2.0, 2.4	0.9, 0.9, 0.9, 0.9, 1.2, 1.6, 2.0, 2.4
5	0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6	0.9, 0.9, 0.9, 0.9, 0.95, 0.95, 0.95, 0.95, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 3.3

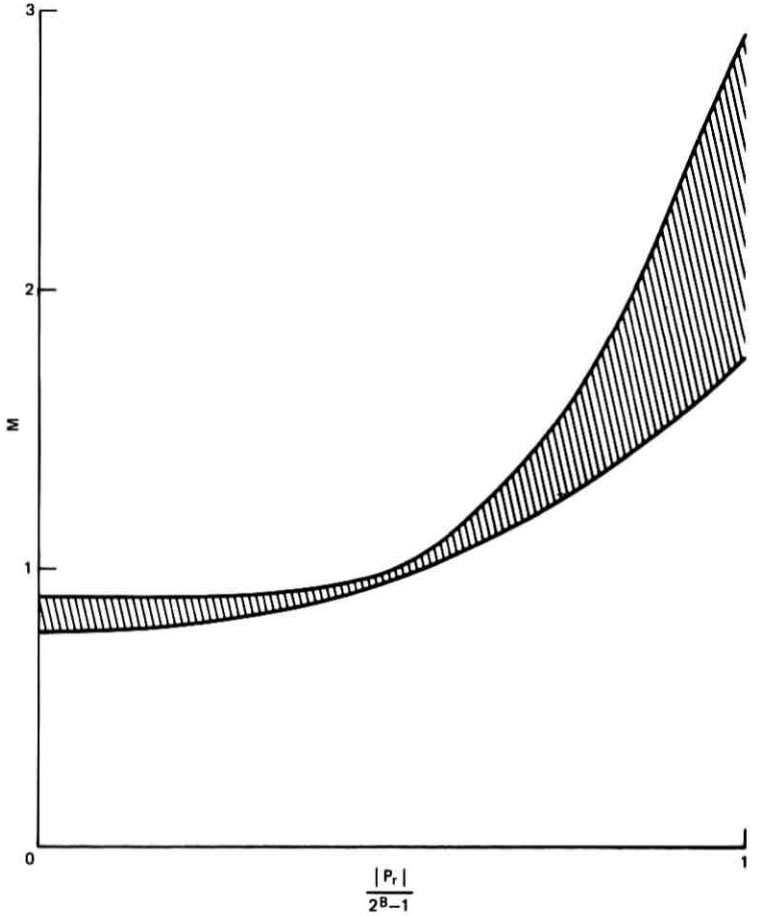

Fig. 9—Desirable form of the multiplier function M for the adaptive quantization of speech signals ($B > 2$) and first-order Gauss-Markov signals which are not highly correlated (say, $C \not> 0.5$).

(*ii*) Differentiation has the effect of decreasing adjacent sample correlation. This seems to explain differences in multipliers as applied to PCM and differential PCM quantizers for speech. Note that the effect is most pronounced for $B = 2$.

(*iii*) Although the input signals are not first-order Markovian, the multipliers have the earlier-mentioned property that step-size increases are relatively more rapid than step-size decreases. Refer to the general diagram in Fig. 9.

TABLE IX—COMPARISON OF SPEECH QUANTIZERS
(ENTRIES ARE SNR VALUES IN dB)

B	Logarithmic PCM with μ-law Quantization	Adaptive PCM with Uniform Quantization	Adaptive DPCM with Uniform Quantization	Adaptive DPCM with Nonuniform Quantization
2	3	9	13	12
3	8	15	18	18
4	15	19	22	24

It may be mentioned that in each of the above simulations, the adaptive techniques also registered an SNR gain of 2 to 4 dB over optimized nonadaptive quantizers. Table IX shows some results pertaining to a band-pass-filtered speech sample. These results were obtained from an independent experiment on coder assessment.[12] The adaptive quantizers (APCM, ADPCM) used the multipliers of Table VIII*, and the nonuniform quantizer characteristics employed in adaptive DPCM are those recommended by Paez and Glisson.[13] Finally, the log-PCM used a $\mu = 100$,[7] and the adaptive quantizers used a maximum-to-minimum-step-size ratio of 100.

Notice from the table that adaptive quantization, as incorporated into PCM, has the potential of outperforming the conventional technique of logarithmic companding. Evidently the advantages over log-PCM are even more impressive in ADPCM, and a companion paper will discuss, at length, the use of 3-bit and 4-bit adaptive quantizers in the DPCM coding of speech.[1]

V. ACKNOWLEDGMENTS

This work was initiated by J. L. Flanagan, encouraged by results on speech quantization due to P. Cummiskey, and sharpened, at a hopelessly blunt point or two, by discussions with D. J. Goodman and comments from D. L. Duttweiler.

REFERENCES

1. Cummiskey, P., Jayant, N. S., and Flanagan, J. L., "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., this issue, pp. 1105–1118.
2. Stroh, R. W., "Optimum and Adaptive Differential Pulse Code Modulation," Ph.D. Thesis, Polytechnic Institute of Brooklyn, 1970.

* Strictly speaking, these values are suboptimal for the nonuniform quantization in the last column of Table IX.

3. Winkler, M. R., "High Information Delta Modulation," IEEE Int. Conv. Record, part 8, 1963, pp. 260–265.
4. Greefkes, J. A., and deJager, F., "Continuous Delta Modulation," Phillips Res. Rep., *23*, No. 2 (1968), pp. 233–246.
5. Abate, J. E., "Linear and Adaptive Delta Modulation," Proc. IEEE, March 1967, pp. 298–307.
6. Jayant, N. S., "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
7. Smith, B., "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 643–709.
8. Wilkinson, R. M., "An Adaptive Pulse Code Modulator for Speech," Proc. Int. Con. Commun., Montreal (June 1971), pp. 1–11 to 1–15.
9. Schlink, W., "A Redundancy Reducing PCM System for Speech Signals," Proc. Int. Zurich Seminar on Integrated Syst. for Speech, Video and Data Commun. (March 1972), pp. F4/1–4.
10. Max, J., "Quantization for minimum distortion," Trans. IRE, *IT-6*, March 1960, pp. 7–12.
11. Goodman, D. J., private communication.
12. Rosenberg, A. E., private communication.
13. Paez, M. D., and Glisson, T. H., "Minimum Mean Squared-error Quantization in Speech PCM and DPCM Systems" IEEE Trans. on Commun. (April 1972), pp. 225–230.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Heuristic Solution of a Signal Design Optimization Problem

By B. W. KERNIGHAN and S. LIN

(Manuscript received February 15, 1973)

This paper discusses a heuristic solution procedure for a combinatorial optimization problem that originates in designing signal constellations for modems.

The design problem is to place m signals in a two-dimensional space to minimize the average error rate under specified noise conditions, using a maximum-likelihood decoding scheme. Intuitively, it amounts (roughly) to spreading the signal points as far apart as possible, according to the distance measurement implied by the noise function.

We show how this problem can be reduced to a discrete one: Given an ℓ by n matrix P, and $m < \ell$, find an m-row subset $M = \{i_1, \cdots, i_m\}$ of the rows of P that maximizes

$$\sum_{j=1}^{n} \max_{i \in M} p_{ij},$$

and then describe an efficient procedure for finding this maximizing set.

Experiments indicate that the procedure is a useful tool, both for analysis of existing and proposed signal constellations and for finding new, near-optimum ones.

I. THE PHYSICAL PROBLEM

This paper discusses a heuristic procedure for solving a combinatorial optimization problem that arises in designing signal constellations for modems. The solution method is also applicable to the covering problem; we will discuss this at the end of this section.

The underlying physical problem is the following: A digital signal s is to be sent through a noisy channel. In general, s may take on only a finite number, m, of distinct values. (In practice, m will be a power of 2.) Since the transmission line is an analog device, any specific s value is encoded for transmission by modulating a carrier wave for a period of time. For instance, s might take on only the two values 0 and

1, in which case the modulation might be to send either of two amplitudes (amplitude modulation) or to send either of two frequencies (frequency modulation).

The modulation considered here is more complex: A specific value of s will be encoded as

$$g(t)[a \sin \omega_c t + b \cos \omega_c t],$$

where ω_c is the carrier frequency, g is an appropriate pulse shape, and a and b are the amplitudes of the sine and cosine components.

The signal s is quantized into one of m levels; for each level, there is a corresponding a and b, so there are m different (a, b) pairs.

The received signal, as always, is corrupted by noise, so a sample value sent as (a, b) is received as (a', b'). At the receiver, a decoder processes this corrupted (a', b') and attempts, according to some criterion for minimizing the error rate, to reconstruct the (a, b) which was sent originally.

The design question is: What (a, b) pairs should be chosen to minimize the error rate for this decoding process?

In the version of the problem we solve, the main constraint is that $a^2 + b^2$ is bounded for all (a, b) pairs, which implies that peak signal power is bounded. A related but more difficult problem requires that the average of $a^2 + b^2$ over the (a, b) pairs is constant; this corresponds to a bound on average power. We discuss this problem in Section VII.

The combinatorial optimization problem is a discrete version of this design question. Let us replace the continuum of points that could represent (a, b) values (everything inside the circle $a^2 + b^2 = 1$) by ℓ discrete points, spread more or less uniformly throughout the region. We call these "allowable" signal values. Since the noise may add to the received amplitude, the received signal can in fact be outside this circle. Let us define additional $n - \ell$ discrete points to represent the additional possible received signals that lie outside the circle $a^2 + b^2 = 1$.

Now define an ℓ by n matrix $P = \{p_{ij}\}$ by

p_{ij} = probability that, if signal i were sent, it would be received as (discrete) point j

$$i = 1, \cdots, \ell, \quad j = 1, \cdots, n.$$

Suppose for convenience that the chosen signal values are points 1 to m (i.e., rows 1 through m of P). The decoding procedure to be used is simply this: If j is the received signal, it is decoded as that i in $1, \cdots, m$ for which p_{ij} is maximum. If $1, \cdots, m$ have equal *a priori* probabilities, this procedure minimizes the error probability.

The probability that a particular signal i is decoded correctly is the probability that it falls into a column j where it is the largest entry. The probability of correct decoding using any particular set of m rows of P is thus the sum of the column maximum elements in those m rows (divided by m). The problem (at last) is to find those m rows that maximize this probability; these will be the signals used. More formally, find an m-row subset $M = \{i_1, \cdots, i_m\}$ of the rows of P that maximizes

$$V_M = \sum_{j=1}^{n} \max_{i \in M} p_{ij}.$$

We call V_M the *value* of the subset M.

In the physical problem, $m < \ell < n$; as an abstract problem, the latter inequality is unnecessary.

Note that the algorithm we will present is essentially insensitive to the characteristics of the matrix P. In practice, this means, for example, that any noise characteristics can be treated effectively. This includes not only the classical additive white noise, but also phase and amplitude jitter components.

As an example of a matrix with quite different characteristics, suppose P has entries which are either 0 or 1. A *covering problem* is "Find a minimum set of rows of P such that these rows together contain a 1 in each column of P." We can use our heuristic procedure to find approximate solutions for the covering problem as follows: Find a maximum value solution of the original problem, using m rows. If the value is less than the number of columns of P, increase m; if it equals the number of columns, decrease m. Find a new solution with the new m. The smallest value of m for which the value equals the number of columns is a minimum cover of P.

II. A HEURISTIC PROCEDURE

The process is based on iterative improvement of random initial solutions. A random set of m rows is chosen from the ℓ possible rows. (In practice, of course, we can also let the procedure try to improve on a specific initial set.) We augment the m initial rows by one row chosen from the $\ell - m$ unused rows, giving us $m + 1$ rows. We then compute which of these $m + 1$ rows contributes the least to the value of the set and remove it. (The row removed might well be the row added.) We then move to the next row in the unused ones and add it to the current m rows. The process terminates when all the $\ell - m$ currently unused rows have been examined without finding a profitable

replacement. This defines a local optimum solution. We then iterate the entire procedure from a new random start.

The resulting solutions are "1-opt" in the sense of Reference 1—that is, no exchange of a single pair of rows can improve the solution. Although 1-opt procedures are among the weaker heuristics, the results are quite acceptable, as we shall see in the next section.

The process is very fast which counteracts the lower effectiveness of 1-opting: a 100 by 100 problem takes about one second on the Honeywell 6070 (in FORTRAN A). The run time for dense matrices is essentially proportional to ℓn and independent of m. For sparse matrices (the situation that occurs in practice), the run time varies only with the number of non-zero elements, which for real problems is proportional to ℓn.

The process is fast, partly because care is taken to do no extra work. In detail, a basic step of the algorithm is as follows (the next section contains a numerical example, which can be followed in parallel):

Initialization. Suppose without loss of generality that the initial rows are $1, \cdots, m$. Call this set M. Let $v(i)$, $i = 1, \cdots, m$, be the *decrease* in value if row i is removed from M. We will compute v. This is done only once per local optimum solution.

We begin by setting $v(\cdot) = 0$. Each column j $(1 \leqq j \leqq n)$ contributes an increment to exactly one component of v, as follows. Find x_j and y_j, the largest and second-largest elements among the first m elements of column j. Record these, and also the rows in which they were found, i_x and i_y $(1 \leqq i_x, i_y \leqq m)$. Now, since x_j is the largest element in column j, it determines the contribution that column makes to the value of M. But if row i_x were removed from M, the contribution of column j would be determined by y_j, the second largest element. Thus, the decrease in value that would result if row i_x were removed from M is $x_j - y_j$, so we add $x_j - y_j$ to $v(i_x)$. This process is done for each column.

Phase 1—Evaluation of a Replacement Row. When the initialization is finished, we evaluate replacement rows. Suppose row r $(m + 1 \leqq r \leqq \ell)$ is the next proposed replacement. We will compute which of the $m + 1$ rows in $M_r = \{1, \cdots, m, r\}$ decreases the value of M_r least when removed. We will do this without examining any of the matrix P except for row r itself.

Let $\Delta(i)$ be the *change* in $v(i)$ that results if row i is removed from M_r. By computing Δ, we do not need to change v unless we are actually going to exchange two rows. Initially, let $\Delta(i) = 0$, $i = 1, \cdots, m, r$. (Let $v(r) = 0$ as well.) The value $\Delta(i)$ will be $\leqq 0$ for i in M while $\Delta(r) \geqq 0$.

For each column j, let $z = p_{rj}$, $x = x_j$, and $y = y_j$; we will do one of (*i*), (*ii*), or (*iii*):

(*i*) If $z \leqq y$, the new element is smaller than second best; no action is necessary.

(*ii*) If $y < z \leqq x$, z is a new second-best element. The Δ value for the row containing x, $\Delta(i_x)$, must be decreased by $z - y$, since the contribution to $v(i_x)$ from column j is now $x - z$ instead of $x - y$.

(*iii*) If $z > x$, we have a new largest element in the column. Add $z - x$ to $\Delta(r)$ (x is now second largest) and subtract $x - y$ from $\Delta(i_x)$, since row i_x no longer contains the largest element in this column.

Phase 2—Determination of Which Row to Remove. We have now determined $\Delta(i)$, the change in value that would result if row i were removed from M_r. Find the minimum among $v(1) + \Delta(1)$, $\cdots$, $v(m) + \Delta(m)$, $v(r) + \Delta(r)$.

If the minimum occurs at row $k \neq r$, let us say, then we must exchange rows r and k, update $v(\cdot)$, and update the records of largest and second-largest elements for each column. Go to Phase 3.

If this minimum occurs at r, there is no profit in replacing one of $1, \cdots, m$ by r. If $\ell - m$ rows have been consecutively examined without profit, we have finished; the set M is the local optimum solution. Otherwise, we must set r to $r + 1$ (wrapping around from ℓ to $m + 1$ if necessary) and go back to the Phase 1 calculation.

Phase 3—Updating After Exchange of Two Rows. As in Phase 1, we must perform one of (*i*), (*ii*), or (*iii*) below for each column. The element x is the largest in M, y is the second largest, and z is the new element from row r. Row k is the row being ejected ($1 \leqq k \leqq m$).

Case (*i*): $z \leqq y$. If $k \neq i_x$ and $k \neq i_y$, then we are not replacing either of the two largest elements in this column, so no updating is necessary. Go to the next column.

If $k = i_x$, we are replacing the largest element with something no better than third largest. We replace x by y (and i_x by i_y) and find a new number two element in M_r—call it w. Since y is now largest, we subtract $w - y$ from $v(i_y)$ and then let $i_y = i_w$.

If $k = i_y$, we are replacing the number two element with something no better than third largest. Again we search for w, the new number two, update $v(i_x)$ by subtracting $w - y$, and update i_y.

Case (*ii*): $y < z \leqq x$. If $k = i_x$, we are replacing the largest element with a new and smaller largest element. The element z replaces x as the largest element, and $\Delta(r)$ is increased by $z - y$.

If $k \neq i_x$, z becomes the new number two element, and $z - y$ is subtracted from $v(i_x)$ to reflect the smaller difference between first and second elements.

Case (*iii*): $x < z$. If $k = i_x$, we are replacing the largest element with a new largest element. The value $\Delta(r)$ is augmented by $x - y$ (it already contains $z - x$ from Phase 1). The element z replaces x.

If $k \neq i_x$, we push x down into second place, since z is now the largest element, and subtract $x - y$ from $v(i_x)$, since x no longer contributes.

After this update has been done for each column, we copy $\Delta(r)$ into $v(k)$ and interchange rows r and k. (In the actual implementation, of course, row movement is just pointer manipulation.) Now go back to Phase 1.

This is the end of the algorithm description. The critical part of this algorithm is evaluating the contribution of a row without doing any of the updating necessary to exchange it, and particularly without scanning the matrix to find any column maxima. This latter operation need be performed only after we have decided upon an exchange; furthermore, it is performed only upon a small set of columns—those for which the element of the row being replaced was first or second largest and for which the replacement element is smaller than both [case (*i*) in Phase 3, above]. In practice, this condition holds for about 10 percent of the columns when we actually do a replacement. Since typically we replace relatively few rows in proportion to the number examined, the time saving is large.

III. AN EXAMPLE

This description may be clarified by one step of an example. Suppose $m = 3$, the cost matrix is

$$P = \begin{bmatrix} 1 & 3 & 3 & 6 & 1 & 6 \\ 2 & 2 & 8 & 7 & 2 & 0 \\ 1 & 5 & 3 & 4 & 4 & 1 \\ 4 & 2 & 4 & 5 & 8 & 2 \\ 7 & 9 & 1 & 1 & 3 & 5 \end{bmatrix},$$

and the first three rows are the current set. (This is obviously not a probability matrix—small integers are better for exposition.) Then the best entries are (2, 5, 8, 7, 4, 6), a value of 32.

Initialization. We find that v_1 is 5 (from column 6), v_2 is 7 (from columns 1, 3, and 4), and v_3 is 4 (from columns 2 and 5). Thus $v = (5, 7, 4)$. Suppose we want to evaluate row 4 as a replacement for one of these rows.

Phase 1. Considering column 1, $4 > 2$ so we are doing case (*iii*). The value Δ_4 is augmented by 2; at the same time, Δ_2 is decreased by 1, because the element in row 2 is no longer largest in column 1.

There is no change in column 2 [case (*ii*)]. In column 3, row 4 represents a new second-largest element, so Δ_2 is decreased by 1. There is no change in column 4. In column 5, add 4 to Δ_4, and subtract 2 from Δ_3. In column 6, decrease Δ_1 by 1. Thus, $\Delta = (-1, -2, -2, 6)$.

Phase 2. The minimum of $v_i + \Delta_i$ is 2, at $i = 3$, implying that row 3 should be ejected. We now commence the updating operation.

Phase 3. For column 1, the largest value is 4 (coincidentally in row 4) and the second largest is 2 (in row 2). Decrease v_2 by 1, since row 2 no longer contributes in this column [case (*iii*)]. In column 2, we are replacing the largest element by a very small one [case (*i*)]; the old number two element becomes the new largest, and we have to search for the new second largest. Add 1 to v_1, since the largest value is a 3 and the second largest a 2. In column 3, $p_{4,3}$ represents a new and bigger second element; decrease v_2 by 1. No change takes place in column 4. In column 5, we are replacing the largest element by a new largest element. The value of v_4 is increased by $2(p_{5,3} - p_{5,2})$. In column 6, we gain a new second element, so v_1 is decreased by 1.

En route, we compute the new value of the solution as 36.

When we have finished, $v = (5, 5, 8)$ for rows 1, 2, and 4, and the process continues by our considering row 5. Notice that out of six columns we only had to search for a second-largest element *once*; the rest of the time, it was immediately at hand.

IV. EXPERIMENTAL RESULTS

We have tried the procedure on several different types of data: matrix entries random on $\{0, \cdots, n\}$, random $0 - 1$ matrices, and various probability matrices based on the physical problem.

For problems formed by generating random entries in the matrix, the fraction of random starts producing the optimum diminishes as m/ℓ increases (except for the trivial cases of very small or very large m), and also as n increases. Although there is significant individual variation, the frequency of obtaining the optimum is close to 100 percent for small problems and still about 20 percent for problems with $m = 10$, $\ell = 60$, $n = 80$, the largest random problems tried. (It should be noted that "optimum" usually means "best solution seen in a large number of trials"; we have strong statistical grounds for believing them optimal, but no proof.)

TABLE I—TYPICAL STATISTICS FOR SMALL RANDOM PROBLEMS (ENTRIES UNIFORM ON $[0, n]$)

m	ℓ	n	Time (ms)	No. of optimums/ Total Trials (range)	No. of Distinct Solutions (range)
5	10	20	21	20/20	1
5	40	40	120	32–39/50	6–9
5	40	80	230	6–47/50	4–19
10	20	20	42	20/20	1
10	40	40	140	25–41/50	3–5
10	40	80	300	11–16/50	5–10
20	60	60	340	18–41/50	2–5
20	60	80	450	12–42/50	7–13

Run time is directly proportional to ℓn; in absolute terms, the run time is about 100 ℓn microseconds per random start.

The procedure almost always makes less than two passes through the $\ell - m$ unused elements; the number of row replacements is roughly equal to m. As we mentioned above, it is only on these occasions that it is necessary to actually update the records of first- and second-best elements, and only for about 10 percent of the columns among the replacements is it necessary actually to scan through the m rows to locate a new second largest. (After initialization, it is never necessary to scan to find the largest.) Table I shows some typical results for these smaller tests.

Limited tests on 0–1 matrices produced similar results, although the run time appears to be slightly lower per case. It is possible in the 0–1 case to make several simplifications that would further decrease run time, and storage requirements could be drastically reduced by

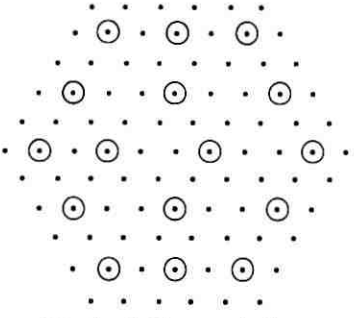

Fig. 1—Optimum solution.

using bit storage. However, we have not experimented extensively on 0–1 matrices. This will be reported in a separate paper.

Several experiments have been performed on various models of the real problem. Since the modems involved have very low error rates, the error transition probability p_{ij}, $i \neq j$, is small. This means that the matrix P has large diagonal elements, a few small elements, and a large number of zeros. For our purposes, the existence of large values is irrelevant; however, many zeros means that a storage organization that does not store zeros is attractive. We will discuss this shortly.

One problem to be faced is how to represent the continuous (a, b) space as a set of discrete points. One crude model we studied uses a "honeycomb" or hexagonal scheme (shown in Fig. 1), since this is a reasonable approximation to circular symmetry. The inner 61 points represent allowable signal locations; the outer 30 are the extra points to take care of the set of received signals that violate the peak power constraint.

For this test, P is defined as

$$\begin{aligned} p_{ij} &= \epsilon \text{ if } j \text{ is a neighbor of } i \ (\epsilon \ll 1) \\ &= 1 - (\epsilon \times \text{number of neighbors}) \text{ if } i = j \\ &= 0 \text{ otherwise.} \end{aligned}$$

This set of probabilities ignores phase jitter, which adds a radially increasing tangential component to the error transitions.

Figure 1 shows an optimum solution (it is easy to prove it optimum); Fig. 2 shows a local optimum differing by ϵ. In both cases, 12 signals are placed symmetrically on the boundary, and the remaining four are placed as well as possible in the interior. For this problem, all solutions were within 9ϵ of optimum and the median within 3ϵ (the mean random start is about 35ϵ away), run times averaged 370 ms per case, and

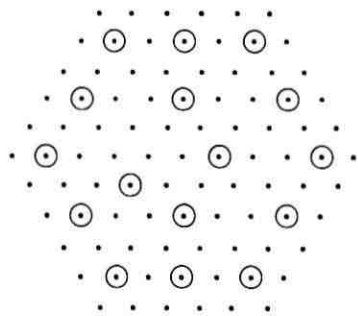

Fig. 2—Suboptimal by one unit.

the frequency of optimum solutions was about 10 percent. This problem appears to be slightly harder than random problems of this size, but run times are smaller.

V. LARGER PROBLEMS

To properly handle larger problems (e.g., to increase the resolution available when discretizing) a version of the program using a sparse matrix representation has been implemented. This involves substantial overhead in accessing elements of the matrix, but it is balanced by the fact that, when most matrix elements are zero, much less processing is required. For example, in the 16/61/91 (i.e., $m/\ell/n$) problem above, average run time increased from 370 to 390 ms, which is not significant. The run time is determined predominantly by the number of non-zero points, which is usually proportional to ℓn.

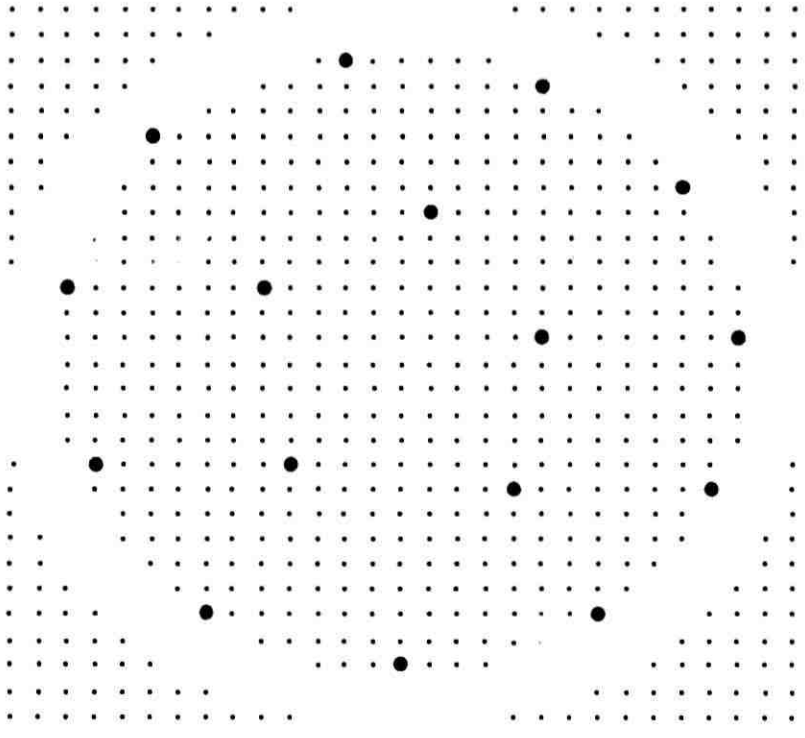

Fig. 3—"5–11" solution in pure Gaussian noise; $\beta = 0.0$; error rate $= 1.07 \times 10^{-7}$.

TABLE II—SOME RESULTS ON DESIGN PROBLEMS
($m = 16$, $N_0 = 0.002$)

	ℓ	n	Non-zero Entries	Run Time (s)	Phase Jitter β (degrees)	Character of Best Solution	Figure
1.	293	421	8,500	6.5	1.5	1–5–10	–
2.	421	577	17,000	14.5	1.5	1–6–9	–
3.	421	577	17,000	14.5	1.5	1–5–10	–
4.	489	665	22,000	19.3	0	5–11	3
5.	489	665	22,500	18.4	1.5	1–5–10	4
6.	489	665	27,000	20.5	3.0	1–6–9	5
7.	489	665	27,000	21.0	3.0	Special design	6
8.	577	749	31,000	24.9	1.5	1–5–10	–

VI. RESULTS ON REAL PROBLEMS

In this section we discuss some of the experimental results obtained from real problems, using formulas for the transition probabilities taken from Reference 2. The probability for the transition from $X = (x_1, x_2)$ to $Y = (y_1, y_2)$ has the general form

$$p(X, Y) = f(X, Y; \beta, N_o) \exp [g(X, Y; \beta, N_o)],$$

where N_o is the noise power of the channel and β is the rms phase jitter in the received signal. The details of f and g do not concern us here; it is sufficient to say that $p(X, Y)$ drops to zero rapidly as Y gets further from X. For example, in the pure Gaussian noise case ($\beta = 0$), we have

$$p(X, Y) = \frac{1}{2\pi N_o} \exp \left[- \frac{\|Y - X\|^2}{2N_o} \right].$$

Thus, the probability matrix is quite sparse when the transition probabilities have been scaled and converted to integers.

The discrete space consists of points on a square lattice, as shown in Fig. 3. The number of rows in P is determined by the number of lattice points within the circle of radius 1. To these are added exterior points approximating all possible additional received points. Since the radial probabilities drop off rapidly, this exterior layer need only be one or two units thick. This extra layer is indicated by the band of omitted points on Fig. 3.

The run time and the storage requirements both grow with the number of lattice points; this limits the resolution we can use. The largest problem tried had 665 rows, 861 columns, and about 40,000

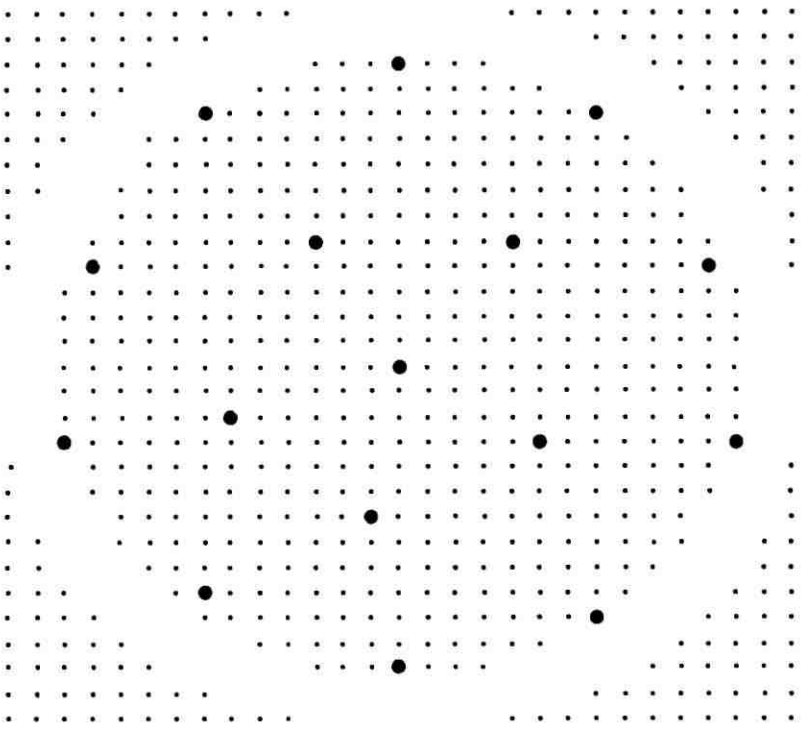

Fig. 4—"1–5–10" solution for 1.5-degree jitter (error rate = 2.97×10^{-7}).

non-zero matrix entries. It should be noted that the matrix has a fourfold symmetry, so, at the price of an increase in computing time (in practice, about 50 percent), only 10,000 entires need be stored.

Two types of experiments were performed. First, several constellations of intrinsic interest were used as initial solutions; the heuristic procedure attempted to improve upon them. Second, the procedure was used to produce good solutions from a large number of random initial configurations. For all solutions, approximate error rates were computed and the constellations displayed.

Table II lists some typical parameters for several experiments at various sizes; Figs. 3 to 5 show the best solutions found for particular parameter settings.

Each signal point is surrounded by a set of points which, when received, will decode into that signal point. This set of points is the "decision region" for that signal point. Because we have quantized a continuous space into small squares, the decision regions surrounding

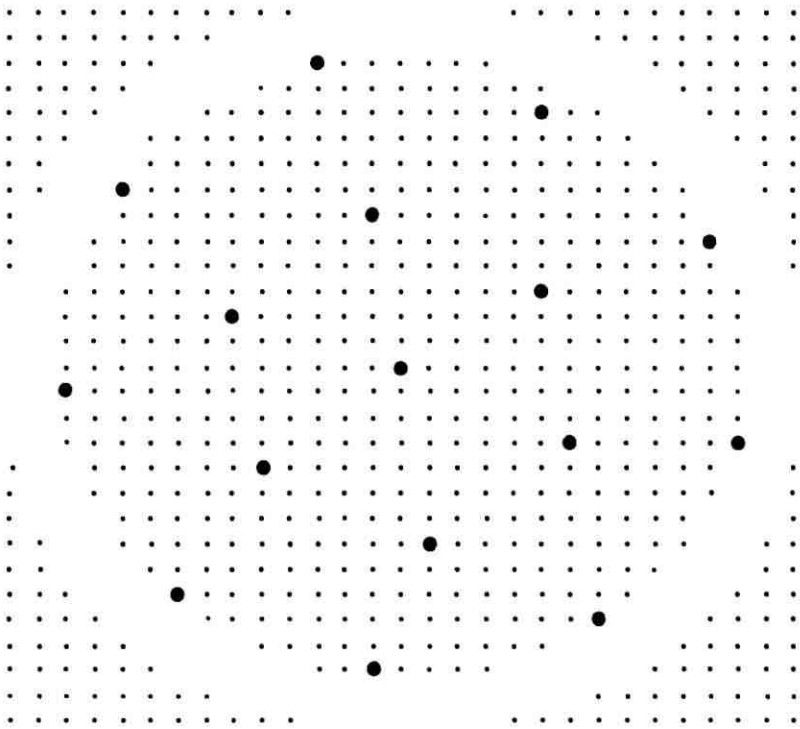

Fig. 5—"1–6–9" solution for 3-degree jitter (error rate = 3.26×10^{-6}).

each signal point have "ragged edges" and are of necessity somewhat arbitrary. As the resolution is made finer, this effect is less serious, and in fact the procedure can make more subtle choices of points and of boundaries, so the apparent error rate decreases with increasing resolution. For this reason, error rate comparisons between different resolutions are not appropriate. However, the rates are internally consistent in that, for any given resolution, the solution character and error rates vary with noise as would be expected.

As predicted by analytic techniques,[2] solutions like "1–5–10" and "1–6–9" are better for high jitter ($\beta > 1.5°$), while "5–11" solutions are better in low jitter cases. (The notation will be evident after examining the figures.) These trends are clearly indicated in Figs. 3 through 5, which show, respectively, the best solution (a 5–11) for zero phase-jitter (pure Gaussian noise) with an error rate of 1×10^{-7}, the best solution (1–5–10) for 1.5 degrees of phase jitter (error rate

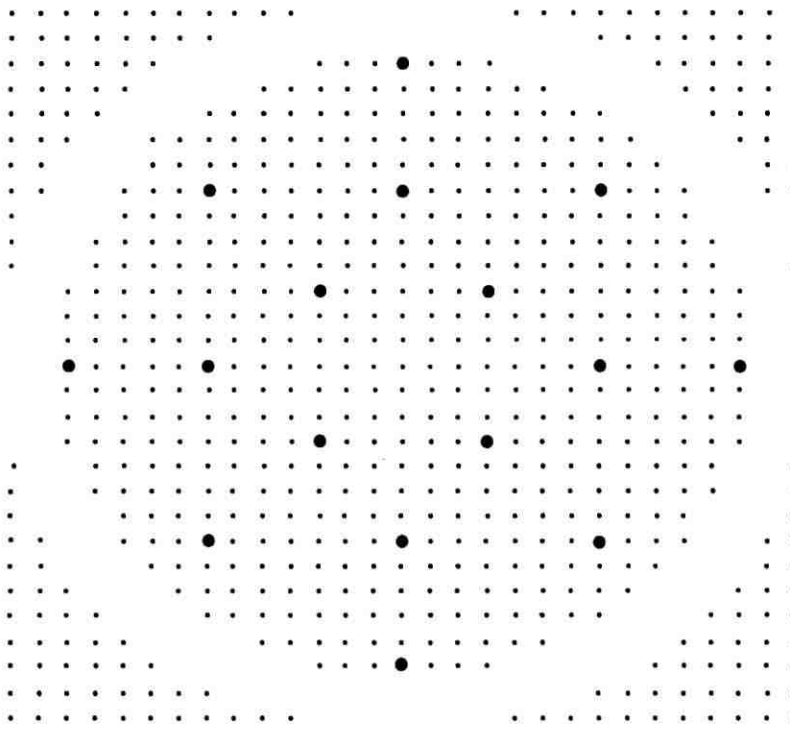

Fig. 6—Competing design, for 3-degree jitter (error rate = 3.47×10^{-5}).

3×10^{-7}), and the best solution (1–6–9) for 3 degrees of jitter (error rate 3×10^{-6}).

For comparison, we experimented with the competing design shown in Fig. 6, at various levels of jitter. This design is intended to be robust over a wide range of jitter. This configuration does degrade less than the others as jitter increases, but its overall performance is very much inferior. Figure 6 shows that, at 3 degrees, its error rate is 3.5×10^{-5}, a factor of 10 worse than the 1–6–9 configuration. These results agree closely with predictions of independent theoretical studies.[2]

VII. CONCLUSIONS

As a solution to a combinatorial optimization problem, the heuristic procedure presented here is quite good for small-to-medium problems, say up to about 100 rows, even for dense matrices. It remains useful, but not strong, for large sparse problems.

As for the original design problem, the number of rows and columns in the matrix both rise with the resolution; thus, highly accurate representations are computationally expensive, so the procedure is generally not appropriate for generating precise answers to specific design questions. Rather, it is most useful in providing quick approximate and comparative optimizations or evaluations, either to furnish insight or to supplement results obtained by analytic techniques.

The extension of this technique to problems with an average power constraint, rather than peak power, appears to be straightforward, although we have not implemented it. [The average power constraint requires that $\sum(a^2 + b^2) \leqq 1$.]

As the simplest solution, start with a random feasible set of m rows. Then, before each possible replace row is selected, test to see if it would violate feasibility; if so, it cannot be used. A more powerful algorithm would permit temporary violations of feasibility in a controlled way. Either of these approaches should serve reasonably well.

VIII. ACKNOWLEDGMENTS

The authors are indebted to G. Foschini for originally suggesting the problem and for providing data and physical insight. We are also indebted to R. Gitlin for several valuable lectures on communication theory.

REFERENCES

1. Lin, S., "Computer Solutions of the Traveling Salesman Problem," B.S.T.J., *44*, No. 10 (December 1965), pp. 2245–2269.
2. Foschini, G., Gitlin, R., and Weinstein, S., "On the Selection of a Two-Dimensional Signal Constellation in the Presence of Phase Jitter and Gaussian Noise," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 927–965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Impulse Response of Fibers With Ring-Shaped Parabolic Index Distribution

By D. GLOGE and E. A. J. MARCATILI

(Manuscript received March 6, 1973)

The index distribution in the cross section of a multimode fiber has an important influence on the modal group velocities and, hence, on the fiber impulse response. In this paper we derive a method for the evaluation of arbitrary circular symmetric index profiles. In particular, we compute the impulse response of a fiber with a ring-shaped parabolic index profile which exhibits useful equalizing properties. The pulse spread is found to be nearly one order of magnitude smaller than that of a fiber with an equal, but abrupt, index decline from core to cladding.

I. INTRODUCTION

Multimode operation of optical fibers relaxes the fabrication tolerances, allows the use of incoherent sources, and can alleviate handling and splicing problems. Modal (group) delay differences are nearly equalized[1,2] if the core index decreases as the square of the fiber radius from a maximum at the axis (Fig. 1). A distribution of this kind is realized in the Selfoc* fiber, which was indeed reported to have very low values of differential mode delay.[3,4]

Since then, the question has been raised whether there are other index profiles which have similar equalizing effects, but are otherwise perhaps more amenable to certain fabrication techniques or have advantages with respect to splicing or bending. Although the latter part of this question is difficult to answer at this time, it is certainly possible to identify at least one profile that has quite effective equalizing properties. Imagine a slab with a square-law index distribution in transverse direction. The group velocities of all its modes are known to be nearly equal.[1] It is then plausible to expect that these properties are approximately preserved if the slab is warped in a way which results in a tube with the cross-sectional index distribution shown in

* Registered trademark of Nippon Electric Co., Ltd. and Nippon Sheet Glass Co., Ltd.

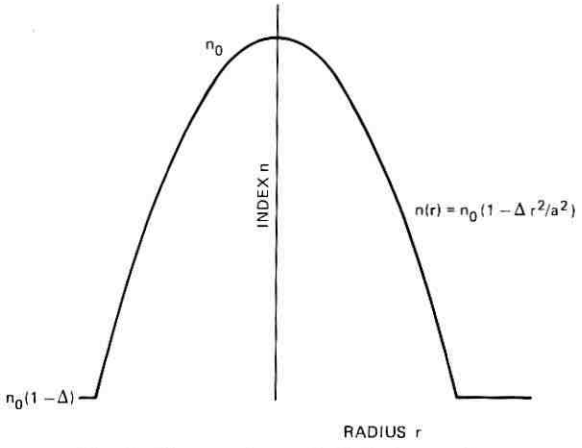

Fig. 1—Concentric parabolic index profile.

Fig. 2. Assume that a cladding material of lower refractive index fills the bore and surrounds the tube to the outside, so that a fiber is formed which guides modes within a tube-like structure with parabolic index distribution.

The purpose of this paper is to identify the modes of this structure, calculate their group velocities, and predict the impulse response to be expected when all modes propagate uncoupled and with equal power. To do this, we employ the WKB description[5] in a form which ignores the anomalies of dielectric waveguide modes near cutoff, assuming that few of all the propagating modes are close to this condition. For the sake of simplicity, we also restrict the following computations

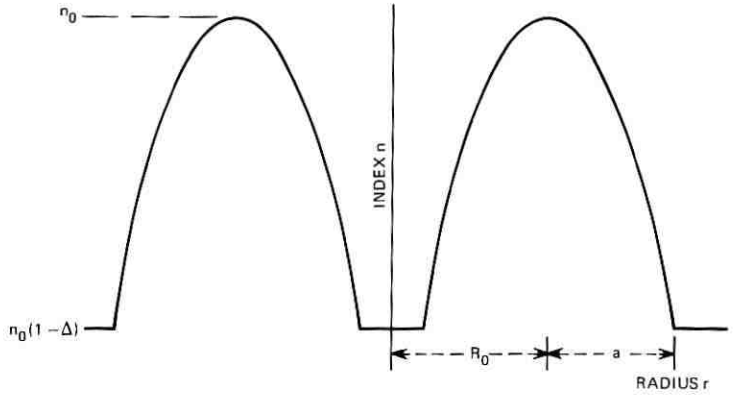

Fig. 2—Cross section through a ring-shaped parabolic index profile. Maximum index n_0 along a circle of radius R_0. Cladding index $n_0(1 - \Delta)$.

to small index variations, so that all propagation directions can be assumed paraxial to the waveguide axis and the corresponding approximations apply.

II. A CHARACTERISTIC EQUATION FOR CIRCULAR SYMMETRIC INDEX DISTRIBUTIONS

Let us adopt a cylindrical coordinate system (r, ϕ, z) and assume that the refractive index n is a function of r only. We define a local wave number

$$k(r) = 2\pi n(r)/\lambda, \tag{1}$$

where λ is the wavelength in free space. Because of the circular symmetry, we can separate the general wave equation and solve for ϕ and z. In doing so, we define an axial propagation constant β and describe the azimuthal periodicity by an azimuthal mode number ν. The remaining partial differential equation for the radial field dependence $E(r)$ has then the form

$$\frac{\partial^2 E}{\partial r^2} + \frac{1}{r}\frac{\partial E}{\partial r} + \left(k^2(r) - \beta^2 - \frac{\nu^2}{r^2}\right) E = 0. \tag{2}$$

Following the usual WKB approach,[5] we substitute

$$E(r) = e^{u(r)}, \tag{3}$$

ignore the second derivative $\partial^2 u/\partial r^2$, and, by solving for $\partial u/\partial r$, we obtain the solution

$$\frac{\partial u}{\partial r} = -\frac{1}{2r} \pm i\sqrt{k^2(r) - \beta^2 - \left(\nu^2 + \frac{1}{4}\right)\Big/ r^2}. \tag{4}$$

Given β and ν, we can find two radii, R_1 and R_2, at which the root in (4) vanishes (Fig. 3). These radii define a ring-shaped region within which eq. (4) has an imaginary part causing the field E to be a periodic function. Outside of the region, E decreases or increases aperiodically.

As in the 2-dimensional case,[5] decreasing (or evanescent) field characteristics outside are obtained if the total phase inside the region is

$$\int_{R_1}^{R_2} \sqrt{k^2(r) - \beta^2 - (\nu^2 + \tfrac{1}{4})/r^2}\, dr = (\mu + \tfrac{1}{4})\pi, \tag{5}$$

where μ is an integer called the meridional mode number. It determines the number of half periods of E in radial direction. The accuracy of (5) improves for large μ, but is in most cases surprisingly good even for small values of μ. Equation (5) permits an evaluation of the propaga-

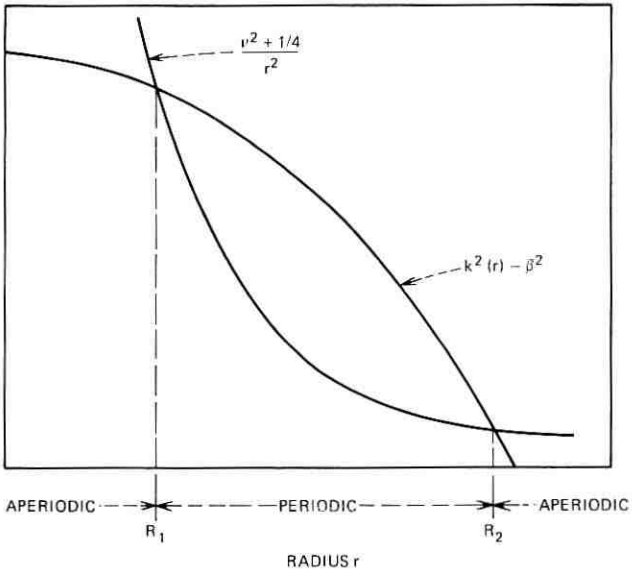

Fig. 3—Sketch defining regions of periodic and aperiodic field characteristics of a mode of azimuthal order ν.

tion constant β for given mode numbers μ and ν. This will now be done for the parabolic ring structure sketched in Fig. 2.

III. GROUP DELAY AND IMPULSE RESPONSE

Figure 2 shows a cross-sectional view of the circular symmetric index distribution. The index has a maximum value n_0 at $r = R_0$, decreases as

$$n(r) = n_0[1 - \Delta(r - R_0)^2/a^2] \quad \text{for} \quad R - a < r < R + a, \quad (6)$$

and has a constant value

$$n(r) = n_0(1 - \Delta) \quad (7)$$

everywhere else. We assume Δ to be small compared to unity, introduce the abbreviations

$$k_0 = 2\pi n_0/\lambda \quad (8)$$

and

$$\rho = r - R_0, \quad (9)$$

and obtain, with the help of (1) and (6),

$$k^2(r) \approx k_0^2(1 - 2\Delta\rho^2/a^2). \quad (10)$$

In order to solve eq. (5) analytically, we assume in addition that $R_0 \gg a$, which permits us to replace r by R_0 in (5). As a result,

$$(\mu + \tfrac{1}{4})\pi = \int_{R_1}^{R_2} \sqrt{k_0^2 - \beta^2 - (\nu^2 + \tfrac{1}{4})/R_0^2 - 2\Delta k_0^2 \rho^2/a^2}\, dr, \quad (11)$$

which has the solution

$$\beta = [k_0^2 - (\nu^2 + \tfrac{1}{4})/R_0^2 - 2\sqrt{2\Delta}(\mu + \tfrac{1}{4})k_0/a]^{\frac{1}{2}}. \quad (12)$$

The phase constant β of a propagating mode must furthermore fulfill the condition

$$\beta \leqq k_0(1 - \Delta) \quad (13)$$

for the cladding field to have evanescent characteristics. This permits us to calculate the total number of propagating modes. Keeping ν fixed, we first determine the number of modes m in a group with the same ν. We do this by solving (12) for μ with $\beta = k_0(1 - \Delta)$. Since $\Delta \ll 1$

$$m = \mu_{\max}(\nu) + 1 = \sqrt{\Delta/2}ak_0 - \frac{a(\nu^2 + 1/4)}{2k_0R_0^2\sqrt{2\Delta}} + \frac{3}{4}. \quad (14)$$

This number decreases as ν increases. The largest possible ν is obtained for $m = 1$. Thus with (14)

$$\nu_{\max} = \left(2k_0^2R_0^2\Delta - \frac{k_0R_0^2}{a}\sqrt{\Delta/2}\right)^{\frac{1}{2}} - \frac{1}{4}. \quad (15)$$

For the following approximations, we ignore the terms $\frac{1}{4}$. In this case, the sum over all m from $\nu = 0$ to $\nu_{\max}$ yields

$$M = \tfrac{2}{3}ak_0^2R_0\Delta, \quad (16)$$

which is the total number of propagating modes. Using the same approximations, we can express m with the help of (15) in the form

$$m = \sqrt{\Delta/2}ak_0(1 - \nu^2/\nu_{\max}^2), \quad (17)$$

an expression which will be used later on.

To calculate the mode delay, we first convince ourselves with the help of (14) and (15) that the μ- and ν-terms in (12) are small (of the order Δ) compared to k_0^2. We therefore approximate β by

$$\beta = k_0 - (\nu^2 + \tfrac{1}{4})/2k_0R_0^2 - \sqrt{2\Delta}(\mu + \tfrac{1}{4})/a. \quad (18)$$

The differentiation of β with respect to the radial frequency

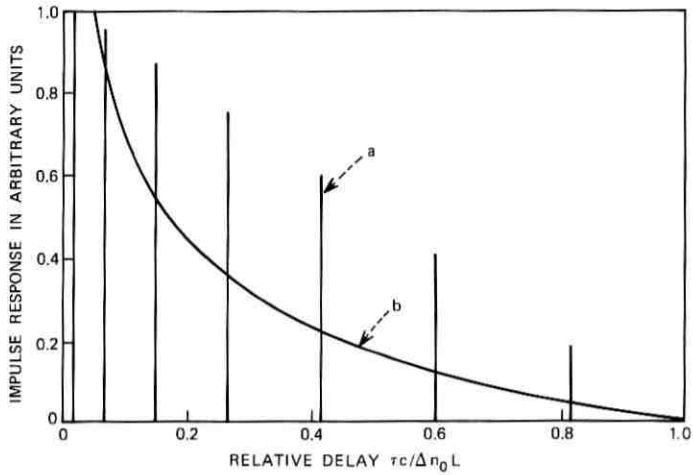

Fig. 4—Impulse response of the thin parabolic tube structure; (*a*) individual mode groups, (*b*) power distribution for large mode numbers ($R_0 \gg a$).

$\omega = ck_0/n_0$ yields the group delay

$$t = \frac{L}{c} n_0 \left[1 + \left(\nu^2 + \frac{1}{4} \right) \Big/ 2k_0^2 R_0^2 \right], \tag{19}$$

where L is the fiber length and c the velocity of light in free space. Since t depends only on ν but not on μ, mode groups with the same ν have the same delay. Consequently, if all modes are excited by equal pulses of unit energy at the fiber input, the output consists of pulses of energy $m(\nu)$ delayed by $t(\nu)$. If we ignore the delay Ln_0/c common to all modes and then insert (15) into (19), again neglecting the terms $\frac{1}{4}$, we can write the delay in the form

$$\tau(\nu) = t - \frac{Ln_0}{c} = \frac{Ln_0\Delta}{c} \frac{\nu^2}{\nu_{\max}^2}. \tag{20}$$

Figure 4 illustrates the output distribution for the case in which the pulses are so narrow that individual groups are resolved. All pulses have the same (very small) width, and their heights correspond to the total energy m in each group.

More meaningful than this plot is a plot of the energy per unit time

$$p(\tau) = m d\nu/d\tau \tag{21}$$

which coincides with the power distribution in the case of very large mode numbers. We find $d\tau/d\nu$ by differentiating (20) and, if the term

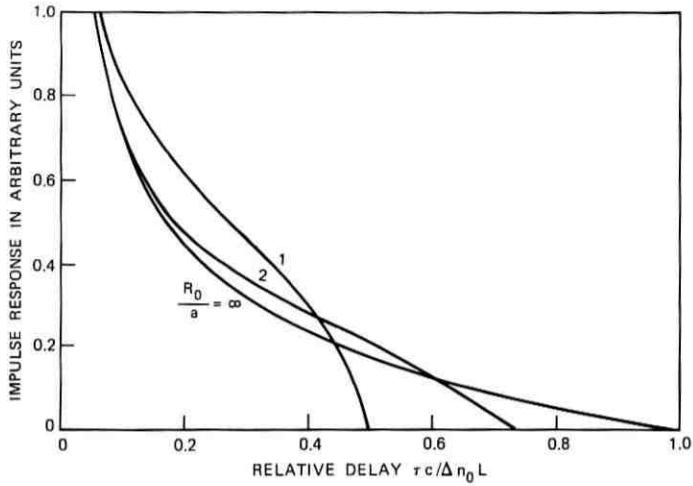

Fig. 5—Impulse response of the parabolic tube structure for various radius-to-thickness ratios.

$\frac{1}{4}$ in (14) is again ignored, we have

$$p(\tau) = \frac{caR_0k_0^2}{\sqrt{2}Ln_0}\left[\left(\frac{\Delta Ln_0}{c\tau}\right)^{\frac{1}{2}} - \left(\frac{c\tau}{\Delta Ln_0}\right)^{\frac{1}{2}}\right]. \tag{22}$$

This function is also plotted in Fig. 4. We find half the total output concentrated in the time interval

$$T = 0.12\Delta n_0 L/c. \tag{23}$$

The remaining power is drawn out in a tail of length $\Delta n_0 L/c$. This tail is caused by modes of high azimuthal order ν, modes which are not present in the 2-dimensional structure and are essentially equalized in the case of the concentric parabolic profile (Fig. 1). In this respect, the parabolic ring structure is inferior to the corresponding concentric profile, yet an effective width of $0.12\ \Delta n_0 L/c$ may be a useful improvement in comparison to a guiding structure with uniform index n_0 which theoretically produces a width $\Delta n_0 L/c$.

The condition $R_0 \gg a$ was necessary for an analytic solution of the integral (11). Exact numerical results for arbitrary ratios R_0/a are shown in Fig. 5. The corrections with respect to (22) are largest for high azimuthal orders. An exact analytical solution can only be found for the maximum delay $\tau(\nu_{max})$ which becomes

$$\tau(\nu_{max}) = \frac{\Delta n_0 L}{c}\left[1 - \frac{1}{8}\left(\sqrt{\frac{R_0^2}{a^2} + 8} - \frac{R_0}{a}\right)^2\right]. \tag{24}$$

As an example, consider a parabolic ring whose half width, a, is equal to the central radius R_0. Let $n_0 = 1.5$ and assume $\Delta = 1$ percent. The total width of the impulse response after 1 km of this fiber would be $\tau(\nu_{max}) = 25$ ns according to (22), but half the power is concentrated within the first 6 ns. Since high-order modes are usually lossier than the low orders, it is likely that much of the pulse tail does not reach the fiber end.

IV. CONCLUSIONS

The WKB approximation yields a simple characteristic equation for the propagating modes in fibers with arbitrary circular symmetric index distribution. We use this method to compute the impulse response of a fiber with a ring-shaped parabolic index distribution. We find that this structure has equalizing properties similar to the concentric parabolic index distribution, except for certain azimuthal mode orders, which lag behind, forming a rather long pulse tail. The rest of the power is concentrated in a time interval which, for a 1-km length and a relative index difference of 1 percent, is only 6 ns.

V. ACKNOWLEDGMENTS

Stimulating discussions with S. E. Miller and Mrs. W. Mammel's help in the numerical computation are gratefully acknowledged.

REFERENCES

1. Miller, S. E., "Light Propagation in Generalized Lens-Like Media," B.S.T.J., *44*, No. 9 (November 1965), pp. 2017–2064.
2. Kawakami S., and Nishizawa, J., "An Optical Waveguide with the Optimum Distribution of the Refractive Index with Reference to Waveform Distortion," IEEE Trans. on Microwave Theory and Tech., *MTT-16* (October 1968), pp. 814–818.
3. Uchida, M., Furukawa, M., Kitano, I., Koizumi, K., and Matsumura, H., "A Light-Focussing Fibre Guide," IEEE J. Quantum Electronics (Digest of Tech. Papers), *QE-5* (June 1969), p. 331.
4. Gloge, D., Chinnock, E. L., and Koizumi, K., "Study of Pulse Distortion in Selfoc Fibers," Electron. Letters, *8* (October 19, 1972), No. 21, pp. 526–527.
5. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics*, New York, McGraw-Hill, 1953, p. 1092.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

The Impulse Response of an Optical Fiber With Parabolic Index Profile

By D. MARCUSE

(Manuscript received March 6, 1973)

To the paraxial approximation there is no difference in the group delay of the modes of a parabolic index fiber. However, the wave optics treatment of the infinitely extended parabolic index medium predicts a slight difference in the group delay of the various modes. This result is used in this paper to predict the shape and width of the impulse response function of a parabolic index fiber with finite radius.

I. INTRODUCTION

The current interest in multimode optical waveguides is related to progress in the fabrication of luminescent diodes which have become cheap and dependable sources of incoherent light. Since incoherent light cannot be injected into a single-mode fiber with high efficiency, multimode waveguides must be used. A disadvantage of using multimode instead of single-mode waveguides is multimode pulse dispersion caused by the fact that the group velocity of the guided modes is not the same. Power injected at one end of the waveguide is shared by many or all of the possible guided modes. As each mode reaches the other end of the guide at a different time, the initial pulse is broadened.

It is the purpose of this paper to calculate the impulse response of a graded-index, multimode fiber.[1-3] The index distribution is assumed to be given by the expression

$$n = n_o\left(1 - \Delta\frac{r^2}{a^2}\right) \qquad 0 \leqq r \leqq a. \tag{1}$$

The parameter a represents the finite radius of the fiber, Δ determines the strength of the index gradient. The analysis is simplified by using the modes of the infinitely extended square-law medium (1) instead of the modes of the actual waveguide of radius a.

The impulse response of a graded-index fiber with parabolic index profile is much more favorable than that of the usual clad fiber with a

rectangular index profile. In the paraxial approximation it is a delta function. The finite width of the impulse response function is attributable to rays that move on trajectories making relatively large angles with the waveguide axis.

II. THE MODES OF THE SQUARE-LAW MEDIUM

The electric or magnetic field components of the modes of the infinitely extended square-law medium are obtained from the reduced wave equation

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + (n^2\mathrm{k}^2 - \beta^2)\Psi = 0, \tag{2}$$

with

$$n^2 = n_o^2\left(1 - 2\Delta\frac{x^2 + y^2}{a^2}\right). \tag{3}$$

The coordinates x and y are oriented in transverse direction to the waveguide axis which points in z direction; $k = 2\pi/\lambda$ is the propagation constant of light in vacuum and β is the propagation constant of the guided modes. The distribution of the square of the refractive index is not simply the square of the index distribution (1). Equation (3) follows from (1) only if we neglect the square of the Δ term. However, we can turn the argument around and consider eq. (3) as correct and (1) as the approximation. The exact solution of (2) with the function (3) is given by[4]

$$\Psi = AH_p\left(\sqrt{2}\,\frac{x}{w}\right)H_q\left(\sqrt{2}\,\frac{y}{w}\right)\exp\left(-\frac{x^2 + y^2}{w^2}\right)e^{-i\beta z}. \tag{4}$$

H_p is the Hermite polynomial of order p. The beam half-width is defined as

$$w = \left(\frac{2}{\Delta}\right)^{\frac{1}{2}}\left(\frac{a}{n_o\mathrm{k}}\right)^{\frac{1}{2}} \tag{5}$$

and the propagation constant is given by

$$\beta = n_o\mathrm{k}\left[1 - 2\frac{\sqrt{2\Delta}}{n_o\mathrm{k}a}(p + q + 1)\right]^{\frac{1}{2}}. \tag{6}$$

However, eqs. (4) through (6) are not an exact solution of Maxwell's equations since an additional term containing the gradient of n^2 has been neglected in (2). Since this additional term does not make a significant contribution in (6)—particularly at large mode numbers p and q—we use the solution in its present form for our discussion of the impulse response of the fiber with parabolic index profile.[5]

III. GROUP DELAY

We are interested in fibers satisfying the inequality

$$\frac{\sqrt{\Delta}}{n_o \mathrm{k} a} \ll 1. \tag{7}$$

We thus approximate (6) in the form

$$\beta = n_o \mathrm{k} - \frac{\sqrt{2\Delta}}{a}(p + q + 1) - \frac{\Delta}{n_o \mathrm{k} a^2}(p + q + 1)^2. \tag{8}$$

The group delay can now be expressed as

$$\tau = \frac{L}{v} = \frac{L}{c}\frac{d\beta}{d\mathrm{k}} = \frac{n_o L}{c}\left[1 + \frac{\Delta}{(n_o \mathrm{k} a)^2}(p + q + 1)^2\right]. \tag{9}$$

v is the group velocity and L is the length of the waveguide. The second term of β in (8) does not contribute to the group delay. In first approximation, if the third term in (8) is neglected, the group delay would be independent of the mode number. The difference in the group delay of the different modes is thus only slight in the square-law medium.

IV. CUTOFF CONDITION

In the infinite square-law medium there is an infinite number of modes; p and q can both assume values from 1 through infinity. The number of guided modes of a fiber with radius a must be finite. It seems reasonable to assume that those modes that interact strongly with the waveguide boundary at $r = a$ lose power at a high rate and become unimportant for the power transport. Low-order modes are concentrated near the waveguide axis while the modes spread out further away from the axis with increasing mode number. At a certain mode number the modes reach into the region of the fiber boundary and thus become very lossy.

It is known from the theory of the WKB approximation[6] that the mode field has an oscillatory behavior in the range

$$n(r)\mathrm{k} > \beta \tag{10}$$

and an exponentially decaying behavior in the range

$$n(r)\mathrm{k} < \beta. \tag{11}$$

It appears logical to let the cutoff point of the guided modes in the fiber with parabolic index profile and finite radius $r = a$ coincide with the condition

$$n(a)\mathrm{k} = \beta. \tag{12}$$

Using the square of (12) and eqs. (3) and (6) results in the cutoff

condition

$$S = (p + q)_c = \sqrt{\frac{\Delta}{2}}\, n_o k a. \tag{13}$$

Using (5) we can also write the cutoff condition in the form

$$S = \frac{a^2}{w^2}. \tag{14}$$

The actual performance of the fiber with parabolic index distribution can now be approximated by assuming that all modes carry equal amounts of power up to the maximum mode number that is determined by (13) or (14).

V. THE IMPULSE RESPONSES

The impulse response of the fiber with parabolic index profile is obtained by counting the number of modes that arrive at the fiber output simultaneously. The power carried by these modes is proportional to their number. The waveguide losses do not influence the shape of the impulse response function if we assume that all modes suffer equal amounts of loss.

Equation (9) shows that modes with constant values of

$$u = p + q \tag{15}$$

arrive simultaneously at the end of the waveguide. The number of modes with equal group delay is obtained by inspection of Fig. 1. The modes of the waveguide occupy the area of the triangle indicated in the figure. Modes with equal transit time lie on the straight line labeled u = const. The area in p, q space is equal to the number of modes contained in it. The number of modes in the interval dh is thus given by the length of the line u = const times dh. The length of the lines u = const is $2h$. We thus have for the number of modes in the interval dh

$$M(h)dh = 2h dh. \tag{16}$$

The total number of modes is

$$N = \int_0^{(1/\sqrt{2})S} M(h)dh = \tfrac{1}{2}S^2. \tag{17}$$

The ratio of $M(h)dh/N$ is equal to the ratio of the power ΔP, that corresponds to the interval dh, divided by the total power P arriving at the end of the waveguide. We thus have

$$\frac{1}{P}\frac{dP}{dh} = \frac{4h}{S^2}. \tag{18}$$

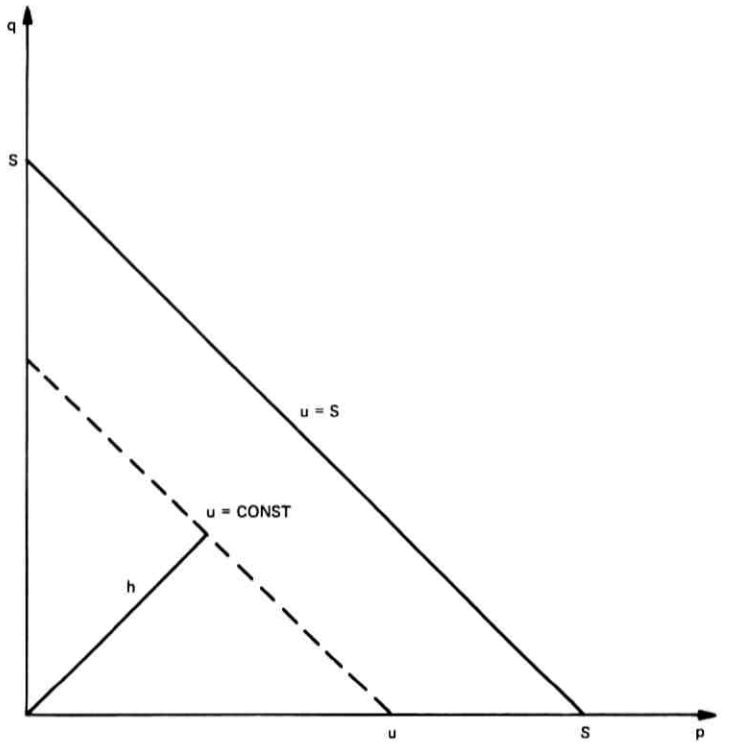

Fig. 1—Mode distribution in p-q space. The dotted line labeled u = const corresponds to modes with equal group delay.

We define the impulse response function $F(\tau)$ as the relative amount of power arriving per unit delay time τ,

$$F(\tau) = \frac{1}{P}\frac{dP}{d\tau} = \frac{1}{P}\frac{dP}{dh}\frac{dh}{d\tau}. \tag{19}$$

All that is left to do is the determination of the function $h(\tau)$. We see from Fig. 1 that

$$h = \frac{1}{\sqrt{2}}u = \frac{1}{\sqrt{2}}(p + q). \tag{20}$$

Neglecting the 1 compared to $p + q$ in (9) we have

$$h = \frac{n_o ka}{\sqrt{2\Delta}}\left(\frac{c\tau}{n_o L} - 1\right)^{\frac{1}{2}}. \tag{21}$$

The impulse response function is now obtained by combining eqs. (13)

and (18) through (21).

$$F(\tau) = \begin{cases} 0 & \frac{c\tau}{n_o L} - 1 < 0 \\ \frac{2c}{n_o L \Delta^2} & 0 < \frac{c\tau}{n_o L} - 1 < \frac{1}{2}\Delta^2. \\ 0 & \frac{c\tau}{n_o^2 L} - 1 > \frac{1}{2}\Delta^2 \end{cases} \tag{22}$$

VI. DISCUSSION

Equation (22) shows that an impulse, shared equally by all the modes at the beginning of the parabolic index fiber, reaches the end as a rectangularly shaped pulse whose width is

$$d\tau = \frac{n_o L}{2c} \Delta^2. \tag{23}$$

The pulse width is thus $d\tau = 0.25$ ns/km for $n_o = 1.5$ and $\Delta = 0.01$. The pulse width increases rapidly with increasing values of Δ. For $\Delta = 0.015$ we have a pulse width of $d\tau = 0.55$ ns/km. However, the impulse response of the parabolic index fiber is much more favorable than the corresponding impulse response of the conventional fiber with discontinuous index distribution whose impulse response width is directly proportional to[7] $n_1/n_2 - 1$ (n_1 = core index, n_2 = cladding index).

VII. ACKNOWLEDGMENT

D. Gloge contributed to this paper through several discussions. His advice is gratefully acknowledged.

REFERENCES

1. Tien, P. K., Gordon, J. P., and Whinnery, J. R., "Focusing of a Light Beam of Gaussian Field Distribution in Continuous and Periodic Lens-Like Media," Proc. IEEE, *53*, No. 2 (February 1965), pp. 129–136.
2. Kawakami, S., and Nishizawa, J., "An Optical Waveguide with the Optimum Distribution of the Refractive Index with Reference to Waveform Distortion," IEEE Trans. Microwave Theory and Techniques, *MTT-16*, No. 10 (October 1968), pp. 814–818.
3. Miller, S. E., "Waveguide for Millimeter and Optical Waves," U. S. Patent No. 3434774, issued March 25, 1969.
4. Marcuse, D., *Light Transmission Optics*, New York: Van Nostrand Reinhold Company, 1972.
5. Marcuse, D., "The Effect of the ∇n^2 Term on the Modes of the Square-Law Medium," IEEE J. Quantum Elec., *QE-9*, No. 9 (September 1973).
6. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics*, vol. II, New York: McGraw-Hill Book Co., 1953.
7. Gloge, D., "Dispersion in Weakly Guiding Fibers," Appl. Opt., *10*, No. 11 (November 1971), pp. 2442–2445.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Baseband Linearity and Equalization in Fiber Optic Digital Communication Systems

By S. D. PERSONICK

(Manuscript received February 21, 1973)

If a sequence of digitally on-off modulated optical pulses is injected into a dielectric waveguide, these pulses may begin to overlap after a sufficient distance of propagation because of material dispersion and/or group delay spreading. In general, the pulses will not add linearly in power, which can complicate the problem of equalization of the square-law (power) detected overlapping output pulses at baseband. This paper illustrates important situations in which the guide may be treated as "pseudo-linear" in power, meaning that the detected guide output pulses appear to add linearly.

I. INTRODUCTION

If a single pulse of optical energy propagates along a dielectric waveguide, pulse broadening can occur for one or more of the following reasons: material dispersion, individual mode waveguide dispersion, or differences in the group delays of different guide modes. In addition, the pulse shape may become only statistically defined because of random mode coupling and/or statistical fluctuations of the optical source.

If a sequence of digitally on-off modulated pulses is injected into a dielectric waveguide, those pulses may begin to overlap after a sufficient distance of propagation. In general, the optical powers in the pulses will not add in a linear manner.† On the other hand, as will be shown below, the guide may be pseudo-linear in power. That is, for the purpose of processing the power received at the output end of the

† The fiber is a medium which is linear *in E field propagation* (from Maxwell's equations). It is usually excited by a power-modulated source, and its output field is detected by a square-law (power sensitive) device. Even if the input pulses are separate, the response of the square-law detector will in general contain cross terms resulting from the overlap in the output pulses.

guide—in order to make decisions as to whether or not each pulse is on or off—we may be able to treat the individual overlapping output pulses as if they added linearly.

If the output power pulses could be considered to add linearly, then, after detection, the resulting current pulses could be separated with a linear equalizer provided the shapes of the pulses are well defined and identical from pulse to pulse. In this paper we consider a number of interesting cases in which the guide can be treated as if it were linear in power (output pulses add linearly) and in which the pulses at the output assume a well-defined shape in spite of random mode coupling and/or source fluctuations.

II. CASE I: MULTIMODE GUIDE, MODE-LOCKED SOURCE, NO MODE COUPLING

The easiest case to visualize in which the guide appears linear in power with overlapping output pulses is that of a multimode guide propagating pulses derived from a mode-locked laser operating in a single spatial mode. It is assumed that there is no mode conversion and that group delay differences among the modes dominate pulse spreading.

Even though the laser puts out a well-defined spatial mode, it may be very difficult to match this to a given fiber mode so that only one fiber mode is excited. We assume that a number of fiber modes are excited by the pulses from the mode-locked laser. The assumption of a mode-locked laser implies that the optical bandwidth being used is small so that material and waveguide dispersion can be neglected, and, in addition, no random fluctuations are present from beating of unlocked source modes. The sequence of nonoverlapping pulses from the laser exciting the guide input will produce identical sequences of nonoverlapping pulses in *each mode* at the guide output. However, the sequences at the output in the various modes will have relative time delays because of the differences in the group delay per unit length associated with the various modes (see Fig. 1). When the fiber output falls upon a detector, the current produced (neglecting shot noise) is proportional to the sum of the powers in all the modes. The sum of the powers in all the modes resulting from a *single input pulse* is shown in Fig. 1. Since the output pulses in a single mode do not overlap and since the detector linearly adds the powers of the various modes, the total detected current will be a sum of pulses modulated on and off, each of which looks like the response to a single input pulse. Thus, the output power produces a detected current which is a filtered version of

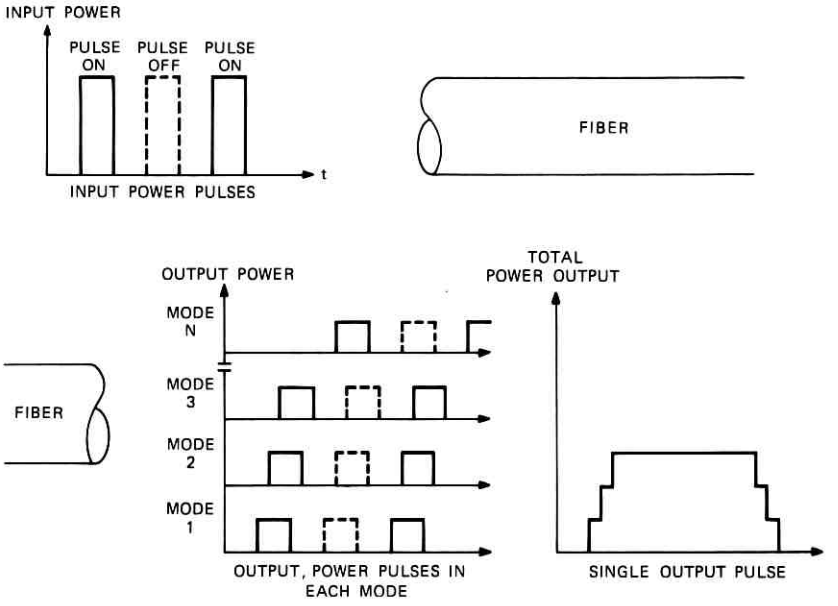

Fig. 1—Multimode propagation without coupling.

the input power. That is, the sequence of nonoverlapping input pulses produces a detected current at the output, which is a sequence of overlapping pulses that add linearly. For the purposes of processing this current, the guide can therefore be considered linear in power or linear at baseband.

III. CASE II: INCOHERENT SOURCE, MULTIMODE GUIDE, NO MODE COUPLING

In this example, we assume that a pulse modulated incoherent source excites one or more modes of a multimode waveguide with no mode coupling. We show that the received output power *in each mode* can be treated as a linearly filtered version of the input power, i.e., that the sequence of nonoverlapping input pulses produces a sequence of output pulses in each mode which add up linearly, whether they overlap (because of material dispersion) or not. Since the detector produces a current which is proportional to the sum of the powers in each mode, the total current will also consist of a sequence of pulses which add linearly, whether they overlap or not. We show that, in order for this effective linearity in power to be valid, it is necessary that the source

bandwidth be sufficiently large compared to the reciprocal of the input pulse duration. How big the ratio of these two quantities must be depends upon how much overlap there is in the output pulses and therefore upon how much equalization is required to separate the pulses at baseband.

We can model the complex amplitude of a given spatial mode at the guide input as follows:

$$\epsilon_{\text{in}}(t) = \sqrt{m(t)}c(t). \tag{1}$$

In the above example, $(m(t))^{\frac{1}{2}}$ represents the modulation and $c(t)$ is a complex Gaussian random process which represents the incoherent carrier. By definition of an incoherent carrier, we have

$$\langle c(t)c(t+u)\rangle = 0, \qquad \langle c(t)c^*(t+u)\rangle = R_c(u). \tag{2}$$

The Fourier transform of $R_c(u)$ is what is called the incoherent source spectrum, shifted to baseband.

The input complex amplitude of (1) produces an output from the guide in the corresponding mode having the following complex amplitude

$$\epsilon_{\text{out}}(t) = \int \epsilon_{\text{in}}(t')h_g(t-t')dt'. \tag{3}$$

In eq. (3), $h_g(u)$ is the guide bandpass impulse response for the mode under consideration. Equation (3) follows from the fact that the guide is linear in voltage.

The average power at the guide input in the given mode is (averaging over the fluctuations in the incoherent carrier)

$$\langle p_{\text{in}}(t)\rangle \triangleq \langle \epsilon_{\text{in}}(t)\epsilon_{\text{in}}^*(t)\rangle = m(t)R_c(0) = R_c(0)[\sum a_k h_{p_{\text{in}}}(t-kT)]. \tag{4}$$

In (4), the modulation $m(t)$ is a sequence of nonoverlapping pulses (modulated on or off) where $h_{p_{\text{in}}}(t)$ is the pulse shape, a_k assumes the value zero or one for each k, and T is the pulse spacing. The average power at the guide output is

$$\langle p_{\text{out}}(t)\rangle = \left\langle \iint \epsilon_{\text{in}}(t')h_g(t-t')\epsilon_{\text{in}}^*(t'')h_g^*(t-t'')dt'dt'' \right\rangle. \tag{5}$$

It is reasonable to assume that the following approximation holds:

$$\langle \epsilon_{\text{in}}(t')\epsilon_{\text{in}}^*(t'')\rangle = \langle \sqrt{m(t')}c(t')\sqrt{m(t'')}c^*(t'')\rangle = \sqrt{m(t')}\sqrt{m(t'')}$$
$$R_c(t'-t'') \approx m(t')R_c(t'-t''). \tag{6}$$

This approximation is valid since the coherence time of a typical

incoherent source such as a GaAs LED is of the order of 10^{-13} seconds, while modulation pulse widths of interest here exceed 10^{-9} seconds.

Substituting (6) into (5) we obtain

$$\langle p_{\text{out}}(t)\rangle = \sum a_k h_{p_{\text{out}}}(t - kT), \tag{7}$$

where

$$h_{p_{\text{out}}}(t) = \int h_{p_{\text{in}}}(t') \left[\int R_c(t' - t'') h_g(t - t') h_g^*(t - t'') dt'' \right] dt'.$$

Thus, the average output power is a linearly filtered version of the average input power. That is, the average output power consists of a sequence of pulses which add linearly even if they overlap.

Thus far, we have considered the average output power, averaging over the fluctuations in the incoherent source power output. We next consider the effect of those fluctuations on the equalized-detected current. We shall show that these source fluctuations will produce negligible deviations in the equalized-detected current from its mean provided that the source bandwidth is sufficiently large.

We can write the power at the guide output in the mode under consideration as the sum of the average power of eq. (7) and a deviation from this average $b(t)$:

$$p_{\text{out}}(t) = \langle p_{\text{out}}(t)\rangle + b(t). \tag{8}$$

If this power falls upon a detector, it produces a current which is proportional to $p_{\text{out}}(t)$ (neglecting shot noise). This current will pass through a filter which performs an equalization function and/or band-limiting. The baseband filter output voltage will therefore be

$$v_{\text{out}}(t) = z \int p_{\text{out}}(t') h_b(t - t') dt', \tag{9}$$

where $h_b(t - t')$ is the baseband filter impulse response and z is an arbitrary proportionality constant.

The mean baseband output voltage is given by

$$\langle v_{\text{out}}(t)\rangle = z \int \langle p_{\text{out}}(t')\rangle h_b(t - t') dt'. \tag{10}$$

The mean square deviation of the baseband voltage from its mean is given by

$$\langle v_{\text{out}}^2(t)\rangle - [\langle v_{\text{out}}(t)\rangle]^2 = \sigma_v^2(t) \triangleq z^2 \int \langle p_{\text{out}}(t') p_{\text{out}}(t'')\rangle \times h_b(t - t') h_b(t - t'') dt' dt'' - [\langle v_{\text{out}}(t)\rangle]^2. \tag{11}$$

Thus, to calculate the ratio of the mean voltage to the rms deviation in order to determine whether or not the deviations are negligible, we need the correlation function of the power $p_{\text{out}}(t)$.

In order to calculate this correlation function, we must recall that $c(t)$ defined in (1) is a complex Gaussian random process and satisfies [in addition to (2)] the following

$$\langle c(t)c^*(t')c(t'')c^*(t''')\rangle = R_c(t-t')R_c(t''-t''') + R_c(t-t''')R_c(t''-t'). \quad (12a)$$

Using (1), (3), (4), and (12a) we obtain

$$\langle p_{\text{out}}(t)p_{\text{out}}(t')\rangle = \langle p_{\text{out}}(t)\rangle\langle p_{\text{out}}(t')\rangle + \left|\int m(\alpha)R_c(\alpha-\beta)h_g(t-\alpha)h_g^*(t'-\beta)d\alpha d\beta\right|^2. \quad (12b)$$

To obtain some numerical results, we assume that the input pulses $h_{p_{\text{in}}}(t)$ defined in (4) are Gaussian in shape and that the guide mode impulse response corresponds to that of a dispersive medium having a group delay τ_0 at the optical source center frequency and a dispersion γ^2 within the optical band of the source. Further, we shall assume that the source spectrum is Gaussian in shape. That is, we shall assume the following:

$$h_{p_{\text{in}}}(t) = \exp(-t^2/2\sigma^2), \quad (2.36\sigma = \text{input pulse width between } \tfrac{1}{2} \text{ points}) \quad (13a)$$

$$\mathfrak{F}\{h_g(t)\} \triangleq H_g(\omega) = \exp\left(-j\frac{\gamma^2\omega^2}{2}\right)\exp(-j\tau_o\omega)\exp(-\omega^2 B_g^{-2}/2) \quad (13b)$$

$$h_g(t) = \frac{1}{\sqrt{2\pi(j\gamma^2 + B_g^{-2})}}\exp[-(t-\tau_o)^2/2(j\gamma^2 + B_g^{-2})] \quad (13c)$$

$$R_c(u) = \exp(-u^2B_s^2/2),$$

where

$$\mathfrak{F}\{\ \} = \text{Fourier transform.}$$

In eq. (13), we have already assumed that the source bandwidth B_s is much greater than the reciprocal of the input pulse width σ. We also shall assume that the guide bandwidth B_g is much larger than the source bandwidth B_s. Since τ_o represents an absolute propagation delay from input to output, we shall neglect it as irrelevant to the problem at hand. Therefore, the only significant parameter in the guide impulse response $h_g(t)$ is the dispersion γ^2.

If we insert the particular functions of (13) into (7) and (12b) using (4), we obtain

$$\langle p_{\text{out}}(t)\rangle = z_2\sum a_k \exp\{-\tfrac{1}{2}(t-kT)^2/[\gamma^4B_s^2 + \sigma^2]\} \quad (14a)$$

(i.e., a sequence of Gaussian-shaped on-off modulated pulses having width $\sqrt{\gamma^4 B_s^2 + \sigma^2}$), where z_2 = an arbitrary proportionality constant which we shall henceforth set to unity.

$$\langle p_{out}(t) p_{out}(t') \rangle = [\langle p_{out}(t) \rangle]^2 \left[1 + \exp \left\{ - \frac{(t - t')^2 B_s^2}{\left[\frac{\gamma^4 B_s^2}{\sigma^2} + 1 \right]} \right\} \right]. \quad (14b)$$

Thus, $h_{p_{out}}(t)$ of (8) is in this case the Gaussian-shaped pulse in (14a). Looking at eqs. (13a) and (14a), we see that $\gamma^4 B_s^2/\sigma^2 \ll 1$ implies that little pulse broadening has occurred in propagation; $\gamma^4 B_s^2/\sigma^2 \gg 1$ implies that considerable pulse broadening has occurred in propagation. Now using (14) in (11), we obtain

$$\frac{[\langle v_{out}(t) \rangle]^2}{\sigma_v^2(t)} = \frac{\left(\int \langle p_{out}(t') \rangle h_b(t - t') dt' \right)^2}{\int [p_{out}(t')]^2 h_b(t - t') h_b(t - t'') e^{-(t' - t'')^2 u^2/2}} dt' dt'', \quad (15)$$

where

$$u^2 = 2B_s^2 [\gamma^4 B_s^2/\sigma^2 + 1]^{-1}.$$

Recall that $\langle v_{out}(t) \rangle$ is the average baseband (detected and equalized) voltage produced by the power output in the mode under consideration, and that it consists of a sum of on-off modulated pulses given in (14a). In addition, $\sigma_v^2(t)$ is the mean squared deviation of this baseband voltage from its mean. Thus, (15) is effectively a signal-to-noise ratio. For a typical broadband source and an equalizer having a bandwidth comparable to 1/(pulse spacing-T), we can treat the Gaussian term of the integrand in the denominator of (15) as a delta function having area $\sqrt{2\pi}/u$. Then (15) becomes

$$\frac{[\langle v_{out}(t) \rangle]^2}{\sigma_v^2(t)} = \frac{\left[\int \langle p_{out}(t') \rangle h_b(t - t') dt' \right]^2}{\frac{\sqrt{2\pi}}{u} \left[\int \langle p_{out}(t') \rangle^2 h_b^2(t - t') dt' \right]}. \quad (16)$$

We can evaluate (16) for particular equalizers, $h_b(t)$, and particular output power pulse widths ($\gamma^4 B_s^2 + \sigma^2$). We recognize from (16) in general that the equivalent noise $b(t)$ of (8) which must be added to the average output power is a signal-dependent noise with correlation function

$$\langle b(t) b(t') \rangle \cong \frac{\sqrt{2\pi}}{u} [\langle p_{out}(t) \rangle]^2 \delta(t - t'), \quad (17)$$

where

$$u^2 = 2B_s^2/[\gamma^4 B_s^2/\sigma^2 + 1].$$

If the power output pulses $h_{p_{out}}(t)$ overlap significantly, then this signal-dependent noise will become stationary when all the pulses are on, which should simplify the calculation of (16). At the other extreme, if the output pulses $h_{p_{out}}(t)$ do not overlap (i.e., $\gamma^4 B_s^2 + \sigma^2 \ll T^2$) and if the baseband equalizer $h_b(t)$ is taken to be a matched filter (matched in shape to the output pulses), then the signal-to-noise ratio is given by

$$\frac{[\langle v_{out}(t)\rangle]^2}{\sigma_v^2(t)} = \sqrt{2}\sigma B_s \tag{18}$$

[if $T^2 \gg \gamma^4 B_s^2 + \sigma^2$, $h_b(t) = h_{p_{out}}(t)$].

In general, for a given desired equalized pulse shape, the equivalent noise will be negligible if the product of the optical source bandwidth B_s and the input pulse width σ is sufficiently large. For practical cases of interest, this product is on the order of 10^4 to 10^5. For reasonable amounts of equalization consistent with other noise considerations (shot noise, thermal noise, etc.), we can treat the guide as being linear in power even if the pulses overlap, i.e., we can neglect the equivalent noise $b(t)$.

When more than one mode is present, we simply add the individual mode output powers, since the detector current is proportional to the sum of the powers in all the modes. If the optical source is spatially incoherent, the equivalent noises $b(t)$ in each mode may be uncorrelated.† In that case, the requirements upon the product of the source bandwidth and the input pulse width are less stringent. This is particularly true if the pulse spreading resulting from dispersion dominates the spreading resulting from the differences in group delay among the various modes.

A simple interpretation which may prove useful follows. It was easy to obtain (7), which showed that the guide was linear in power if we averaged out the source fluctuations. Since the source is very broadband, we can think of it as a sum of independently fluctuating sources separated by a frequency spacing equal to the bandwidth of the modulation pulses at the input. Thus, the output power is the sum of the fluctuating powers associated with each equivalent independently fluctuating optical source. The average output powers associated with these equivalent sources add systematically, while the indepen-

† That is, if the optical source is close to the fiber, each fiber mode effectively sees an independent carrier, and therefore we obtain averaging of the fluctuations in these carriers when the mode powers at the output of the guide are added.

dent fluctuations about the average add at random. Thus, for a sufficiently large number of equivalent sources—corresponding to a large product of optical source bandwidth and input pulse width—the total fluctuations become small compared to the average power. In effect, one has a frequency diversity system.

IV. CASE III: MODE-LOCKED COHERENT SOURCE, MULTIMODE GUIDE, MODE COUPLING

In this example, we consider a spatially coherent mode-locked laser source and a multimode guide. Unlike example I, we assume considerable mode coupling. Once again, the rationale of using a multimode guide with a single mode source may be the inability to stably match the source to a single-mode guide. Before proceeding, we must model the transmission properties of a multimode guide with random coupling.

Very little is known about the complete statistical properties of the guide under consideration here. Any particular guide, which is linear in voltage (field), can be characterized as having a set of modes associated with an ideal guide having no geometry perturbations (which are the source of coupling). The input and output complex envelopes of the corresponding input and output optical fields can be expanded using the orthogonal guided modes, which together with the continuum of radiating modes form a complete orthonormal series. We can relate the complex amplitudes in each input and output mode by a matrix impulse response. That is, calling the complex amplitude in mode k at the input $\epsilon_{k_{in}}(t)$, and the complex amplitude in mode j at the output $\epsilon_{j_{out}}(t)$, we have

$$\epsilon_{j_{out}}(t) = \sum_k \int \epsilon_{k_{in}}(t')h_{jk}(t - t')dt', \tag{19}$$

where $h_{jk}(t - t')$ is the bandpass impulse response from input mode k to output mode j.

In order to proceed in the analysis to follow, we need at least the fourth-order joint statistics of the random processes $h_{jk}(t)$. (The reader is cautioned that the term *random process* refers to the fact that the actual $h_{jk}(t)$ for each j and k will be different for different guides because of the random mode coupling. For any particular guide which does not change its physical parameters in time, $h_{jk}(t)$ is a fixed but *a priori* unpredictable function of time. All averaging and references to statistical properties refer to ensembles of guides whose gross physical properties are alike.)

As mentioned above, little is known about the statistics of the $h_{jk}(t)$. Rowe and Young,[1] Personick,[2] and Marcuse[3] have shown in

various analyses that, for a particular j and k, the Fourier transform of $h_{jk}(t)$, $H_{jk}(\omega)$ can be considered a stationary random process under various restricted conditions which include the assumption that the optical bandwidth being used is not too large. That is, one may argue under restricted conditions that

$$\langle H_{jk}(\omega)H^*_{jk}(\omega + \sigma)\rangle = S_{jk}(\sigma) \tag{20}$$

(where the averaging is over an ensemble of guides having identical gross properties).

In another analysis, Marcuse[4] has shown that, for a particular j and k and for a sufficiently long guide (so that enough mode coupling has taken place), one has

$$\langle |H_{jk}(\omega)|^4\rangle \approx 2\langle |H_{jk}(\omega)|^2\rangle^2.$$

This last result is consistent with (but certainly not a sufficient condition for) the possibility that $H_{jk}(\omega)$ is a complex Gaussian random process.

Based on this admittedly scanty evidence which should certainly be explored in more depth, we shall *assume* that the Fourier transforms $H_{jk}(\omega)$ of the $h_{jk}(t)$ satisfy the following conditions which would be satisfied if the $H_{jk}(\omega)$ were joint complex Gaussian random processes

$$\begin{aligned} &\langle H_{jk}(\omega)H_{lm}(\omega + \sigma)\rangle = 0, \qquad \langle H_{jk}(\omega)H^*_{lm}(\omega + \sigma)\rangle = S_{jklm}(\sigma) \\ &\langle H_{jk}(\omega)H^*_{lm}(\omega + \sigma)H_{no}(\omega + \sigma')H^*_{pq}(\omega + \sigma'')\rangle \\ &\qquad = S_{jklm}(\sigma)S_{nopq}(\sigma'' - \sigma') + S_{jkpq}(\sigma'')S_{nolm}(\sigma - \sigma'). \end{aligned} \tag{21}$$

It is hoped that the results which we shall next derive will be qualitatively valid for actual multimode guides with random mode coupling.

The guide input optical field complex amplitude is given by

$$\epsilon_{\text{in}}(t) = \epsilon_p(\rho)\epsilon_\tau(t), \tag{22}$$

where $\epsilon_p(\rho)$ represents the spatial variation of the field over the guide input plane and $\epsilon_\tau(t)$ represents the time variation of the field and includes the modulation. We shall assume that the modulation consists of a sum of nonoverlapping on-off modulated pulses with spacing T

$$\begin{aligned} \epsilon_\tau(t) &= \sum a_k h_{\text{in}}(t - kT), \\ a_k &= 0 \quad \text{or} \quad 1 \quad \text{for each } k. \end{aligned} \tag{23}$$

We can expand the input field in the guide modes as follows:

$$\epsilon_p(\rho) = \sum \epsilon_k\phi_k(\rho) + \text{unguided remainder}, \tag{24}$$

where $\phi_k(\rho)$ is guided mode k.

From (20) and (24), we obtain the complex amplitude in mode j at the guide output in terms of the input complex envelope and an impulse response

$$\epsilon_{j_{\text{out}}}(t) = \sum_k \int \epsilon_k \epsilon_\tau(t') h_{jk}(t - t')dt'$$

$$= \int \epsilon_\tau(t') h_j(t - t')dt', \qquad \text{where } h_j(t) \triangleq \sum_k \epsilon_k h_{jk}(t). \quad (25)$$

Since $h_j(t)$ is a weighted sum of individual responses $h_{jk}(t)$, it follows from (21) and the linearity properties of the Fourier transform that the transform $H_j(\omega)$ of $h_j(t)$ must also satisfy

$$\langle H_j(\omega) H_j^*(\omega + \sigma)\rangle = S_j(\sigma), \qquad \langle H_j(\omega) H_j(\omega + \sigma)\rangle = 0$$
$$\langle H_j(\omega) H_j^*(\omega + \sigma) H_j(\omega + \sigma') H_j^*(\omega + \sigma'')\rangle \quad (26)$$
$$= S_j(\sigma) S_j(\sigma'' - \sigma') + S_j(\sigma'') S_j(\sigma - \sigma').$$

We next show that the guide may be considered under restricted circumstances to be linear in power with a well-defined output pulse shape even though the output pulses overlap and even though there is unpredictable mode coupling.

First, we can write down the power at the guide input and at the guide output in mode j:

$$p_{\text{in}}(t) = \left[\int |\epsilon_p(\rho)|^2 d_\rho^2\right] \sum a_k |h_{\text{in}}(t - kT)|^2 \quad (27)$$

(since $h_{\text{in}}(t) h_{\text{in}}^*(t - kT) = 0$ for $k \neq 0$, and $a_k = 0$ or 1).

$$p_{j_{\text{out}}}(t) = \int \epsilon_\tau(t') \epsilon_\tau^*(t'') h_j(t - t') h_j^*(t - t'')dt'dt''.$$

Next, we can ensemble average $p_{j_{\text{out}}}(t)$ over the ensemble of similar guides to find the average power response.

In Reference 2, it is shown that (26) implies that

$$\langle h_j(t - t') h_j^*(t - t'')\rangle = s_j(t - t')\delta(t' - t''), \quad (28)$$

where $s_j(t) = \mathfrak{F}^{-1}\{S_j(\omega)\}$ (inverse Fourier transform).

Using (28) and (27) we obtain

$$\langle p_{j_{\text{out}}}(t)\rangle = \sum a_k h_{j_{\text{out}}}(t - kT),$$

where

$$h_{j_{\text{out}}}(t) = \int |h_{\text{in}}(t')|^2 s_j(t - t')dt'. \quad (29)$$

We thus see that the input power consists of a sum of on-off modu-

lated overlapping pulses, and the average output power (averaged over an ensemble of guides with identical gross physical properties) consists of a similar sum of on-off modulated pulses which in general may overlap. Thus, in this ensemble-averaged sense, the guide is linear in power with impulse response $s_j(t)$. We must next investigate how an individual guide in the ensemble can deviate from this average and under what conditions these deviations can be neglected for communications purposes.

In order to study these deviations we must consider them in the context of a detector followed by an equalizing (or simply band-limiting) filter. Since the detector produces a current which is proportional to the linear sum of the powers in each guide mode at the output (neglecting shot noise), we consider the response to one mode only for the moment. The voltage at the equalizer output is related to the output power in mode j as follows:

$$v_{\text{out}}(t) = z \int p_{j_{\text{out}}}(t') h_{\text{det.-filt.}}(t - t')dt', \tag{30}$$

where $h_{\text{det.-filt.}}(t)$ is the detector-filter impulse response and z is an arbitrary constant.

The average (over an ensemble of guides) voltage produced is a sum of on-off modulated pulses, since we have already shown that the average output power in mode j is a sum of on-off modulated pulses

$$\langle v_{\text{out}}(t) \rangle = \sum a_k h_{v_{\text{out}}}(t - kT), \tag{31}$$

where

$$h_{v_{\text{out}}}(t) = \int h_{j_{\text{out}}}(t') h_{\text{det.-filt.}}(t - t')dt'.$$

The mean squared deviation from this average voltage is given as follows:

$$\begin{aligned}\sigma_v^2(t) &\triangleq \langle v_{\text{out}}^2(t) \rangle - \langle v_{\text{out}}(t) \rangle^2 \\ &= z^2 \iint \langle p_{j_{\text{out}}}(t') p_{j_{\text{out}}}(t'') \rangle h_{\text{det.-filt.}}(t - t') \\ &\qquad \times h_{\text{det.-filt.}}(t - t'')dt'dt'' - \langle v_{\text{out}}(t) \rangle^2. \end{aligned} \tag{32}$$

To calculate the mean squared deviation, we need the correlation function of the output power in mode j. From (27) we obtain

$$\begin{aligned}\langle p_{j_{\text{out}}}(\alpha) p_{j_{\text{out}}}(\beta) \rangle &= \iiiint [\epsilon_\tau(t') \epsilon_\tau^*(t'') \epsilon_\tau(t''') \epsilon_\tau^*(t'''')] \\ &\times \langle h_j(\alpha - t') h_j^*(\alpha - t'') h_j(\beta - t''') h_j^*(\beta - t'''') \rangle dt'dt''dt'''dt''''. \end{aligned} \tag{33}$$

In order to obtain numerical results, we must make some assumptions to facilitate the products and convolutions of (33). We assume that the guide power impulse response $s_j(t)$ for mode j is Gaussian in shape. It has been shown[1,2] that this should be the case under restricted conditions for long guides. We assume that the input pulses, $h_{in}(t)$, are Gaussian in shape with a width less than $\frac{1}{3}$ the pulse spacing, T, so that the previous assumption that they do not overlap is not violated for practical purposes. That is, we assume

$$\begin{aligned} h_{in}(t) &= e^{-t^2/2\sigma^2}, \qquad 3\sigma < T = \text{pulse spacing} \\ s_j(t) &= e^{-t^2/2\gamma^2}, \qquad \gamma > 3\sigma. \end{aligned} \tag{34}$$

From (33), (26), and (34) we obtain

$$\begin{aligned} \langle p_{j_{out}}(\alpha)\rangle &\cong z \sum a_k \exp -\left\{(\alpha - kT)^2 \Big/ 2\left[\frac{\sigma^2}{2} + \gamma^2\right]\right\} \\ \langle p_{j_{out}}(\alpha)p_{j_{out}}(\beta)\rangle &\cong \Bigg[z^2 \sum_{kml}\sum\sum a_k a_{k-m} a_l a_{l+m} \\ &\quad \times \exp -\left\{\frac{(\alpha - kT)^2}{2(\sigma^2/2 + \gamma^2)}\right\} \exp -\left\{\frac{(\beta - lT)^2}{2(\sigma^2/2 + \gamma^2)}\right\} \\ &\quad \times \exp -\left\{\frac{[(\alpha - \beta) - mT]^2\gamma^2}{2(\sigma^2/2 + \gamma^2)\sigma^2}\right\}\Bigg] \\ &\quad + \langle p_{j_{out}}(\alpha)\rangle\langle p_{j_{out}}(\beta)\rangle. \end{aligned} \tag{35}$$

The approximations of (35) become equalities when the width of the average power impulse response γ becomes large compared to the input pulse width, σ, i.e., when there is a lot of pulse spreading in propagation. We shall soon see that this will be the case of interest in this example. Before attempting to use (35) in (32) to evaluate the magnitude of the deviations of a particular guides power from the ensemble average, we can make some simplifications and comments upon (35). Using the assumptions $\gamma \gg \sigma$, we have

$$\langle p_{j_{out}}(\alpha)\rangle = z \sum a_k \exp\{-(\alpha - kT)^2/2\gamma^2\} \tag{36}$$

$$\begin{aligned} R_b(\alpha, \beta) &\triangleq \langle p_{j_{out}}(\alpha)p_{j_{out}}(\beta)\rangle - \langle p_{j_{out}}(\alpha)\rangle\langle p_{j_{out}}(\beta)\rangle \\ &= z^2 \sum_{kml}\sum\sum a_k a_{k-m} a_l a_{l+m} e^{[-(\alpha - kT)^2/2\gamma^2]} e^{[-(\beta - lT)^2/2\gamma^2]} \\ &\quad \times e^{-(\alpha - \beta - mT)^2/2\sigma^2}. \end{aligned}$$

If we assume that σ is small compared to the pulse spacing T and if we assume that the detector-equalizer combination passes frequencies only up to the inverse of the pulse spacing T, then we can make the

further approximation

$$R_b(\alpha, \beta) = z^2 \sum_{klm}\sum\sum a_k a_{k-m} a_l a_{l+m} \times [e^{(\alpha-kT)^2/2\gamma^2} e^{-(\beta-lT)^2/2\gamma^2}] \delta(\alpha - \beta - mT)\sqrt{2\pi\sigma^2}. \quad (37)$$

What we have shown so far is that the power in mode j at the guide output can be considered to be of the form

$$p_{j_{out}}(t) = \langle (p_{j_{out}}(t) \rangle + b(t), \quad (38)$$

where

$$\langle b(\alpha) b(\beta) \rangle = R_b(\alpha, \beta),$$

where $b(t)$ represents the deviations of the power in a particular guide in mode j from the ensemble average.

Combining (36) and (38) with (32), we obtain the ratio of the (average voltage)^2 at the detector filter output to the mean square deviation from this average voltage

$$\frac{S}{N} \triangleq \frac{\langle v_{out}(t) \rangle^2}{\sigma_v^2(t)} = \frac{\left[\int \sum a_k e^{-(t'-kT)^2/2\gamma^2} h_{det.-filt.}(t - t') dt' \right]^2}{(\sqrt{2\pi}\sigma) \int \sum\sum\sum a_k a_l a_{k-m} a_{l+m} e^{[-(t'-kT)/2\gamma^2]} e^{[-(t''-lT)/2\gamma^2]} \times \delta(t' - t'' - mT) h_{det.-filt.}(t - t') h_{det.-filt.}(t - t'') dt' dt''}. \quad (39)$$

It is clear that, whatever the equalizing filter is, this ratio increases with decreasing σ. Thus, as the input power pulses to the guide become narrow, two things happen: the average output pulse widths become independent of the input pulse width and the deviations of the output power in mode j from the ensemble average become negligible. Exactly how small σ has to be depends upon how much the average output power pulses overlap and how much equalization we are therefore using. In the extreme case of no output pulse overlap, assuming that the equalizer response $h_{det.-filt.}(t)$ is matched to the average pulse power, we obtain

$$T > 3\gamma \text{ (no pulse overlap at output)}$$

$$\frac{S}{N} = \frac{\gamma\sqrt{2}}{\sigma} \quad \text{for } \sigma \ll \gamma \text{ (output pulse width} \gg \text{input pulse width)} \quad (40)$$

$$h_{det.-filt.}(t) = h_{j_{out}}(-t) \text{ (matched filter equalizer).}$$

Obviously, the guide acts linearly in power if the output pulses don't overlap, but (40) shows that the output pulses take on a well-

defined shape (i.e., the deviations from the ensemble average are negligible) in spite of the random mode coupling. For cases where the output pulses overlap, the conditions for the deviations from the ensemble average to be negligible are more stringent, i.e., the signal-to-noise ratio (39) is less than the special case (40).

Summarizing, we started with a mode-locked laser putting out pulses which were much narrower than the spacing between them and on-off modulated. This optical field excited a guide with random mode coupling. We modeled the transfer function relating the input field complex envelope to the complex amplitude of a particular mode at the output as having specific properties (26) associated with a complex Gaussian random process. This model was justified only in the sense that it was consistent with available but scanty analytical results on guides with random coupling. We showed that the ensemble average output power in the mode under consideration looked like a linearly filtered version of the input power. That is, the average output power was a linear sum of pulses which could in general overlap. Thus, on the average, the guide looked linear in power for digital communication applications. We showed that the deviations in a particular guide from this ensemble average linearity behavior would be negligible provided the input pulses were very narrow compared to the width of the guide average power impulse response. How narrow the input pulses had to be depended upon how much equalization was required to separate the output pulses.†

It is clear that, since the detector adds the powers in each output mode, the total power will be a linear sum of pulses if the individual mode powers are. In addition, we may suspect that the deviations from the ensemble average in each mode may add randomly while the average powers add systematically. Thus, some improvement in the signal to "noise" ratio may accrue from this spatial diversity.

An interpretation of what is happening to make the deviations negligible is the following: Since the guide average power impulse response for output mode j, $s_j(t)$, is the Fourier transform of the two-frequency correlation function $S_j(\omega)$ defined in (26), we can interpret the reciprocal of the width of $s_j(t)$ as the bandwidth difference over which the guide transfer function between the input field and the output mode j becomes uncorrelated. When we use narrow input pulses compared to the width of the average power impulse response, we use a lot of bandwidth compared to this correlation bandwidth and

† Remember that the output pulse shape becomes independent of the input pulse shape as the input pulses get narrow.

thus obtain frequency diversity. As the input pulses become very narrow, the output pulses become fixed in shape equal to $s_j(t)$, but the diversity keeps increasing, resulting in averaging out of the deviations from one guide to the next.

V. CASE IV: INCOHERENT SOURCE, MULTIMODE FIBER WITH MODE COUPLING

In this example, we consider an incoherent intensity-modulated source exciting a multimode fiber with random mode coupling. We assume that material dispersion is negligible. (Since we shall be considering a wideband source, we may question the physical reality of neglecting material dispersion. On the other hand, the qualitative results to follow may provide insight into more general cases.) To simplify what will prove to be a somewhat complicated analysis, we shall consider the response in a particular guide mode, j, at the output, to the field in a particular mode, k, at the guide input. Extension to consideration of the total input field and the total response should be straightforward, using the techniques outlined below.

The input field complex amplitude in mode k is of the form

$$\epsilon_{k_{\text{in}}}(t) = \sqrt{m(t)}c(t), \tag{41}$$

where $m(t)$ is the modulation and $c(t)$ is the optical incoherent carrier which satisfies

$$\langle c(t)c^*(u)\rangle = R_c(t-u) \tag{42a}$$

$$\langle c(t)c^*(u)c(t')c^*(u')\rangle = R_c(t-u)R_c(t'-u') + R_c(t-u')R_c(t'-u). \tag{42b}$$

The complex amplitude in mode j at the output due to the input field in mode k is given by

$$\epsilon_{j_{\text{out}}}(t) = \int \epsilon_{k_{\text{in}}}(t')h_{jk}(t-t')dt', \tag{43}$$

where the impulse response coupling mode k to j is assumed to satisfy the same statistics as were outlined in the first paragraphs of Section IV, i.e.,

$$\begin{aligned}
&(i)\ H_{jk}(\omega) = \mathfrak{F}\{h_{jk}(t)\}\\
&(ii)\ \langle H_{jk}(\omega)H^*_{jk}(\omega+\sigma)\rangle = S_{jk}(\sigma)\\
&(iii)\ \langle H_{jk}(\omega)H^*_{jk}(\omega+\sigma)H_{jk}(\omega+\sigma')H^*_{jk}(\omega+\sigma'')\rangle\\
&\qquad = S_{jk}(\sigma)S_{jk}(\sigma''-\sigma') + S_{jk}(\sigma'')S_{jk}(\sigma-\sigma')\\
&(iv)\ \langle h_{jk}(t)h^*_{jk}(t')\rangle = s_{jk}(t)\delta(t-t'),
\end{aligned} \tag{44}$$

where

$$s_{jk}(t) = \mathfrak{F}^{-1}\{S_{jk}(\omega)\}.$$

The average input power (averaging over the source fluctuations) is

$$\langle p_{k_{\text{in}}}(t)\rangle = m(t)R_c(0) = [\sum a_k h_{\text{in}}(t - kT)]R_c(0), \tag{45}$$

i.e., a sequence of on-off modulated pulses where $a_k = 0$ or 1, T = pulse spacing.

The output power is given by

$$p_{j_{\text{out}}}(t) = \iint \sqrt{m(t')}\sqrt{m(t'')}c(t')c^*(t'')h_{jk}(t - t')h^*_{jk}(t - t'')dt'dt''. \tag{46}$$

If we average over the source fluctuations and the guide statistics we obtain, using (44) and (6)

$$\begin{aligned}\langle p_{j_{\text{out}}}(t)\rangle &= \int m(t')R_c(t' - t'')\langle h_{jk}(t - t')h^*_{jk}(t - t'')\rangle dt''dt' \\ &= \int m(t')R_c(0)s_{jk}(t - t')dt' = \sum a_k h_{j_{\text{out}}}(t - kT),\end{aligned} \tag{47}$$

where

$$h_{j_{\text{out}}}(t) = R_c(0)\int h_{\text{in}}(t')s_{jk}(t - t')dt'.$$

Thus, on the average, the output power looks like a linearly filtered version of the input power with impulse response $s_{jk}(t)$.

This average output power will produce an average detected-equalized voltage which is also a linearly filtered version of the input power. We next show that, if the optical bandwidth of the incoherent source is sufficiently large, then the deviations of the detected equalized voltage from its average, because of fluctuations in the source and deviations of the impulse response of a particular guide from the ensemble average, can be neglected.

As before, the detected-equalized voltage is given by

$$v_{\text{out}}(t) = z\int p_{\text{out}}(t')h_{\text{det.-filt.}}(t - t')dt', \tag{48}$$

where z is an arbitrary constant.

The mean detected-equalized current and the mean square deviation from the mean are given by

$$\langle v_{\text{out}}(t)\rangle = \sum a_k h_{v_{\text{out}}}(t - kT),$$

where

$$h_{v_{out}}(t) = \int h_{j_{out}}(t')h_{det.-filt.}(t - t')dt'$$

$$\sigma_v^2(t) = \langle v_{out}^2(t)\rangle - \langle v_{out}(t)\rangle^2 = \int \langle p_{j_{out}}(t')p_{j_{out}}(t'')\rangle \times h_{det.-filt.}(t - t')h_{det.-filt.}(t - t'')dt'dt'' - \langle v_{out}(t)\rangle^2 \quad (49)$$

To calculate the mean square deviation we need the correlation function of $p_{j_{out}}(t)$. Using (6), (42b), and (44) we obtain the following complicated expression:

$$\langle p_{j_{out}}(t')p_{j_{out}}(t'')\rangle = \langle p_{j_{out}}(t')\rangle\langle p_{j_{out}}(t'')\rangle\left[1 + \frac{|R_c(t'' - t')|^2}{R_c^2(0)}\right] + \int [m(\tau)]^4 s_{jk}(t' - \tau)s_{jk}(t'' - \tau)d\tau \int |R_c(u)|^2 du + \int [m(\tau)]^2[m(t'' - t' - \tau)]^2 |s_{jk}(t' - \tau)|^2 d\tau \times \int |R_c(u)|^2 du. \quad (50)$$

In order to obtain numerical results we must make some assumptions as to the shapes of $s_{jk}(t)$, $h_{in}(t)$, and $R_c(t)$. We shall assume the following:

$$\begin{aligned} h_{in}(t) &= \exp -t^2/2\sigma^2 && \text{(Gaussian-shaped input pulses)} \\ \mathfrak{F}\{R_c(t)\} &= \exp -\omega^2/2B_s^2 && \text{(Gaussian-shaped source spectrum)} \\ s_{jk}(t) &= \exp -t^2/2\gamma^2 && \text{(Gaussian-shaped power impulse response—appropriate for long guides).} \end{aligned} \quad (51)$$

Using (51) in (50) we obtain the expressions

$$\langle p_{j_{out}}(t)\rangle = \sum a_k e^{[-(t-kT)^2/2(\sigma^2+\gamma^2)]} \frac{\sigma\gamma B_s}{\sqrt{\sigma^2 + \gamma^2}}$$

$$\langle p_{j_{out}}(t')p_{j_{out}}(t'')\rangle = \langle p_{j_{out}}(t')\rangle\langle p_{j_{out}}(t'')\rangle\{1 + \exp[-(t'' - t')^2B_s^2]\} + \frac{B_s\sigma\gamma}{2\sqrt{\sigma^2 + \gamma^2}} \sum \sum a_k a_j \{e^{[-(t''-t')^2/4\gamma^2]}e^{-[(k-j)T]^2/4\sigma^2} \times e^{[-1/\sigma^2+\gamma^2][t'+[(t''-t')/2]-(k+j)(T/2)]^2} + e^{-[(k-j)T+(t''-t')]^2/4\sigma^2} \times e^{[-1/\sigma^2+\gamma^2][t'+[(t''-t')/2]-(k+j)(T/2)]^2}\}. \quad (52)$$

Equation (52) is still fairly complicated, but we can make some general observations. The mean output power is proportional to the source bandwidth B_s, and therefore the mean output power squared

will be proportional to B_s^2. Looking at the mean squared output power, we see that the portions involving the double sum are proportional to B_s. Those terms will become negligible compared to the mean output power squared as B_s gets large. The term $\exp -(t'' - t')^2 B_s^2$ can be approximated by an impulse of area $\sqrt{2\pi/B_s}$. Thus, this term is also proportional to B_s^{-1} and will be negligible for large B_s. Thus, for large B_s we can approximate the mean square power by the mean power squared. Equivalently, $\sigma_v^2(t)$ will become negligible compared to $\langle v_{\text{out}}(t)\rangle$ as the source bandwidth becomes very large. Just how large the source bandwidth has to be depends upon how much overlap there is to average output pulses and therefore upon how much equalization has to be done. In the simple case where the output pulses do not overlap and where the equalizer is matched in shape to the output pulses, we obtain

$$\frac{S}{N} = \frac{\langle v_{\text{out}}(t)\rangle^2}{\sigma_v^2(t)} \cong \frac{\sqrt{2}B_s}{\dfrac{1}{\sqrt{\gamma^2+\sigma^2}} + \dfrac{\sqrt{\gamma^2+\sigma^2}}{\sigma\sqrt{2\gamma^2+\sigma^2}} + \dfrac{\sqrt{\gamma^2+\sigma^2}}{\gamma\sqrt{2\sigma^2+\gamma^2}}} \tag{53}$$

for

$$T > 3\sqrt{\sigma^2+\gamma^2} = 1.5 \times \text{output pulse width}$$

$$h_{\text{det.-filt.}}(t) = h_{j_{\text{out}}}(-t) \text{ (matched filter).}$$

From (53) we see that, in this special case, the ratio of the mean detected equalized (filtered) voltage squared to the mean squared deviation, because of source fluctuations and guide random coupling, will be greater than 1000 if the optical source bandwidth is more than 1000 times the reciprocals of both the input pulse duration and the average power impulse response duration.

Typically, the source bandwidth is more than 10^{13} radians per seconds. Thus, the deviations are negligible for input and output pulse widths larger than 10^{-10} seconds. Of course, more bandwidth is required to make the deviations negligible when there is considerable pulse overlap.

An interpretation is similar to previous cases. The requirement that the source bandwidth be large compared to the reciprocal of the duration of the input pulses allows averaging out of the fluctuations in the source. The requirement that the source bandwidth be large compared to the reciprocal of the guide average power impulse response duration gives the frequency diversity that averages out the deviations between guides resulting from random mode coupling.

VI. CONCLUSIONS

We conclude that there are a number of interesting circumstances in which a fiber system, normally linear in voltage, can be considered linear in power for digital communication purposes. The input power consisting of a sum of on-off modulated pulses produces an output power (and thus a detected current) which is also a sequence of on-off modulated pulses, possibly overlapping. The output pulse shape is well-defined under the conditions described above, which usually amount to using enough optical bandwidth to have sufficient frequency diversity to average out source fluctuations and/or random mode coupling differences between guides. This power linearity allows baseband equalization (with the usual noise penalties) to allow use of the guide at higher bit rates than would be associated with the criterion that the output pulses must not overlap. Many assumptions made above may not be completely applicable to particular guides, but it is hoped that, qualitatively, some insight as to when power linearity may occur will be derived from these results.

REFERENCES

1. Rowe, H. E., and Young, D. T., "Transmission Distortion in Multimode Random Waveguides," IEEE Trans. on Microwave Theory and Techniques, *MTT-20*, No. 6 (June 1972), pp. 349–364.
2. Personick, S. D., "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
3. Marcuse, D., "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1199–1232.
4. Marcuse, D., "Fluctuations of the Power of Coupled Modes," B.S.T.J., *51*, No. 8 (October 1972), pp. 1793–1800.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Optimum Network Call-Carrying Capacity

By R. L. FRANKS and R. W. RISHEL

(Manuscript received September 8, 1972)

A telephone network with switching and trunk congestion is considered. An optimization problem expressed in terms of mean numbers of calls and mean rates of flow of calls in various categories of service throughout the network is formulated. The maximum mean number of talking calls given by this optimization problem is an upper bound on the mean number of talking calls which could be carried by the network using theoretically optimum network management. Examples are given suggesting that the upper bound is close to values which actually can be attained.

The optimum of the problem is achieved by controls which (i) restrict the number of calls coming into the network from the end offices and (ii) route appropriate fractions of the remaining calls over the various possible routes.

I. INTRODUCTION

Telephone communication facilities are designed to adequately handle peak traffic loads of an average day. In many instances the system is subject to higher loads. Classic examples of situations in which overloads occur are during holidays such as Christmas or Mother's Day, after disasters such as earthquakes or hurricanes, and during facility failures. Because of the time lag necessary to install new equipment, high overloads can also occur in normal operation in cases in which predicted traffic growth is greatly exceeded by actual traffic growth. An interesting observed phenomenon is that under certain high-load situations fewer calls may be completed than during normal load periods. Recognition of this gave rise to the subject of network management. One objective of network management is to control the handling of calls so that the maximum number of calls is put through the network.

An interesting discussion of the network management problem and early work to understand the phenomena involved in it is contained in Ref. 1. In response to the problems mentioned in Ref. 1, a simulation

study of the network management problem was carried out in Ref. 2. Simulations[2] of a modest-sized telephone network were developed and were used to evaluate the performance of various control techniques.

A large number of control techniques have been suggested for managing the network under overload conditions. Some papers which are representative of these controls are Refs. 3 through 5.

For given point-to-point calling rates, and given network management controls, let us measure the steady-state performance of the network by the expected total number of talking calls carried by the network. Define the capacity of the network (with the given calling rates) to be the maximum performance that could be obtained by any network management controls. Our objective is to set up techniques for computing the capacity of the network. The capacity of a network can be used as a benchmark in evaluating network management controls. That is, the performance of a given network management control system could be computed and compared against the capacity to tell how effective the given control system is.

The calls carried on a telephone network can be considered as a stochastic process described by a very large number of variables. There are techniques which apply to optimization of stochastic systems;[6–9] however, these would not give practical methods for computation of the optimum capacity.

For simplicity, we will consider only the steady-state situation. Our method will be to establish inequalities and equations which must be satisfied by the mean steady-state values of numbers of calls in various categories of being set up and mean rates of flow to calls into and out of these categories. From these equations and inequalities an optimization problem will be formulated. It will not be claimed that the set of inequalities and equations obtained is an exhaustive set. Hence the optimum value of the criterion of the optimization problem will only be an upper bound on the optimum value for a corresponding criterion for the real system. Later, examples will be given to show that in these cases the optimum value of the optimization problem can be nearly obtained by an appropriate choice of controls incorporated in a simulation of the network.

While the situations are quite different, our treatment of this problem follows in spirit a corresponding technique in optimal control theory, called "relaxation" of the problem.[10,11] However, the interest there is in finding an optimal control, while we are not after an optimal control, but merely want a good upper bound for the message carrying capacity of the network.

II. BACKGROUND

The factors underlying the decrease in the number of calls carried by the telephone network as it became highly overloaded were already well understood in the early work of Ref. 1. As a call is being set up, it uses equipment in one switching machine until the next switching machine on its route accepts the call and receives the destination of the call from the previous machine. If a switching machine becomes overloaded, machines adjacent to it will have to wait longer to have their calls accepted and the destinations passed on. This causes an increase in the service time for putting a call through these machines. This in turn may cause the adjacent machines to become congested. A current device to relieve this congestive phenomenon is that calls that must wait longer than a fixed time-out time for a subsequent machine are given a no-circuit announcement. However, even with time-outs, switching machine congestion can back up throughout the network.

Calls being set up occupy trunks on the partial route over which they have progressed. If a large number of calls are attempting on routes which are blocked, a portion of the capacity of certain links could be used by these ineffective attempts trying to set up. This would use capacity that could be used by talking calls. These ineffective attempts also use switching capacity in machines preceding the blockage. Most blocked attempts try again. These retrials increase the congestion.

The model which will be set up will incorporate the features mentioned above. Based on these observations, the model must take into account both trunking congestion and switching machine congestion. Since the route a call may take can be controlled and calls which are given busy signals free the entire partial route they occupied, the model should keep track of the partial routes over which calls have progressed.

Throughout the entire paper we will be interested only in the expected values of the various variables in the steady state. We will assume throughout that the processes are ergodic.

III. SWITCHING MACHINE MODEL

A block diagram of the operations of the switching machine which will be modeled is given in Fig. 1. This is a simplified model of the Bell System No. 4A-ETS switching machine. In the model a call coming into a switching machine enters a queue to wait for a vacant sender. When a call gets a sender it inpulses its destination information to the sender and releases the sender it had in the previous machine.

The call then gets into a queue for a decoder-marker combination. This decoder-marker decides on the machine the call should be routed

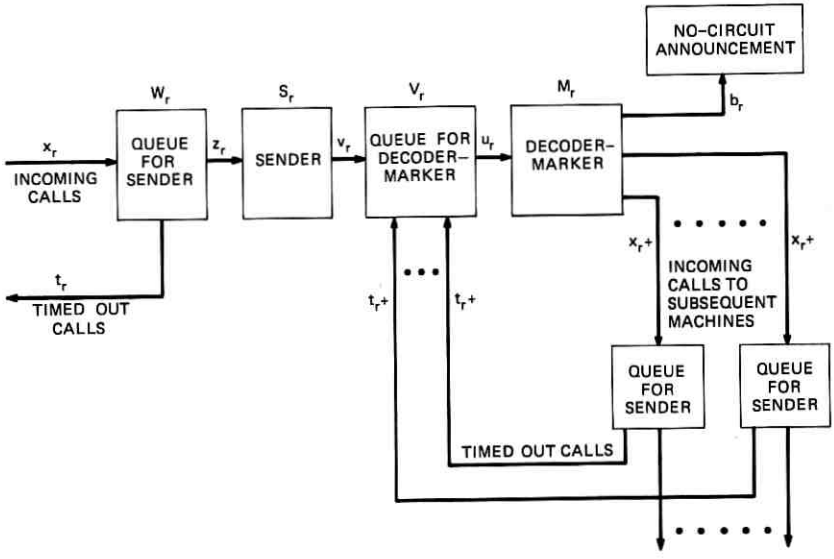

Fig. 1—Switching machine block diagram.

to next, tests for a vacant trunk, and sets up the connection, if possible. If there are no vacant trunks to appropriate subsequent machines, a no-circuit announcement is given. After a no-circuit announcement, the trunks on all the links that the call had progressed through become vacant again. If it is routed to a subsequent machine, it enters a queue for a sender in that machine.

The sender in the current machine is occupied by the call from the time it begins processing the call until the call has transmitted its destination information to the sender which processes the call in the subsequent machine. If a call waits longer than a fixed time to get a sender in the subsequent office, it is timed out. If it is timed out, the call is sent back to the marker-decoder which then connects it to a no-circuit announcement.

The process of a call being connected for service to a sender is accomplished by a sender link controller and a sender link connector. These devices test for an idle sender and an idle path to the sender. Then a path through the switch from the incoming trunk to the sender is selected and the connection established by closing appropriate switches.

The process of connecting a call for service to a marker is carried out in a similar fashion by a marker connector.

IV. COMPLETE AND PARTIAL ROUTES

A telephone network can be thought of as a collection of call switching machines connected by communication links. In the operation of the network, a talking call which hangs up frees a trunk on all the links of the route it occupied. A call which is in some state of being set up will occupy a partial route. If it is given a no-circuit announcement, it will free trunks on all the links of the partial route it occupied.

To model routes and partial routes, we will make the following definitions. Assume there are k switching machines of the network labeled by the integers $1, \cdots, k$. We will use the letter R to denote a complete route. A complete route

$$R = (i_1, \cdots, i_n)$$

is a succession of switching machines. A call occupying a complete route is always considered to be a talking call and it is understood that the call is occupying trunks on the links connecting switching machines in adjacent positions in the expression R.

A partial route describes the route occupied by a call in the process of being set up and the destination of the call. Let r designate a partial route. If

$$r = (i_1 \cdots i_{n-1}, i_n),$$

it is understood that the call has passed through machines $i_1 \cdots i_{n-2}$, is waiting to get into i_{n-1} or is in i_{n-1} being processed, and its destination is machine i_n. It occupies trunks on links connecting adjacent machines from i_1 to i_{n-1}. When a partial route r is of the above form, we shall say the partial route "terminates" at machine i_{n-1}.

The symbol r^+ will be used to designate a route subsequent to r. If $r = (i_1 \cdots i_{n-1}, i_n)$, r^+ may be the complete route

$$R = (i_1 \cdots i_{n-1}, i_n)$$

in the case in which there is a link between machine i_{n-1} and the destination i_n; or it may be a route of the type

$$r^+ = (i_1, \cdots, i_{n-1}, i_m, i_n),$$

that is, a partial route in which the call has passed to one further machine i_m on its way to its destination i_n than the call on partial route r had.

V. MODEL DESCRIPTION

In a given switching machine, calls can be distinguished by the partial route they occupy and the stage of processing they are in.

Completed talking calls can be distinguished by the complete route which they occupy. To describe the operation of the switching machines, we define the variables:

W_r = mean number of calls on partial route r waiting for a sender
S_r = mean number of calls on partial route r which are inpulsing into a sender
V_r = mean number of calls on partial route r waiting for service by a decoder-marker
M_r = mean number of calls on partial route r being serviced by a decoder-marker
N_R = mean number of talking calls on complete route R
x_r = mean rate of flow of calls on partial route r into the sender queue
z_r = mean rate of flow of calls on partial route r into the sender
v_r = mean rate of flow of calls on partial route r from the sender into the decoder-marker queue
u_r = mean rate of flow of calls on partial route r into decoder-marker
t_r = mean rate at which calls in sender queue on partial route r are timed out
b_r = mean rate at which calls on partial route r are blocked
z_R = mean rate at which calls are being completed through the network on complete route R.

We shall show that the following statements must be satisfied by these mean rates and mean numbers.

1. The mean rate at which calls flow into senders in a given machine is less than or equal to a constant times the mean number of senders which are not currently processing calls.
2. The mean rate at which calls arrive at the marker queue from the senders is equal to a constant times the mean number of senders which are processing calls which have not yet entered a marker queue.
3. The mean rate at which calls flow into the marker-decoder is less than or equal to a constant times the mean number of markers not currently processing calls.
4. The mean rate at which calls leave the markers is equal to a constant times the mean number of markers which are processing calls.

Consider statement 1. Let us define a call to be in the process of connecting to a sender from the time an idle sender is found until the connection has been completed to that sender. Defining the process this way, no time-outs occur during it. While a call is undergoing this connection process, we will say it is in the connector. We use this definition:

number of calls connecting to some sender
$\leqq$ number of free senders.

The above inequality also must hold for the mean values of both quantities.

Now the number connecting to some sender is the number in a queuing system (the connector) whose service time is the time required to make a connection. Applying Little's Theorem[12,13] gives:

E {number of calls connecting to some sender}
= (mean rate at which calls are flowing into the connector)
$\times$ (mean time to make a connection).

Since we are considering a steady-state situation, the mean rate at which calls are flowing into the connector equals the mean rate at which calls are connected to some sender. Hence we obtain statement 1. The constant in statement 1 is the reciprocal of the mean work time required to connect a call to a sender.

Consider statement 2. Calls attach to a sender, impulse their destination information, and then enter a marker queue. Applying Little's law to the number of calls inpulsing into a sender gives,

mean number of calls inpulsing into a sender
= (mean rate at which calls arrive at senders)
$\times$ (mean inpulsing time).

Since the system is in steady state, the mean rate at which calls arrive at senders equals the mean rate at which calls arrive at the marker queue. The mean number of calls inpulsing into a sender is the mean number of calls in the sender which have not yet entered the marker queue. Hence statement 2 is established.

Statements 3 and 4 are statements concerning markers similar to statements 1 and 2 concerning senders. They can be established by using Little's law in a similar manner to statements 1 and 2.

Next, statements 1 through 4 will be expressed more formally as equations and inequalities in the variables defined previously. For a

given switching machine, let

I_i = all partial routes terminating in machine i
O_i = all partial routes immediately subsequent to a partial route terminating in machine i
$\mathcal{S}_i$ = total number of senders in machine i
$\mathfrak{M}_i$ = total number of decoder-markers in machine i.

The total rate at which calls are entering senders in machine i is the sum over all the partial routes which terminate at machine i of the rates at which calls on those partial routes are entering senders. In symbols this is

$$\sum_{r \in I_i} z_r.$$

The mean number of calls occupying senders in machine i is

$$\sum_{r \in I_i} (S_r + V_r + M_r) + \sum_{r^+ \in O_i} W_r{}^+.$$

Hence, the mean number of free senders is

$$\mathcal{S}_i - \left\{ \sum_{r \in I_i} (S_r + V_r + M_r) + \sum_{r^+ \in O_i} W_r{}^+ \right\}.$$

Hence, statement 1 may be written as the inequality

$$\sum_{r \in I_i} z_r \leqq C_1 \left\{ \mathcal{S}_i - \sum_{r \in I_i} (S_r + V_r + M_r) - \sum_{r^+ \in O_i} W_r{}^+ \right\}. \tag{1}$$

A similar interpretation of statement 3 yields

$$\sum_{r \in I_i} u_r \leqq C_3 \left\{ \mathfrak{M}_i - \sum_{r \in I_i} M_r \right\}. \tag{3}$$

Statements 2 and 4 are expressed by

$$v_r = C_2 S_r, \tag{2}$$

$$u_r = C_4 M_r. \tag{4}$$

The number of calls on a given link either talking or being processed in some switching machine must be less than or equal to the number of trunks in that link. Hence, the expected number of calls of these types must be less than or equal to the number of trunks on the link. The inequalities expressing this are given by

$$\sum_{R \supset i,j} N_R + \sum_{r \supset i,j} [W_r + S_r + V_r + M_r] \leqq C_{ij}. \tag{5}$$

In this notation there is one inequality for each link connecting a machine i and a machine j, C_{ij} is the number of trunks on this link, and the notations $R \supset i,j$ and $r \supset i,j$ indicate that the sums are respectively over all complete routes or all partial routes which pass through link i,j.

In the steady state the expected rate at which calls flow into any category of processing in a switching machine must equal the rate at which they flow out. Thus the following flow-in equal flow-out equations must hold.

$$x_r = z_r + t_r \tag{6}$$

$$z_r = v_r \tag{7}$$

$$v_r + \sum_{r^+ \in 0_i} t_r^+ = u_r \tag{8}$$

$$u_r = \sum_{r^+ \in 0_i} x_r^+ + b_r. \tag{9}$$

Let A_R denote the probability that a call that completes through the network on route R will be answered by the customer. The rate at which calls are completing to talking calls on route R is $A_R z_R$. If the mean length of a talking call is $1/\nu$, calls will be hanging up at rate ν. To be in steady state, the equation

$$A_R z_R = \nu N_R \tag{10}$$

must hold.

Let λ_{ij} denote the mean rate at which calls wish to enter the network originating at machine i with destination machine j. Suppose that, if a call is placed and receives a no-circuit announcement, the customer decides to retry with probability P or to give up placing the call with probability $1 - P$. Suppose this is true independently for every call irrespective of how many times the customer may have tried previously to place the call and failed.*

Let a_{ij} denote the rate at which calls including retrials are being placed from i to j. Let r_{ij} denote the rate at which no-circuit announcements are being given and s_{ij} the rate at which calls are being completed from i to j. Then

$$a_{ij} = s_{ij} + r_{ij} = \lambda_{ij} + P r_{ij}. \tag{11}$$

* Customer retrial behavior is discussed by Wilkinson in Ref. 14. The model considered here can be considered as an idealized approximation to the more complicated behavior reported in Ref. 14.

Solving for r_{ij},

$$r_{ij} = \frac{\lambda_{ij} - s_{ij}}{1 - P}. \tag{12}$$

Now the rate at which calls are being completed between i and j is the sum over all the complete routes joining i and j of the rate at which calls are flowing onto these complete routes. Hence,

$$s_{ij} = \sum_{R=(i,\cdots,j)} A_R z_R. \tag{13}$$

Using (11), (12), and (13) gives

$$a_{ij} = \frac{\lambda_{ij} - P \sum_{R=(i,\cdots,j)} A_R z_R}{1 - P}. \tag{14}$$

Calls entering the network at machine i will be assumed to originate through an end office which leads into machine i. Since this is so, there

---

$$\sum_{r\in I_i} z_r \leqq C_1\{S_i - \sum_{r\in I_i}(S_r + V_r + M_r) - \sum_{r^+\in O_i} W_{r^+}\} \tag{2.1}$$

$$\sum_{r\in I_i} u_r \leqq C_3\{\mathfrak{M}_i - \sum_{r\in I_i} M_r\} \tag{2.2}$$

$$x_{(i,j)} = \frac{\lambda_{ij} - P \sum_{R=(i,\cdots,j)} A_R z_R}{1 - P} - b_{(i,j)} \tag{2.3}$$

$$\sum_{R\supset i,j} N_R + \sum_{r\supset i,j}(W_r + S_r + V_r + M_r) \leqq C_{ij} \tag{2.4}$$

$$v_r = C_2 S_r \tag{2.5}$$

$$u_r = C_4 M_r \tag{2.6}$$

$$x_r = z_r + t_r \tag{2.7}$$

$$z_r = v_r \tag{2.8}$$

$$A_R z_R = \nu N_R \tag{2.9}$$

$$u_r = v_r + \sum_{r^+\in O_i} t_{r^+} = \sum_{r^+\in O_i} x_{r^+} + b_r \tag{2.10}$$

Fig. 2—Equations and inequalities described in Section V.

will be a possibility that a call which wishes to enter the network at machine i with destination machine j may be blocked in the end office prior to its getting into the network. Let $b_{(i,j)}$ denote this rate at which calls from i to j are blocked in the originating end office.

If $r = (i, j)$, that is, r is the partial route of a call just starting at i whose destination is j, then the rate of flow onto r is given by

$$x_r = a_{ij} - b_{(i,j)}.$$

Since all the variables of the problem are mean numbers of calls or mean rates at which calls are flowing, all variables must be non-negative.

The equations and inequalities which have been described so far are gathered together in Fig. 2.

Rewriting the equations of Fig. 2, it can be seen that x_r, v_r, u_r, N_R, S_r, and M_r can be expressed in terms of z_r, t_r, and b_r. Eliminating these variables from the problem, we arrive at the equations expressed in Fig. 3.

$$z_r = b_r + \sum_{r^+ \in O_i} z_{r^+} \tag{3.1}$$

$$\sum_{r \in I_i} \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_4} \right) z_r + \sum_{r \in I_i} V_r + \sum_{r \in O_i} W_{r^+} + \frac{1}{C_4} \sum_{r \in I_i} \sum_{r^+ \in O_i} t_{r^+} \leqq S_i \tag{3.2}$$

$$\left(\frac{1}{C_3} + \frac{1}{C_4} \right) \left[\sum_{r \in I_i} z_r + \sum_{r \in I_i} \sum_{r^+ \in O_i} t_{r^+} \right] \leqq \mathfrak{M}_i \tag{3.3}$$

$$\sum_{R \supset i,j} \frac{A_R}{\nu} z_R + \sum_{r \supset i,j} \left(W_r + \frac{1}{C_2} z_r + V_r + \frac{1}{C_4} Z_r + \frac{1}{C_4} \sum_{r^+ \in O_i} t_{r^+} \right) \leqq C_{ij} \tag{3.4}$$

$$z_{(ij)} + t_{(ij)} + b_{(ij)} = \left[\lambda_{ij} - P \sum_{R = (i, \cdots, j)} A_R z_R \right] [1 - P]^{-1} \tag{3.5}$$

Fig. 3—Formulas of the relaxed optimization problem.

VI. AN OPTIMIZATION PROBLEM

Up to this point we have been considering the very complex stochastic process of calls progressing through a telephone network. For any such process, the constraints in Fig. 3 must be satisfied by the appropriate mean values. Next we will consider the relaxed optimization problem mentioned in Section I.

The total expected number of talking calls at a given time is

$$\sum_R N_R. \tag{15}$$

Suppose it is desired to maximize this quantity. From (10) this is equivalent to maximizing

$$\sum_R A_R z_R. \tag{16}$$

Let us first assume that it is possible to control the variables

$$z_R,\ z_r,\ V_r,\ W_r,\ b_r,\ t_r.$$

We formulate the optimization problem of maximizing (16) subject to the formulas of Fig. 3.

This is the relaxed optimization problem mentioned in Section I. Notice that while we have argued that the formulas of Fig. 3 must hold, it seems apparent that other relationships than those given in Fig. 2 will have to be satisfied for corresponding variables in the real network. However, we shall treat this relaxed optimization problem as a "model" for the network and investigate the optimum mean number of calls carried by this model. Later, for an example, the calls carried by the model and those carried by a call-by-call simulation will be compared to show that these call-carrying capacities are close. In this sense, this indicates that the model mirrors the major features of the call-carrying capacity of the network.

Notice that the variables which are to be controlled contain mean rates of flow of calls throughout the network. Implicit in this is the assumption that the routing of calls is to be chosen in the relaxed optimization problem. This will result in a routing different than conventional alternate routing. The relaxed optimization problem as formulated is a linear programming problem. If rules for alternate routing were imposed in addition to the formulas of Fig. 2, this linearity would be destroyed. For this reason the more flexible type of routing will be allowed rather than insisting on conventional alternate routing.

Notice that it might be felt that the variables to be controlled by the program are not really subject to control in a real network. It will

be shown that the optimum solution only adjusts a subset of those variables and it does appear that the variables which must be adjusted in the optimum solution are subject to adjustment in a real network.

VII. GRADE OF SERVICE CONSTRAINT

In the relaxed optimization problem that has just been formulated, no provision has been made to assure maintenance of an appropriate level of service. For instance, it is conceivable that the solution of the optimization problem would deny service completely to some point-to-point pair if the facilities on the route it used could be better utilized by other traffics. A provision should be made to insure at least a certain level of calling between each point-to-point pair.

Since the situation to be considered is one in which the network is already overloaded, it will not be possible to keep the blocking for each point-to-point pair below desired levels. However, it is possible to require that each point-to-point pair has the capability of completing through the network a fixed minimal number of calls per unit time. The total mean rate at which calls are completed through the network between point-to-point pair (i, j) is

$$\sum_{R=(i,\cdots,j)} z_R.$$

The inequalities

$$\sum_{R=(i,\cdots,j)} z_R \geqq K_{ij} \tag{17}$$

assure that calls between each pair (i, j) will be completed through the network at a mean rate greater than or equal to K_{ij}. In the future inequalities, (17) will always be added to the formulas of the relaxed optimization problem given in Fig. 3.

VIII. REDUCTION OF THE OPTIMIZATION PROBLEM

Notice that in this problem if z_R, z_r, b_r, t_r, V_r, W_r is an optimal solution, then there is another optimal solution with the same z_R and z_r, the same b_r for r not of the form $r = (ij)$, new $b(ij)$ equal to the previous $b(ij) + t(ij)$, and with $V_r = W_r = t_r = 0$. This is so since the equations and inequalities are still satisfied and the quantity (16) to be maximized is unchanged.

Notice also that eq. (3.1) in Fig. 3 implies that

$$z_r \geqq \sum_{r \supset R} z_R. \tag{18}$$

In this, the notation $r \supset R$ means that the sum is to be taken over all

complete routes which have the partial route r as their common beginning and destination.

Let

$$z_R,\ z_r,\ b_r,\ 0,\ 0,\ 0$$

be any optimal solution of the type described above. If

$$b_r' = 0, \qquad r \neq (i, j),$$

$$b_{(ij)} = \left[\lambda_{ij} - P \sum_{R=(i,\cdots,j)} A_R z_R\right][1 - P]^{-1} - \sum_{R=(i,\cdots,j)} z_R, \tag{19}$$

$$z_r' = \sum_{r \supseteq R} z_R. \tag{20}$$

Then

$$z_R,\ z_r',\ b_r',\ 0,\ 0,\ 0$$

is an optimal solution since (20) implies equation (3.1) in Fig. 3 is satisfied, (18) and (19) imply that the equations or inequalities (3.2) through (3.5) are satisfied, and (16) and (17) are unchanged. This implies that the optimization problem may be rewritten in terms of only the variables $b_{(i,j)}$ and z_R through using eqs. (19) and (20). The optimization problem is restated in this form in Fig. 4.

IX. CONSEQUENCES OF THE OPTIMIZATION PROBLEM

Notice that the variables of the final optimization problem are

$b_{(i,j)}$ = the fraction of calls given a no-circuit announcement in their originating office

z_R = rate at which calls are being completed through the network on complete route R

and that it has been shown that if $b_{(ij)}$ and z_R are optimal values for the final relaxed optimization problem, then

$$z_R,\ z_r = \sum_{R \supseteq r} z_R, \quad V_r = 0, \quad W_r = 0, \quad t_r = 0$$

$$b_r = \begin{cases} b_{(i,j)} & \text{if } r = (i, j) \\ 0 & \text{if } r \neq (i, j) \end{cases}$$

are optimal values for the original optimization problem.

This implies that the solution of the relaxed optimization problem may be taken to be of the following form: An appropriate fraction of point-to-point attempts are blocked in their originating end office. The remainder is appropriately divided among the various complete routes

Maximize

$$\sum_R A_R Z_R \tag{4.1}$$

subject to

$$\sum_{r \in I_i} \sum_{r \supset R} z_R$$

$$\leqq \text{Min} \left[S_i \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_4} \right)^{-1}, \mathfrak{M}_i \left(\frac{1}{C_3} + \frac{1}{C_4} \right)^{-1} \right] \tag{4.2}$$

$$\sum_{R \supset i,j} \frac{A_R}{\nu} z_R + \left(\frac{1}{C_2} + \frac{1}{C_4} \right) \sum_{r \supset i,j} \sum_{r \supset R} Z_R \leqq C_{ij} \tag{4.3}$$

$$\sum_{R=(i,\cdots,j)} (1 - P + PA_R) z_R = \lambda_{ij} - (1 - P) b_{(ij)} \tag{4.4}$$

$$\sum_{R=(i,\cdots,j)} z_R \geqq K_{ij} \tag{4.5}$$

Fig. 4—Reduction of the optimization problem.

joining the origin and the destination. No further blocking takes place. The fractions of traffic assigned to each complete route are chosen so that no queues build up in senders or markers and no blocking occurs on trunks. They also maximize the total expected number of talking calls. The fraction of each point-to-point traffic which is blocked at its source and the fractions which are routed over each complete route joining a point-to-point pair can be considered to be the optimal controls chosen by the relaxed optimization problem.

Of course, the above paragraph refers to the solution of the relaxed optimization problem and not to an optimal control for a real network. Recall that the relaxed optimization problem contained some, but not all, of the constraints of a real network. As a result, its optimal solution may violate some of those additional constraints. However, the message-carrying capacity found by the relaxed optimization problem must be at least as large as the capacity of the corresponding real network.

The optimal solution of the relaxed problem is suggestive of a good, but suboptimal, control for a real network: Code block calls in the end offices by the same amount as used in the relaxed optimization problem

and route the calls as indicated. Do not intentionally block calls at any other point. Since the real system is stochastic, this will result in some queues forming and, depending on the method of implementation, in some blocking internally in the network. Because of this, the control might be improved by choosing slightly different values of code blocking.

X. JUSTIFICATION OF THE OPTIMIZATION PROBLEM

It is natural and crucial to ask if the capacity of the model will approximate the capacity of the stochastic network. It might be thought that the optimal solution of the model is an "Alice in Wonderland" solution due to the relaxed nature of this model, especially since it chooses just the right amounts of traffic to route over each complete route so that there is never any internal blocking or any queues in the switching mathines. We will try to demonstrate that this is not so by the following argument.

It has been shown in establishing the equations and inequalities of Fig. 4 that corresponding variables for any real network must satisfy these conditions. Hence, these variables for a real network are a feasible set for the optimization problem. Thus the capacity of the model is an upper bound on the capacity of a real network. We will show in typical examples that controls similar to those of the optimization problem when incorporated in a stochastic simulation of these networks give a carried load close to this upper bound.

The first example is shown in Fig. 5. The small network shown there was subjected to severe overloads. An overload factor of one corre-

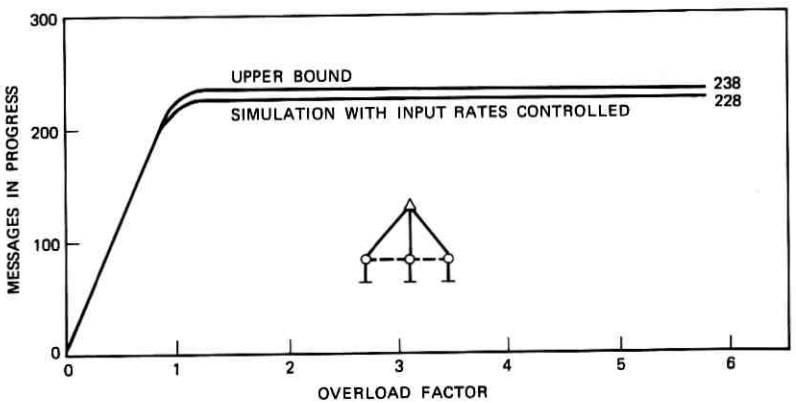

Fig. 5—Comparison of upper bound and achieved carrier loads with severe overloads.

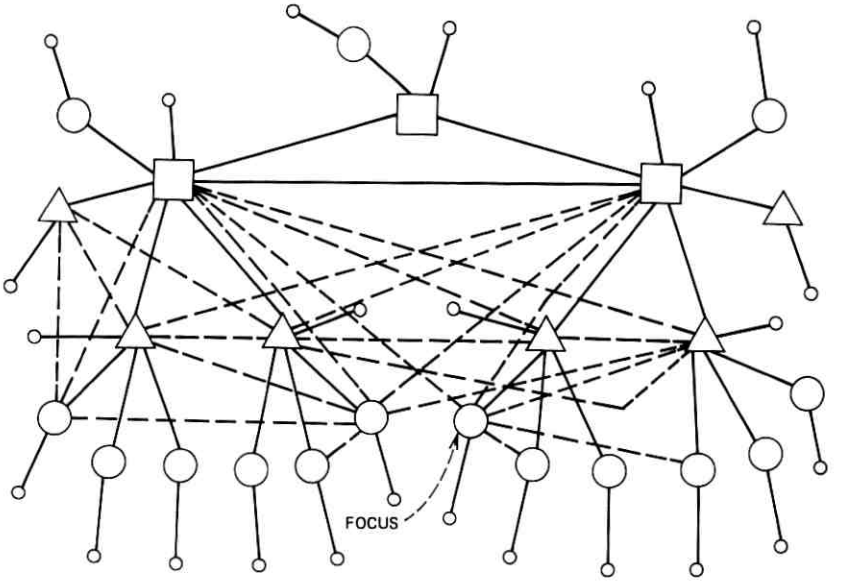

Fig. 6—Network used in focused overload example.

sponds to the traffic load for which the network was designed, an overload factor of two corresponds to twice the design load, and so forth. For all the overloads shown the upper bound on the network capacity was 238 messages. A Monte Carlo simulation of the network was run restricting the offered traffic at the end offices. The steady-state average number of messages carried was 228.

The second example was run on the considerably larger network shown in Fig. 6. This network received its design offered load except that all offered traffics destined for the node marked "Focus" were eight times their design levels and all traffics originating at that node were twice their design levels. Focused overloads of this type occur in the toll network. A similar pattern might occur following a natural disaster in the vicinity of the node marked "Focus."

Four cases were run on this network with this offered load. In all four cases only in-chain routing was used, i.e., only message paths which could exist under the current hierarchial routing scheme were allowed. The results are shown in Fig. 7. First, a Monte Carlo simulation was run using short sender timing as the only network management control. The average number of messages carried in steady state was 250. Second, the simulation was run using the network manage-

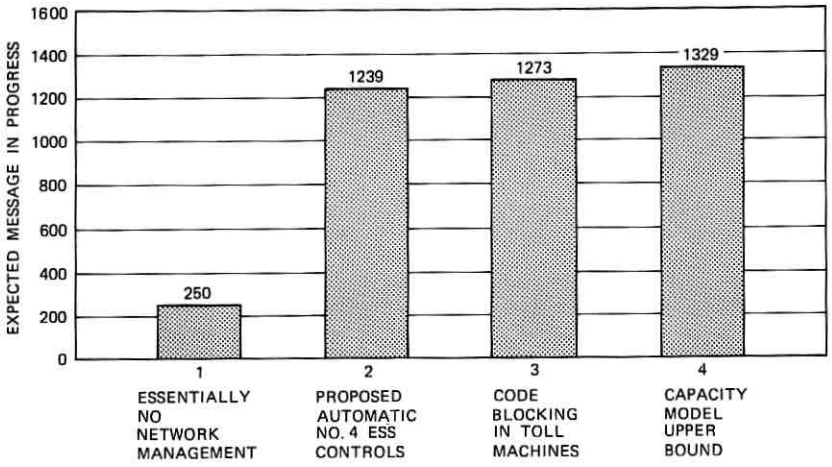

Fig. 7—Comparison of upper bound and achieved carrier loads with focused overload.

ment controls proposed for the No. 4 ESS switching machine, the next generation of Bell System toll switching machines. The steady-state average number of messages carried by the network was 1239. Third, an analytic model[15] was run with code blocking, not in the end offices but in their associated toll machines. The steady-state average number of messages carried was 1273. Finally, the upper bound on the network capacity was 1329.

In this example the proposed, economically feasible No. 4 ESS controls achieve most of the improvement between the essentially uncontrolled case and the upper bound on network capacity. The upper bound is more nearly attained by choosing code blocking in the toll centers, based on knowing the mean offered loads. Presumably the upper bound could be even more closely approached if code blocking was done in the end offices and some routing controls were included.

In summary, not only is it likely that controls based on complete knowledge of the underlying distributions can nearly achieve the upper bound on capacity, but also economically feasible control schemes can approach it.

XI. SOLUTION INFORMATION

Computer programs have been set up to solve the final linear programming problem. The program consists of two parts. The first part takes given data on the network, such as number of toll centers, which toll centers are connected by links, number of trunks on a link, and machine operating constants. Then it writes the equations given in

Fig. 4 in a form suitable for use in a linear programming algorithm. The second part uses a linear programming algorithm to solve the linear programming problem. These programs have been used to compute optimal controls for moderate-sized networks.

For a network with 17 nodes, 57 links, and 272 point-to-point offered traffics, it cost about $45 to run both programs on an IBM 360. For 21 nodes, 204 links, and 441 point-to-point offered traffics, it cost about $100 to run both programs on the same machine. By modifying the standard linear programming algorithm, taking into account the special structure of the model, it appears that it will be feasible to compute capacities for very large networks.

The output of the solution algorithm for the linear programming problem contains much additional information which can be important in evaluating the network's operation. The amount of slackness in the inequality constraints and the optimal variables for the dual linear program are printed out.

In Fig. 4, the left side of inequality (4.2) is the rate at which calls are processed in switching machine i. The left side of inequality (4.3) is the average number of calls on the trunk group connecting machines i and j. The left side of inequality (4.5) is the completion rate for calls originating at machine i and destined for machine j.

The dual variables for inequalities (4.2) and (4.3) are related to the incremental increase in the optimal carried load which could be achieved by adding one sender or marker or one trunk in the indicated place. The dual variables for inequality (4.5) give the incremental increase in the carried load which could be achieved by relaxing the minimum service constraint for a particular source-destination pair, i-j.

The amount of slackness in inequalities (4.2) and (4.3) gives information on how efficiently the toll centers and trunks are being utilized in carrying the given traffic.

XII. CONCLUSIONS

A steady-state, mean-flow-rate model of an overload telephone network with trunk and switching congestion was set up and optimized. The expected number of calls carried by this model is an upper bound for the number of calls which can be carried in the real network. The model takes into account the progression of calls along their routes as they are being set up and the amount of trunk space and switching machine capacity used by the calls which are in the process of being set up. The form of the model was such that the optimization could be carried out using conventional linear programming.

The calls carried by the optimized model can be used as a standard against which the calls carried by various network management control systems can be compared. The form of the controls of the optimized model should provide insight for network management. Computations indicate that there can be a significant difference between the number of calls carried by an unmanaged system and the optimally managed system. Some information which may be valuable in assessing the need for additional facilities in the network is available from peripheral information supplied by the linear program.

XIII. ACKNOWLEDGMENTS

The authors would like to express their appreciation for helpful discussions with Richard Ellis, Jack Holtzman, Sheldon Horing, Edwin Messerli, Patrick Spagon, Irvin Yavelberg, and Roger Wets. We also thank John Kohut and Pamela Vaughn for the use of programs they have written, and Terry Leventhal for finding the network response with No. 4 ESS controls used in Fig. 7.

REFERENCES

1. Bader, J. A., unpublished work.
2. Burke, P. J., unpublished work.
3. Burke, P. J., "Automatic Overload Controls in a Circuit-Switched Communications Network," Proc. Nat. Elec. Conf., *24* (December 1968), pp. 667–672.
4. Laude, J. A., "Local Network Management Overload Administration of Metropolitan Switched Networks," Proc. Nat. Elec. Conf., *24* (December 1968), pp. 673–678.
5. Weber, J. H., "A Simulation Study of Routing and Control in Communications Networks," B.S.T.J., *43*, No. 6 (November 1964), pp. 2639–2676.
6. Beneš, V. E., "Optimal Routing in Connecting Networks Over Finite Time Intervals," B.S.T.J., *46*, No. 10 (December 1967), pp. 2341–2352.
7. Kushner, H. J., *Introduction to Stochastic Control*, New York: Holt, Rinehart & Winston, 1971.
8. Ross, S. M., *Applied Probability Models with Optimization Applications*, New York: Holden Day, 1970.
9. Rishel, R. W., "Necessary and Sufficient Dynamic Programming Conditions for Continuous Time Stochastic Optimal Control," SIAM J. Cont., *8*, No. 4 (1970), pp. 559–572.
10. Gamkrelidze, R. D., "On Sliding Optimal States," Sov. Math.—Dokl. (1962), pp. 559–562.
11. Warga, J., "Relaxed Variational Problems," J. Math. Anal. & Applic., *4* (1962), pp. 111–128.
12. Little, J. D. C., "A Proof of the Queueing Formula: $L = \lambda W$," Oper. Res., *9*, (1961), pp. 383–387.
13. Jewell, W. S., "A Simple Proof of: $L = \lambda W$," Oper. Res., *15* (1967), pp. 1109–1116.
14. Wilkinson, R. I., "Some Characteristics of Telephoning Behavior," Advanced Tel. Traffic Eng. Conf., Michigan State University (July 23, 1970).
15. Franks, R. L., and Rishel, R. W., "Overload Model of Telephone Network Operation," to be published in November 1973 B.S.T.J.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Analysis of First-Come First-Served Queuing Systems With Peaked Inputs

By H. HEFFES

(Manuscript received March 8, 1973)

This paper treats the problem of analyzing a first-come first-served queuing system, in equilibrium, when subjected to a peaked input (e.g., traffic overflowing a trunk group with Poisson input). The basic GI/M/N (renewal input to N exponential servers) queuing result is used, together with each of two models for representing peaked traffic, the Equivalent Random (E-R) model and the Interrupted Poisson Process (IPP) model. The equilibrium virtual delay distribution is derived and compared with the equilibrium distribution of delays seen by arriving calls. Numerical examples are presented, along with comparisons of results using both the above models. The results show that delays can be quite sensitive to peakedness.

I. INTRODUCTION

It is well known, from analysis of blocking in trunk groups, that the blocking seen by peaked traffic (e.g., traffic overflowing a first-choice trunk group with Poisson input) can be significantly larger than blocking seen by Poisson traffic with the same intensity. In this paper, we are interested in determining the effect peaked traffic has on delays in queuing systems. The analysis was motivated by a study of sender attachment delay in Crossbar Tandem switching machines receiving alternate routed (peaked) traffic.

We treat the problem of analyzing a first-come first-served queuing system with peaked input for the situation where there is no idle server if there is a waiting customer. The basic tool is the GI/M/N queuing result which requires a characterization of the input process in terms of the Laplace-Stieltjes transform of the interarrival time distribution. This characterization is provided using Wilkinson's Equivalent Random[1] (E-R) model where the peaked input is modeled as an overflow process from a finite trunk group with Poisson input.

The size of the finite trunk group and intensity of the Poisson input are chosen so that the overflow process produces the desired mean and peakedness (variance to mean ratio of trunks up on an infinite trunk group). The E-R method has been widely used in analyzing blocking in trunking networks and is considered our basic model when the source of the traffic peakedness is from alternate routing.

A second model, which gives a much simpler characterization of the peaked input (both computationally and analytically), is the Interrupted Poisson Process.[2] Here the peaked traffic is considered to be the output process of a switch, with Poisson input, where the switch is opened and closed for independent, exponentially distributed time intervals. The parameters of the switch are chosen to match either the first two or three moments of trunks up on an infinite trunk group with the corresponding moments obtained for the E-R model.

Comparisons between results using each of the above models are presented along with a set of numerical results which show that delays can be quite sensitive to peakedness.

For Poisson traffic, the virtual delay distribution (time congestion)* is identical with the delay distribution seen by arriving calls (call congestion). This is not the case for peaked traffic. Since in some applications measurements form estimates of time congestion (e.g., SADR measurements), it is of interest to relate the time and call congestion quantities. Numerical results show the sensitivity of the relationship to peakedness.

II. GI/M/N QUEUING RESULTS†

Consider a recurrent input‡ to an N server queuing system. The service times are independent, exponentially distributed with the mean service time given by $1/\mu_2$. The customers are served in their order of arrival, and there is no idle server if there is a waiting customer.

Let $F(x)$ denote the distribution function of the interarrival times. The mean interarrival time, $1/\lambda_2$, is given by

$$\frac{1}{\lambda_2} = \int_0^\infty x dF(x) \tag{1}$$

* Time and call congestion are commonly used in trunking analysis (BCC system; i.e., blocked calls cleared). In the delay case (BCD system), we use call congestion for the delays seen by arriving calls and time congestion for the virtual delay. See appendix for precise definitions.

† The reader is referred to chapter 2 of Ref. 3 for a more detailed mathematical description of these results.

‡ The peaked input will be modeled by recurrent processes. For these processes, $F(0^+) = 0$.

and the Laplace-Stieltjes transform of $F(x)$ is given by

$$\Phi(s) = \int_0^\infty e^{-sx} dF(x). \tag{2}$$

We define the load/server (sometimes called occupancy) as

$$\rho = \frac{\lambda_2}{N\mu_2}. \tag{3}$$

The following result is available to us:[3] If $\rho < 1$, then the equilibrium delay distribution (as seen by arriving calls, i.e., call congestion) exists and is given by

$$Pr[\text{delay} > T] = \frac{A}{1-\omega} \exp\left[-N\mu_2(1-\omega)T\right]. \tag{4}$$

The exponential delay distribution is seen to be a function of two parameters, ω and A. The parameter ω is the solution, in (0, 1), of the equation

$$\omega = \Phi[N\mu_2(1-\omega)]. \tag{5}$$*

For $\rho < 1$, this equation is known to have a unique solution in (0, 1); furthermore, the solution can be found by successively iterating on (5); i.e.,

$$\omega_{i+1} = \Phi[N\mu_2(1-\omega_i)] \tag{6}$$

with $\omega_0 \in [0, 1)$. If we define

$$\Phi_j = \Phi(j\mu_2) \tag{7a}$$

and

$$C_j = \prod_{\nu=1}^{j} \frac{\Phi_\nu}{1-\Phi_\nu}, \tag{7b}$$

then the parameter A is given by

$$A = \frac{1}{\left[\frac{1}{1-\omega} + \sum_{j=1}^{N} \frac{\binom{N}{j}[N(1-\Phi_j)-j]}{C_j(1-\Phi_j)[N(1-\omega)-j]}\right]}. \tag{8}$$

Thus, given the characterization of the input in terms of $\Phi(s)$, eqs. (6), (7), and (8) provide the means to compute the equilibrium delay distribution as seen by arriving calls (4). The mean of the equilibrium

* In Poisson traffic, the solution of (5) is $\omega = \rho$. In some cases, to be treated later, (5) can be solved in closed form.

delay distribution is given by

$$E[\text{delay}] = \frac{A}{N\mu_2(1-\omega)^2}. \tag{9}$$

In the next two sections we discuss the Equivalent Random and Interrupted Poisson models for generating peaked traffic and characterizing $\Phi(s)$.

III. EQUIVALENT RANDOM MODEL

The E-R model treats peaked traffic as an overflow process from a finite trunk group with Poisson input. The holding time for the trunk group is assumed to be exponential with mean $1/\mu_1$.* The number of trunks and the intensity of the Poisson traffic are chosen so the mean and variance to mean ratio (peakedness) of trunks up on an infinite overflow group closely match the desired mean and peakedness. If we denote the desired mean and peakedness by m_d and z_d respectively, then Rapp's formulas[4] for the Poisson load (a_{eq}) and number of trunks (c),

$$a_{eq} = \frac{\lambda_{eq}}{\mu_1} = m_d z_d + 3z_d(z_d - 1) \tag{10}$$

$$c = a_{eq}\left(\frac{m_d + z_d}{m_d + z_d - 1}\right) - m_d - 1, \tag{11}$$

yield an overflow process which approximates the desired process. It should be noted that use of (10) and (11), or truncation of c from (11), will often lead to overflow traffic with mean and peakedness different from (but usually close to) the desired values. To quantify this effect, the actual mean and peakedness should be computed using

$$m = a_{eq}B(c, a_{eq}) \tag{12}$$

$$z = \left[1 - m + \frac{a_{eq}}{c + 1 + m - a_{eq}}\right], \tag{13}$$

where B is the Erlang B function.

Takács[3], Chapter 4, shows that the Laplace-Stieltjes transform of the interarrival time distribution of the overflow traffic is given by

$$\Phi(s) = \sum_{j=0}^{c}\binom{c}{j}\frac{1}{\lambda_{eq}^{j}}\prod_{i=0}^{j-1}(s + i\mu_1)\Big/\sum_{j=0}^{c+1}\binom{c+j}{j}\frac{1}{\lambda_{eq}^{j}}\prod_{i=0}^{j-1}(s + i\mu_1), \tag{14}$$†

* Note that we can have $\mu_1 \neq \mu_2$. For example, if overflow traffic from trunks is offered to a group of senders, $1/\mu_1$ is of the order of minutes, whereas $1/\mu_2$ is of the order of seconds.

† The value c is considered an integer. In practice, we round c given by (11) up and down and choose the one that gives an actual z closest to z_d. The actual m and z are computed from (12) and (13) using the rounded values of c.

where the empty product is unity. Note that $\Phi(s)$, given by (14), has to be repeatedly evaluated in (6) for the solution of (5) and in (8). In its present form, (14) is unsuitable for computation (for large c) because of numerical problems. In order to avoid these problems, we use a recursive method for the evaluation of $\Phi(s)$ developed by A. Descloux.[5] If we use the notation $\Phi^c(s)$ to denote the dependence of $\Phi(s)$ on c, then $\Phi^c(s)$ satisfies

$$[\Phi^k(s)]^{-1} = \frac{s}{\lambda_{eq}} + 1 + \frac{\mu_1}{\lambda_{eq}} k - \frac{\mu_1}{\lambda_{eq}} k\Phi^{k-1}(s) \tag{15}$$

with initial condition

$$\Phi^o(s) = \frac{\lambda_{eq}}{s + \lambda_{eq}}. \tag{16}$$

The solution is thus obtained as follows: Using (15) to evaluate $\Phi(s)$, we iterate on (6) to find the ω parameter and subsequently the A parameter (8). Having evaluated the ω and A parameters we can compute $Pr[\text{delay} > T]$ from (4).

We now turn our attention to the interrupted Poisson process model for generating peaked traffic, the resulting simplifications, and comparisons with the E-R model.

IV. INTERRUPTED POISSON PROCESS MODEL

This model, suggested by W. S. Hayward and analyzed by A. Kuczura,[2] treats peaked traffic as a Poisson process modulated by a random switch where the switch is opened and closed for independent, exponentially distributed time intervals. The importance of this process is that it can provide a simple and accurate approximation to overflow traffic.

This model contains three parameters, the intensity of the Poisson process into the switch, λ_s calls per second; the mean open time of the switch, $1/\bar{\omega}_s$ seconds; and the mean closed time of the switch, $1/\bar{\gamma}_s$ seconds. If we choose

$$\frac{\lambda_s}{\mu_1} = A_s = mz + 3z(z-1), \tag{17}$$

$$\frac{\bar{\omega}_s}{\mu_1} = \omega_s = \frac{m}{A_s}[m + 3z - 1], \tag{18}$$

and

$$\frac{\bar{\gamma}_s}{\mu_1} = \gamma_s = \left[\frac{A_s}{m} - 1\right]\omega_s, \tag{19}$$

then the mean and variance to mean ratio of trunks up on an infinite

trunk group, with mean holding time $1/\mu_1$, will be m and z respectively.[2] This corresponds to the two-parameter match of Ref. 2. The so-called three-parameter match is obtained by matching the first three moments of trunks up on an infinite trunk group with the corresponding moments that would be obtained using the E-R model. In this case, c and a_{eq} are computed from (10) and (11) and the switch parameters are obtained from

$$\frac{\lambda_s}{\mu_1} = A_s = a_{eq}\left[\frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)}\right], \tag{20}$$

$$\frac{\tilde{\omega}_s}{\mu_1} = \omega_s = \frac{\delta_0}{A_s}\left[\frac{A_s - a_{eq}\delta_1}{\delta_1 - \delta_0}\right], \tag{21}$$

and

$$\frac{\tilde{\gamma}_s}{\mu_1} = \gamma_s = \frac{\omega_s}{a_{eq}}\left[\frac{A_s - a_{eq}\delta_0}{\delta_0}\right], \tag{22}$$

where the δ_k, defined in Ref. 2, are given by

$$\delta_0 = B(c, a_{eq}) \tag{23}*$$

and

$$\delta_k^{-1} = \frac{c + k - a_{eq}}{k} + \frac{a_{eq}}{k}\,\delta_{k-1}. \tag{24}\dagger$$

For a given mean m and peakedness z, it has been shown[2] that the Laplace-Stieltjes transform of the interarrival time distribution of the output process of the switch is given by

$$\Phi(s) = \frac{k_1 r_1 \mu_1}{s + r_1\mu_1} + \frac{k_2 r_2 \mu_1}{s + r_2\mu_1}, \tag{25}$$

where the parameters r_1, r_2, k_1, and k_2 are given by

$$r_1 = \tfrac{1}{2}[A_s + \omega_s + \gamma_s + \sqrt{(A_s + \omega_s + \gamma_s)^2 - 4A_s\omega_s}] \tag{26}$$

$$r_2 = \tfrac{1}{2}[A_s + \omega_s + \gamma_s - \sqrt{(A_s + \omega_s + \gamma_s)^2 - 4A_s\omega_s}] \tag{27}$$

$$k_1 = \frac{A_s - r_2}{r_1 - r_2} \tag{28}$$

$$k_2 = 1 - k_1 \tag{29}$$

and A_s, ω_s, and γ_s are given by (17), (18), and (19) for the two-parameter match and by (20), (21), and (22) for the three-parameter match.

---

* B is the Erlang B function.

† This equation can be simply obtained from equation (1.15) in the appendix to Ref. 1.

With (25) defining $\Phi(s)$, eq. (5) becomes a cubic in the ω parameter of the delay distribution. Dividing the cubic by the known root at unity gives the desired root in (0, 1)

$$\omega = \left(\frac{1 + \alpha_1 + \alpha_2}{2}\right) - \sqrt{\left(\frac{1 + \alpha_1 + \alpha_2}{2}\right)^2 - \frac{\lambda_s}{N\mu_2}\left(1 + \frac{\bar{\omega}_s}{N\mu_2}\right)} \quad (30)$$

with

$$\alpha_1 = \frac{r_1\mu_1}{N\mu_2} \quad (31)$$

and

$$\alpha_2 = \frac{r_2\mu_1}{N\mu_2}. \quad (32)$$

Thus ω, given by (30), together with (8) specify the equilibrium delay distribution. Note that the iteration procedure (6) has been eliminated and that Φ_j, defined by (7a), is simple to compute using (25).

V. TIME CONGESTION

For Poisson traffic, the virtual delay distribution (time congestion) is identical with the delay distribution as seen by arriving calls (call congestion). This is not the case for peaked traffic. Since in some applications measurements form estimates of time congestion (e.g., SADR measurements), it is of interest to relate the time and call congestion quantities.

We define time congestion as the delay in being serviced experienced by a fictitious call arriving at an arbitrary time t when the system is in equilibrium.* It is shown in the appendix that, in equilibrium, the relationship between call congestion (CC) and time congestion (TC) is given by

$$Pr[\text{TC} > T] = \frac{\rho}{\omega} Pr[\text{CC} > T] \quad (33)^\dagger$$

if $\rho < 1$ and the interarrival time distribution, $F(x)$, is not a lattice distribution.‡ $Pr[\text{CC} > T]$ is the delay distribution seen by arriving calls and is given by (4).

VI. NUMERICAL RESULTS AND DISCUSSION

One measure of system performance of possible interest is the mean delay experienced by arriving calls, given by (9), versus CCS offered

* See appendix for a more precise definition of time congestion.
† For Poisson traffic, $\omega = \rho$ giving $Pr[\text{TC} > T] = Pr[\text{CC} > T]$.
‡ The E-R model and Interrupted Poisson model clearly satisfy the nonlattice hypothesis.

to the servers where

$$\#\,\mathrm{CCS} = 36\,\frac{\mu_1}{\mu_2}\,m \tag{34}$$

with m given by (12) in the E-R model and

$$m = A_s\,\frac{\omega_s}{\omega_s + \gamma_s} \tag{35}$$

for the interrupted Poisson model. Here A_s, γ_s, and ω_s are given by (17), (18), and (19) for the two-parameter match and by (20), (21), and (22) for the three-parameter match. Figures 1(a) and 1(b) are plots of the mean delay characteristics for each of the three models of interest. The values of peakedness z ranging from 1 to 3 are presented on Fig. 1(a), and the $z = 4$ results are plotted on Fig. 1(b). The parameters of this example are $N = 18$, $1/\mu_1 = 180$ seconds, and $1/\mu_2 = 7.6$ seconds.

It is seen that, while the two-parameter results tend to overestimate both the E-R and three-parameter results, the differences are indistinguishable (for the entire CCS range shown) up to $z = 1.5$ and small

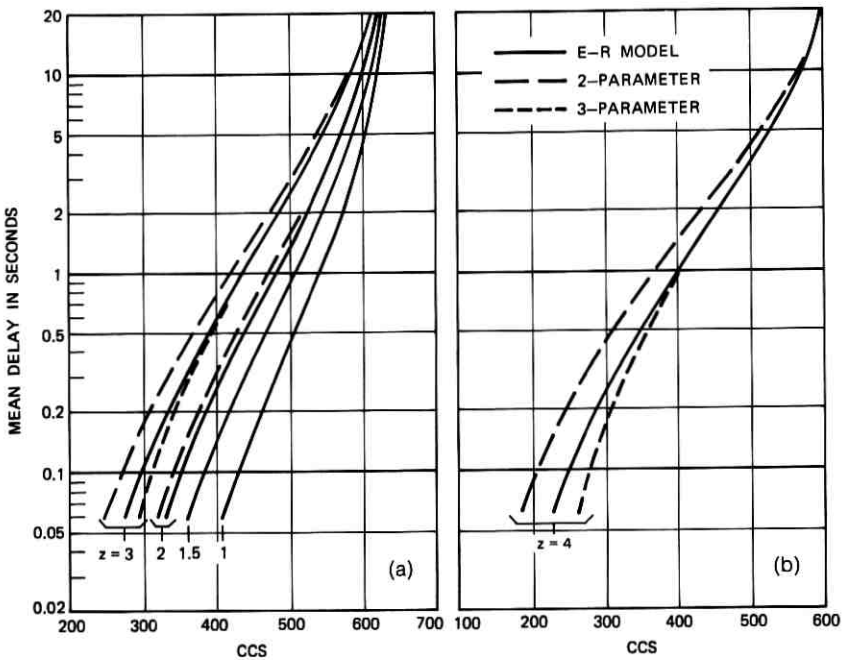

Fig. 1—Mean delay characteristics.

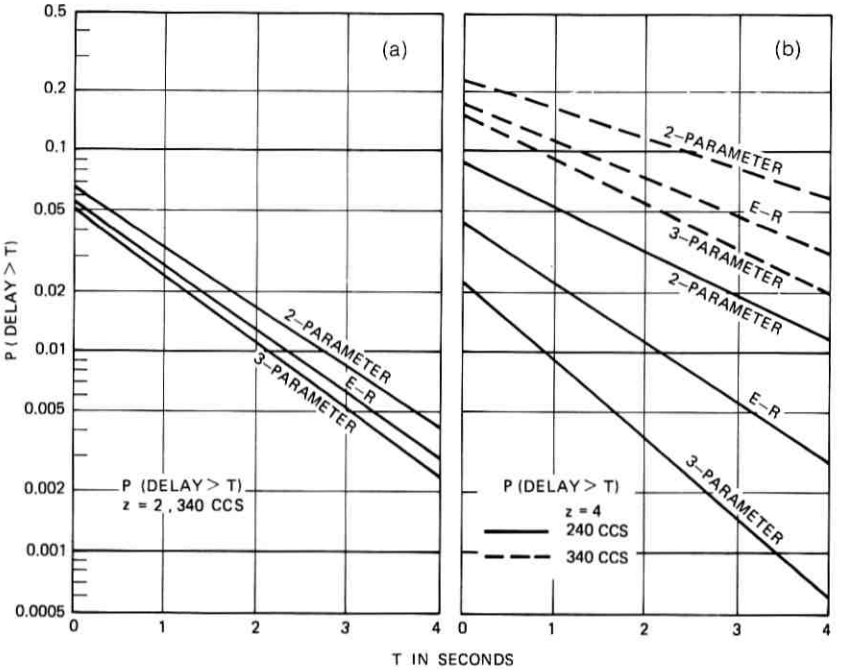

Fig. 2—Delay distribution.

up to about $z = 2$. The three-parameter results* are seen to be indistinguishable from the E-R results up to $z = 2.0$ and close up to about $z = 3$. In all cases, the E-R results tend to lie between the two- and three-parameter results. We see in Fig. 1(b) that the results differ greatly for $z = 4$. The results of Ref. 6 show that fixing the equivalent-random mean and variance for a renewal process does not necessarily tightly tie down the blocking in the BCC case. We are observing the same phenomenon here. In fact, nonnegligible discrepancies between the E-R and three-parameter results for larger z's can occur despite a matching of three moments.

Specific delay distributions are shown in Figs. 2(a) and 2(b). An observed property of the results are that the ω parameter (5) for the two-parameter case exceeded the ω parameter of the E-R model which, in turn, exceeded the ω parameter for the three-parameter case. This explains the slope differences. This also tends to order the $T = 0$ results as shown. In many cases, the A parameter of the distribution

* In order to avoid severe numerical problems in computing the three-switch parameters, double precision was used in eqs. (20) through (24).

(8) for the two-moment match exceeded the A parameter for the E-R model which, in turn, exceeded the A parameter for the three-parameter match. From these figures, we again observe the closeness of results for $z = 2$ and the large differences for $z = 4$.

Figures 3(a) and 3(b) plot the load service relationship $P(\text{delay} > 2.5$ seconds) versus offered load for each of the three models. The comparisons again exhibit the same characteristics that were seen for the mean delay results [Figs. 1(a) and 1(b)].

These results show the extreme sensitivity of the queuing system performance to the peakedness of the input process. They also give some insight into the region where two- and three-parameter results are expected to be closest, i.e., low peakedness and high congestion. When seeing the discrepancies between the two- and three-parameter matches and the E-R model, we may question which is the bench mark. If the peakedness arises from alternate routing, the E-R model seems basic since it is an overflow model which has been shown to accurately describe superposition of overflows.[1] This has led to its wide use in analyzing blocking in trunking networks. It should also be noted that the parameters of the interrupted Poisson process have been chosen to match the moments of the E-R model.

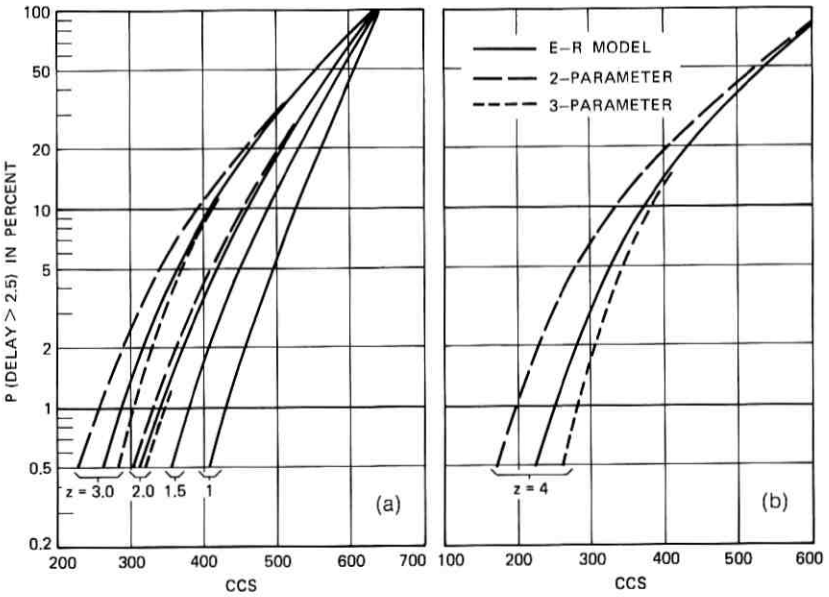

Fig. 3—Load service characteristics.

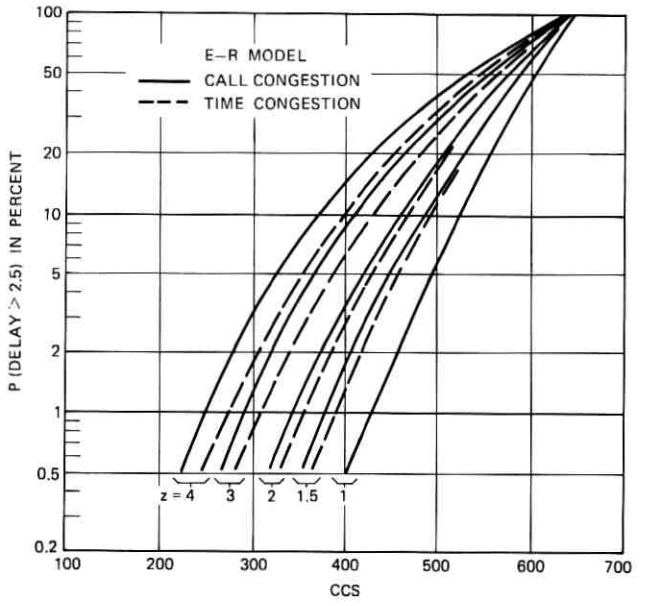

Fig. 4—Call and time congestion.

Although there are discrepancies between the two- and three-moment matches and the E-R model for high z's, the IPP is a close approximation to the E-R model for a wide range of practical z's. Furthermore, it should be emphasized that it provides a convenient method of analysis using birth and death equations (and simplified Laplace-Stieltjes transform) in many cases where the E-R model is intractable.

Figures 4 and 5 result from applying (33) to the example under consideration using the E-R model. Figure 4 shows the load service relationship for both call and time congestion. The call congestion results, taken from Figs. 3(a) and 3(b), are reproduced here for comparison. It is seen that the time congestion (TC) results consistently fall below the call congestion (CC) results.* Figure 5 shows the CC-to-TC ratio (ω/ρ) as a function of peakedness. While for a given load the time and call congestion can differ substantially, the decrease in load that makes up the difference may be relatively small.

* It has been shown by R. P. Marzec that for the Interrupted Poisson Process, $Pr[TC > 0] \leq Pr[CC > 0]$. This implies $\rho/\omega \leq 1$, which in turn implies $Pr[TC > T] \leq Pr[CC > T]$.

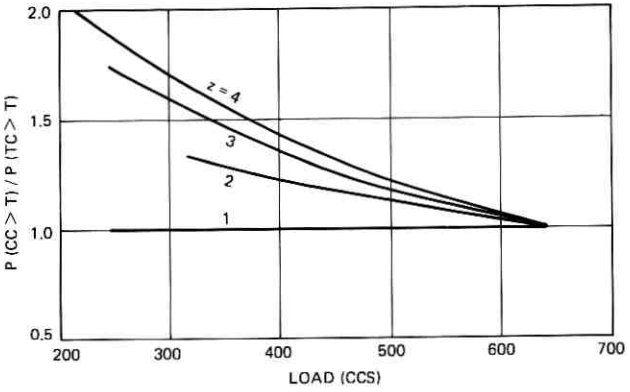

Fig. 5—Call-to-time-congestion ratio.

VII. ACKNOWLEDGMENTS

The author expresses many thanks to J. M. Holtzman for his helpful suggestions and comments. He also thanks S. E. Miller for her excellent job in programming.

APPENDIX

Time Congestion

Let $\xi(t)$ denote the state of the system (number of customers waiting or being served at time t) and let $P_k(t) = Pr[\xi(t) = k]$. Moreover, define $\xi_n = \xi(t_n^-)$ (i.e., the number of customers in the system just prior to the arrival of the nth customer). Takács shows (Ref. 3, Theorem 1, p. 148) if $\rho < 1$ then the limiting distribution

$$\lim_{n\to\infty} P[\xi_n = k] = P_k \qquad (k = 0, 1, \cdots) \tag{36}$$

exists and is independent of the initial distribution. Furthermore, he shows that

$$P_k = A\omega^{k-N} \qquad (k = N, N+1, \cdots), \tag{37}$$

where A is given by (8) and ω by (5). It is also true that the above holds for $k = N - 1$, i.e.,

$$P_k = A\omega^{k-N} \qquad (k = N-1, N, N+1, \cdots). \tag{38}$$

If we denote by η_n the waiting time of the nth customer, then the equilibrium delay distribution,

$$W(x) = \lim_{n\to\infty} W_n(x) = \lim_{n\to\infty} P[\eta_n \leqq x],$$

exists and is given by

$$W(x) = \sum_{j=0}^{N-1} P_j + \sum_{j=N}^{\infty} P_j \int_0^x e^{-N\mu_2 y} \frac{(N\mu_2 y)^{j-N}}{(j-N)!} N\mu_2 dy, \tag{39}$$

which reduces to (4). We have

$$W(x) = \sum_{j=0}^{N-1} P_j + \sum_{j=N}^{\infty} P_j I_j(x),$$

where $I_j(x)$ represents the integral in (39). The complementary distribution, $\bar{W}(x)$, is given by

$$\bar{W}(x) = 1 - W(x) = \sum_{j=N}^{\infty} P_j[1 - I_j(x)]. \tag{40}$$

If $\rho < 1$ and the interarrival distribution $F(x)$ is nonlattice, then the equilibrium time congestion probabilities given by

$$P_k^* = \lim_{t\to\infty} P_k(t) \qquad (k = 0, 1, \cdots) \tag{41}$$

exist and are independent of the initial state. Furthermore,

$$P_k^* = \rho P_{k-1} \qquad (k = N, N+1, \cdots). \tag{42}$$*

At this point, we are able to evaluate the equilibrium time congestion distribution under the hypothesis $\rho < 1$, and $F(x)$ is not a lattice distribution.

Denote $\eta(t)$ as the waiting time of a fictitious arrival at time t and let

$$W(t, x) = Pr[\eta(t) \leqq x]. \tag{43}$$

We have

$$W(t, x) = \sum_{j=0}^{N-1} P_j(t) + \sum_{j=N}^{\infty} P_j(t) \int_0^x e^{-N\mu_2 y} \frac{(N\mu_2 y)^{j-N}}{(j-N)!} N\mu_2 dy \tag{44}$$

since $\eta(t) = 0$ if $\xi(t) < N$. And if $\xi(t) = j \geqq N$, then fictitious arrival must wait for $j + 1 - N$ successive departures. These departures follow a Poisson process with intensity $N\mu_2$. Taking the limit and using (41) we have

$$W^*(x) = \lim_{t\to\infty} W(t, x) = \sum_{j=0}^{N-1} P_j^* + \sum_{j=N}^{\infty} P_j^* I_j(x). \tag{45}$$

Using (42), the complementary distribution is given by

$$\bar{W}^*(x) = \rho \sum_{j=N}^{\infty} P_{j-1}[1 - I_j(x)]. \tag{46}$$

* See Theorem 2, p. 153 of Ref. 3.

From (38), (40), and (46):

$$\bar{W}^*(x) = \frac{\rho}{\omega} \bar{W}(x) . \qquad (47)^*$$

This result is given by Ref. 7, p. 229, for a single-server queue; the multiserver case is given here for completeness.

REFERENCES

1. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U. S. A." (Appendix, J. Riordan), B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
2. Kuczura, A., "The Interrupted Poisson Process As An Overflow Process," B.S.TJ., *52*, No. 3 (March, 1973), pp. 437–448.
3. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
4. Rapp, Y., "Planning of Junction Network in a Multi-Exchange Area, I. General Principles," Ericsson Tech., *20* (1964), No. 1, pp. 77–130.
5. Descloux, A., "On Overflow Processes of Trunk Groups With Poisson Inputs and Exponential Service Times," B.S.T.J., *42*, No. 2 (March, 1963), pp. 383–397.
6. Holtzman, J. M., "The Accuracy of the Equivalent Random Method with Renewal Inputs," Seventh International Teletraffic Congress, Stockholm, 1973.
7. Cohen, J. W., *The Single Server Queue*, New York: Wiley, 1969.

---

* Note that, for Poisson traffic, $\omega = \rho$, which gives $\bar{W}^*(x) = \bar{W}(x)$.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Low-Loss Splices in Optical Fibers

By R. M. DEROSIER and J. STONE

(Manuscript received March 20, 1973)

Several methods are reported for making splices in optical fibers. The methods have application to both liquid-core and solid-core fibers and have been demonstrated for liquid-core fibers. The lowest-loss splice consists of an inserted glass pin and an outer sleeve. Best repeatable results are 0.4 dB loss in the splice. A splicing device has been constructed which provides automatic alignment of the components and automatic assembly for several fibers at once. The technique may be directly extended to multiple splicing as for fiber cables.

I. INTRODUCTION

Several methods have been demonstrated for making splices in optical fibers. These have included fused butt joints,[1,2] sandwiching of the fiber ends between grooved lucite blocks,[3,4] and mounting in a snug-fitting sleeve.[5] These techniques have been applied to multimode and single-mode fibers. The most desirable splice in an optical fiber would be one which can be made quickly and simply, causes negligible loss at the junction, and does not have a bulky connector at the junction. Furthermore, it should be made with a technique which is applicable to simultaneous splicing of a number of fibers in a bundle.

As a step toward this goal, we have constructed a splicing device which has been used to make several types of splices in multimode optical fibers. It can be used to make splices in both solid-core and liquid-core optical fibers, although we have demonstrated its use here only for liquid-core fibers. Its extension to simultaneously splicing a number of fibers is obvious and a feature to permit this has been included, although not demonstrated. Our best results, which have been obtained repeatedly, are 0.4 dB loss in the splice.

II. DESCRIPTION OF SPLICING DEVICE

The splicing device was built to be used in several different ways:

(*i*) To make temporary butt joints for a number of fibers.

(*ii*) To make temporary splices for a number of liquid-core fibers by simultaneously inserting transparent index-of-refraction-matched glass pins in each splice.[6]

(*iii*) To make permanent splices for a number of fibers by sealing a snug-fitting glass sleeve on each splice.

(*iv*) To make permanent splices for a number of liquid-core fibers by inserting glass pins and sealing on snug-fitting glass sleeves.

In all of the above applications it was intended that the device be capable of carrying out its function by aligning the ends to be spliced without any detailed alignment or separate adjustment for each fiber.

A photograph of the splicing device is shown in Fig. 1. It consists of two sliding carriers, shown in Fig. 2, which travel on snug-fitting slide bearings on precisely aligned drill rod tracks. Each of the carriers has several grooves in it into which fibers may be dropped. The cover clamp is slid in on top to hold the fibers snugly in place without crushing them.

A pan which can be slid up and down is filled with index-matching liquid. In the case of liquid-core fibers this is the core liquid. When the pan is raised up the fibers are immersed in liquid, and when it is dropped the fibers are above the liquid surface. In the center of the

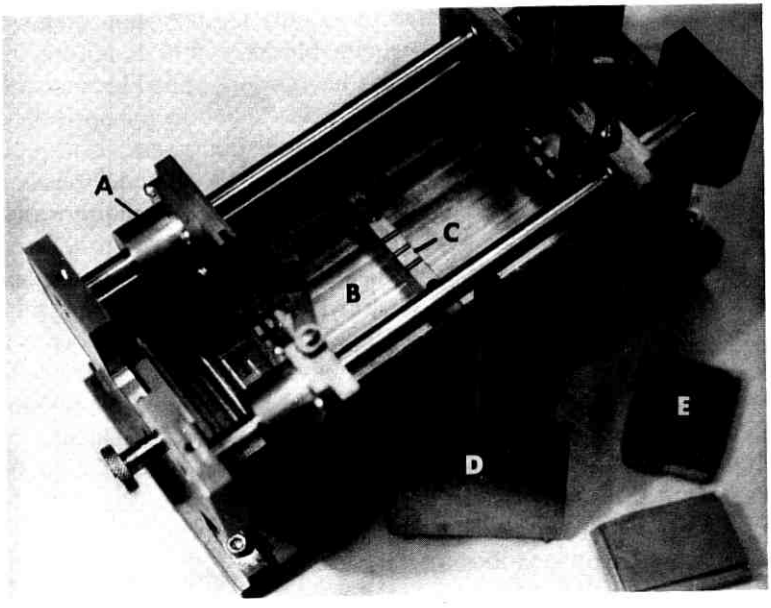

Fig. 1—Fiber splicing device. A, sliding carrier; B, immersion pan; C, pin holder; D, removable clamp for pins; E, removable clamp for fibers.

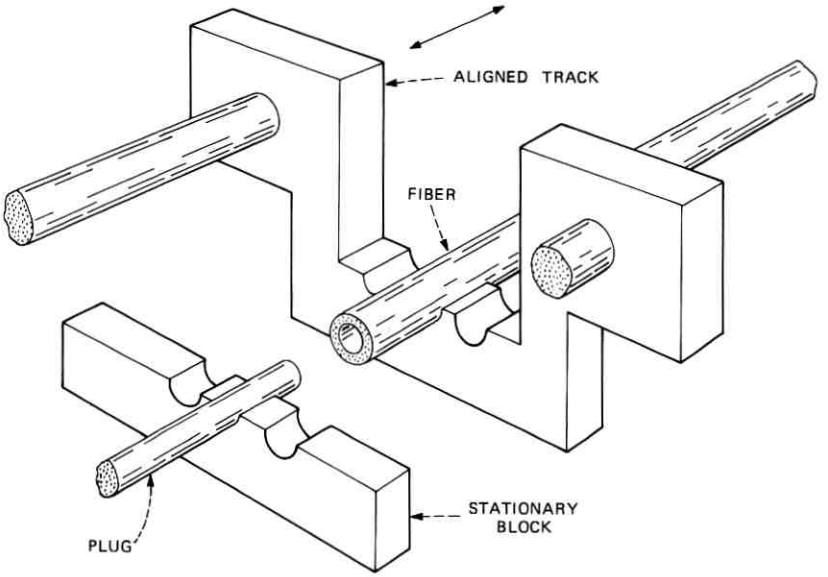

Fig. 2—Detail of fiber splicer sliding carrier and pin holder. Components are not to scale.

pan is a pedestal with grooves in it. The glass pins to be inserted into the fibers sit in these grooves and are aligned by them in such a manner that, with the pan raised to its highest position, when the two sliding carriers are brought together the pins are coaxial with the fibers to be spliced. A cover clamp is then dropped in to hold the pins in place. If pins are not used, this portion of the device is ignored. As can be seen from the construction of the device, when a splice has been completed, the spliced fiber can be lifted out of the splicing device by simply lifting off the clamps and then raising the fiber. If a number of fibers have been spliced simultaneously, each of the fibers is now an independently spliced fiber, with only a slender sleeve adding to its bulk.

III. FABRICATION OF A SPLICE

The discussion here will be given in terms of a single splice including a pin and sleeve. The two fiber ends to be joined are positioned in the aligning grooves of the sliding carriers and clamped in place. The grooves are 200 microns wide at the bottom and can be used with fibers of outside diameter of about 175 to 200 microns. An excess length of fiber is left extending from the clamp. The splice end is then obtained by scoring the fiber at the point where it is to be broken and

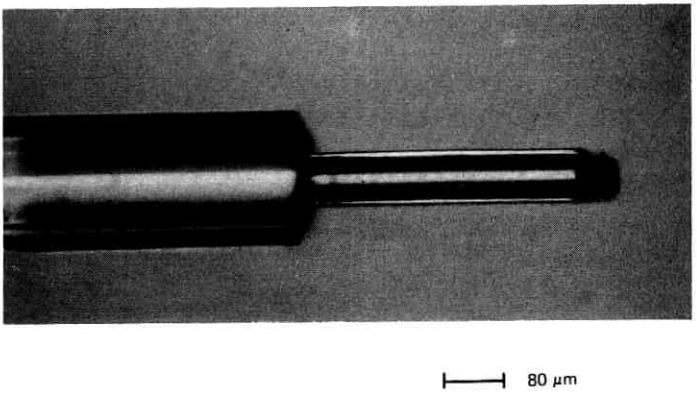

Fig. 3—Intermediate stage of splicing showing glass pin inserted into fiber core. The exposed end of the fiber is more irregular than what is used in an actual splice.

then pulling to break it. A satisfactory end is obtained in this manner, as can be seen in Fig. 3. The fiber is broken when immersed in liquid and kept immersed until the fiber has been pinned in order to avoid formation of a bubble at the junction.

The fiber pin is made from glass rod. Its index of refraction must be higher than that of the fiber cladding to provide guidance. A small index mismatch between liquid and pin will give a negligible reflection loss (a mismatch of 1 percent gives a reflection at normal incidence of 10^{-4}). However, it is better if the pin index of refraction is slightly

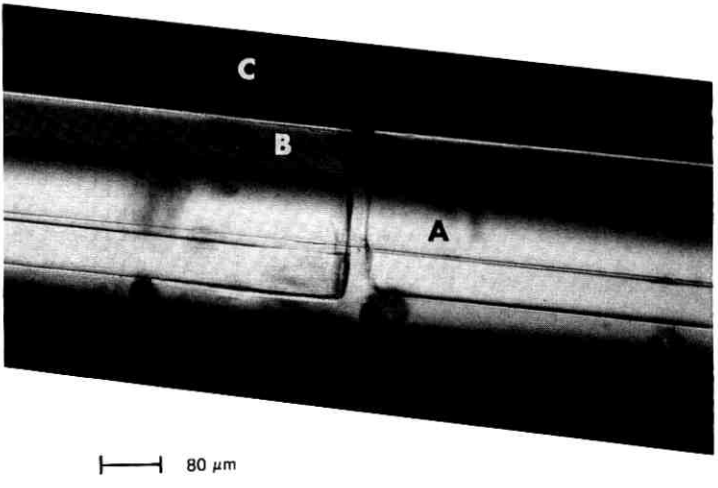

Fig. 4—Assembled splice including: A, glass pin; B, fiber cladding; C, outer sleeve.

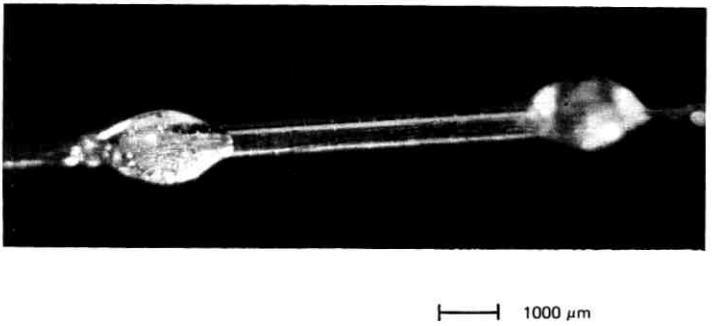

Fig. 5—Finished splice with epoxy-sealed sleeve.

higher rather than slightly lower than that of the liquid, since for the latter case there will be some mode conversion transferring energy to the cladding. The fibers used had a quartz cladding index of refraction, $n_D = 1.468$, and a tetrachloroethylene core, $n_D = 1.505$, and the glass plug was made from an ordinary commercial glass, $n_D = 1.52$. The pins were made by stretching the rod after softening in a flame. The core diameter of the fiber was about 100 microns, and loose-fitting cylindrical plugs of about 80 microns diameter were obtained in lengths of about 6 mm. It was found that loose-fitting plugs did not give extra loss and were easier to insert than tight-fitting plugs. (Any attempt to make a permanent splice by jamming a plug into the fiber ends was abandoned because the fiber broke from the pressure of the plug.) Furthermore, keeping the plugs undersized avoids the necessity for making them tapered. The plugs were cleaned with "Windex D"* before using. The pin was placed on the center pedestal of the splicing device and clamped in place. The two sliding carriers were then brought together and the pin entered both fiber ends.

Due to a small amount of residual play and misalignment (≈ 25 microns) it was necessary to manipulate the device slightly in order to insert the pin. It is this limitation which prevents the making of simultaneous splices. Further shop work should eliminate the residual errors. Figure 3 shows a pin which has been inserted into a fiber.

The outside sleeve is made of glass. Its composition and optical quality are unimportant, and other materials could be used. The sleeve is made by stretching glass tubing after softening in a flame.

A length of this tubing is slipped over one of the fiber ends and slid out of the way but kept between the sliding carriers before insertion of the fiber end in the splicing device. When the pin connection has

* Copyright 1970, The Drackett Products Co., Cincinnati, Ohio.

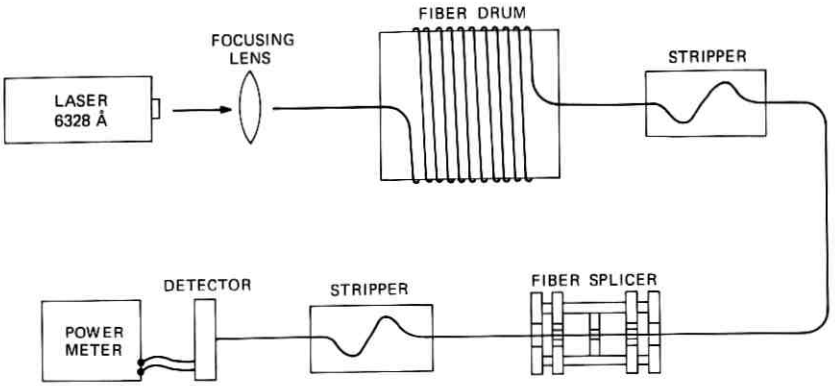

Fig. 6—Test setup for measuring splice loss.

been completed, the sleeve is slid over the splice, and then the outer clamp and the clamp on the sliding carrier blocking the sleeve are removed. The third clamp is then removed and the completed but unsealed splice is lifted out. The splice at this stage is shown in Fig. 4. The final step is to seal the sleeve onto the fiber. This is done by placing a drop of quick-curing epoxy at each end of the sleeve. The epoxy consists of equal parts of Shell Epon 828* and Minit-Cure† hardener. It cures in several minutes. The completed splice is shown in Fig. 5.

IV. MEASUREMENT OF SPLICE LOSS

The test setup used to measure the loss due to the splice is shown in Fig. 6. Light from a He-Ne laser at 6328 Å is injected into the fiber. Since, as is to be expected, the loss in the splice is mode-dependent, a realistic measurement is obtained only for a mode distribution characteristic of that existing in the fiber in its "steady state," i.e., a long distance from the beginning of the fiber, such that the mode distribution is independent of light-launching conditions. Therefore, the splices were made after about 100 m of fiber since previous measurements have shown that the steady-state condition has been obtained. Furthermore, it is necessary to strip cladding energy both before and after the fiber.

Splice loss was then obtained by measuring the fiber output without a splice and comparing it with the output with an intervening splice.

V. RESULTS

In Table I we summarize results obtained for a simple butt joint and two types of permanent splices. It can be seen that the best results

* Manufactured by Shell Petroleum Co., New York.
† Manufactured by Allaco Products, Inc., Braintree, Massachusetts.

TABLE I—SPLICING LOSS OBTAINED FOR VARIOUS TYPES OF SPLICES

Method	Two fiber ends with no plug (butt joint)	Two fiber ends aligned by an outer glass sleeve	Two fiber ends aligned with a plug and outer sleeve
Average results	1.8 dB	1.2 dB	0.53 dB (19 splices)
Best results	0.86 dB	0.70 dB	0.40 dB (6 splices)

are obtained using a plug and sleeve combination. The best results of 0.40 dB are repeatable, and compare favorably with other reported techniques. They are slightly better than any reported for multimode fibers, and are exceeded only by Someda's best results for single-mode fibers.[4] While Someda's technique presumably should give as good or better results for multimode fibers, this has not actually been demonstrated.

The epoxied splices have been found to be quite strong and, under tension, are stronger than the fiber itself which will break first when stretched. The splices have been observed over several weeks with no observable increase in loss.

VI. SUMMARY

A simple mechanical device has been constructed and used to make splices in liquid-core optical fibers. This device is also suitable for splicing solid-core fibers. The device makes a prealigned splice consisting of a transparent glass plug and an outside sealed-on sleeve. Repeatable results of 0.4 dB loss have been obtained. These compare favorably with other reported results. The device may also be used to make splices either with the outer sleeve or pin alone. However, in either case the losses are higher. The method is directly extendable to making multiple splices as, for example, in a cable. The resultant splices are compact and independent of each other.

REFERENCES

1. Bisbee, D. L., "Optical Fiber Joining Technique," B.S.T.J., *50*, No. 10 (December 1971), pp. 3153–3158.
2. Dyott, R. B., Stern, J. R., and Stewart, J. H., "Fusion Junction for Glass-Fiber Waveguides," Elec. Letters, *8*, No. 11 (June 1, 1972), pp. 290–292.
3. Standley, R. D., and Braun, F. A., "Some Results on Fiber Optic Connectors," unpublished work.
4. Someda, C. G., "Simple, Low-Loss Joints Between Single-Mode Optical Fibers," B.S.T.J., *52*, No. 4 (April 1973), pp. 583–596.
5. Astle, H. W., "Optical Fiber Connector with Inherent Alignment Feature," unpublished work.
6. Marcatili, E. A. J., patent applied for.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

A Note on Optimal Approximating Manifolds of a Function Class

By ARUN NETRAVALI

(Manuscript received March 30, 1973)

Concept of n-width and extremal subspaces, first introduced by Kolmogorov, plays an important part in mathematical problems of approximation of classes of functions and in engineering problems of signal representation and reconstruction. In this short paper, explicit expressions for n-width and extremal subspaces are obtained for a class which is of some engineering importance.

I. INTRODUCTION

Kolmogorov[1] first introduced the concept of n-width as a measure of the degree of approximation of a set Ω of a normed linear space X by linear manifolds of finite dimension. If $\mathfrak{M}$ is an n-dimensional subspace of X, then its deviation $E(\Omega, \mathfrak{M})$ from Ω is defined as

$$E(\Omega, \mathfrak{M}) = \sup_{x \in \Omega} \{ \inf_{y \in \mathfrak{M}} \|x - y\| \} \tag{1}$$

and the n-width

$$d_n(\Omega) = \inf \{E(\Omega, \mathfrak{M}) : \mathfrak{M} \subset X, \dim \mathfrak{M} = n\}. \tag{2}$$

If the lower bound in (2) is attained by $\mathfrak{M}_*$, then $\mathfrak{M}_*$ is called the n-dimensional extremal subspace.

Since Kolmogorov, several authors, including Lorentz,[2] Tihomirov,[3] Mitjagin,[4] Golomb,[5] and Jerome,[6] have obtained results on n-widths and extremal subspaces of several important function classes. Golomb has obtained expressions for n-widths and the narrowest subspaces for ellipsoids determined by a linear, nonnegative, self-adjoint operator in a Hilbert space. In this paper, these results are extended to the case of a class formed by the intersection of an ellipsoid and a unit sphere. These classes are important in signal representation and reconstruction problems whenever the magnitude of signals has to be constrained because of certain physical reasons.

II. THE MAIN RESULT

Following Golomb, if $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are subspaces of a complex Hilbert space, H, with the inner product $\langle \cdot, \cdot \rangle$, let us call $\mathfrak{M}_1$ narrower than $\mathfrak{M}_2$ (written: $\mathfrak{M}_1 \ll \mathfrak{M}_2$) whenever $\mathfrak{M}_1^{\perp} \cap \mathfrak{M}_2 \neq \phi$, ϕ being the null space and $\mathfrak{M}_1^{\perp}$ the orthogonal complement of $\mathfrak{M}_1$ in H. The class Ω under consideration is defined by

$$\Omega_1 = \{f: f \in \mathfrak{D}(A), \langle Af, f \rangle \leqq K\}, \tag{3}$$

$$\Omega_2 = \{f: \langle f, f \rangle = 1\}, \tag{4}$$

and

$$\Omega = \Omega_1 \cap \Omega_2, \tag{5}$$

where A is a linear, nonnegative [i.e., $\langle Af, f \rangle \geqq 0$ for all $f \in \mathfrak{D}(A)$], not necessarily bounded, self-adjoint operator with domain $\mathfrak{D}(A)$ which is dense in H. Let $\lambda \rightarrow E_\lambda$ be the spectral family of A which is continuous from the left and $\mathcal{E}_\lambda$ be the range of E_λ.

The case $\Omega_1 \subset \Omega_2$ is not possible, and the case $\Omega_2 \subset \Omega_1$ leads to the following trivial result:

$$\begin{aligned} E(\Omega, \mathfrak{M}) &= 1, \quad \text{if} \quad \mathfrak{M} \neq H \\ &= 0, \quad \text{if} \quad \mathfrak{M} = H. \end{aligned} \tag{6}$$

Therefore, these two cases are not considered below. It is also assumed that[†] $\sigma_{\min} < \delta^{-2}$, $\sigma_{\min} < K < \delta^{-2}$, where $\sigma_{\min}$ is the minimum value of the spectrum of A.

Theorem 1:

$$(i) \quad E(\Omega, \mathfrak{M}) \geqq \left\{\frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}\right\}^{\frac{1}{2}}, \quad \text{if} \quad \mathfrak{M} \ll \mathcal{E}_{\delta^{-2}}, \tag{7}$$

$$(ii) \quad E(\Omega, \mathfrak{M}) = \left\{\frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}\right\}^{\frac{1}{2}}, \quad \text{if} \quad \mathfrak{M} = \mathcal{E}_{\delta^{-2}}. \tag{8}$$

Proof: By definition,

$$\mathfrak{M} \ll \mathcal{E}_{\delta^{-2}} \Rightarrow \mathfrak{M}^{\perp} \cap \mathcal{E}_{\delta^{-2}} \neq \phi.$$

Then let $\mathfrak{M}^{\perp} \cap \mathcal{E}_{\delta^{-2}} = G$ and consider the following two cases:

(*A*) $\mathcal{E}_{\sigma_{\min}}$ is not empty.

In this case, take $f_0 \in \mathcal{E}_{\delta^{-2}}$ such that

$$f_0 = f_1 + f_2, \qquad f_1 \in G, \; f_2 \in \mathcal{E}_{\sigma_{\min}}, \tag{9}$$

$$\|f_1\|^2 = \frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}, \tag{10}$$

[†] If $K < \sigma_{\min}$, Ω will be empty.

and

$$\left.\begin{aligned} f_2 &= \ell \cdot \frac{P_{\mathcal{E}_{\sigma_{\min}}} f_1}{\|P_{\mathcal{E}_{\sigma_{\min}}} f_1\|} \\ &= \ell \cdot e \end{aligned}\right\} \quad \begin{aligned} &\text{if} \quad P_{\mathcal{E}_{\sigma_{\min}}} f_1 \neq 0 \\ &\text{otherwise,} \end{aligned} \tag{11}$$

where e is any unit vector in $\mathcal{E}_{\sigma_{\min}}$ and

$$\ell = \sqrt{1 - \|P_{\mathcal{E}_{\sigma_{\min}}^{\perp}} f_1\|^2} - \|P_{\mathcal{E}_{\sigma_{\min}}} f_1\|. \tag{12}$$

This construction of f_0 is illustrated geometrically in Fig. 1, where OA is a vector in G and OB is along the subspace spanned by the projection $P_{\mathcal{E}_{\sigma_{\min}}} f_1$. Then f_0 is the vector sum of f_1 and a vector f_2 along OB such that f_0 has unit magnitude.

It is easy to see that

$$\begin{aligned} \|f_0\|^2 &= \|f_1 + f_2\|^2 \\ &= \|P_{\mathcal{E}_{\sigma_{\min}}} f_1 + f_2\|^2 + \|P_{\mathcal{E}_{\sigma_{\min}}^{\perp}} f_1\|^2 \\ &= 1 \end{aligned} \tag{13}$$

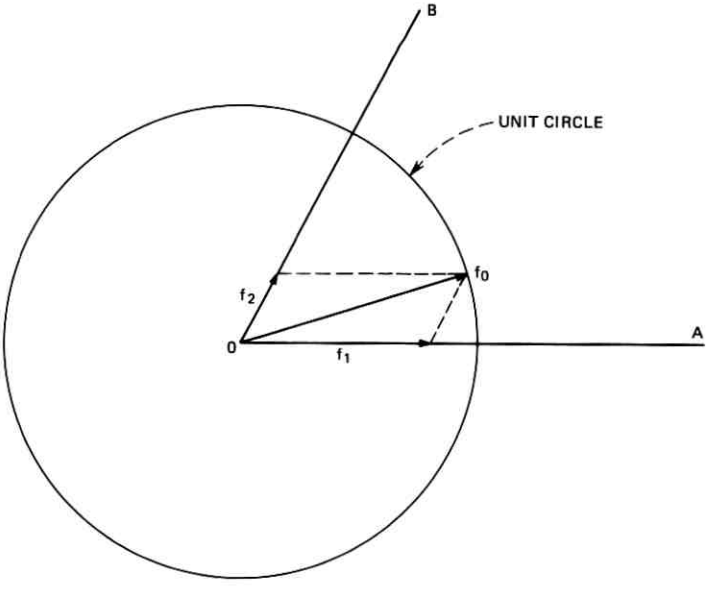

Fig. 1—Geometrical construction of f_0.

and

$$\begin{aligned}\langle Af_0, f_0\rangle &= \int_{\sigma_{\min}}^{\delta^{-2}} \lambda d[\|E_\lambda f_0\|^2] \\ &\leqq \sigma_{\min}[\|P_{\mathcal{E}_{\sigma_{\min}}} f_1 + f_2\|^2] + \delta^{-2}[\|P_{\mathcal{E}^{\perp}_{\sigma_{\min}}} f_1\|^2] \\ &\leqq \sigma_{\min}[1 - \|P_{\mathcal{E}^{\perp}_{\sigma_{\min}}} f_1\|^2] + \delta^{-2}[\|P_{\mathcal{E}^{\perp}_{\sigma_{\min}}} f_1\|^2] \\ &\leqq \|P_{\mathcal{E}^{\perp}_{\sigma_{\min}}} f_1\|^2[\delta^{-2} - \sigma_{\min}] + \sigma_{\min} \\ &\leqq K - \sigma_{\min} + \sigma_{\min} = K. \qquad (14)\end{aligned}$$

Hence $f_0 \in \Omega$. Now

$$\begin{aligned}&\|P_G f_0\|^2 \\ &= \langle f_1 + P_G f_2, f_1 + P_G f_2\rangle \\ &= \langle f_1, f_1\rangle + \langle P_G f_2, P_G f_2\rangle + 2\langle f_1, P_G f_2\rangle \\ &= \langle f_1, f_1\rangle + \langle P_G f_2, P_G f_2\rangle \\ &\qquad + \frac{2\ell}{\|P_{\mathcal{E}_{\sigma_{\min}}} f_1\|} \langle f_1, P_G P_{\mathcal{E}_{\sigma_{\min}}} f_1\rangle, \qquad \text{if} \quad P_{\mathcal{E}_{\sigma_{\min}}} f_1 \neq 0 \qquad (15) \\ &\geqq \|f_1\|^2 = \frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}. \qquad (16)\end{aligned}$$

Also, if $P_{\mathcal{E}_{\sigma_{\min}}} f_1 = 0$, then the third term in the RHS of eq. (15) is equal to zero and (16) still holds. Equations (13), (14), and (16) together imply that

$$E^2(\Omega, \mathfrak{M}) \geqq E^2(f_0, \mathfrak{M}) \geqq \frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}. \qquad (17)$$

(*B*) $\mathcal{E}_{\sigma_{\min}}$ is empty.

Since $\sigma_{\min}$ is a point in the spectrum, a sequence exists of nonincreasing and nonnegative real numbers $\{\delta_i\}$ such that each δ_i is in the spectrum of A and the following conditions are met:

$$\lim_i \delta_i = \sigma_{\min}, \qquad (18a)$$

$$\delta_i < K, \qquad \text{for all } i, \qquad (18b)$$

and $\mathcal{E}_{\delta_i}$ is not empty (i.e., it contains more than the zero element) for any δ_i. Then, for each δ_i using the same construction as in (*A*) above (using $\mathcal{E}_{\delta_i}$ instead of $\mathcal{E}_{\sigma_{\min}}$), it can be shown that

$$E^2(\Omega, \mathfrak{M}) \geqq \frac{K - \delta_i}{\delta^{-2} - \delta_i}. \qquad (19)$$

However, since $(K - \delta_i)/(\delta^{-2} - \delta_i)$ is a nondecreasing sequence of positive numbers tending to $(K - \sigma_{\min})/(\delta^{-2} - \sigma_{\min})$,

$$E^2(\Omega, \mathfrak{M}) \geqq \frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}. \qquad (20)$$

This proves (*i*).

(*ii*) Consider now the case in which $\mathfrak{M} = \mathcal{E}_{\delta^{-2}}$. For any $f \in \Omega$,

$$E^2(f, \mathcal{E}_{\delta^{-2}}) = \|f - E_{\delta^{-2}} f\|^2$$

$$= \int_{\delta^{-2}}^{\infty} d[\|E_\lambda f\|^2]$$

$$\leqq \frac{1}{\delta^{-2} - \sigma_{\min}} \int_{\delta^{-2}}^{\infty} (\lambda - \sigma_{\min}) d(\|E_\lambda f\|^2)$$

$$\leqq \frac{1}{\delta^{-2} - \sigma_{\min}} \int_{\sigma_{\min}}^{\infty} (\lambda - \sigma_{\min}) d(\|E_\lambda f\|^2) \quad (21)$$

$$\leqq \frac{K - \sigma_{\min}}{\delta^{-2} - \delta_{\min}}. \quad (22)$$

Equation (22) follows from (21) because, for all $f \in \Omega$,

$$\int_{\sigma_{\min}}^{\infty} \lambda d[\|E_\lambda f\|^2] \leqq K \quad (23)$$

and

$$\int_{\sigma_{\min}}^{\infty} d[\|E_\lambda f\|^2] = \|f\|^2 = 1. \quad (24)$$

It follows from (22) that

$$E(\Omega, \mathcal{E}_{\delta^{-2}}) \leqq \left\{\frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}\right\}^{\frac{1}{2}}. \quad (25)$$

Theorem 2:

$$E(\Omega, \mathcal{E}_{\delta^{-2}}) = \left\{\frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}\right\}^{\frac{1}{2}} \quad (26)$$

if and only if δ^{-2} *is in the spectrum of* A *and, in this case,*

$$E(\Omega, \mathfrak{M}_{\delta^{-2}}) > \left\{\frac{K - \sigma_{\min}}{\delta^{-2} - \sigma_{\min}}\right\} \quad (27)$$

if $\mathfrak{M} \ll \mathcal{E}_{\delta^{-2}}$.

Proof of this theorem is similar to the proof of Theorem 2 in Golomb[5] and is therefore omitted.

III. ACKNOWLEDGMENT

The author would like to thank H. J. Landau for helpful criticism of the original draft.

REFERENCES

1. Kolmogorov, A. N., "Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse," Ann. Math., *37*, No. 1 (January 1936), pp. 107–111.
2. Lorentz, G. G., "Metric entropy, widths and superpositions of functions," Amer. Math. Monthly, *69* (1962), pp. 469–485.
3. Tihomirov, V. M., "The widths of sets in functional spaces and the theory of best approximations," Uspehi Mat. Nauk, *15* (1960), No. 3 (93), pp. 81–120.
4. Mitjagin, B. S., "Approximation of functions in L^p- and C-spaces on torus," Mat. Sb., *58* (1962), pp. 397–414.
5. Golomb, M., "Optimal Approximating Manifolds in L_2-spaces," J. Math. Anal. Appl., *12* (1965), pp. 505–512.
6. Jerome, J. W., "On the L_2 n-width of certain classes of functions of several variables," J. Math. Anal. Appl., *20* (1967), pp. 110–113.

Contributors to This Issue

FRANCIS J. BROPHY, B.S. (Mathematics), 1968, St. Joseph's College; M.S. (Mathematics), 1971, Stevens Institute of Technology; Bell Laboratories, 1968—. Mr. Brophy has been involved with software simulation of various data transmission techniques and most recently with the design of digital filters.

PETER CUMMISKEY, B.S. (Electrical Engineering), 1963, Fairleigh Dickinson University; M.S. (Electrical Engineering), 1966, and Dr.Sc. (Electrical Engineering), 1973, Newark College of Engineering; Bell Laboratories, 1962—. Mr. Cummiskey has designed experimental hardware for use in speech research. He has also been called upon to interface equipment to minicomputers and to program them. Mr. Cummiskey's most recent work has been with delta modulation and DPCM coding of speech signals. The work contributed to the paper in this issue was done in partial fulfillment of the requirements for the Doctor of Science degree at Newark College of Engineering.

RICHARD M. DEROSIER, A.A.S.E.E., 1967, Hudson Valley Community College; Bell Laboratories, 1967—. Initially, Mr. Derosier's work concerned the fabrication and development of GaAs injection laser diodes. He is also associated with studies of mode conversion and radiation losses from various dielectric waveguides. Currently, he is working with fiber joining techniques.

JAMES L. FLANAGAN, B.S., 1948, Mississippi State University; S.M., 1950, and Sc.D., 1955, Massachusetts Institute of Technology. Faculty of Electrical Engineering, Mississippi State University, 1950–1952; Air Force Cambridge Research Center, 1954–1957. Bell Laboratories, 1957—. Mr. Flanagan has worked in speech and hearing research, computer simulation and digital encoding, and acoustics research. He is Head, Acoustics Research Department. Fellow, IEEE; Fellow Acoustical Society of America; Tau Beta Pi; Sigma Xi; member of several government and professional society boards, including committees of the National Academy of Sciences and the National Academy of Engineering.

GERARD J. FOSCHINI, B.S.E.E., 1961, Newark College of Engineering; M.E.E., 1963, New York University; Ph.D. (Mathematics), 1967,

Stevens Institute of Technology; Bell Laboratories, 1961—. Mr. Foschini initially worked on real-time program design. Since 1965, he has been mainly engaged in analytical work concerning the transmission of signals. Currently, he is working in the area of data communication theory. Member, Sigma Xi, Mathematical Association of America, American Men of Science, New York Academy of Sciences.

Richard L. Franks, B.S.E.E., 1963, University of Washington; M.S. (E.E.), 1969, and Ph.D. (E.E.), 1970, University of California, Berkeley; Bell Laboratories, 1970—. Mr. Franks has done work in control theory and algorithms. His current interest is in the modeling and analysis of telephone traffic systems. Member, IEEE, Tau Beta Pi, Sigma Xi.

Richard D. Gitlin, B.E.E., 1964, City College of New York; M.S., 1965, and D.Eng.Sc., 1969, Columbia University; Bell Laboratories, 1969—. Mr. Gitlin is presently concerned with problems in data transmission. Member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

D. Gloge, Dipl. Ing., 1961, Dr. Ing., 1964, Technical University of Braunschweig, Germany; Bell Laboratories, 1965—. Mr. Gloge's work has included the design and field testing of various optical transmission media and the application of ultra-fast measuring techniques to optical component studies. He is presently engaged in transmission research related to optical fiber communication systems.

Harry Heffes, B.E.E., 1962, City College of New York; M.E.E., 1964, and Ph.D., 1968, New York University; Bell Laboratories, 1962—. Mr. Heffes' work was previously in the areas of control theory, filtering theory, guidance and navigation problems, trajectory optimization, and error analyses. More recently, he has been concerned with modeling and analysis of telephone traffic systems. He has also been an Adjunct Associate Professor of Electrical Engineering at New York University. Member, Tau Beta Pi, Eta Kappa Nu.

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; and on the detection of fading signals. His current interests include source encoding and pattern discrimination. Member, IEEE.

BRIAN W. KERNIGHAN, B.A.Sc., 1964, University of Toronto; Ph.D., 1969, Princeton University; Bell Laboratories, 1969—. Mr. Kernighan has worked primarily on developing efficient heuristic procedures for combinatorial optimization problems. Member, ACM, IEEE.

S. LIN, B.A., 1951, University of the Philippines; M.A., 1953, and Ph.D., 1963, The Ohio State University; Assistant Professor of Mathematics, Ohio University, 1959–1962; Lecturer and Research Associate, The Ohio State University, 1962–1963; Visiting Lecturer, Princeton University, 1972; Bell Laboratories, 1963—. Since joining Bell Laboratories, Mr. Lin has been engaged in research on heuristic techniques to solve large combinatorial problems by computer and development of computer-oriented algorithms to solve problems in number theory and combinatorial analysis. Member, American Mathematical Society, Mathematical Association of America, SIAM, AAAS, Phi Kappa Phi.

ENRIQUE A. J. MARCATILI, Aeronautical Engineer, 1947, and E.E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–54; Bell Laboratories, 1954—. Mr. Marcatili has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated on optical transmission media. Fellow, IEEE.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of two books. Fellow, IEEE; member, Optical Society of America.

ARUN NETRAVALI, B.Tech.(Hons.), 1967, Indian Institute of Technology; M.S.(E.E.), 1969, and Ph.D., 1971, Rice University; Bell

Laboratories, 1972—. Mr. Netravali is concerned with systems engineering for switched video communication systems, including the *Picturephone*® system. Member, Tau Beta Pi, Sigma Xi.

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.

Stephen O. Rice, B.S. (Electrical Engineering), 1929, and D.Sc. (Hon.), 1961, Oregon State College; Bell Laboratories, 1930–1972. Mr. Rice has been concerned with theoretical problems related to electromagnetic wave propagation, signal modulation, and noise. At the time of his retirement from Bell Laboratories, he was head of the Communications Analysis Research Department. In 1965, Mr. Rice received the Mervin J. Kelly Award from the Institute of Electrical and Electronic Engineers. Fellow, IEEE.

Raymond W. Rishel, B.S., 1952, M.S., 1953, and Ph.D., 1959, University of Wisconsin; Department of Mathematics, Brown University, 1959–1960; Boeing Company, Seattle, Washington, 1960–1968; Department of Mathematics, Washington State University, 1968–1969; Bell Laboratories, 1969–1972; Department of Mathematics, University of Kentucky, 1972—. At Bell Laboratories, Mr. Rishel did research on stochastic control, and on the application of stochastic control to network management and to control of queuing systems involved in switching machines. Member, SIAM, American Mathematical Society.

David Slepian, University of Michigan, 1941–1943; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Laboratories, 1950—. Mr. Slepian has been engaged in mathematical research in communication theory and noise theory, as well as in a variety of aspects of applied mathematics. During the academic year 1958–59, he was a Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley and during the spring semesters of 1967 and 1970 he was a Visiting Professor of Electrical Engineering at the University of Hawaii. He now is Professor of Electrical Engineering at the University of Hawaii and shares his time between that institute and Bell Laboratories. He was Editor of the Proceedings of the IEEE

during 1969 and 1970. Fellow, IEEE, Institute of Mathematical Statistics. Member, AAAS, SIAM.

JULIAN STONE, B.S. (Physics), 1950, The City College, New York; M.S. (Physics), 1951, and Ph.D. (Physics), 1958, New York University; Bell Laboratories, 1969—. Mr. Stone taught at The City College from 1952 to 1953 and 1956 to 1958. He was at The Hudson Laboratories of Columbia University from 1953 to 1969 where he was Associate Director for General Physics and was active in underwater acoustics and optics. At Bell Laboratories, Mr. Stone has been working on problems in optical transmission. Member, American Physical Society.

JACK K. WOLF, B.S.E.E., 1956, University of Pennsylvania; M.S.E., 1957, M.A., 1958, and Ph.D., 1960, Princeton University. From 1963 to 1965, Mr. Wolf was a member of the Department of Electrical Engineering, New York University. From 1965 until June 1973 he served on the faculty of the Polytechnic Institute of Brooklyn, since 1969 in the capacity of Professor of Electrical Engineering. He is currently Professor and Chairman of the Electrical and Computer Engineering Department at the University of Massachusetts. During the academic year 1968–1969, Mr. Wolf was on leave of absence from the Polytechnic as a member of the Mathematics and Statistics Research Center, Bell Laboratories. He spent the year 1971–72 as a National Science Foundation Senior Postdoctoral Fellow at the University of Hawaii. Fellow, IEEE; member, AAAS, AAUP.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

B. S. T. J. BRIEF

A Finline Radiator

By D. C. HOGG and W. E. LEGG

(Manuscript received May 11, 1973)

I. INTRODUCTION

Tapered finline might constitute a radiating element compatible with planar technology and therefore be suitable for antenna arrays fed by integrated circuitry. We have constructed such an element, not by photoetching a substrate but by mounting a tapered fin in a circular waveguide such that it is fed in Robertson's traditional way,[1] and then applying a dielectric "substrate" to the fin. Measurements of radiation patterns at 18.5 GHz are given for the prototype element using various "substrate" thicknesses. It is found that the element has considerable gain, and that performance improves with the thickness of the simulated substrate.

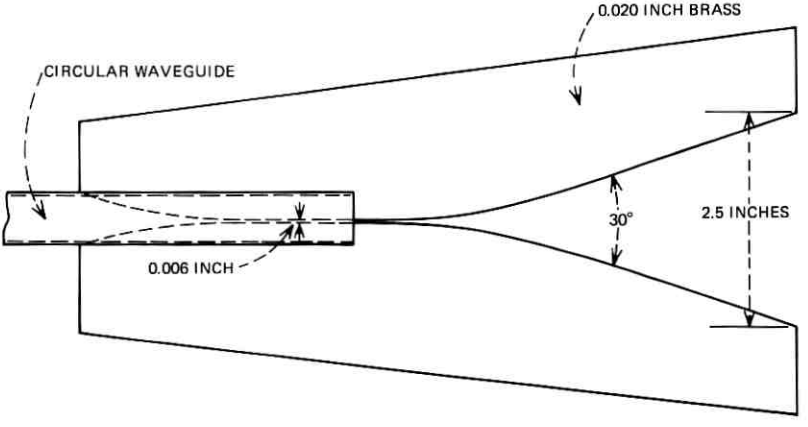

Fig. 1—Finline radiator.

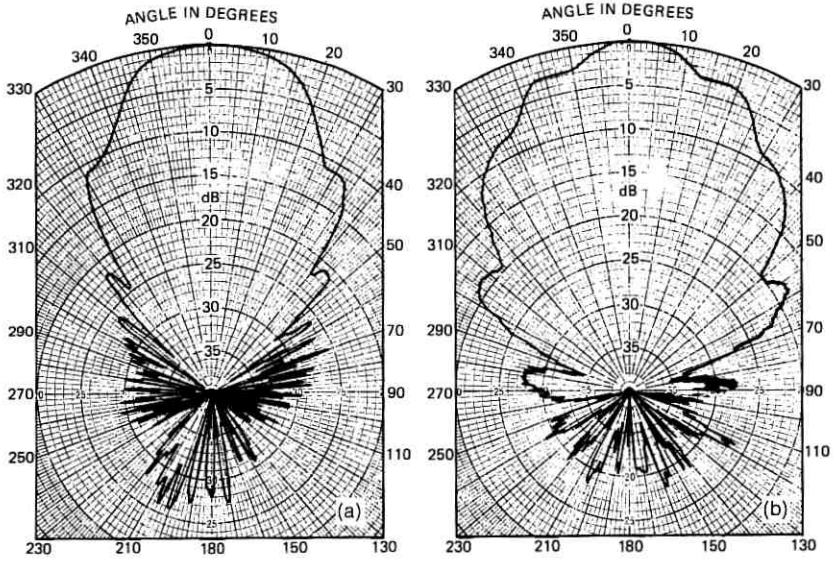

Fig. 2—Radiation patterns at 18.5 GHz with no dielectric applied. (a) H-plane pattern. (b) E-plane pattern.

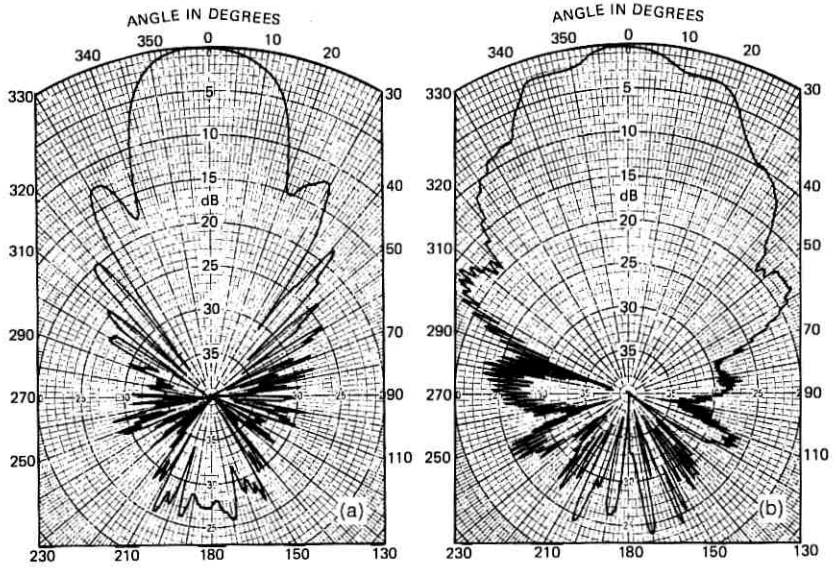

Fig. 3—Radiation patterns at 18.5 GHz with dielectric thickness of 10 mils. (a) H-plane pattern. (b) E-plane pattern.

II. DESIGN

The impedance and power flow along a fin in circular waveguide has been calculated by S. P. Morgan[2] but, to the author's knowledge, no information is available regarding radiation properties. Therefore, the simple device shown in Fig. 1 was built for initial tests. The dominant-mode circular waveguide is matched into the finline gap via a taper. After exit from the waveguide the finline is gently expanded to a $2\frac{1}{2}$ inch "aperture" which amounts to about four wavelengths at 18.5 GHz. The fin is 20 mils thick and the finline *per se* has a gap of 6 mils, resulting in an impedance of about 70 ohms. The taper is such that the phase error over the aperture, in the plane of the fin, is about $\pi/2$; the flare angle is 30 degrees. The return loss exceeds 20 dB over, at least, a ten-percent band.

III. RADIATION PATTERNS

Figure 2 shows the H- and E-plane radiation patterns (orthogonal to and in the plane of the fin, respectively); in this case no dielectric (substrate) is applied.

The E-plane pattern (in the plane of the fin) has several shoulders; but these are not unexpected[3] for phase errors of the order $\pi/2$. Figures 3, 4, and 5 show patterns where polyethylene substrates of thickness 10, 15, and 27 mils, respectively, were applied over an entire side of the fin structure, i.e., also within the waveguide. Note that the E-plane patterns remain essentially constant with increasing substrate thickness.

The H-plane patterns (in the plane orthogonal to the fin) are interesting. In the first place, the pattern of Fig. 2a is more well-behaved and somewhat narrower than the E-plane pattern, Fig. 2b. This means that the field spreads considerably in the transverse plane resulting in a sizeable effective aperture dimension. Moreover, as shown in Figs. 3a, 4a, and 5a, the addition of substrate material narrows the beamwidth of the H-plane pattern. Table I lists the 3- and 10-dB beamwidths of the H-plane pattern as a function of substrate thickness.

The gain of the 27-mil version is about 15 dB.

IV. CONCLUSION

The tapered finline radiator appears to be a reasonable candidate for an array element where planar technology is involved. Element beamwidths appropriate for certain tasks (such as illumination of the contiguous United States from synchronous orbit where beamwidths of the order of 10 degrees are desirable) appear feasible by suitable choice

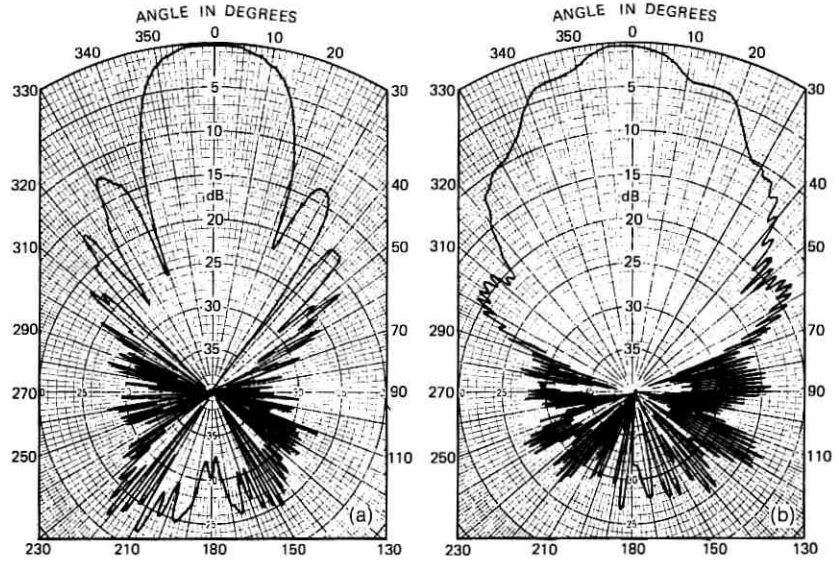

Fig. 4—Radiation patterns at 18.5 GHz with dielectric thickness of 15 mils. (a) H-plane pattern. (b) E-plane pattern.

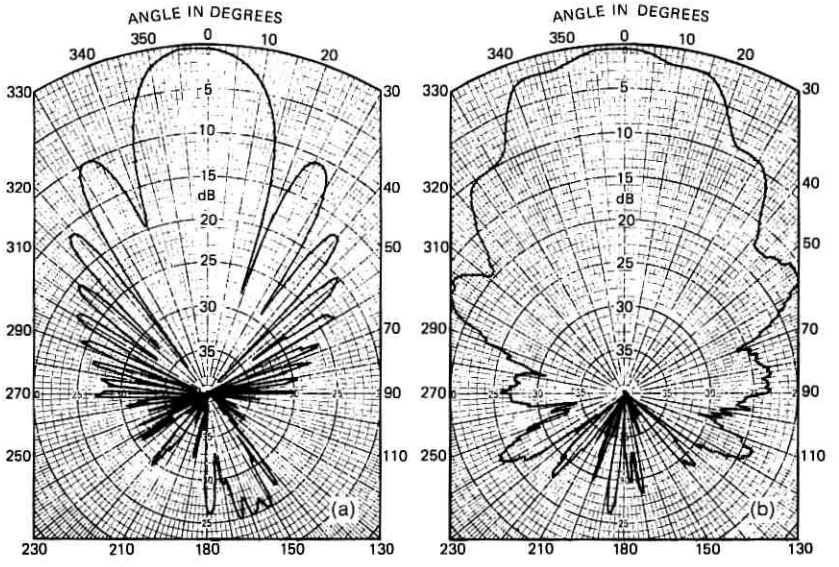

Fig. 5—Radiation patterns at 18.5 GHz with dielectric thickness of 27 mils. (a) H-plane pattern. (b) E-plane pattern.

TABLE I—BEAMWIDTHS OF H-PLANE PATTERN

Substrate (Polyethylene) Thickness-Mils	H-Plane Beamwidths Degrees	
	3 dB	10 dB
0	30	48
10	23	36
15	20	34
27	20	32

of flare angle and aperture. Of course, an objective of the design is that the radiator be integrable with microstrip circuitry; this requires a microstrip-to-finline (slotline) coupler, perhaps of the type described by S. B. Cohn.[4]

V. ACKNOWLEDGMENTS

We thank R. A. Desmond for attachment of the dielectric layers and M. V. Schneider for his kind suggestions.

REFERENCES

1. Robertson, S. D., "Ultra-Bandwidth Finline Coupler," Proc. IRE, *43*, June 1955, pp. 739–741.
2. Morgan, S. P., "Theoretical Properties of Fin Waveguide," unpublished work, 1954.
3. See for example Fig. 10-3 of *Antenna Engineering Handbook*, Henry Jasik, ed., New York: McGraw-Hill.
4. Cohn, S. B., "Slot Line on a Dielectric Substrate," IEEE Trans. Microwave Theory and Techniques, *MTT-17*, No. 10 (October 1969), p. 768.