THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 54 October 1975 Number 8

The Wire-Tap Channel

By A. D. WYNER

(Manuscript received May 9, 1975)

We consider the situation in which digital data is to be reliably transmitted over a discrete, memoryless channel (DMC) that is subjected to a wire-tap at the receiver. We assume that the wire-tapper views the channel output via a second DMC. Encoding by the transmitter and decoding by the receiver are permitted. However, the code books used in these operations are assumed to be known by the wire-tapper. The designer attempts to build the encoder-decoder in such a way as to maximize the transmission rate R, and the equivocation d of the data as seen by the wire-tapper. In this paper, we find the trade-off curve between R and d, assuming essentially perfect ("error-free") transmission. In particular, if d is equal to H_S, the entropy of the data source, then we consider that the transmission is accomplished in perfect secrecy. Our results imply that there exists a $C_s > 0$, such that reliable transmission at rates up to C_s is possible in approximately perfect secrecy.

I. INTRODUCTION

In this paper we study a (perhaps noisy) communication system that is being wire-tapped via a second noisy channel. Our object is to encode the data in such a way that the wire-tapper's level of confusion will be as high as possible. To fix ideas, consider first the simple special case depicted in Fig. 1 (in which the main communication system is noiseless). The source emits a data sequence $S_1, S_2, \cdots$, which consists of independent copies of the binary random variable S, where $\Pr\{S = 0\} = \Pr\{S = 1\} = \frac{1}{2}$. The encoder examines the first K source bits $\mathbf{S}^K = (S_1, \cdots, S_K)$ and encodes $\mathbf{S}^K$ into a binary N vector

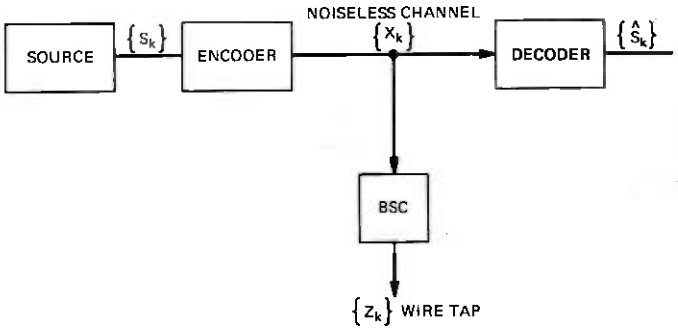


Fig. 1—Wire-tap channel (special case).

$\mathbf{X}^N = (X_1, \cdots, X_N)$. $\mathbf{X}^N$ in turn is transmitted perfectly to the decoder via the noiseless channel and is transformed into a binary data stream $\hat{\mathbf{S}}^K = (\hat{S}_1, \cdots, \hat{S}_K)$ for delivery to the destination. The "error probability" is defined as

$$P_e = \frac{1}{K} \sum_{k=1}^{K} \Pr \{S_k \neq \hat{S}_k\}. \tag{1}$$

The entire process is repeated on successive blocks of K source bits. The transmission rate is K/N bits per transmitted channel symbol.

The wire-tapper observes the encoded vector $\mathbf{X}^N$ through a (memoryless) binary symmetric channel (BSC) with crossover probability $p_0 (0 < p_0 \leqq \frac{1}{2})$. The corresponding output at the wire-tap is $\mathbf{Z}^N = (Z_1, \cdots, Z_N)$, so that for $x, z = 0, 1$ $(1 \leqq n \leqq N)$,

$$\Pr \{Z_n = z | X_n = x\} = (1 - p_0)\delta_{x,z} + p_0(1 - \delta_{x,z}).$$

We take the equivocation

$$\Delta \triangleq \frac{1}{K} H(\mathbf{S}^K | \mathbf{Z}^N) \tag{2}$$

as a measure of the degree to which the wire-tapper is confused. The logarithms in H are, as are all logarithms in this paper, taken to the base 2. The system designer would like to have P_e close to zero, with K/N and Δ as large as possible.

Consider the following schemes:

(*i*) Set $K = N = 1$, and let $X_1 \equiv S_1$. This results in $P_e = 0$, $K/N = 1$, and $\Delta = H(X_1 | Z_1) = h(p_0)$, where

$$h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log (1 - \lambda), \quad 0 \leqq \lambda \leqq 1, \tag{3}$$

(take $0 \log 0 = 0$).

(*ii*) Set $K = 1$, and let N be arbitrary. Let C_0 be the subset of binary N space, $\{0, 1\}^N$, consisting of those N vectors with even parity (i.e., an even number of 1's). Let $C_1 \subseteq \{0, 1\}^N$ be the subset of vectors with odd parity. The encoder works as follows. When $S_1 = i$, $(i = 0, 1)$, the encoder output $\mathbf{X}^N$ is a randomly chosen vector in C_i. Thus, the encoder is a *channel* with transition probability

$$\Pr\{\mathbf{X}^N = \mathbf{x} | S_1 = i\} = \begin{cases} 2^{-(N-1)}, & \mathbf{x} \in C_i, \\ 0, & \mathbf{x} \notin C_i, \end{cases}$$

for $i = 0, 1$. Clearly, the decoder can recover S_1 from $\mathbf{X}^N$ perfectly, so that $P_e = 0$. We now turn to the wire-tapper who observes $\mathbf{Z}^N$, the output of the BSC corresponding to the input $\mathbf{X}^N$. Let $\mathbf{z} \in \{0, 1\}^N$ be a vector of, say, even parity. Then

$$\begin{aligned} \Pr\{S_1 = 0 | \mathbf{Z}^N = \mathbf{z}\} &= \Pr\left\{\begin{array}{l}\text{the BSC makes an} \\ \text{even number of errors}\end{array}\right\} \\ &= \sum_{\substack{j=0 \\ j \text{ even}}}^{N} \binom{N}{j} p_0^j (1 - p_0)^{N-j} = \tfrac{1}{2} + \tfrac{1}{2}(1 - 2p_0)^N. \end{aligned}$$

The last equality can be verified by applying the binomial formula to

$$[(1 - p_0) \pm x p_0]^N = \sum_{j=0}^{N} \binom{N}{j} p_0^j (1 - p_0)^{N-j} (\pm x)^j.$$

Then

$$\begin{aligned} 2 \sum_{j \text{ even}} \binom{N}{j} p_0^j (1 - p_0)^{N-j} &= (1 - p_0 + 1 \cdot p_0)^N + (1 - p_0 - 1 \cdot p_0)^N \\ &= 1 + (1 - 2p_0)^N \end{aligned}$$

(S. P. Lloyd). Similarly, for $\mathbf{z} \in \{0, 1\}^N$ of odd parity,

$$\begin{aligned} \Pr\{S_1 = 0 | \mathbf{Z}^N = \mathbf{z}\} &= \Pr\left\{\begin{array}{l}\text{the BSC makes an} \\ \text{odd number of errors}\end{array}\right\} \\ &= \tfrac{1}{2} - \tfrac{1}{2}(1 - 2p_0)^N. \end{aligned}$$

Therefore, for all $\mathbf{z} \in \{0, 1\}^N$,

$$H(S_1 | \mathbf{Z}^N = \mathbf{z}) = h[\tfrac{1}{2} - \tfrac{1}{2}(1 - 2p_0)^N],$$

so that

$$\begin{aligned} \Delta = H(S_1 | \mathbf{Z}^N) &= h[\tfrac{1}{2} - \tfrac{1}{2}(1 - 2p_0)^N] \\ &\to 1 = H(S_1), \quad \text{as } N \to \infty. \end{aligned}$$

Thus, as $N \to \infty$, the equivocation at the wire-tap approaches the unconditional source entropy, so that communication is accomplished in perfect secrecy. The "catch" is that, as $N \to \infty$, the transmission rate $K/N = 1/N \to 0$.

A central question to which this paper is addressed is whether or not it is possible to transmit at a rate bounded away from zero, and yet achieve approximately perfect secrecy, i.e., $\Delta \approx H(S_1)$. Before giving the answer to this question, we shall describe the more general problem that is addressed in the sequel.

Refer to Fig. 2. The source is discrete and memoryless with entropy H_S. The "main channel" and the "wire-tap channel" are discrete memoryless channels with transition probabilities $Q_M(\cdot \mid \cdot)$ and $Q_W(\cdot \mid \cdot)$, respectively. The source and the transition probabilities Q_M and Q_W are given and fixed. The encoder, as in the above example, is a channel with the K vector $\mathbf{S}^K$ as input and the N vector $\mathbf{X}^N$ as output. The vector $\mathbf{X}^N$ is in turn the input to the main channel. The main channel output and the wire-tap channel input is $\mathbf{Y}^N$. The wire-tap channel output is $\mathbf{Z}^N$. The decoder associates a K vector $\hat{\mathbf{S}}^K$ with $\mathbf{Y}^N$, and the error probability P_e is given by (1). The equivocation Δ is given by (2), and the transmission rate is KH_S/N source bits per channel input symbol. Roughly speaking, a pair (R, d) is achievable if it is possible to find an encoder-decoder with arbitrarily small P_e, and KH_S/N about R, and Δ about d (with perhaps N and K very large). Our main problem is the characterization of the family of achievable (R, d) pairs, and such a characterization is given in Theorem 2. It turns out (Theorem 3) that, in nearly every case, there exists a "secrecy capacity," $C_s > 0$, such that (C_s, H_S) is achievable [while, for $R > C_s$, (R, H_S) is not achievable]. Thus, it is possible to reliably transmit information at the positive rate C_s in essentially perfect secrecy.

For the special case of our introductory example ($H_S = 1$, Q_M corresponding to a noiseless channel and Q_W to a BSC), the conclusion of Theorem 2 specializes to the assertion that (R, d) is achievable if and only if $0 \leqq R \leqq 1$, $0 \leqq d \leqq 1$, and $Rd \leqq h(p_0)$. Note that scheme (i) suggested above for this special case asserts that $R = 1$, $d = h(p_0)$

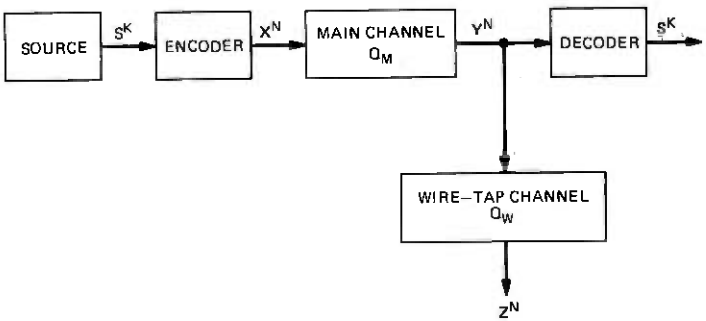


Fig. 2—Wire-tap channel (general case).

is achievable. From Theorem 2, this value of $d = h(p_0)$ is the maximum achievable d, if $R = 1$. Scheme (*ii*) above asserts that $R = 0$, $d = 1$ is achievable, but this is distinctly suboptimal since from Theorem 2, $R = h(p_0)$, $d = 1$ is achievable. Thus, reliable transmission at a rate $h(p_0)$ is possible with perfect secrecy, and $C_s = h(p_0)$.

An outline of the remainder of this paper now follows. In Section II, we give a formal statement of the problem and state the main results (Theorems 2 and 3). In Section III we give a proof of Theorem 2 for the special case discussed above (main channel noiseless, wire-tap channel a BSC). In Section IV, we prove the converse half of Theorem 2, and in Section V the direct half of that theorem.

II. FORMAL STATEMENT OF THE PROBLEM AND SUMMARY OF RESULTS

In this section we give a precise statement of the problem that we stated informally in Section I. We then summarize our results.

First, a word about notation. Let $\mathfrak{U}$ be an arbitrary finite set. Denote its cardinality by $|\mathfrak{U}|$. Consider $\mathfrak{U}^N$, the set of N vectors with components in $\mathfrak{U}$. The members of $\mathfrak{U}^N$ will be written as

$$\mathbf{u}^N = (u_1, u_2, \cdots, u_N),$$

where subscripted letters denote the components and boldface superscripted letters denote vectors. A similar convention applies to random vectors and random variables, which are denoted by upper-case letters. When the dimension N of a vector is clear from the context, we omit the superscript.

For random variables X, Y, Z, etc., the notation $H(X)$, $H(X|Y)$, $I(X;Y)$, $I(X;Y|Z)$, etc., denotes the standard information quantities as defined in Gallager.[1] The logarithms in these quantities are, as are all logarithms in this paper, taken to the base 2. Finally, for $n = 3, 4, 5, \cdots$, we say that the sequence of random variables $\{X_i\}_{i=1}^n$ is a "Markov chain" if $(X_1, X_2, \cdots, X_{j-1})$ and $(X_{j+1}, \cdots, X_n)$ are conditionally independent, given $X_j (1 < j < n)$. We make repeated use of the fact that, if X_1, X_2, X_3 is a Markov chain, then

$$H(X_3|X_1, X_2) = H(X_3|X_2). \tag{4}$$

At this point we call attention to Appendix A, in which the data-processing theorem and Fano's inequality are given in several forms.

We now turn to the description of the communication system. We assume that the system designer is given a source and two channels that are defined as follows.

(*i*) The *source* is defined by the sequence $\{S_k\}_1^\infty$, where the S_k are independent, identically distributed random variables that take

values in the finite set $\mathcal{S}$. We assume that the probability law that defines the $\{S_k\}$ is known. Let the entropy $H(S_k) = H_S$. In Appendix C we show how to extend the results of this paper to arbitrary stationary finite alphabet ergodic sources.

(*ii*) The *main channel* is a discrete memoryless channel with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and transition probability $Q_M(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Since the channel is memoryless, the transition probability for N vectors is

$$Q_M^{(N)}(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} Q_M(y_n|x_n). \tag{5}$$

Denote the channel capacity of the main channel by C_M.

(*iii*) The *wire-tap channel* is also a discrete memoryless channel with input alphabet $\mathcal{Y}$, finite output alphabet $\mathcal{Z}$, and transition probability $Q_W(z|y)$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$. The cascade of the main channel and the wire-tap channel is another memoryless channel with transition probability

$$Q_{MW}(z|x) = \sum_{y \in \mathcal{Y}} Q_W(z|y) Q_M(y|x). \tag{6}$$

Occasionally, when there is no ambiguity, we use the transition probability of a channel to denote the channel itself. Let C_{MW} be the capacity of channel Q_{MW}.

With the source statistics and channels Q_M and Q_W given, the designer must specify an encoder and a decoder, defined as follows.

(*iv*) The *encoder* with parameters (K, N) is another channel with input alphabet $\mathcal{S}^K$, output alphabet $\mathcal{X}^N$, and transition probability $q_E(\mathbf{x}|\mathbf{s})$, $\mathbf{s} \in \mathcal{S}^K$, $\mathbf{x} \in \mathcal{X}^N$. When the K source variables $\mathbf{S}^K = (S_1, \cdots, S_K)$ are the input to the encoder, the output is the random vector $\mathbf{X}^N$. Let $\mathbf{Y}^N$ and $\mathbf{Z}^N$ be the output of channels $Q_M^{(N)}$ and $Q_{MW}^{(N)}$, respectively, when the input is $\mathbf{X}^N$. The equivocation of the source at the output of the wire-tap channel (corresponding to a particular encoder) is

$$\Delta \triangleq \frac{1}{K} H(\mathbf{S}^K|\mathbf{Z}^N). \tag{7}$$

We take Δ as our criterion of the wire-tapper's confusion. From the system designer's point of view, it is, of course, desirable to make Δ large.

(*v*) The *decoder* is a mapping

$$f_D: \mathcal{Y}^N \to \mathcal{S}^K. \tag{8a}$$

Let $\hat{\mathbf{S}} = (\hat{S}_1, \cdots, \hat{S}_K) = f_D(\mathbf{Y})$. Corresponding to a given encoder and

decoder, the *error-rate* is

$$P_e = \frac{1}{K} \sum_{k=1}^{N} \Pr \{S_k \neq \hat{S}_k\}. \tag{8b}$$

We refer to the above as an encoder-decoder (K, N, Δ, P_e).* The applicability of the above to the system in Fig. 2 should be obvious.

Next, we say that the pair (R, d) (where $R, d > 0$) is *achievable* if, for all $\epsilon > 0$, there exists an encoder-decoder (N, K, Δ, P_e) for which

$$\frac{(H_S K)}{N} \geqq R - \epsilon, \tag{9a}$$

$$\Delta \geqq d - \epsilon, \tag{9b}$$

$$P_e \leqq \epsilon. \tag{9c}$$

Our problem is to characterize the set $\mathfrak{R}$ of achievable (R, d) pairs. Let us remark here that it follows immediately from the definition that $\mathfrak{R}$ is a closed subset of the first quadrant of the (R, d) plane. Before stating our characterization of $\mathfrak{R}$, we digress to discuss a certain information-theoretic quantity that plays a crucial role in our solution.

Consider the channels Q_M, Q_W, and Q_{MW} defined above. Let $p_X(x)$, $x \in \mathfrak{X}$, be a probability mass function and let X be the random variable defined by

$$\Pr \{X = x\} = p_X(x), \quad x \in \mathfrak{X}.$$

Let Y, Z be the outputs of channels Q_M and Q_{MW}, respectively, when X is the input. For $R \geqq 0$, let $\mathcal{P}(R)$ be the set of p_X such that $I(X; Y) \geqq R$. Of course, $\mathcal{P}(R)$ is empty for $R > C_M$, the capacity of channel Q_M. Finally, for $0 \leqq R \leqq C_M$, define

$$\Gamma(R) \triangleq \sup_{p_X \in \mathcal{P}(R)} I(X; Y|Z). \tag{10}$$

We remark here that, for any distribution p_X on $\mathfrak{X}$, the corresponding X, Y, Z forms a Markov chain, so that the definition of mutual information and (4) yield

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(X|Z) - H(X|Y) = I(X; Y) - I(X; Z). \end{aligned} \tag{11}$$

Thus, we can write (10) as

$$\Gamma(R) = \sup_{p_X \in \mathcal{P}(R)} I(X; Y|Z) = \sup_{p_X \in \mathcal{P}(R)} [I(X; Y) - I(X; Z)]. \tag{12}$$

* This should be read as "... an encoder-decoder with parameters (K, N, Δ, P_e)."

As an example, suppose that $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$. Let Q_M be a noiseless (binary) channel, and let Q_W be a binary symmetric channel (BSC) with crossover probability p_0. Then for arbitrary p_X,

$$\begin{aligned} I(X;Y) - I(X;Z) &= H(X) - [H(Z) - H(Z|X)] \\ &= h(p_0) + H(X) - H(Z) \leqq h(p_0), \end{aligned}$$

where $h(\cdot)$ is defined in (3). The inequality follows from the well-known fact (see, for example, Ref. 2) that the entropy of the output of a BSC, i.e., $H(Z)$, is not less than the entropy of the input, $H(X)$. Further, $H(X) = H(Z)$ if and only if $p_X(0) = p_X(1) = \frac{1}{2}$. Since this distribution belongs to $\mathcal{P}(R)$, for all R, $0 \leqq R \leqq C_M = 1$, we conclude that, in this case,

$$\Gamma(R) = h(p_0), \quad 0 \leqq R \leqq C_M. \tag{13}$$

In Appendix B, we establish the following lemma concerning $\Gamma(R)$.

Lemma 1: The quantity $\Gamma(R)$, $0 \leqq R \leqq C_M$, satisfies the following:

(i) *The "supremum" in the definition of Γ[(10) or (12)] is, in fact, a maximum—i.e., for each R, there exists a $p_X \in \mathcal{P}(R)$ such that $I(X;Y|Z) = \Gamma(R)$.*
(ii) *$\Gamma(R)$ is a concave function of R.*
(iii) *$\Gamma(R)$ is nonincreasing in R.*
(iv) *$\Gamma(R)$ is continuous in R.*
(v) *$C_M \geqq \Gamma(R) \geqq C_M - C_{MW}$, where C_M and C_{MW} are the capacities of channels Q_M and Q_{MW}, respectively.*

We can now state our main result, the proof of which is given in the remaining sections.

Theorem 2: The set $\mathcal{R}$, as defined above, is equal to $\bar{\mathcal{R}}$, where

$$\bar{\mathcal{R}} \triangleq \{(R, d): \ 0 \leqq R \leqq C_M, \ \ 0 \leqq d \leqq H_S, \ \ Rd \leqq H_S\Gamma(R)\}. \tag{14}$$

Remarks:

(1) A sketch of a typical region $\bar{\mathcal{R}}$ is given in Fig. 3. In the above example (Q_M noiseless and Q_W a BSC), $\Gamma(R) = h(p_0)$, a constant, so that the curve $Rd = H_S\Gamma(R)$ is a hyperbola. Observe that in this case the region $\bar{\mathcal{R}}$ is not convex. This is in contrast to the up-to-now essentially universal situation in multiple-user Shannon theory problems, where the solution is nearly always a convex region. Whether or not $\Gamma(R)/R$ is always convex, as it appears in Fig. 3, is an open question.

(2) The points in $\bar{\mathcal{R}}$ for which $R = C_M$ correspond to data rates of about the capacity of Q_M. This is clearly the maximum rate at which reliable transmission over Q_M is possible. An equivocation at the wire-tap of about $H_S\Gamma(C_M)/C_M$ is achievable at this rate. An increase in equivocation requires a reduction of transmission rate.

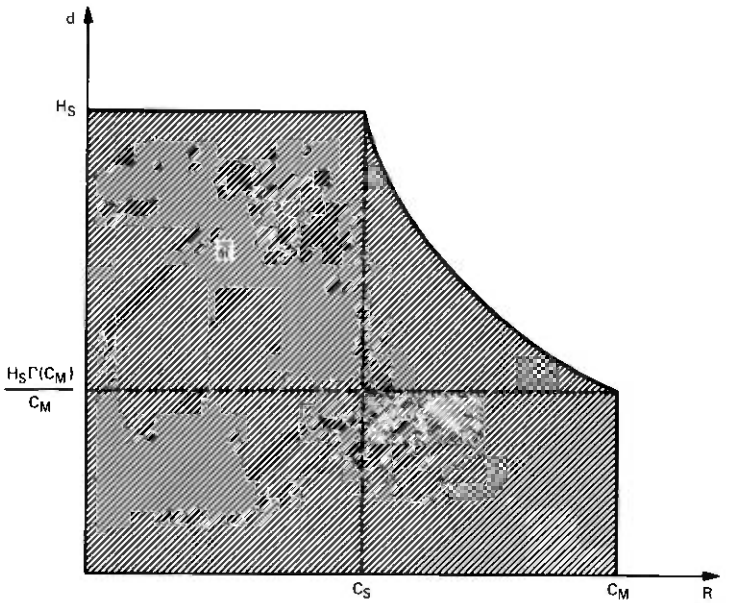


Fig. 3—Region $\overline{\mathcal{R}}$.

(3) The points in $\overline{\mathcal{R}}$ for which $d = H_S$ are of considerable interest. These correspond to an equivocation for the wire-tapper of about H_S—i.e., perfect secrecy. A transmission rate of

$$C_s = \max_{(R,\, H_S) \in \overline{\mathcal{R}}} R$$

is therefore achievable in perfect secrecy. We call C_s the "secrecy capacity" of the channel pair (Q_M, Q_W). The following theorem clarifies this remark.

Theorem 3: If $C_M > C_{MW}$, there exists a unique solution C_s of

$$C_s = \Gamma(C_s). \tag{15}$$

Further, C_s satisfies

$$0 < C_M - C_{MW} \leqq \Gamma(C_M) \leqq C_s \leqq C_M, \tag{16}$$

and C_s is the maximum R such that $(R, H_S) \in \mathcal{R}$.

Proof: Define $G(R) = \Gamma(R) - R$, $0 \leqq R \leqq C_M$. From Lemma 1 (*v*),

$$G(C_M) = \Gamma(C_M) - C_M \leqq 0,$$

and

$$G(0) = \Gamma(0) \geqq C_M - C_{MW} > 0.$$

Since by Lemma 1, (*iii*) and (*iv*), $G(R)$ is continuous and strictly

decreasing in R, a unique $C_s \in (0, C_M]$ exists such that $G(C_s) = \Gamma(C_s) - C_s = 0$. This is the unique solution to (15). Inequality (16) follows from $C_s \in (0, C_M]$ and Lemma 1, (*iii*) and (*v*). Finally, from (15) and (16) we have $(C_s, H_S) \in \bar{\mathcal{R}} = \mathcal{R}$. Also, if $(R_1, H_S) \in \mathcal{R}$, then $H_S R_1 \leqq H_S \Gamma(R_1)$ so that $G(R_1) \geqq 0$. Since $G(R)$ is strictly decreasing in R, we conclude that $R_1 \leqq C_s$. Thus, C_s is the maximum of those R for which $(R_1, H_S) \in \mathcal{R}$, completing the proof.

(4) It is clear that the source statistics enter into the solution only via the source entropy H_S. We also remind the reader that the fairly simple extension of Theorems 2 and 3 to a stationary, ergodic source is given in Appendix C.

(5) If we define P_{ew}, the "wire-tapper's" error probability, as the error rate at a decoder built by the wire-tapper [defined analogously to (8)], then it follows from Fano's inequality (see Appendix A) that

$$\Delta \leqq h(P_{ew}) + P_{ew} \log |\mathcal{S}|.$$

Thus, a large value of the equivocation Δ implies a large value of P_{ew} (which the system designer will find desirable).

III. PROOF OF THEOREM 2 FOR A SPECIAL CASE

In this section we prove Theorem 2 for the very special case discussed in Section I. All alphabets $\mathcal{S}$, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ are equal to $\{0, 1\}$. The source $\{S_k\}$ satisfies $\Pr\{S_k = 0\} = \Pr\{S_k = 1\} = \frac{1}{2}$. Channel Q_M is noiseless, i.e., $Q_M(y|x) = \delta_{x,y}$; and channel Q_W is a BSC with crossover probability p_0 ($0 \leqq p_0 \leqq \frac{1}{2}$), i.e.,

$$Q_W(z|y) = (1 - p_0)\delta_{y,z} + p_0(1 - \delta_{y,z}). \tag{17}$$

We show here that (R, d) is achievable if and only if

$$R \leqq C_M = 1, \quad d \leqq H_S = 1, \quad Rd \leqq h(p_0). \tag{18}$$

Since, for this case, $\Gamma(R) = h(p_0)$, this result is a special case of the as-yet-unproven Theorem 2. We begin with the converse ("only if") part of the result. Let $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Z}^N$ correspond to an encoder-decoder (N, K, Δ, P_e) (note that $\mathbf{Y}^N = \mathbf{X}^N$). Then, making repeated use of the identity $H(U, V) = H(U) + H(V|U)$, we can write (dropping the superscript on vectors)

$$\begin{aligned} K\Delta &= H(\mathbf{S}^K|\mathbf{Z}^N) = H(\mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &= H(\mathbf{S}, \mathbf{X}, \mathbf{Z}) - H(\mathbf{X}|\mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &= H(\mathbf{Z}|\mathbf{X}, \mathbf{S}) + H(\mathbf{X}, \mathbf{S}) - H(\mathbf{X}|\mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &\overset{(a)}{=} H(\mathbf{Z}|\mathbf{X}) + H(\mathbf{S}|\mathbf{X}) + H(\mathbf{X}) - H(\mathbf{X}|\mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &\overset{(b)}{=} Nh(p_0) + H(\mathbf{S}|\mathbf{X}) + [H(\mathbf{X}) - H(\mathbf{Z})] - H(\mathbf{X}|\mathbf{S}, \mathbf{Z}). \end{aligned} \tag{19}$$

These steps are justified as follows.

(a) From the fact that $(\mathbf{S}, \mathbf{X}, \mathbf{Z})$ is a Markov chain and (4), so that $H(\mathbf{Z}|\mathbf{X}, \mathbf{S}) = H(\mathbf{Z}|\mathbf{X})$.

(b) Since $\mathbf{X}$, $\mathbf{Z}$ are the input and output, respectively, of a BSC, $H(\mathbf{Z}|\mathbf{X}) = Nh(p_0)$, regardless of the distribution for $\mathbf{X}$.

Now from Fano's inequality [use ineq. (78) with $V = \mathbf{X}$], we have $H(\mathbf{S}|\mathbf{X}) \leqq Kh(P_e)$. Further, the entropy of the output of a BSC $\geqq$ the entropy of the input [this follows from Mrs. Gerber's lemma (Ref. 2, Theorem 1)], so that $H(\mathbf{X}) - H(\mathbf{Z}) \leqq 0$. Finally, $H(\mathbf{X}|\mathbf{S}, \mathbf{Z}) \geqq 0$. Thus, (19) yields for any encoder-decoder (K, N, Δ, P_e),

$$K\Delta \leqq Nh(p_0) + Kh(P_e),$$

or

$$\frac{K}{N}[\Delta - h(P_e)] \leqq h(p_0). \tag{20}$$

Now suppose that (R, d) is achievable. It follows from the ordinary converse to the coding theorem (Ref. 1, Th. 4.3.4, p. 81) that $R \leqq C_M = 1$. Further, since $\Delta \leqq H_S = 1$, we conclude that $d \leqq 1$. Finally, if we apply (20) to an encoder-decoder (N, K, Δ, P_e) that satisfies (9) with $\epsilon > 0$ arbitrary, we have

$$(R - \epsilon)[(d - \epsilon) - h(\epsilon)] \leqq h(p_0).$$

Letting $\epsilon \to 0$ yields $Rd \leqq h(p_0)$. Thus, we have established the converse of Theorem 2, i.e., that an achievable (R, d) must satisfy (18).

We begin the proof of the direct half of Theorem 2 with a digression about group codes for the BSC. Let $G \subseteq \{0, 1\}^N$ be a group code (i.e., a parity check code) as defined for example in Ref. 1, Chapter 6, or Ref. 3, Chapter 4. The group code G has $M = 2^N/|G|$ cosets. Denote the cosets by $C_0 = G$, C_1, C_2, $\cdots$, C_{M-1}. Of course, the cosets are disjoint and

$$\bigcup_{i=0}^{M-1} C_i = \{0, 1\}^N.$$

Let λ be the word error probability when group code G (or for any of the cosets) is used on a BSC with crossover probability p_0, with maximum-likelihood (minimum distance) decoding. Thus, for each coset C_i, $0 \leqq i \leqq M - 1$, there exists a decoder mapping D_i: $\{0, 1\}^N \to C_i$, such that if $\mathbf{X}^N$ is the input to a BSC with crossover probability p_0, and $\mathbf{Z}^N$ is the corresponding output, then for all $\mathbf{x} \in C_i$, $0 \leqq i \leqq M - 1$,

$$\Pr\{D_i(\mathbf{Z}^N) \neq \mathbf{X}^N | \mathbf{X}^N = \mathbf{x}\} = \lambda.$$

Thus, regardless of the probability distribution for $\mathbf{X}^N$,

$$\Pr\{D_i(\mathbf{Z}^N) \neq \mathbf{X}^N | \mathbf{X}^N \in C_i\} = \lambda.$$

Letting $\psi(\mathbf{x}) = i$, for $\mathbf{x} \in C_i$, $0 \leqq i \leqq M-1$, we have, from Fano's inequality [use ineq. (76) with $U = \mathbf{X}^N$, $V = \mathbf{Z}^N$, $\hat{U} = D_i(Z^N)$],

$$H(\mathbf{X}^N | \mathbf{Z}^N, \psi = i) \leqq h(\lambda) + \lambda \log |C_i|.$$

Therefore, for any $\mathbf{X}$ distribution (which induces a distribution of ψ),

$$H(\mathbf{X}^N | \mathbf{Z}^N, \psi) \leqq h(\lambda) + \lambda \log |G|. \tag{21}$$

We conclude this digression by stating as a lemma the well-known result of Elias that there exists a group code for transmitting reliably over a BSC at any rate up to capacity. A proof of this result can be found in Ref. 1, Section 6.2.

Lemma 4: Let $\epsilon_1 > 0$, $r < 1 - h(p_0)$ be arbitrary. Then, provided N is sufficiently large, there exists a group code G of block length N with $|G| \geqq 2^{Nr}$, such that, on the BSC *with crossover probability p_0, the error probability $\lambda \leqq \epsilon_1$.*

We now prove the direct half of Theorem 2 for our special case by showing that any (R, d), where R is rational, which satisfies

$$R \cdot d = h(p_0), \tag{22a}$$

$$0 \leqq d < 1, \tag{22b}$$

$$0 \leqq R \leqq 1 \tag{22c}$$

is achievable. Thus, for (R, d) satisfying (22), and arbitrary $\epsilon > 0$, we must show the existence of an encoder-decoder (N, K, Δ, P_e) that satisfies (9). We now proceed to this task.

Let K, N satisfy

$$\frac{K}{N} = R. \tag{23}$$

Let G be a binary group code with block length N and with $|G| = 2^{(N-K)}$. Thus, G has $M = 2^K$ cosets $\{C_i\}_{i=0}^{M}$. We can assume that the set $\mathcal{S}^K = \{0, 1\}^K$ is the set of integers $\{0, 1, \cdots, M-1\}$. We construct the encoder such that when the source vector $\mathbf{S}^K = i$,* the encoder output $\mathbf{X}^N$ is a randomly chosen member of coset C_i—i.e.,

$$\Pr\{\mathbf{X}^N = \mathbf{x} | \mathbf{S} = i\} = \begin{cases} \dfrac{1}{|C_i|} = \dfrac{1}{|G|} = 2^{-(N-K)}, & \text{for } \mathbf{x} \in C_i, \\ 0, & x \notin C_i, \end{cases}$$

$0 \leqq i \leqq M-1$. Since $\mathbf{S}^K$ is uniformly distributed on $\{0, 1, \cdots, M-1\}$, $\mathbf{X}^N$ is uniformly distributed on $\mathcal{X}^N = \{0, 1\}^N$. Thus, in particular,

$$H(\mathbf{X}^N) = H(\mathbf{Z}^N) = N, \tag{24}$$

* This is an abuse of notation. A more precise statement is that $\mathbf{S}^K$ is a binary representation of i.

where, as always, $\mathbf{Z}^N$ is the output of the wire-tap channel when $\mathbf{X}^N$ is the input. Also let us observe here that the quantity $\psi(\mathbf{X}^N)$, defined in the above digression, is identical to $\mathbf{S}^K$. Thus, (21) yields

$$H(\mathbf{X}^N|\mathbf{Z}^N, \mathbf{S}^K) \leqq h(\lambda) + \lambda(N - K), \tag{25}$$

where λ is the error probability for the group code G.

We now turn to the decoder. Letting $D(\mathbf{y}) = i$, when $\mathbf{y} \in C_i$, we conclude (since the channel Q_M is noiseless) that

$$P_e = 0. \tag{26}$$

Since (23) and (26) imply (9a) and (9c), it remains to show that a G exists such that the resulting encoder-decoder will satisfy (9b).

We now invoke (19), which is valid for any encoder-decoder. Substituting (24) and (25) into (19), and invoking (26), which implies $H(\mathbf{S}|\mathbf{X}) = 0$, we obtain

$$\Delta \geqq \left(\frac{N}{K}\right) h(p_0) - \frac{h(\lambda)}{K} - \lambda\left(\frac{N}{K} - 1\right). \tag{27}$$

Now, from (22a) and (23), we have

$$\frac{N}{K}\, h(p_0) = \frac{h(p_0)}{R} = d,$$

and from (23),

$$\lambda\left(\frac{N}{K} - 1\right) = \lambda\left(\frac{1}{R} - 1\right).$$

Thus, (27) yields

$$\Delta \geqq d - \left[\frac{h(\lambda)}{K} + \lambda\left(\frac{1}{R} - 1\right)\right]. \tag{28}$$

Finally, since from (23) and (22a) we have

$$|G| = 2^{N-K} \leqq 2^{N[1-h(p_0)/d]},$$

we can invoke Lemma 4 with $r = 1 - h(p_0)/d < 1 - h(p_0)$ [from (22b)] to assert the existence of a group code G with λ sufficiently small to make the term in brackets in (28) $\leqq \epsilon$. Then $\Delta \geqq d - \epsilon$, which is (9b). This completes the proof of the direct half.

IV. CONVERSE THEOREM

In this section, we establish the converse theorem that the family of achievable rates $\mathfrak{R}$ is contained in $\overline{\mathfrak{R}}$ as defined in (14). Suppose that

$(R, d) \in \mathcal{R}$. That $R \leqq C_M$ follows from the ordinary converse to the coding theorem (Ref. 1, Theorem 4.3.4, p. 81). That $d \leqq H_S$ follows from

$$\Delta = \frac{1}{K} H(\mathbf{S}^K | \mathbf{Z}^N) \leqq \frac{1}{K} H(\mathbf{S}^K) = H_S.$$

Thus, it remains to show that $Rd \leqq H_S \Gamma(R)$. We do this via a lemma, the proof of which is given at the conclusion of this section.

Lemma 5: *Let* $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$ *correspond to an encoder-decoder* (N, K, Δ, P_e). *Then*

$$(i) \quad \frac{K}{N} [\Delta - \delta(P_e)] \leqq \frac{1}{N} \sum_{n=1}^{N} I(X_n; Y_n | Z_n, \mathbf{Y}^{n-1}), \tag{29a}$$

$$(ii) \quad \frac{K}{N} [H_S - \delta(P_e)] \leqq \frac{1}{N} \sum_{n=1}^{N} I(X_n; Y_n | \mathbf{Y}^{n-1}), \tag{29b}$$

where

$$\delta(P_e) = h(P_e) + P_e \log |\mathcal{S}|, \tag{29c}$$

and where the $n = 1$ *term in the summations of* (29*a*, *b*) *is given the obvious interpretation—i.e., that* $I(X_1; Y_1 | Z_1, \mathbf{Y}^0) = I(X_1; Y_1 | Z_1)$, *etc.*

Now for $n = 2, 3, \cdots, N$, any $\mathbf{y} \in \mathcal{Y}^{n-1}$, set

$$\alpha_n(\mathbf{y}) = I(X_n; Y_n | \mathbf{Y}^{n-1} = \mathbf{y}). \tag{30a}$$

Also let

$$\alpha_1 = I(X_1; Y_1). \tag{30b}$$

It follows from the definition of $\mathcal{P}(R)$ in Section II that the distribution p_1, defined by

$$p_1(x) \triangleq \Pr \{X_1 = x\}, \quad x \in \mathcal{X},$$

belongs to $\mathcal{P}(\alpha_1)$. Similarly, for $2 \leqq n \leqq N$, with $\mathbf{y} \in \mathcal{Y}^{n-1}$ fixed, define

$$p_{n,\mathbf{y}}(x) \triangleq \Pr \{X_n = x | \mathbf{Y}^{n-1} = \mathbf{y}\}, \quad x \in \mathcal{X}.$$

Then $p_{n,\mathbf{y}} \in \mathcal{P}[\alpha_n(\mathbf{y})]$. Thus, from (10) and the fact that channels $Q_M^{(N)}$ and $Q_W^{(N)}$ are memoryless,

$$\Gamma(\alpha_1) \geqq I(X_1; Y_1 | Z_1), \tag{31a}$$

and for $2 \leqq n \leqq N$, $\mathbf{y} \in \mathcal{Y}^{n-1}$,

$$\Gamma[\alpha_n(\mathbf{y})] \geqq I(X_n; Y_n | Z_n, \mathbf{Y}^{n-1} = \mathbf{y}). \tag{31b}$$

It follows that the right member of (29a) is (giving the $n = 1$ term the obvious interpretation)

$$\frac{1}{N}\sum_{n=1}^{N} I(X_n; Y_n | Z_n, \mathbf{Y}^{n-1})$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{\mathbf{y}\in\mathcal{Y}^{n-1}} \Pr\{\mathbf{Y}^{n-1} = \mathbf{y}\} I(X_n; Y_n | Z_n, \mathbf{Y}^{n-1} = \mathbf{y})$$

$$\overset{(a)}{\leqq} \frac{1}{N}\sum_{n}\sum_{\mathbf{y}} \Pr\{\mathbf{Y}^{n-1} = \mathbf{y}\}\Gamma[\alpha_n(\mathbf{y})] \tag{32}$$

$$\overset{(b)}{\leqq} \Gamma\left[\frac{1}{N}\sum_{n}\sum_{\mathbf{y}} \Pr\{\mathbf{Y}^{n-1} = \mathbf{y}\}\alpha_n(\mathbf{y})\right]$$

$$\overset{(c)}{=} \Gamma\left(\frac{1}{N}\sum_{n} I(X_n Y_n | \mathbf{Y}^{n-1})\right)$$

$$\overset{(d)}{\leqq} \Gamma\left(\frac{K}{N} H_S - \delta(P_e)\right).$$

Step (a) follows from (31), step (b) from the concavity of Γ [Lemma 1(*ii*)], step (c) from the definition of α_n, and step (*d*) from (29b) and the monotonicity of Γ [Lemma 1(*iii*)]. Applying (29a) to (32) yields

Corollary 6: *For any encoder-decoder* (N, K, Δ, P_e),

$$\frac{K}{N}[\Delta - \delta(P_e)] \leqq \Gamma\left[\frac{K}{N} H_S - \delta(P_e)\right]. \tag{33}$$

We now show that, if $(R, d) \in \mathcal{R}$, then $Rd \leqq H_S\Gamma(R)$. Let $(R, d) \in \mathcal{R}$, and let $\epsilon > 0$ be arbitrary. Apply Corollary 6 to the encoder-decoder (N, K, Δ, P_e) that satisfies (9). Inequalities (33) and (9) yield

$$(R - \epsilon)[(d - \epsilon) - \delta(\epsilon)] \leqq H_S\Gamma[(R - \epsilon) - \delta(\epsilon)]. \tag{34}$$

Letting $\epsilon \to 0$ and invoking the continuity of Γ [Lemma 1(*iv*)] yield $Rd \leqq H_S\Gamma(R)$, completing the proof of the converse. It remains to prove Lemma 5.

Proof of Lemma 5:

(*i*) Let $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$ correspond to an encoder-decoder (N, K, Δ, P_e). First observe that

$$\frac{1}{K} H(\mathbf{S}^K | \mathbf{Z}^N, \mathbf{Y}^N) \leqq \frac{1}{K} H(\mathbf{S}^K | \mathbf{Y}^N)$$

$$\overset{(a)}{\leqq} h(P_e) + P_e \log(|\mathcal{S}| - 1) = \delta(P_e). \tag{35}$$

Inequality (a) follows from Fano's inequality [use (78) with $V = \mathbf{Y}^N$].

Next, using the definition of Δ (7) and (35), write

$$K\Delta = H(\mathbf{S}^K | \mathbf{Z}^N) \leqq H(\mathbf{S}^K | \mathbf{Z}^N) - H(\mathbf{S}^K | \mathbf{Z}^N, \mathbf{Y}^N) + K\delta(P_e)$$

$$= I(\mathbf{S}^K; \mathbf{Y}^N | \mathbf{Z}^N) + K\delta(P_e)$$

$$\leqq I(\mathbf{X}^K; \mathbf{Y}^N | \mathbf{Z}^N) + K\delta(P_e). \tag{36}$$

The last inequality in (36) follows from the data-processing theorem, since given $\mathbf{Z}^N = \mathbf{z}$, $(\mathbf{Y}^N, \mathbf{X}^N, \mathbf{S}^K)$ is a Markov chain (Appendix A). Transposing the $K\delta(P_e)$ term in (36) and continuing:

$$
\begin{aligned}
K[\Delta - \delta(P_e)] &\leqq I(\mathbf{X}^N; \mathbf{Y}^N | \mathbf{Z}^N) \\
&= H(\mathbf{X}^N | \mathbf{Z}^N) - H(\mathbf{X}^N | \mathbf{Z}^N, \mathbf{Y}^N) \\
&\overset{(a)}{=} H(\mathbf{X}^N | \mathbf{Z}^N) - H(\mathbf{X}^N | \mathbf{Y}^N) \\
&= I(\mathbf{X}^N; \mathbf{Y}^N) - I(\mathbf{X}^N; \mathbf{Z}^N) \\
&= H(\mathbf{Y}^N) - H(\mathbf{Z}^N) + H(\mathbf{Z}^N | \mathbf{X}^N) - H(\mathbf{Y}^N | \mathbf{X}^N) \\
&\overset{(b)}{=} \sum_{n=1}^{N} [H(Y_n | \mathbf{Y}^{n-1}) - H(Z_n | \mathbf{Z}^{n-1}) \\
&\qquad + H(Z_n | X_n) - H(Y_n | X_n)] \\
&\overset{(c)}{\leqq} \sum_{n=1}^{N} [H(Y_n | \mathbf{Y}^{n-1}) - H(Z_n | \mathbf{Z}^{n-1}, \mathbf{Y}^{n-1}) \\
&\qquad + H(Z_n | X_n) - H(Y_n | X_n)] \\
&\overset{(d)}{=} \sum_{n=1}^{N} [H(Y_n | \mathbf{Y}^{n-1}) - H(Z_n | \mathbf{Y}^{n-1}) + H(Z_n | X_n, \mathbf{Y}^{n-1}) \\
&\qquad + H(Y_n | X_n, \mathbf{Y}^{n-1})] \\
&= \sum_{n=1}^{N} [I(X_n, Y_n | \mathbf{Y}^{n-1}) - I(X_n; Z_n | \mathbf{Y}^{n-1})] \\
&= \sum_{n=1}^{N} [H(X_n | Z_n, \mathbf{Y}^{n-1}) - H(X_n | Y_n, \mathbf{Y}^{n-1})] \\
&\overset{(e)}{=} \sum_{n=1}^{N} [H(X_n | Z_n, \mathbf{Y}^{n-1}) - H(X_n | Y_n, Z_n, \mathbf{Y}^{n-1})] \\
&= \sum_{n=1}^{N} I(X_n; Y_n | Z_n, \mathbf{Y}^{n-1}). \qquad (37)
\end{aligned}
$$

The steps in (37) that require explanation are:

(a) that follows from the fact that $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$ is a Markov chain and (4);

(b) that follows from the standard identity

$$H(\mathbf{U}^N) = \sum_{n=1}^{N} H(U_n | \mathbf{U}^{n-1}),$$

and the fact that channels $Q_M^{(N)}$ and $Q_{MW}^{(N)}$ are memoryless;

(c) that follows from the fact that conditioning decreases entropy;

(d) that follows on applying (4) to the Markov chains $(\mathbf{Z}^{n-1}, \mathbf{Y}^{n-1}, Z_n)$, $(\mathbf{Y}^{n-1}, X_n, Y_n, Z_n)$;

(e) that follows from the fact that, given $\mathbf{Y}^{n-1}$, (X_n, Y_n, Z_n) is a Markov chain.

Since (37) is (29a), we have established part (i) of Lemma 5.

(ii) With $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$, as in part (i) write

$$\begin{aligned} H(\mathbf{S}^K) &= I(\mathbf{S}^K; \mathbf{Y}^N) + H(\mathbf{S}^K | \mathbf{Y}^N) \\ &\leqq I(\mathbf{X}^N; \mathbf{Y}^N) + K\delta(P_e), \end{aligned} \tag{38}$$

where the inequality follows from the data-processing theorem (since $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, is a Markov chain) and from Fano's inequality as in (35). Since $H(\mathbf{S}^K) = KH_S$, (38) yields

$$\begin{aligned} K[H_S - \delta(P_e)] &\leqq I(\mathbf{X}^N; \mathbf{Y}^N) \\ &\overset{(a)}{=} \sum_{n=1}^{N} [H(Y_n | \mathbf{Y}^{n-1}) - H(Y_n | X_n)] \\ &\overset{(b)}{=} \sum_{n=1}^{N} [H(Y_n | \mathbf{Y}^{n-1}) - H(Y_n | X_n, \mathbf{Y}^{n-1})] \\ &= \sum_{n=1}^{N} I(X_n; Y_n | \mathbf{Y}^{n-1}). \end{aligned} \tag{39}$$

Step (a) follows on application of $H(\mathbf{Y}^N) = \sum_n H(Y_n | \mathbf{Y}^{n-1})$, and the memorylessness of channel $Q_M^{(N)}$, and step (b) from the fact that $\mathbf{Y}^{n-1}$, X_n, Y_n is a Markov chain. Inequality (39) is (29b), so that the proof of Lemma 5 is complete.

V. DIRECT HALF OF THEOREM 2

In this section we establish the direct (existence) part of Theorem 2, that is, $\overline{\mathfrak{R}} \subseteq \mathfrak{R}$. The first step is to establish two lemmas that are valid for any encoder-decoder as defined in Section II.

Lemma 7: *Let* $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$ *correspond to an arbitrary encoder-decoder* (N, K, Δ, P_e). *Then*

$$K\Delta \triangleq H(\mathbf{S}^K | \mathbf{Z}^N) = H(\mathbf{S}^K) + I(\mathbf{X}^N; \mathbf{Z}^N | \mathbf{S}^K) - I(\mathbf{X}^N; \mathbf{Z}^N). \tag{40}$$

Proof: By repeatedly using the identity $H(U, V) = H(U) + H(V | U)$, we obtain (we have omitted superscripts)

$$\begin{aligned} K\Delta &= H(\mathbf{S} | \mathbf{Z}) = H(\mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &= H(\mathbf{S}, \mathbf{Z}, \mathbf{X}) - H(\mathbf{X} | \mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &= H(\mathbf{Z} | \mathbf{X}, \mathbf{S}) + H(\mathbf{X}, \mathbf{S}) - H(\mathbf{X} | \mathbf{S}, \mathbf{Z}) - H(\mathbf{Z}) \\ &= H(\mathbf{Z} | \mathbf{X}, \mathbf{S}) + H(\mathbf{S}) + [H(\mathbf{X} | \mathbf{S}) - H(\mathbf{X} | \mathbf{S}, \mathbf{Z})] - H(\mathbf{Z}) \\ &= H(\mathbf{S}) + I(\mathbf{X}; \mathbf{Z} | \mathbf{S}) - [H(\mathbf{Z}) - H(\mathbf{Z} | \mathbf{X}, \mathbf{S})]. \end{aligned} \tag{41}$$

Now, since $\mathbf{S}$, $\mathbf{X}$, $\mathbf{Z}$ is a Markov chain, $H(\mathbf{Z}|\mathbf{X}, \mathbf{S}) = H(\mathbf{Z}|\mathbf{X})$ [by (4)]. Thus, the term in brackets in the right member of (41) is $I(\mathbf{X}; \mathbf{Z})$, completing the proof.

We now give some preliminaries for the second of the two lemmas. For the remainder of this section we take the finite set $\mathfrak{X}$ to be $\{1, 2, \cdots, A\}$. Let X^* be a random variable that takes values in $\mathfrak{X}$ with probability distribution

$$\Pr\{X^* = i\} = p_X^*(i), \quad 1 \leqq i \leqq A.$$

Let Y^* and Z^* be the output of channels Q_M, and Q_{MW}, respectively, when X^* is the input. As always, Q_{MW} is the cascade of Q_M and Q_W, so that X^*, Y^*, Z^* is a Markov chain. Next, for $1 \leqq i \leqq A$, and $\mathbf{x} \in \mathfrak{X}^N$ define

$$\begin{aligned}\#(i, \mathbf{x}) &\triangleq \text{card}\,\{n: x_n = i\} \\ &= \text{number of occurrences of the symbol } i \text{ in the} \\ &\qquad N\text{-vector } \mathbf{x}. \end{aligned} \tag{42}$$

For $N = 1, 2, \cdots$, define the set of "typical" X sequences as the set

$$T^* = T^*(N) = \left\{\mathbf{x} \in \mathfrak{X}^N: \left|\frac{\#(i, \mathbf{x})}{N} - p_X^*(i)\right| \leqq \delta_N, 1 \leqq i \leqq A\right\}, \tag{43a}$$

where

$$\delta_N \triangleq N^{-\frac{1}{4}}. \tag{43b}$$

Let us remark in passing that the random N-vector $\mathbf{X}^{*N}$ consisting of N independent copies of X^* satisfies $E\#(i, \mathbf{X}^{*N}) = Np_X^*(i)$, and $\text{Var}[\#(i, \mathbf{X}^{*N})] = Np_X^*(i)[1 - p_X^*(i)]$, for $1 \leqq i \leqq A$. Thus, by Chebyshev's inequality

$$\begin{aligned}\Pr\{\mathbf{X}^{*N} \notin T^*(N)\} &\leqq \sum_{i=1}^{A} \Pr\{|\#(i, \mathbf{X}^*) - Np_X^*(i)| > N\delta_N\} \\ &\leqq \sum_{i=1}^{A} \text{Var}[\#(i, \mathbf{X}^*)]/N^2\delta_N^2 = 0\left(\frac{1}{\sqrt{N}}\right) \to 0, \end{aligned} \tag{44}$$

as $N \to \infty$.

We can now state the second of our lemmas. We give the proof at the conclusion of this section.

Lemma 8: Let $\mathbf{X}^N$, $\mathbf{Z}^N$ correspond to an arbitrary encoder and let X^*, Z^*, T^* correspond to an arbitrary p_X^* as above. Then

$$\frac{1}{N} I(\mathbf{X}^N; \mathbf{Z}^N) \leqq I(X^*, Z^*) + (\log A) \Pr\{\mathbf{X}^N \notin T^*(N)\} + f_1(N),$$

where $f_1(N) \to 0$, as $N \to \infty$.

Lemma 8 implies that, if the encoder is such that with high probability $\mathbf{X}^N \in T^*$, then $(1/N)I(\mathbf{X}^N; \mathbf{Z}^N)$ cannot be much more than $I(X^*, Z^*)$.

Lemmas 7 and 8 hold for any encoder-decoder. Our next step is to describe a certain ad-hoc encoder-decoder and deduce several of its properties. We then show that when the parameters of the ad-hoc scheme are properly chosen, the direct half of Theorem 2 will follow easily.

We begin the discussion of the ad-hoc scheme by reviewing some facts about source coding. With the source given as in Section II, for $K = 1, 2, \cdots$, there exists a ("source encoder") mapping F_E: $\mathcal{S}^K \to \{1, 2, \cdots, M\}$, where

$$M = 2^{KH_S(1+\delta_K)}, \tag{45}$$

and $\delta_K = K^{-\frac{1}{2}}$. Let F_D: $\{1, 2, \cdots, M\} \to \mathcal{S}^K$ be a ("source decoder") mapping, and let

$$P_{es}^{(K)} = \Pr\{F_D \circ F_E(\mathbf{S}^K) \neq \mathbf{S}^K\}$$

be the resulting error probability. It is very well known that there exists (for each K) a pair (F_E, F_D) such that, as $K \to \infty$,

$$P_{es}^{(K)} = \Pr\{F_D(W) \neq \mathbf{S}^K\} \to 0, \tag{46a}$$

where

$$W = F_E(\mathbf{S}^K). \tag{46b}$$

We will design our system to transmit W using an (F_E, F_D) that satisfies (46).

We now turn to our ad-hoc system. (Refer to Fig. 4.) The source output is the vector $\mathbf{S}^K$, and the output of the source decoder is $W = F_E(\mathbf{S}^K)$. Let

$$q_i \triangleq \Pr\{W = F_E(\mathbf{S}^K) = i\}, \quad 1 \leqq i \leqq M. \tag{47}$$

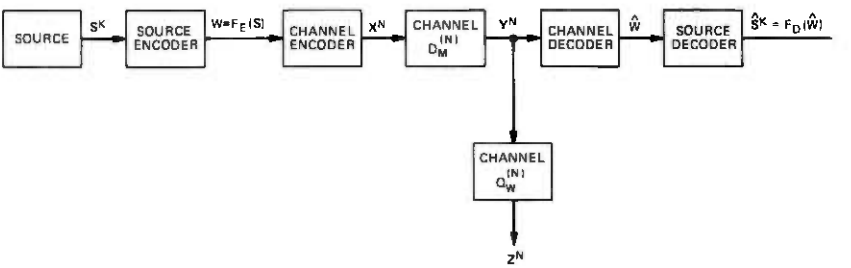


Fig. 4—Ad-hoc encoder-decoder.

Next, let $M_1 = M_2M$ be a multiple of M to be specified later. Let

$$\{\mathbf{x}_m\}_1^{M_1}$$

be a subset of $\mathcal{X}^N$. Clearly, $\{\mathbf{x}_m\}$ can be viewed as a channel code for channel $Q_M^{(N)}$ or channel $Q_{MW}^{(N)}$. The channel encoder and decoder in Fig. 4 work as follows. The channel encoder and decoder each contains a partition of $\{\mathbf{x}_m\}_1^{M_1}$ into M subcodes $C_1, C_2, \cdots, C_M$, each with cardinality M_2. Assume that

$$C_i = \{\mathbf{x}_{(i-1)M_2+1}, \cdots, \mathbf{x}_{iM_2}\}, \quad 1 \leqq i \leqq M. \tag{48}$$

When the random variable $W = i$, then the channel encoder output $\mathbf{X}^N$ is a (uniformly) randomly chosen member of the subcode C_i. Thus, for $1 \leqq i \leqq M$, $1 \leqq j \leqq M_2$,

$$\Pr\{\mathbf{X}^N = \mathbf{x}_{(i-1)M_2+j} | W = i\} = \frac{1}{M_2}, \tag{49a}$$

and

$$\Pr\{\mathbf{X}^N = \mathbf{x}_{(i-1)M_2+j}\} = \frac{q_i}{M_2}. \tag{49b}$$

Now the set $\{\mathbf{x}_m\}_1^{M_1}$ can be thought of as a channel code for channel $Q_M^{(N)}$ with prior probability distribution on the code words given by (49b). A decoder for the code is a mapping $G: \mathcal{Y}^N \to \{\mathbf{x}_m\}_1^{M_1}$ and the (word) error probability is

$$\lambda = \Pr\{G(\mathbf{Y}^N) \neq \mathbf{X}^N\}, \tag{50}$$

where $\mathbf{Y}^N$ is the output of $Q_M^{(N)}$, when the input $\mathbf{X}^N$ has distribution given by (49b). We assume that the channel decoder in Fig. 4 has stored the mapping G. When the channel output is $\mathbf{y} \in \mathcal{Y}^N$, the channel decoder computes $G(\mathbf{y})$. When $G(\mathbf{y}) \in C_i$, the channel decoder output is i, $1 \leqq i \leqq M$. Letting $\hat{W}$ be the output of the channel decoder, we have

$$\Pr\{W \neq \hat{W}\} \leqq \lambda.$$

The final step in the system of Fig. 4 is the emission by the source decoder of $\hat{\mathbf{S}}^K = F_D(\hat{W})$, where F_D: $\{1, 2, \cdots, M\} \to \mathcal{S}^K$ is chosen so that (46) holds. We have

$$\begin{aligned}\Pr\{\mathbf{S} = \hat{\mathbf{S}}\} &= \Pr\{\mathbf{S} = F_D(\hat{W})\} \\ &\geqq \Pr\{S = F_D(W); W = \hat{W}\}.\end{aligned}$$

Thus,

$$\begin{aligned}P_e \leqq \Pr\{\mathbf{S} \neq \hat{\mathbf{S}}\} &\leqq \Pr\{\mathbf{S} \neq F_D(\mathbf{W})\} \\ &\quad + \Pr\{W \neq \hat{W}\} \leqq P_{es}^{(K)} + \lambda. \end{aligned} \tag{51}$$

Next, let us observe that each of the subcodes C_i can be considered a code for channel $Q_{MW}^{(N)}$ with M_2 code words and uniform prior distribution on the code words. Let λ_i be the resulting (word) error probability for code C_i $(1 \leqq i \leqq M)$ with an optimal decoder, and let

$$\bar{\lambda} = \sum_{i=1}^{M} q_i \lambda_i. \tag{52}$$

We now establish

Lemma 9: For the ad-hoc encoder-decoder defined above

$$I(\mathbf{X}^N; \mathbf{Z}^N | \mathbf{S}^K) \geqq \log M_2 - [h(\bar{\lambda}) + \bar{\lambda} \log M_2].$$

Proof: Let $\mathbf{S}^K$ be such that $W = F_E(\mathbf{S}^K) = i$. Then the channel input $\mathbf{X}^N$ given $W = i$ has distribution given by (49a), i.e., $\mathbf{X}^N$ is a randomly chosen member of C_i. Since λ_i is the error probability for code C_i used on channel $Q_{MW}^{(N)}$, Fano's inequality [use (76) with $U = \mathbf{X}^N$, $V = \mathbf{Z}^N$, $\hat{U} =$ the decoded version of $\mathbf{Z}^N$ when code C_i is used] yields

$$H(\mathbf{X}^N | \mathbf{Z}^N, W = i) \leqq h(\lambda_i) + \lambda_i \log M_2,$$

and, since $H(\mathbf{X}^N | W = i) = \log M_2$, we have

$$I(\mathbf{X}^N; \mathbf{Z}^N | W = i) \geqq \log M_2 - h(\lambda_i) - \lambda_i \log M_2.$$

Averaging over i using the weighting $\{q_i\}$, and using the concavity of $h(\cdot)$, we have

$$I(\mathbf{X}^N; \mathbf{Z}^N | W) \geqq \log M_2 - [h(\bar{\lambda}) + \bar{\lambda} \log M_2]. \tag{53}$$

Finally, since $\mathbf{S}$, W, $\mathbf{X}$, $\mathbf{Z}$ is a Markov chain, (4) yields

$$\begin{aligned} I(\mathbf{X}^N; \mathbf{Z}^N | W) &= H(\mathbf{Z} | W) - H(\mathbf{Z} | \mathbf{X}W) \\ &= H(\mathbf{Z} | W, \mathbf{S}) - H(\mathbf{Z} | \mathbf{X}) \\ &= H(\mathbf{Z} | W, \mathbf{S}) - H(\mathbf{Z} | \mathbf{X}, \mathbf{S}) \\ &\leqq H(\mathbf{Z} | \mathbf{S}) - H(\mathbf{Z} | \mathbf{X}, \mathbf{S}) = I(\mathbf{X}^N; \mathbf{Z}^N | \mathbf{S}). \end{aligned} \tag{54}$$

Inequalities (53) and (54) imply Lemma 9.

We are now ready to combine the above lemmas as:

Corollary 10: Let p_X^* be an arbitrary probability distribution on $\mathfrak{X}$, and let $T_X^*(N)$, X^*, Y^*, Z^* be as defined above (corresponding to p_X^*). Assume that $\mathbf{S}^K$, $\mathbf{X}^N$, $\mathbf{Y}^N$, $\mathbf{Z}^N$ correspond to the above ad-hoc encoder-decoder with parameters N, K, M, M_1, M_2, λ, $\bar{\lambda}$. Let P_e and Δ correspond to this ad-hoc scheme. Then

$$P_e \leqq P_{es}^{(K)} + \lambda \tag{55a}$$

and

$$\frac{K}{N}\Delta \geqq \frac{K}{N}H_S + \frac{1}{N}\log M_2 - I(X^*, Z^*) - \frac{h(\bar{\lambda})}{N} - \frac{\bar{\lambda}\log M_2}{N} - (\log A)\Pr\{\mathbf{X}^N \notin T_X^*(N)\} - f_1(N), \quad (55b)$$

where $f_1(N) \to 0$ *as* $N \to \infty$.

Proof: Inequality (55a) is the same as (51). Inequality (55b) is obtained by substituting the results of Lemmas 8 and 9 into (40) and using $H(\mathbf{S}^K) = KH_S$.

Finally, we are ready to prove the direct half of Theorem 2. We do this by showing that any pair (R, d), which satisfies

$$R \cdot d = H_S\Gamma(R), \quad (56a)$$

$$0 \leqq R \leqq C_M, \quad (56b)$$

$$0 \leqq d \leqq H_S, \quad (56c)$$

is achievable. Thus, for (R, d) satisfying (56) and for arbitrary $\epsilon > 0$, we show that our ad-hoc scheme with appropriately chosen parameters satisfies (9). To begin with, choose K, N to satisfy

$$\frac{K}{N} = \frac{R}{H_S}. \quad (57)$$

(Assume that R/H_S is rational.) Note that (57) implies (9a). Also, let p_X^* be a distribution on $\mathfrak{X}$ that belongs to $\mathcal{P}(R)$ and achieves $\Gamma(R)$—that is,

$$I(X^*; Y^*) \geqq R,$$
$$I(X^*; Y^*) - I(X^*; Z^*) = I(X^*; Y^*|Z^*) = \Gamma(R), \quad (58)$$

where X^*, Y^*, Z^* correspond to p_X^*. We now assume that an encoder-decoder is constructed according to the above ad-hoc scheme with the parameter*

$$M_1 = \exp_2\left\{N\left[I(X^*; Y^*) - \frac{\epsilon R}{2H_S}\right]\right\}, \quad (59)$$

where X^*, Y^* correspond to the above choice of p_X^*. With this choice of M_1, and with M given by (45), we have

$$M_2 = \frac{M_1}{M} = \exp_2\left\{N\left[I(X^*; Y^*) - \frac{K}{N}H_S - \frac{K}{N}H_S\delta_K - \frac{\epsilon R}{2H_S}\right]\right\}. \quad (60)$$

Note that, from (57),

* Assume that the right member of (59) is an integer. If not, a trivial modification of the sequel is necessary.

$$\begin{aligned}
\frac{1}{N}\log M_2 &= I(X^*;Y^*) - \frac{K}{N}H_S - \frac{K}{N}H_S\delta_K - \frac{\epsilon R}{2H_S} \\
&\overset{(a)}{=} I(X^*;Y^*) - R - R\delta_K - \frac{\epsilon R}{2H_S} \\
&= I(X^*;Y^*) - \frac{(Rd/H_S)}{(d/H_S)} - R\delta_K - \frac{\epsilon R}{2H_S} \\
&\overset{(b)}{\leqq} I(X^*;Y^*) - \Gamma(R) - R\delta_K - \frac{\epsilon R}{2H_S} \\
&= I(X^*;Y^*) - I(X^*;Y^*|Z^*) - R\delta_K - \frac{\epsilon R}{2H_S} \\
&\overset{(c)}{=} I(X^*;Z^*) - R\delta_K - \frac{\epsilon R}{2H_S}. \qquad (61)
\end{aligned}$$

Step (a) follows from (57), step (b) from (56a) and (56c), and step (c) from the fact that X^*, Y^*, Z^* is a Markov chain—see (11).

Let us now apply Corollary 10 to the ad-hoc scheme with the above choice of M_1, M_2, and with the above choice of p_X^*. Inequality (55a) remains

$$P_e \leqq P_{es}^{(K)} + \lambda, \qquad (62)$$

and substituting (60) into (55b) yields

$$\begin{aligned}
(R\Delta)/H_S &\geqq I(X^*;Y^*) - I(X^*;Z^*) - f_2(N) \\
&= \Gamma(R) - f_2(N), \qquad (63a)
\end{aligned}$$

where

$$f_2(N) = \frac{\epsilon R}{2H_S} + R\delta_K + \frac{h(\bar{\lambda})}{N} + \frac{\bar{\lambda}\log M_2}{N} + (\log A)\Pr\{\mathbf{X}^N \notin T^*(N)\} + f_1(N). \qquad (63b)$$

Now observe $f_2(N)$ and $\bar{\lambda}$ depend on the choice of the set $\{\mathbf{x}_m\}_1^{M_1}$. The following lemma asserts the existence of a $\{x_m\}$ such that these quantities are small. Its proof is given at the end of this section.

Lemma 11: With p_X^* and M_1, M_2 as given above, there exists for arbitrary N a set

$$\{\mathbf{x}_m\}_{m=1}^{M_1}$$

such that

$$\left.\begin{matrix}\Pr\{\mathbf{X}^N \notin T^*(N)\}, \\ \lambda, \\ \bar{\lambda}\end{matrix}\right\} \leqq f_3(N), \qquad (64)$$

where $f_3(N) \to 0$, as $N \to \infty$.

Now let the set $\{\mathbf{x}_m\}_1^{M_1}$ in the ad-hoc scheme be chosen to satisfy (64). Then, from (62) and (64) [using the fact that $P_{es}^{(K)} \to 0$, as $K \to \infty$ (46)], we can choose N (and $K = NR/H_S$) sufficiently large so that

$$P_e \leqq \epsilon,$$

this is (9c). It remains to establish (9b). But from (64) with N sufficiently large, we can make

$$R\delta_K + \frac{h(\bar{\lambda})}{N} + \frac{\bar{\lambda} \log M_2}{N} + (\log A) \Pr \{\mathbf{X}^N \notin T^*(N)\} + f_1(N) \leqq \frac{\epsilon R}{2H_S}.$$

Then (63) and (56a) yield

$$\Delta \geqq \frac{H_S \Gamma(R)}{R} - \epsilon = d - \epsilon,$$

which is (9b). Thus, (R, d) is achievable and the proof of the direct half of Theorem 2, i.e., $\bar{\mathfrak{R}} \subseteq \mathfrak{R}$, is complete. It remains to prove Lemmas 11 and 8.

Proof of Lemma 11: We begin with some notation. For $\mathbf{x} \in \mathfrak{X}^N$, let

$$\mu(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \notin T^*(N), \\ 0, & \text{otherwise.} \end{cases} \tag{65}$$

Also for a given set $\{\mathbf{x}_m\}_1^{M_1}$, let $\lambda^{(m)}(\mathbf{x}_1, \cdots, \mathbf{x}_{M_1})$ be the error probability that results when $\{\mathbf{x}_m\}$ is used as a channel code for channel $Q_M^{(N)}$ with prior probabilities (49b) when code word $\mathbf{x}_m$ is transmitted and when maximum liklihood decoding is used. Thus,

$$\lambda = \sum_{i=1}^{M} \sum_{m=(i-1)M_2+1}^{iM_2} \frac{q_i}{M_2} \lambda^{(m)}(\mathbf{x}_1, \cdots, \mathbf{x}_{M_1}).$$

Further, with λ_i defined as above as the error probability for code C_i on $Q_{MW}^{(N)}$, write $\lambda_i = \lambda_{MW}(\mathbf{x}_{(i-1)M_2+1}, \cdots, \mathbf{x}_{iM_2}) = \lambda_{MW}(C_i)$, so that the dependence of λ_i on C_i is explicit. We have

$$\bar{\lambda} = \sum_{i=1}^{M} q_i \lambda_i = \sum q_i \lambda_{MW}(C_i).$$

Finally, define

$$\begin{aligned} \Phi(\mathbf{x}_1, \cdots, \mathbf{x}_{M_1}) &\triangleq \Pr \{\mathbf{X}^N \notin T_X^*(N)\} + \lambda + \bar{\lambda} \\ &= \sum_{i=1}^{M} \sum_{m=(i-1)M_2+1}^{iM_2} \frac{q_i}{M_2} [\mu(\mathbf{x}_m) + \lambda^{(m)}(\mathbf{x}_1, \cdots, \mathbf{x}_{M_2})] \\ &\qquad + \sum_{i=1}^{M} q_i \lambda_{MW}(C_i). \end{aligned} \tag{66}$$

Now suppose that the set $\{x_m\}_1^{M_1}$ is chosen at random, with each $\mathbf{x}_m$ chosen independently from $\mathfrak{X}^N$, with probability distribution $p_{X^*}^{(N)}(\mathbf{x})$

$= \prod_{n=1}^{N} p_X^*(x_n)$. We establish the lemma by showing that $E\Phi \leqq F_3(N)$. Now observe that, from (59), $(1/N) \log M_1$ is bounded below $I(X^*, Y^*)$. Also from (61), $(1/N) \log M_2$ is bound below $I(X^*; Z^*)$. It follows from the standard random channel-coding theorem (see, for example, Ref. 1, Theorem 5.6.2) that $E\lambda^{(m)}$, $E\lambda_{MW} \leqq f_4(N) \to 0$, as $N \to \infty$. Further, $E\mu = \Pr\{\mathbf{X}^* \notin T_X^*(N)\} \leqq f_5(N) \to 0$, by (44). Thus, $E\Phi \leqq 2f_4(N) + f_5(N) \triangleq f_3(N) \to 0$. Hence the lemma.

Proof of Lemma 8: Here too we begin with some notation. Let p be a probability distribution on $\mathfrak{X}$, and let $\mathcal{I}(p)$ be the mutual information between the input and output of channel Q_{MW} when the input has distribution p. It is known (Ref. 1, Theorem 4.4.2) that $\mathcal{I}(p)$ is a concave function of p. Let $\mu(\mathbf{x})$ be as in (65), and write (for any encoder-decoder)

$$\begin{aligned}\frac{1}{N} I(\mathbf{X}^N; \mathbf{Z}^N) &= \frac{1}{N} I[\mathbf{X}^N, \mu(\mathbf{X}^N); \mathbf{Z}^N] \\ &= \frac{1}{N} I[\mathbf{X}^N; \mathbf{Z}^N | \mu(\mathbf{X}^N)] + \frac{1}{N} I[\mu(\mathbf{X}^N); \mathbf{Z}^N] \\ &= \frac{1}{N} \sum_{j=0}^{1} \Pr\{\mu(\mathbf{X}^N) = j\} I(\mathbf{X}^N; \mathbf{Z}^N | \mu(\mathbf{X}^N) = j) \\ &\qquad + \frac{1}{N} I[\mu(\mathbf{X}^N); \mathbf{Z}^N]. \quad (67)\end{aligned}$$

Now

$$\begin{aligned}\frac{1}{N} \Pr\{\mu(\mathbf{X}^N) = 1\} I[\mathbf{X}^N; \mathbf{Z}^N | \mu(\mathbf{X}^N) = 1] \\ \leqq (\log A) \Pr\{\mathbf{X}^N \notin T^*(N)\}, \quad (68)\end{aligned}$$

and

$$\frac{1}{N} I[\mu(\mathbf{X}^N); \mathbf{Z}^N] \leqq \frac{1}{N} H[\mu(\mathbf{X}^N)] \leqq \frac{1}{N}. \quad (69)$$

One term remains in (67). Using the memoryless property of channel $Q_{MW}^{(N)}$ (Ref. 1, Theorem 4.2.1), we have

$$\begin{aligned}\frac{1}{N} I(\mathbf{X}^N; \mathbf{Z}^N | \mu = 0) &\leqq \frac{1}{N} \sum_{n=1}^{N} I(X_n; Z_n | \mu = 0) \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathcal{I}(p_n) \leqq \mathcal{I}\left(\frac{1}{N} \sum_{n=1}^{N} p_n\right), \quad (70a)\end{aligned}$$

where p_n is the probability distribution for X_n given $\mu = 0$, i.e., for $1 \leqq i \leqq A$,

$$p_n(i) = \sum_{\mathbf{x} \in T^*} \delta_{x_n, i} \Pr\{\mathbf{X}^N = \mathbf{x} | \mathbf{X}^N \in T^*\}. \quad (70b)$$

The last inequality in (70a) follows from the concavity of $\mathcal{I}$. From

(70b),

$$\tilde{p}(i) \triangleq \frac{1}{N} \sum_{n=1}^{N} p_n(i) = \sum_{\mathbf{x} \in T^*} \Pr\{\mathbf{X}^N = \mathbf{x} | \mathbf{X} \in T^*\} \frac{\#(i, \mathbf{x})}{N}. \quad (71)$$

The definition of T^* (43) and eq. (71) yields

$$|\tilde{p}(i) - p_X^*(i)| \leqq \delta_N \to 0, \quad \text{as } N \to \infty.$$

Since $g(p)$ is a continuous function of p, we have

$$|g(\tilde{p}) - g(p_X^*)| \leqq g(N) \to 0, \quad \text{as } N \to \infty. \quad (72)$$

Substituting (72) into (70a), we obtain

$$\frac{1}{N} \Pr\{\mu = 0\} I(\mathbf{X}^N; \mathbf{Z}^N | \mu = 0) \leqq g(p_X^*) + g(N)$$
$$= I(X^*; Z^*) + g(N). \quad (73)$$

Finally, setting $f_1(N) = (1/N) + g(N)$, and substituting (68), (69), and (73) into (67) we have Lemma 8.

VI. ACKNOWLEDGMENTS

I would like to acknowledge helpful discussions with my colleagues D. Slepian, H. S. Witsenhausen, and C. Mallows that contributed to this paper. In particular, the problem was originally formulated in collaboration with Mr. Witsenhausen, and the coding scheme described above for the special case (main channel noiseless, wire-tap channel a BSC) is based on an idea of Mr. Mallows. I also wish to thank M. Hellman of Stanford University, whose recent paper[4] stimulated this research. Furthermore, the pioneering work of C. E. Shannon[5] on relating information theoretic ideas to cryptography should be noted.

APPENDIX A

The Data-Processing Theorem and Fano's Inequality

Let U, V, $\hat{U}$ be discrete random variables that form a Markov chain. Then the *data-processing theorem* can be stated as

$$H(U|V) \leqq H(U|\hat{U}), \quad (74a)$$

or equivalently

$$I(U; V) \geqq I(U; \hat{U}). \quad (74b)$$

Inequality (74a) follows on writing

$$H(U|V) \overset{(a)}{=} H(U|V, \hat{U}) \overset{(b)}{\leqq} H(U|\hat{U}),$$

where step (a) follows from (4), and (b) from the fact that conditioning decreases entropy [Ref. 1, eq. (2.3.13)].

Next, let U, V, $\hat{U}$ be a Markov chain as above, but now assume that U, $\hat{U}$ take values in $\mathfrak{U}(|\mathfrak{U}| \leqq \infty)$. Let

$$\lambda = \Pr\{U \neq \hat{U}\}. \tag{75}$$

Fano's inequality is

$$H(U|V) \leqq h(\lambda) + \lambda \log(|\mathfrak{U}| - 1) \leqq h(\lambda) + \lambda \log |\mathfrak{U}|. \tag{76}$$

To verify (76), define the random variable

$$\Phi(U, \hat{U}) = \begin{cases} 0, & U = \hat{U}, \\ 1, & U \neq \hat{U}, \end{cases}$$

and then write

$$\begin{aligned} H(U|V) &\overset{(a)}{\leqq} H(U|\hat{U}) \leqq H(U, \Phi|\hat{U}) \\ &= H(\Phi|\hat{U}) + H(U|\hat{U}, \Phi) \\ &\leqq H(\Phi) + H(U|\hat{U}, \Phi) \\ &= H(\Phi) + \Pr\{\Phi = 0\}H(U|\hat{U}, \Phi = 0) \\ &\quad + \Pr\{\Phi = 1\}H(U|\hat{U}, \Phi = 1) \\ &\overset{(b)}{=} h(\lambda) + (1 - \lambda)\cdot 0 + \lambda H(U|\hat{U}, \Phi = 1) \\ &\overset{(c)}{\leqq} h(\lambda) + \lambda \log(|\mathfrak{U}| - 1) \leqq h(\lambda) + \lambda \log |\mathfrak{U}|, \end{aligned}$$

which is (76). Step (a) is (74a), and step (b) follows from the fact that, given $\Phi = 0$, then $U = \hat{U}$, so that $H(U|\hat{U}, \Phi = 0) = 0$, and step (c) from the fact that, given $\Phi = 1$, U takes one of the $|\mathfrak{U}| - 1$ values in $\mathfrak{U}$ excluding $\hat{U}$.

A variation of Fano's inequality is the following. Let $\mathbf{S}^K$, V, $\hat{\mathbf{S}}^K$ be a Markov chain where the coordinates of $\mathbf{S}^K$ and $\hat{\mathbf{S}}^K$ take the values in the set $\mathfrak{s}$. Let

$$P_{ek} = \Pr\{S_k \neq \hat{S}_k\} \tag{77a}$$

and

$$P_e = \frac{1}{K}\sum_{k=1}^{K} P_{ek}. \tag{77b}$$

We will show that Fano's inequality implies

$$\frac{1}{K} H(\mathbf{S}^K|V) \leqq h(P_e) + P_e \log(|\mathfrak{s}| - 1) \triangleq \delta(P_e). \tag{78}$$

To verify (78), write

$$\begin{aligned} \frac{1}{K} H(\mathbf{S}^K|V) &\overset{(a)}{\leqq} \frac{1}{K}\sum_{k=1}^{N} H(S_k|V) \\ &\overset{(b)}{\leqq} \frac{1}{K}\sum_{k=1}^{N} \delta(P_{ek}) \overset{(c)}{\leqq} \delta(P_e), \end{aligned}$$

which is (78). Step (a) is a standard inequality, step (b) follows on applying (76) to the Markov chain S_k, V, $\hat{S}_k$, and step (c) from the concavity of $\delta(\cdot)$.

APPENDIX B

Proof of Lemma 1

(*i*) With no loss of generality, let $\mathfrak{X} = \{1, 2, \cdots, A\}$. Any probability distribution p_X can be thought of as an A-vector $\mathbf{p} = (p_1, p_2, \cdots, p_A)$. Since $I(X; Y)$ is a continuous function of p_X, the set $\mathcal{P}(R)$ is a compact subset of Euclidean A-space. Since $I(X; Y|Z)$ is also a continuous function of p_X, we conclude that $I(X; Y|Z)$ has a maximum on $\mathcal{P}(R)$. This is part (*i*).

(*ii*) Let $0 \leqq R_1, R_2 \leqq C_M$, and $0 \leqq \theta \leqq 1$. We must show that

$$\Gamma[\theta R_1 + (1 - \theta)R_2] \geqq \theta\Gamma(R_1) + (1 - \theta)\Gamma(R_2). \tag{79}$$

For $i = 1, 2$, let $\mathbf{p}_i \in \mathcal{P}(R_i)$ achieve $\Gamma(R_i)$. In other words, letting X_i, Y_i, Z_i correspond to $\mathbf{p}_i$, $i = 1, 2$, then

$$I(X_i, Y_i) \geqq R_i, \quad I(X_i, Y_i|Z_i) = \Gamma(R_i). \tag{80}$$

Now let the random variable X be defined as in Fig. 5. For $i = 1, 2$, the box labeled "$\mathbf{p}_i$" generates the random variable X_i that has probability distribution "$\mathbf{p}_i$." The switch takes upper position ("position 1") with probability θ and the lower position ("position 2") with probability $1 - \theta$. Let V denote the switch position. In the figure, $V = 1$. Assume that V, X_1, X_2 are independent. As indicated in the figure, $X = X_i$, when $V = i$, $i = 1, 2$. Now

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \overset{(a)}{=} H(Y) - H(Y|X, V) \\ &\geqq H(Y|V) - H(Y|X, V) = I(X; Y|V) \\ &= \theta I(X; Y|V = 1) + (1 - \theta)I(X; Y|V = 2) \\ &= \theta I(X_1; Y_1) + (1 - \theta)I(X_2; Y_2) \\ &\overset{(b)}{\geqq} \theta R_1 + (1 - \theta)R_2. \end{aligned} \tag{81}$$

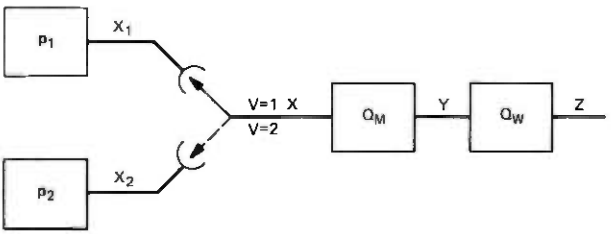


Fig. 5—Defining the random variable X.

Step (a) follows from the fact that V, X, Y is a Markov chain and (4). Step (b) follows from (80). Inequality (81) implies that the distribution defining X belongs to $\mathcal{P}[\theta R_1 + (1-\theta)R_2]$. Thus, from the definition of Γ,

$$\Gamma[\theta R_1 + (1-\theta)R_2] \geqq I(X; Y|Z). \tag{82}$$

Continuing (82) and paralleling (81), we have

$$\begin{aligned}
\Gamma[\theta R_1 + (1-\theta)R_2] &\geqq H(Y|Z) - H(Y|XZ) \\
&= H(Y|Z) - H(Y|XZV) \\
&\geqq H(Y|ZV) - H(Y|XZV) \\
&= I(X; Y|ZV) = \theta I(X; Y|Z, V=1) \\
&\quad + (1-\theta) I(X; Y|Z, V=2) \\
&= \theta I(X_1; Y_1|Z_1) + (1-\theta) I(X_2; Y_2|Z_2) \\
&= \theta\Gamma(R_1) + (1-\theta)\Gamma(R_2),
\end{aligned}$$

which is (79). This is part (*ii*).

(*iii*) This part follows immediately from the definition of $\Gamma(R)$ (10), since $\mathcal{P}(R)$ is a nonincreasing set.

(*iv*) Since $\Gamma(R)$ is concave on $[0, C_M]$, and nonincreasing, it must be continuous for $0 \leqq R < C_M$. Thus, we need only verify the continuity of $\Gamma(R)$ at $R = C_M$. Let $\mathbf{p}$ be a probability distribution on $\mathfrak{X}$ viewed as a vector in Euclidean A-space, as in the proof of part (*i*). Let $g(\mathbf{p})$ and $\hat{g}(\mathbf{p})$ be the values of $I(X; Y)$ and $I(X; Y|Z)$, respectively, which correspond to $\mathbf{p}$. $g(p)$ and $\hat{g}(p)$ are continuous functions of $\mathbf{p}$.

Now let $\{R_j\}_1^\infty$ be a monotone increasing sequence such that $R_j \to C_M$, and $R_j \leqq C_M$. We must show that, as $j \to \infty$,

$$\Gamma(R_j) \to \Gamma(C_M). \tag{83}$$

Now from the monotonicity of $\Gamma(R)$, $\lim_{j\to\infty} \Gamma(R_j)$ exists and

$$\lim_{j\to\infty} \Gamma(R_j) \geqq \Gamma(C_M). \tag{84}$$

It remains to verify the reverse of ineq. (84). Let $\{\mathbf{p}_j\}_1^\infty$ satisfy

$$g(\mathbf{p}_j) \geqq R_j, \quad \hat{g}(\mathbf{p}_j) = \Gamma(R_j), \tag{85}$$

for $1 \leqq j < \infty$. Since the set of probability A-vectors is compact, there exists a probability distribution $\mathbf{p}^*$ on $\mathfrak{X}$ such that for some subsequence $\{\mathbf{p}_{j_k}\}_{k=1}^\infty$

$$\lim_{k\to\infty} \mathbf{p}_{j_k} = \mathbf{p}^*.$$

It follows from the continuity of $g(\cdot)$, and (85) that $g(\mathbf{p}^*) \geqq C_M$, so that $\mathbf{p}^* \in \mathcal{P}(C_M)$. Therefore, from the continuity of $\hat{g}(\cdot)$, and (85), we have

$$\lim_{j\to\infty} \Gamma(R_j) = \lim_{k\to\infty} \Gamma(R_{jk}) = \lim_{k\to\infty} \hat{g}(\mathbf{p}_{jk}) = \hat{g}(\mathbf{p}^*) \overset{(a)}{\leqq} \Gamma(C_M), \quad (86)$$

where step (a) follows from $\mathbf{p}^* \in \mathcal{P}(C_M)$. Inequalities (84) and (86) yield (83) and part (*iv*).

(*v*) From (12),

$$\begin{aligned} \Gamma(R) &= \sup_{p_X \in \mathcal{P}(R)} [I(X;Y) - I(X;Z)] \\ &\leqq \sup_{p_X \in \mathcal{P}(R)} I(X;Y) \leqq C_M, \end{aligned}$$

which is the first inequality in part (*v*). Also, using (12),

$$\begin{aligned} \Gamma(C_M) &= \sup_{p_X \in \mathcal{P}(C_M)} [I(X;Y) - I(X;Z)] \\ &\geqq \sup_{p_X \in \mathcal{P}(C_M)} [I(X;Y) - C_{MW}] = C_M - C_{MW}. \end{aligned} \quad (87)$$

Since $\Gamma(R)$ is nonincreasing, (87) yields $\Gamma(R) \geqq \Gamma(C_M) \geqq C_M - C_{MW}$, completing the proof of part (*v*).

APPENDIX C

Source with Memory

In this appendix, we show how to modify our definitions and results for a source with memory. We will take the source output sequence $\{S_k\}$ to be a stationary, ergodic sequence (where S_k takes values in $\mathcal{S}$) with entropy (as defined in Ref. 1, Section 3.5) of H_S. As in Section II, we continue to assume that $|\mathcal{S}| < \infty$, and that the source statistics are known.

The channels Q_M and Q_W remain as in Section II, as does the definition of an encoder-decoder with parameters N and K. The definition of P_e also remains unchanged, but a new definition for Δ is necessary. To see this, let us suppose that the source was binary, i.e., $\mathcal{S} = \{0, 1\}$, with entropy H_S, and with $H(S_1) > H_S$. Suppose also that the channel Q_M is a noiseless binary channel, and that Q_W has zero capacity. A possible encoder-decoder has $K = N = 1$ and takes $X_1 = S_1$. Such a scheme has $P_e = 0$, but with Δ as defined in (7) given by $\Delta = H(S_1) > H_S$. Using (9), this would lead us to accept the pair $[H_S, H(S_1)]$ as achievable, which would not be reasonable. Accordingly, we give a new definition of Δ.

Let $\mathbf{S}^K$, $\mathbf{Z}^N$ correspond to an encoder with parameters K, N as defined in Section II. Let $\mathbf{S}^K(j)$, $\mathbf{Z}^N(j)$, $j = 1, 2, \cdots, \nu$, correspond to

the ν successive repetitions of the encoding process. Then define the equivocation at the wire-tap as

$$\begin{aligned}\Delta &= \lim_{\nu\to\infty} \frac{1}{K\nu} H[\mathbf{S}^K(1), \cdots, \mathbf{S}^K(\nu) | \mathbf{Z}^N(1), \cdots, \mathbf{Z}^N(\nu)] \\ &= \lim_{\nu\to\infty} \frac{1}{K\nu} H(\mathbf{S}^{K\nu} | \mathbf{Z}^{N\nu}).\end{aligned} \tag{88}$$

With Δ as defined by (88), we define the sets $\mathfrak{R}$ and $\bar{\mathfrak{R}}$ as in Section II. We claim that Theorem 2 remains valid.

The proof of the converse-half of Theorem 2 given in Section IV goes over to the case where the source has memory with only trivial changes. Further, the results in Section V are all valid exactly for the source with memory. They yield that, if (R, d) satisfies (56), then we can for $\epsilon > 0$ arbitrary find an encoder-decoder with parameters N, K, and P_e which satisfies

$$\frac{KH_S}{N} \geqq R - \epsilon, \tag{89a}$$

$$P_e \leqq \epsilon, \tag{89b}$$

$$\frac{1}{K} H(\mathbf{S}^K | \mathbf{Z}^N) \geqq d - \epsilon. \tag{89c}$$

Further, we can do this for arbitrarily large K. We show below that there exists a function $f(K)$, $K = 1, 2, \cdots$, such that for any code with parameters K, N

$$\Delta = \lim_{\nu\to\infty} \frac{1}{K\nu} H(\mathbf{S}^{K\nu} | \mathbf{Z}^{N\nu}) \geqq \frac{1}{K} H(\mathbf{S}^K | \mathbf{Z}^N) - f(K), \tag{90}$$

where $\lim_{K\to\infty} f(K) = 0$, and $f(K)$ depends only on the source statistics. Combining (90) with (89c), we have

$$\Delta \geqq d - \epsilon - f(K).$$

Since $f(K) \to 0$, we conclude that (R, d) is achievable. This is the direct half of Theorem 2. It remains to verify (90).

First, imagine that the encoder-decoder begins operation infinitely far in the past. Let $[\mathbf{S}(j), \mathbf{Z}(j)]$ be the $(\mathbf{S}^K, \mathbf{Z}^K)$ corresponding to the jth encoding operation, $-\infty < j < \infty$. Thus, $\mathbf{S}^{K\nu} = (S_1, \cdots, S_{K\nu}) = [\mathbf{S}(1), \cdots, \mathbf{S}(\nu)]$ and $\mathbf{Z}^{K\nu} = [\mathbf{Z}(1), \cdots, \mathbf{Z}(\nu)]$, $\nu = 1, 2, \cdots$. Let $\mathbf{Z}^* = [\cdots, \mathbf{Z}(-1), \mathbf{Z}(0), \mathbf{Z}(+1), \cdots]$. Of course,

$$H(\mathbf{S}^{K\nu} | \mathbf{Z}^{N\nu}) \geqq H(\mathbf{S}^{K\nu} | \mathbf{Z}^*). \tag{91}$$

Further,

$$H(\mathbf{S}^{K_\nu}|\mathbf{Z}^*) = H[\mathbf{S}(1), \cdots, \mathbf{S}(\nu)|\mathbf{Z}^*]$$

$$\overset{(a)}{=} \sum_{j=1}^{\nu} H[\mathbf{S}(j)|\mathbf{Z}^*, \mathbf{S}(j+1), \cdots, \mathbf{S}(\nu)]$$

$$\overset{(b)}{=} \sum_{j=1}^{\nu} H[\mathbf{S}(1)|\mathbf{Z}^*, \mathbf{S}(2), \cdots, \mathbf{S}(j)]$$

$$\overset{(c)}{\geqq} \nu H[\mathbf{S}(1)|\mathbf{Z}^*, \mathbf{S}(2), \cdots, \mathbf{S}(\nu)] \geqq \nu H[\mathbf{S}(1)|\mathbf{Z}^*, \mathbf{S}'], \quad (92)$$

where $\mathbf{S}' = [\mathbf{S}(2), \mathbf{S}(3), \cdots]$. Step (a) is a standard identity, step (b) follows from the stationarity of the sequence $\{S_k\}$ and the memorylessness of the channel Q_{MW}, and step (c) follows from the fact that conditioning decreases entropy. Now, let

$$\mathbf{S} = \mathbf{S}^K = \mathbf{S}(1), \quad \mathbf{S}' = [\mathbf{S}(2), \mathbf{S}(3), \cdots],$$
$$\mathbf{Z} = \mathbf{Z}^N = \mathbf{Z}(1), \quad \mathbf{Z}' = [\cdots, \mathbf{Z}(-1), \mathbf{Z}(0), \mathbf{Z}(+2), \cdots].$$

Thus, (91) and (92) become

$$\frac{1}{K_\nu} H(\mathbf{S}^{K_\nu}|\mathbf{Z}^{N_\nu}) \geqq \frac{1}{K} H(\mathbf{S}|\mathbf{Z}, \mathbf{Z}', \mathbf{S}')$$

$$= \frac{1}{K}[H(\mathbf{SZ}|\mathbf{Z}'\mathbf{S}) - H(\mathbf{Z}|\mathbf{Z}'\mathbf{S}')]$$

$$= \frac{1}{K}[H(\mathbf{S}|\mathbf{Z}'\mathbf{S}') + H(\mathbf{Z}|\mathbf{SZ}'\mathbf{S}') - H(\mathbf{Z}|\mathbf{Z}'\mathbf{S}')]$$

$$\overset{(a)}{=} \frac{1}{K}[H(\mathbf{S}|\mathbf{S}') + H(\mathbf{Z}|\mathbf{S}) - H(\mathbf{Z}|\mathbf{Z}'\mathbf{S}')]$$

$$\geqq \frac{1}{K}[H(\mathbf{S}|\mathbf{S}') + H(\mathbf{Z}|\mathbf{S}) - H(\mathbf{Z})]. \quad (93)$$

Step (a) follows from the fact that $\mathbf{Z}'$, $\mathbf{S}'$, $\mathbf{S}$ and $(\mathbf{S}', \mathbf{Z}')$, $\mathbf{S}$, $\mathbf{Z}$ are Markov chains, and (4). Now

$$\frac{1}{K} H(\mathbf{S}|\mathbf{S}') = \frac{1}{K} \sum_{k=1}^{K} H(S_k|\mathbf{S}', S_{k+1}, \cdots, S_K)$$

$$= \frac{1}{K} \sum_{k=1}^{K} H_S = H_S. \quad (94)$$

Also,

$$\left|\frac{1}{K} H(\mathbf{S}) - H_S\right| \leqq f(K) \to 0, \quad \text{as } K \to \infty. \quad (95)$$

Substituting (95) and (94) into (93), we have

$$\frac{1}{K_\nu} H(\mathbf{S}^{K_\nu}|\mathbf{Z}^{N_\nu}) \geqq \frac{1}{K}[H(\mathbf{S}) + H(\mathbf{Z}|\mathbf{S}) - H(\mathbf{Z})] - f(K)$$

$$= \frac{1}{K} H(\mathbf{S}|\mathbf{Z}) - f(K),$$

which is (90).

REFERENCES

1. R. G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley, 1968.
2. A. D. Wyner and J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences and Applications: Part I," IEEE Transactions on Information Theory, *IT-19* (Nov. 1973), pp. 769–772.
3. R. B. Ash, *Information Theory*, New York: Interscience, 1965.
4. Martin E. Hellman, "The Information Theoretic Approach to Cryptography," Stanford University, Center for Systems Research, April 1974.
5. C. E. Shannon, "Communication Theory of Secrecy Systems," B.S.T.J., *28*, No. 4 (October 1949), pp. 656–715.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

Optimum Direct Detection for Digital Fiber-Optic Communication Systems

By G. J. FOSCHINI, R. D. GITLIN, and J. SALZ

(Manuscript received January 23, 1975)

We report on optimum direct detection of digital data signals that are transmitted over optical fibers. Direct detection is provided by a photodetector whose output current is modeled as a noisy filtered Poisson stream of pulses. In this model, the time-varying pulse arrival rate is proportional to a linearly distorted version of the modulating signal. We show how the photodetector output is processed to derive the minimum probability-of-error receiver. Special attention is given to certain practical limiting cases.

When the average energy in the response of the photodetector to an individual photon is small compared to the additive thermal noise, the optimum detector is shown to be linear except for the use of precomputed bias terms. At the other extreme are the photomultiplier and the avalanche photodiode where the average energy in the response of the photodetector to a single photon is large compared with the additive noise. In this situation, we show that the optimum detector estimates the photon arrival times and then uses these estimates in a weighted counter. In both limiting cases, the detectors are specialized to one-shot M-ary and synchronous multilevel pulse-amplitude modulated (PAM) signals with intersymbol interference. For PAM signaling, we demonstrate that finite system memory allows application of dynamic programming to provide a detector implementation whose computational complexity does not increase with time.

I. INTRODUCTION

In recent years much attention has been focused on communication over optical channels.[1,2] Most early work was concerned with the physics of the electromagnetic transmission phenomena associated with various optical media and with the devices needed to change electrical signals to optical ones, and vice versa. In this paper, we are concerned with the optimum (maximum likelihood) reception of digital data transmitted over the fiber-optic channel. Our work was motivated by the many invaluable discussions we have had with S. D. Personick on this subject.

We shall not dwell on the quantum mechanical limitations imposed on the measurements of signals in the optical frequency range. Instead, we adopt a practical approach and assume at the outset that direct detection is used to convert optical energy to an electrical signal. This is accomplished by using a photodetector prior to any signal processing. Thus, we study a classical optical reception problem with the understanding that the photodetector output can be examined in every detail so as to extract all relevant information.

In a fiber-optic communication system, information is conveyed by modulating the intensity of a light source, such as a light-emitting diode. This is manifested in a photon stream whose arrival times form a Poisson process with a time-varying intensity function. The photodetector output current can then be modeled as a noisy filtered Poisson process whose intensity function is the sum of a dispersed version of the modulating wave and a background dark current. Thus, the central problem in communication systems employing a fiber-optic medium is the detection of the intensity function. Bar-David[3] and Gagliardi and Karp[4] have considered the optimal reception problem in the absence of dispersion (intersymbol interference) and additive thermal noise, while Personick[5–7] and Messerschmitt[8,9] have considered linear suboptimum receivers to combat these deleterious effects.

Section II reviews the communication theoretic model of the fiber-optic channel. Section III presents two simple examples that are intended to focus on certain system essentials and to illustrate some fundamental ideas involved in subsequent work. Section IV develops a general representation for the likelihood functional. Sections V and VI consider reception when the energy in the response of the photodetector to an individual photon is much smaller than the thermal noise, while Sections VII and VIII consider the complementary situation of large average energy per pulse-to-thermal noise.

II. A REVIEW OF THE MATHEMATICAL MODEL

In the past few years, a pragmatic communication theoretic model for data transmission over the fiber-optic channel has evolved. The papers by Personick[5–7,10] contain an up-to-date account of this model as well as provide more complete references on the physical aspects of fiber-optic communication. For the purpose of this investigation, it will suffice to think of the optical modulation process as providing a proportionate variation in the rate of photon arrivals at the photodetector. This device, of which there are several types, is a transducer that converts optical to electrical signals. The photodetector output current is illustrated in Fig. 1, and can be described as the sum of a

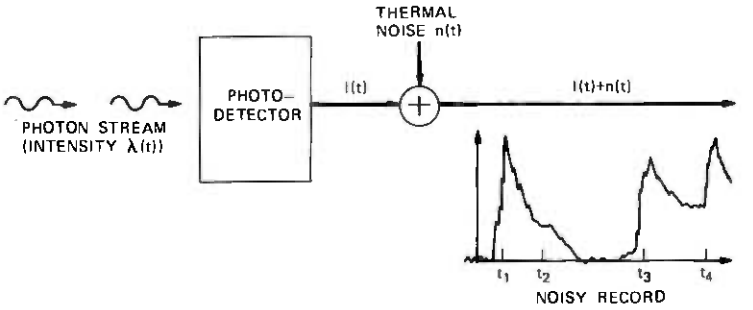


Fig. 1—Photodetection.

filtered Poisson process

$$I(t) = \sum_{k=1}^{\nu(t)} g_k w(t - t_k) \tag{1}$$

and white gaussian noise, $n(t)$, with spectral density N_0. The photon arrival times $t_1, t_2, \cdots$ are a family of independent, identically distributed, random variables, as are the positive gains $g_1, g_2, \cdots$. Moreover, these two families of random variables are independent of each other. The pulse $w(t)$ is square-integrable and is the convolution of two pulse shapes. The first pulse is the response of the photodetector circuitry to the generation of a single charge-carrier (i.e., an electron or a hole), while the second pulse is included for mathematical expediency so as to whiten the noise at the photodetector output.† We distinguish between two types of photodetectors, those that provide avalanche gain and those that do not. In the latter category is the photodiode that operates with $g_i = 1$, $i = 1, \cdots, \nu$ and results in a pulse energy-to-noise ratio $\int w^2(t)dt/N_0$, which is typically -20 dB. In other words, the response of the photodetector to an individual photon is masked by the additive background noise. This is in contrast to the photomultiplier and the avalanche photodiode where the gains possess a (discrete) probability distribution whose mean, $\bar{g}$, can be rather large and whose variance is a power ($\geqq 1$) of the mean.[11] For these devices, the average pulse energy-to-noise ratio $\overline{g^2}\int w^2(t)dt/N_0$ can be on the order of 20 dB.

The stochastic process $\nu(t)$, which is the number of pulses generated at the photodetector output in the interval $(0, t)$, is a Poisson process with intensity $\lambda(t)$, and therefore

$$\Pr[\nu(t) = N] = \exp\{-\Lambda(t)\}\frac{[\Lambda(t)]^N}{N!}, \tag{2}$$

† Note that the inclusion of a reversible operation, such as a whitening filter, does not affect the performance of an optimum detector.

where

$$\Lambda(t) = \int_0^t \lambda(t')dt'. \tag{3}$$

Moreover, each photon arrival time t_k possesses the probability density

$$p(t_k) = \frac{\lambda(t_k)}{\int \lambda(t')dt'}, \tag{4}$$

where the integral is over the observation time.†

In the digital fiber-optic communication system under discussion here, the positive intensity function $\lambda(t)$ is the information-bearing signal and is the average rate of electrons produced by the photodetector. The manner in which $\lambda(t)$ is manifest in the received optical signal (the photodetector input) is through the relation

$$\lambda(t) = k\mathcal{P}(t) + \lambda_0, \tag{5}$$

where $\mathcal{P}(t)$ is the received optical power, k is a constant conversion factor, and λ_0 is the average dark, or ambient, current in "counts" per second.‡ Thus, information is transmitted by modulating the optical power and must be recovered by processing the noisy photodetector output, $I(t) + n(t)$. As a result of transmitting the optical signal through the fiber-guide medium, the intensity function at the photodetector output will be the sum of a linearly distorted version of the transmitter intensity and the dark current. In the sequel, $\lambda(t)$ will be understood to mean the intensity function at the receiver.

Statistical averages of $I(t)$ are found by elementary calculations. For example,

$$E[I(t)] = E(g) \int_{-\infty}^{\infty} \lambda(\tau)w(t - \tau)d\tau \tag{6}$$

and

$$\sigma^2_{I(t)} = E(g^2) \int_{-\infty}^{\infty} \lambda(\tau)w^2(t - \tau)d\tau, \tag{7}$$

where $E(g)$ and $E(g^2)$ are the average and average square of the avalanche gain g. Higher moments can also be readily evaluated.

A linear channel model with additive "noise" is suggested by (6) and (7). In such a model, the desired signal is taken to be the average value of $I(t)$, namely $\lambda(t)$ passed through a filter with impulse response $E[g]w(t)$. One component of the added noise can be thought of as the signal dependent process $I(t) - E[I(t)]$, which has mean zero and

† Note that the arrival times are not assumed to be ordered.

‡ In free-space optical communication systems, $\lambda(t)$ must be regarded as having a noisy component.

variance given by (7). In addition to this noise, the gaussian noise must also be included before processing. While this linear model is a convenient approximation in some situations,[5-10] for purposes of this investigation we work with the process $I(t)$ directly.

Now that all the physical parameters have been defined, the optimum detection problem can be stated as follows:

Given that the intensity function can assume one of M equiprobable positive functions $\lambda_m(t)$, $0 \leqq t \leqq T$, $m = 1, \cdots, M$, the task of the detector is to decide which one of the M intensities has been transmitted after processing $I(t)$ plus gaussian noise for T seconds. Of particular interest is the synchronous pulse-amplitude modulated (PAM) signal

$$\lambda(t) = \sum_k a_k f(t - kT) + \lambda_0,$$

where each data bit, a_k, assumes the value 0 or 1, $1/T$ is the data rate in bits/s, and $f(t)$ is a positive time-dispersed pulse.

The subject of our investigation is summarized by the question: How should the photodetector output, $I(t) + n(t)$, be processed so as to minimize the probability of error?

III. A MOTIVATING SIMPLIFIED DISCRETE MODEL—TWO EXAMPLES

To preview, in an elementary way, some ideas that are more fully developed in the sequel and also to serve as a motivation to the reader, we present a simplified version of the model discussed in the last section.

In a simplified theoretical model, the time index t is assumed to take on the discrete set of values $t_1, t_2, \cdots, t_J$, where $t_j = j\Delta$. Thus, instead of writing

$$I(t) = \sum_{k=1}^{\nu(t)} g_k w(t - t_k)$$

for the photodetector response to a photon stream, we write

$$I(t_j) = \sum_{k=1}^{J} g_k q_k w(t_j - t_k) \qquad j = 1, 2, \cdots, J. \tag{8}$$

In the above expression, $\{q_k\}_1^J$ can be regarded as an independent Bernoulli sequence with probabilities†

$$\Pr\{q_k = 1\} = \lambda_k \qquad \text{and} \qquad \Pr\{q_k = 0\} = 1 - \lambda_k,$$

† For convenience, we have taken $\Delta = 1$, and so we have written λ_k and $1 - \lambda_k$ instead of $\lambda_k\Delta$ and $1 - \lambda_k\Delta$.

where we have in mind that $0 < \lambda_k \ll 1$. Thus, $q_k = 1$ (or 0) represents the arrival (or nonarrival) of a photon at time t_k. We make the further simplifying assumption that $w(t_j - t_k) = A\delta_{jk}$ (A a positive constant), where δ_{jk} is the Kronecker delta and is nonzero only when $j = k$. This corresponds to assuming that the pulses $w(t)$ and $w(t - \Delta)$ do not overlap. Within this simplified framework, the received time-discrete signal is of the form

$$I(t_j) = g_j q_j A, \qquad j = 1, 2, \cdots, J. \tag{9}$$

We recall that $\{\lambda_j^{(m)}\}_{j=1}^{J}$ is the intensity function associated with the mth hypothesis. The particular intensity which is active is, of course, unknown at the receiver beyond the knowledge of the finite set from which it was chosen. The last ingredient of our model is to include the fact that the observation $I(t_j)$ is noisy and is given by

$$y(t_j) = g_j q_j A + n_j, \tag{10}$$

where the noise samples are assumed to be gaussian, independent, and zero-mean and have variance N_0. In relation to the more accurate model of the previous section, σ can be thought of as the standard deviation corresponding to $\int_{t_{j-1}}^{t_j} n(t)dt$. As is well known, the optimum detector computes the likelihood (the a posteriori probability density of the received signal conditioned on each hypothesis—in this case, the intensity) and selects the maximum. In statistical parlance, this is a standard multihypothesis testing problem. We now develop the form of the likelihood for two different assumptions on the nature of generation of secondary electrons:

(*i*) No avalanche gain ($g_j \equiv 1$).
(*ii*) Discrete avalanche gain (g_j takes on values 1, 2, $\cdots$, G, with probabilities $\rho_1, \rho_2, \cdots, \rho_G$).

In each case, we first obtain the likelihood for one observation. Owing to the nonoverlapping assumption on the pulses and the independent noise samples, the likelihood for J independent observations is given as a product. Our goal is to obtain a simple representation for the effective† likelihood $L^{(m)}(\lambda_1, \lambda_2, \cdots, \lambda_J; y_1, y_2, \cdots, y_J)$, where the superscript m denotes which intensity is assumed active. Given the received samples $y_1, \cdots, y_J$, the maximum likelihood (optimum) receiver selects the index m^* that maximizes $L^{(m)}$ and declares that intensity $\lambda^{(m^*)}$ is present. We shall find that, if N_0 is small, then the

† "Effective" refers to the fact that constants common to all hypotheses are dropped.

likelihood assumes an especially simple form. Specifically, in the high signal-to-noise ratio case, the likelihood is of the form

$$L^{(m)} \sim \prod_{j=1}^{J} (\lambda_j^{(m)})^{\bar{q}_j} (1 - \lambda_j^{(m)})^{1-\bar{q}_j}, \tag{11}$$

where $\bar{q}_j = 1$ if $y_j \geqq y_T$ (and zero otherwise). The quantity y_T is a threshold value that we shall derive for each example. Alternatively, the log-likelihood is expressible as the *weighted counter*

$$\sum_{j=1}^{J} \bar{q}_j \log \lambda_j^{(m)} + [(1 - \bar{q}_j) \log (1 - \lambda_j^{(m)})], \tag{12}^{\dagger}$$

where $\bar{q}_j$ is an estimated photon arrival process. In the complementary case of small signal-to-noise ratio ($N_0 \to \infty$), the detector is of the matched-filter or correlator type. The effective likelihood in this case is

$$L^{(m)} \sim c \sum_{j=1}^{J} \lambda_j^{(m)} y_j - b^{(m)}, \tag{13}$$

where c is a constant and $b^{(m)}$ is a hypothesis-sensitive bias term. We now turn to the specific examples.

(*i*) *The Photodiode (No Avalanche Gain)*

The single observation y_j is defined as

$$\begin{aligned} y_j &= n_j, && \text{with probability } 1 - \lambda \\ &\text{and} && \\ y_j &= A + n_j, && \text{with probability } \lambda. \end{aligned} \tag{14}$$

We temporarily drop the subscripts dealing with time (j) and hypothesis (m) while investigating this single observation. The likelihood is the mixture probability density

$$p(y) = (2\pi N_0)^{-\frac{1}{2}} \exp\left\{-\frac{y^2}{2N_0}\right\}\left[(1 - \lambda) + \lambda \exp\left\{\frac{Ay}{N_0} - \frac{A^2}{2N_0}\right\}\right]. \tag{15}$$

Noticing the hypothesis (λ) insensitivity of the first term, the effective likelihood becomes

$$L(y) = (1 - \lambda) + \lambda \exp\left\{\frac{Ay}{N_0} - \frac{A^2}{2N_0}\right\}. \tag{16}$$

A simple calculation shows that the two terms in (16) are equal when

---

† The reader familiar with Ref. 3 might expect an additional $-A$ term in (12). Owing to the simplified Bernoulli model employed above, this is not the case. However, the more refined analysis in the sequel will include this term.

$y = y_T$, where

$$y_T = \frac{A}{2} + \frac{N_0}{A} \log \left(\frac{1 - \lambda}{\lambda} \right). \tag{17}$$

For small N_0, $y_T \approx A/2$ and the graph of $L(y)$ converges to the solid line shown in Fig. 2. So, for $N_0/A^2 |\log \lambda|$ small, the effective likelihood can be approximated as

$$\hat{L}(y) = \begin{cases} 1 - \lambda, & y \leqq y_T \\ \lambda \exp \left\{ \dfrac{Ay}{N_0} - \dfrac{A^2}{2N_0} \right\}, & y > y_T. \end{cases} \tag{18}$$

The sense of the approximation is expressed by the following easily proven statement: For each $\delta > 0$, one can find an $N_0 > 0$ so that

$$\Pr \left\{ \left| \frac{L(y)}{\hat{L}(y)} - 1 \right| > \delta \right\} = 0. \tag{19}$$

To simplify the likelihood, note that $\exp \{(Ay/N_0) - (A^2/2N_0)\}$ and y_T are hypothesis-insensitive and can be deleted from the effective likelihood, and since we are assuming that λ is extremely small, $1 - \lambda$ can be treated as 1. The effective likelihood is then simply

$$\hat{L}^{(m)} = [\lambda_j^{(m)}]^{\bar{q}_j}, \tag{20}$$

where $\bar{q}_j = 1$ if $y_j > y_T$ and zero otherwise. Because of the independence of the noise samples and the nonoverlapping property of the

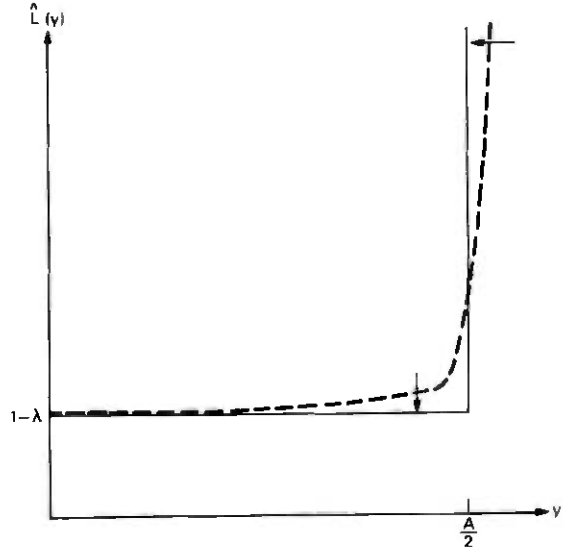


Fig. 2—Convergence of graph of $\hat{L}(y)$ to the asymptotic form ($N_0 \to 0$).

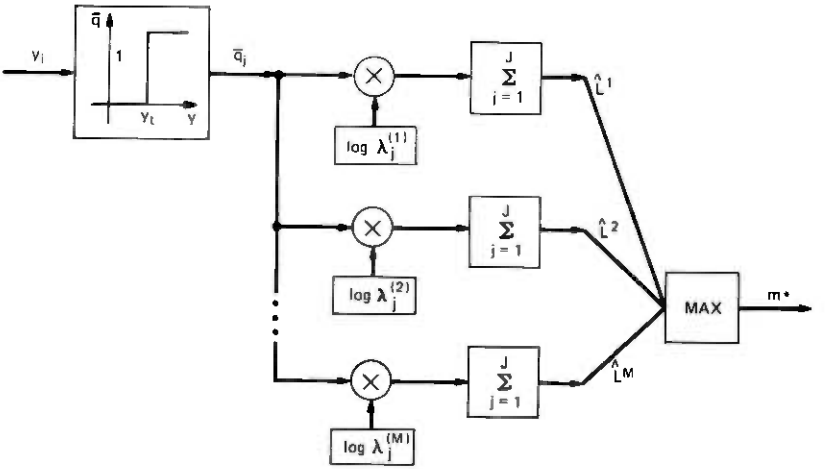


Fig. 3—Threshold-based weighted counter.

pulses, the likelihood for J observations is the product

$$\hat{L}^{(m)} = \prod_{j=1}^{J} [\lambda_j^{(m)}]^{\bar{q}_j}, \tag{21}$$

which yields the weighted counter

$$\log \hat{L}^{(m)} = \sum_{j=1}^{J} \bar{q}_j \log \lambda_j^{(m)} \tag{22}$$

shown in Fig. 3. The receiver selects the index that maximizes (22) and declares that the corresponding intensity was transmitted.

In the complementary case of low signal-to-noise ratio ($N_0 \to \infty$), we expand the likelihood function in a Taylor series and retain the dominating terms. This step must be done with care, since the numerator of the exponent has variance N_0, while N_0 also appears in the denominator. By normalizing the exponent, it is seen that the variance of the exponent is proportional to $1/N_0$; thus, the exponent will be small and a series expansion is useful. Keeping the first two terms in such an expansion of (16) gives

$$\hat{L}(y_j) = 1 - \lambda_j \left(\frac{Ay_j}{N_0} - \frac{A^2}{2N_0} \right), \tag{23}$$

and the likelihood for J observations becomes[†] the digital correlator

[†] We have used the fact that $\lambda_j/N_0[Ay_j - (A^2/2)] \ll 1$, and with $\epsilon_i \ll 1$ that $\Pi(1 + \epsilon_j) \sim 1 + \sum \epsilon_j$.

(matched-filter)

$$\hat{L}^{(m)} = \prod_{j=1}^{J} \left[1 - \lambda_j \left(\frac{Ay_j}{N_0} - \frac{A^2}{2N_0}\right)\right] \sim \sum_{j=1}^{J} \lambda_j^{(m)} \left(Ay_j - \frac{A^2}{2}\right), \quad (24)$$

which is shown in Fig. 4.

(*ii*) *The Photomultiplier or Avalanche Photodiode (Discrete Avalanche)*

Again, we start with the single observation case but now, because of the avalanche mechanism, a single primary gives rise to 1 or 2 or $\cdots$, G secondaries with probabilities ρ_1, ρ_2, $\cdots$, ρ_G, respectively ($\sum_1^G \rho_i = 1$). So the measurement y is modified as

$$y = \begin{cases} n, & \text{with probability } 1 - \lambda \\ A + n, & \text{with probability } \lambda\rho_1 \\ \vdots & \\ GA + n, & \text{with probability } \lambda\rho_G. \end{cases} \quad (25)$$

The likelihood is the mixture density

$$p(y) = \frac{(1-\lambda)}{\sqrt{2\pi N_0}} \exp\left\{-\frac{y^2}{2N_0}\right\} + \sum_{l=1}^{G} \frac{\lambda\rho_l}{\sqrt{2\pi N_0}} \exp\left\{-\frac{(y - lA)^2}{2N_0}\right\}. \quad (26)$$

Factoring out hypothesis-insensitive terms, the effective likelihood becomes

$$L(y) = (1 - \lambda) + \lambda \sum_{l=1}^{G} \rho_l \exp\left\{\frac{lAy}{N_0} - \frac{l^2A^2}{2N_0}\right\}. \quad (27)$$

As $N_0 \to 0$, we notice that $\hat{L}(y) \sim (1 - \lambda)$ for $y < A/2$. When $y > A/2$, let $\bar{l}A$ denote the number A, $2A$, $\cdots$, or GA that is closest

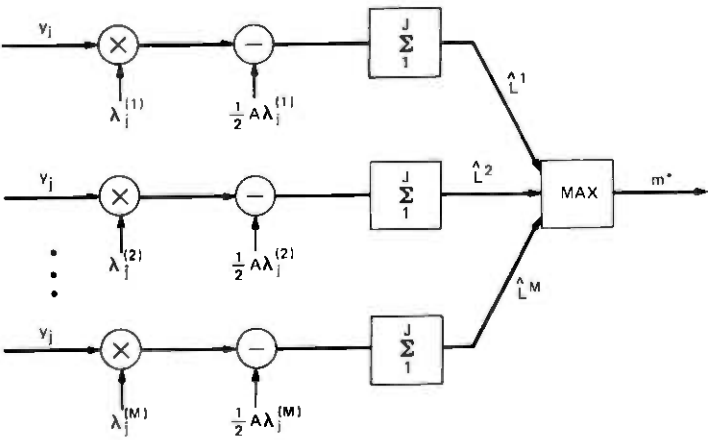


Fig. 4—Elementary version of digital correlator.

to y. Then the series appearing in (27) will be dominated by one term, and the likelihood becomes

$$\hat{L}(y) \sim \rho_l \lambda \exp\left\{ \frac{lAy}{N_0} - \frac{l^2A^2}{2N_0} \right\}, \qquad \text{as } N_0 \to 0.$$

Proceeding as in the previous example, we consider both N_0 and λ small and drop hypothesis-insensitive terms from the approximate likelihood to obtain

$$\hat{L}(y) = [\lambda]^{\bar{q}}, \tag{28}$$

where $\bar{q} = 1$ when $y \geq A/2$ and zero otherwise.† Moreover, note that the threshold is the same as in the nonavalanche case. This is because the detector is only interested in ascertaining whether or not a photon has arrived and need not estimate the magnitude of the avalanche gain. Again, for J measurements, the corresponding log-likelihood expression is simply the weighted counter

$$\log \hat{L} = \sum_{j=1}^{J} \bar{q}_j \log \lambda_j. \tag{29}$$

As $N_0 \to \infty$, we again expand the likelihood (27) in a Taylor series to obtain

$$\hat{L}(y) = 1 - \lambda \left\{ 1 - \sum_{l=1}^{G} \rho_l \left[1 + \frac{lAy}{N_0} - \frac{l^2A^2}{2N_0} \right] \right\}, \tag{30}$$

which, for J measurements, becomes

$$\log \hat{L}^{(m)} = \sum_{j=1}^{J} \lambda_j^{(m)} \left\{ 1 - \sum_{l=1}^{G} \rho_l \left[1 + \frac{lAy_j}{N_0} - \frac{l^2A^2}{2N_0} \right] \right\}. \tag{31}$$

The above is again interpreted as a correlator where $\lambda_j^{(m)}$ is correlated with $Ay_j/N_0 \cdot \sum_{l=1}^{G} l\rho_l = (Ay_j/N_0)E[g]$.

IV. THE MAXIMUM LIKELIHOOD DETECTOR

Here, we begin to answer the question posed at the end of Section II by presenting a derivation of the likelihood function associated with the received signal. The likelihood function is the probability measure of the photodetector output, given that a particular intensity is active. It is well known[12] that, when one of M equally likely signals $\lambda_m(t)$ is transmitted, the optimum (minimum probability error) detector computes the M values of the likelihood function evaluated at the received waveform and declares that the jth signal was sent, where the jth likelihood function is the largest.

† As expected when $N_0 \to 0$, the avalanche gain provides no essential benefit. A more interesting asymptotic evaluation and one that is more akin to reality is obtained by parameterizing the gain distribution such that $E[g]/N_0 \to \infty$.

We denote the received signal by

$$y(t) = I_m(t) + n(t), \qquad 0 \leqq t \leqq \mathcal{T}, \tag{32}$$

where $I_m(t)$ is the information-carrying, filtered, Poisson process

$$I_m(t) = \sum_{k=1}^{\nu(t)} g_k w(t - t_k), \tag{33}$$

and where the index m [corresponding to $\lambda_m(t)$] is hidden in the statistics of $\{t_k\}$ and $\nu(t)$. These statistics are described by (2) to (4) with $\lambda(t)$ replaced by $\lambda_m(t)$.

The task of the optimum receiver is thus to process the photodetector output $y(t)$ for $\mathcal{T}$ seconds and then decide which intensity function $\lambda_m(t)$, $m = 1, 2, \cdots, M$ is in effect. As we have mentioned earlier, the random variables $\{g_k\}$ represent the avalanche gains, and the pulse shape $w(t)$ is so far arbitrary with the only requirement being finite energy. Although in actual practice the noise at the output of the photodetector is not white, it can be whitened by a filter before additional processing and the effect of this filter will be manifest in the shape of $w(t)$.

The conditional likelihood function [when $I_m(t)$ is fixed] has the standard form[13]

$$L_m[y \mid I_m] = \exp\left\{\frac{1}{N_0}\int_0^{\mathcal{T}} I_m(t)y(t)dt - \frac{1}{2N_0}\int_0^{\mathcal{T}} I_m^2(t)dt\right\}. \tag{34}$$

The desired likelihood is the expectation of (34) with respect to $I_m(t)$ for fixed m, i.e.,

$$L_m(y) = E_I\{L_m[y \mid I_m]\}. \tag{35}$$

Once the intensity $\lambda_m(t)$ is specified, the above expectation is taken with respect to the number of arrivals, the arrival times, and the avalanche gain values. The detailed evaluation of this expectation and the interpretation of the resulting structures, in terms of implementable physical operations on $y(t)$, is our objective. The exact structure is sufficiently complex that many judicious approximations will have to be made to glean the essential nature of the operations.

We remark that a representation of (35) in terms of an estimator-correlator structure has recently been treated in the literature.[12,14–16] The optimum detector has been shown to be a correlation detector and the deterministic signal in the classical correlator is replaced by its least-squares estimate. This is a reformulation of the detection problem in terms of an estimation problem. Proponents of this method have taken the viewpoint that various suboptimum detectors are suggested by this formulation. A typical approach might be to replace the least-

squares estimate by the linear least-squares estimate or some other approximation, and to approximate the resulting stochastic integral by conventional integrals. While this might be reasonable, it does not indicate the direction of the approximation. We prefer an approach that, to be sure, has many approximations and makes use of estimates in place of the true quantities, but that can be explicitly related to the optimum detector under the asymptotic conditions of large and small signal-to-noise ratio.

Toward this end, we proceed by writing (35) in more detail. Neglecting edge effects on the integrals and assuming that the observation time $\mathcal{T}$ is much larger than the effective duration of a single pulse $w(t)$, we can express the inner product and the square term indicated in (34) as

$$\int_0^{\mathcal{T}} I_m(t)y(t)dt = \sum_{k=1}^{\nu} g_k P(t_k), \tag{36}$$

where

$$P(t_k) = \int_0^{\mathcal{T}} w(t - t_k)y(t)dt.$$

The square term is written as

$$\int_0^{\mathcal{T}} I_m^2(t)dt = \sum_{k,j=1}^{\nu} g_k g_j R(t_k - t_j), \tag{37}$$

where $R(t) = \int_0^{\mathcal{T}} w(\tau)w(t - \tau)d\tau$ is defined as the pulse correlation function.

Substituting (36) and (37) into (35), we obtain

$$L_m(y) = E_I \exp\left\{\frac{1}{N_0}\sum_{k=1}^{\nu} g_k P(t_k) - \frac{1}{2N_0}\sum_{k,j=1}^{\nu} g_k g_j R(t_k - t_j)\right\}. \tag{38}$$

Employing the vector notation $\mathbf{g}_\nu = (g_1, g_2, \cdots, g_\nu)$ and $\mathbf{t}_\nu = (t_1, t_2, \cdots, t_\nu)$ gives the expression

$$L_m(y) = E_{\mathbf{t}_\nu, \mathbf{g}_\nu, \nu}\left[\exp\left\{\frac{1}{N_0}\sum_{k=1}^{\nu} g_k P(t_k) - \frac{1}{2N_0}\sum_{k,j=1}^{\nu} g_k g_j R(t_k - t_j)\right\}\right], \tag{39}$$

and after performing the indicated expectations we obtain a detailed representation of the likelihood function

$$L_m(y) = \exp\left[-\Lambda_m(\mathcal{T})\right]\sum_{n=0}^{\infty}\frac{1}{n!}\sum_{\mathbf{g}_n}\int_0^{\mathcal{T}} dt_n \prod_{k=1}^{n}\lambda_m(t_k)\prod_{k=1}^{n}\rho(g_k)$$
$$\times \exp\left\{\frac{1}{N_0}\sum_{k=1}^{n} g_k P(t_k) - \frac{1}{2N_0}\sum_{k,j=1}^{n}\sum_{k,j=1}^{n} g_k g_j R(t_k - t_j)\right\}, \tag{40}$$

where $\rho(g_i)$ is the (discrete) probability density function of the ava-

lanche gains and where it is understood that, when $n = 0$, the summand is taken to be unity.

To more easily interpret and/or mechanize the likelihood calculations, it will be convenient in some applications to assume that the photon arrivals can only occur at discrete instants of time $\{j\Delta\}$, where Δ is some fixed (small) interval and $j = 0, 1, 2, \cdots, J$. The integer J will be defined as the closest integer to T/Δ. This assumption is easily accommodated in (40) by replacing $\int dt_n$ with a multidimensional sum $\sum_{t_n}$ over the lattice $\{t_k = j\Delta: k = 1, 2, \cdots, n; j = 0, 1, \cdots, J\}$, and by replacing $\lambda(t_k = j\Delta)$ with the probability that† $j\Delta - \Delta/2 \leqq t_k \leqq j\Delta + \Delta/2$. The likelihood function under this set of assumptions then becomes

$$L_m(y) = \exp\left[-\lambda_m(T)\right] \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{t_n=0}^{J} \sum_{g_n} \prod_{k=1}^{n} \lambda(t_k) \prod_{i=1}^{n} \rho(g_i) \times \exp\left\{\frac{1}{N_0}\left[\sum_{k=1}^{n} g_k P(t_k) - \tfrac{1}{2} \sum_{k,m=1}^{n} g_k g_m R(t_k - t_m)\right]\right\}, \quad (41)$$

which will be referred to as the (time) discrete likelihood function.

The two infinite functional series, (40) and (41), are not of much use as they stand. However, under a variety of physically realistic situations and by making suitable physical approximations as well as asymptotic expansions, we shall be able to deduce from these representations real-time implementable signal-processing algorithms.

By suitably normalizing the likelihood functions, (40) and (41), $1/N_0$ can be replaced by the (pulse) signal-to-noise ratio. This parameter α^2 will play a central role in our subsequent treatment, and its relative size will dictate our particular approach. The normalization entails replacing $R(t)$ by $R(t)/R(0)$, $P(t_k)$ by $P(t_k)/R(0)\bar{g}$, and the random variables g_k by $g_k/\bar{g}$, where $\bar{g} = Eg_k$; consequently,

$$\alpha^2 = \frac{\bar{g}^2 R(0)}{N_0}$$

and may be viewed as an average pulse signal-to-noise ratio. As we have discussed in the preceding section, in some applications this parameter is small, while in others it is large. Thus, our investigations in the sequel will focus on these two ranges. Additionally, different treatments of the likelihood ratio are also required, depending upon the presence or absence of avalanche gain.

It is instructive to give a still different representation for the likelihood, which will be found useful in the sequel. Towards this end, we introduce a zero-mean, stationary gaussian process $x(t)$ with correlation

† This probability is given by

$$\int_{j\Delta-\Delta/2}^{j\Delta+\Delta/2} \lambda(t)dt \approx \lambda(j\Delta)\cdot\Delta.$$

function,

$$E[x(t)x(t+\tau)] = R(\tau),$$

and can then write (39) in the form

$$L_m(y) = E_{t_\nu, g_\nu, \nu}\left[\exp\left\{\alpha^2 \sum_{k=1}^{\nu} g_k P\{t_k\}\right\} E_x \exp\left\{-i\alpha \sum_{k=1}^{\nu} g_k x(t_k)\right\}\right], \tag{42}$$

where we have used the elementary identity for gaussian processes

$$\exp\left\{-\alpha/2 \sum_{k,j=1}^{\nu} g_k g_j R(t_k - t_j)\right\} = E_x \exp\left\{i\alpha \sum_{k=1}^{\nu} g_k x(t_k)\right\}.$$

Since, over the observation interval, (42) is absolutely integrable, the expectation with respect to x and the other random variables may be interchanged. By noting that

$$E_{t_n, g_n, \nu} \exp\left\{\sum_{k=1}^{\nu} g_k x(t_k)\right\} = \exp(-\Lambda_m) \cdot \left\{\exp \sum_{g_n} \int_0^T \rho(\mathbf{g})\lambda_m(t) \exp[i\alpha x(t)]dt\right\}, \tag{43}$$

we can write (42) in the form

$$L_m(y) = \exp(-\Lambda_m) E_x \left\{\exp\left(\sum_{\mathbf{g}} \int_0^T \rho(\mathbf{g})\lambda_m(t) \times \exp[\alpha^2 g_j P(t) + i\alpha g_j x(t)]dt\right)\right\}. \tag{44}$$

In particular, in the absence of avalanche gain, i.e., $\rho(g) = \delta(g - 1)$ (44) assumes the compact form

$$L_m(y) = \exp(-\Lambda_m) E_x \left\{\exp\left(\int_0^T \lambda_m(t) \exp[\alpha^2 P(t) + i\alpha x(t)]dt\right)\right\}. \tag{45}$$

It may appear that the introduction of the process $x(t)$ did not simplify matters, since the explicit evaluation of the expectations again leads to an infinite functional series without adding insight into the nature of the processor. We shall nevertheless find this representation useful. As will be seen, when suitable approximations are made and asymptotic behaviors explored, a great deal of insight can be gained from the alternative representations for the likelihood† (40), (44), and (45), as well as the discrete likelihood (41).

† By normalizing the exponent, i.e., introducing α^2, we should actually use new symbols to denote $g_m/\bar{g}$ and $R/R(0)$. To avoid introducing extra notation, we retain the symbols g_m and $R(0)$, but we realize that, whenever α^2 is present, these variables have been normalized.

V. SMALL SIGNAL-TO-NOISE RATIO ($\alpha^2 \to 0$)

Here we consider the physical situation corresponding to small s/n ($\alpha^2 \to 0$). This occurs when a photodiode is used for direct detection. In this application, the response to an individual photon is masked by the background noise, and we do not expect the receiver to make explicit use of the information supplied by an individual pulse. Rather, the aggregate effect will be important. This is in contrast to the "counting" receivers (for large α^2), where individual counts contribute explicitly to the final decision. Since the avalanche gains are unity in this application, the likelihood function takes the form of (45). Two signaling situations of interest are examined next.

5.1 M-ary signaling

Since $\alpha^2 \ll 1$ (typically, $\alpha = -20$ dB), our approach will be to expand (45) in a power series in α^2 and retain the first two terms.† Consider the following Taylor series approximation to the argument of the exponent in (45). Again dropping the index m, let

$$\xi(\alpha, x) = \exp\left\{\int_0^T \lambda(t) \exp\left[\alpha^2 P(t) + i\alpha x(t)\right] dt\right\}$$

$$\sim e^{\Lambda} + \xi'(0, x)\alpha + \xi''(0, x)\frac{\alpha^2}{2}. \quad (46)$$

Evaluating the derivatives, the asymptotic likelihood function becomes

$$L_m(y) \sim E_x\left[1 + \alpha\int_0^T \lambda_m(t)x(t)dt + \frac{\alpha^2}{2}\left(\int_0^T 2\lambda_m(t)P(t)dt - \int_0^T\int_0^T \lambda_m(t_1)\lambda_m(t_2)x(t_1)x(t_2)dt_1dt_2\right) - \frac{\alpha^2}{2}\int_0^T \lambda_m(t)x^2(t)dt\right]. \quad (47)$$

Recalling that the exponent has been normalized such that $Ex = 0$, $Ex^2 = 1$, and $Ex(t_1)x(t_2) = R(t_1 - t_2)$, we get, after performing the averages,

$$L_m(y) \sim 1 + \alpha^2\left[\int_0^T \lambda_m(t)P(t)dt - \frac{1}{2}\int_0^T\int_0^T \lambda_m(t_1)\lambda_m(t_2)R(t_1 - t_2)dt_1dt_2 - \frac{1}{2}\int_0^T \lambda_m(t)dt\right] \quad (48)$$

or

$$\bar{L}_m(y) = \log L_m(y) \sim \int_0^T \lambda_m(t)P(t)dt - \frac{1}{2}\int_0^T\int_0^T \lambda_m(t_1)\lambda_m(t_2)R(t_1 - t_2)dt_1dt_2 - \tfrac{1}{2}\Lambda_m. \quad (49)$$

The detector involves linear operations on the filtered received signal $P(t)$, addition of constants, and a maximization. As shown in Fig 5,

† Of course, the same answer would be obtained by working with (40).

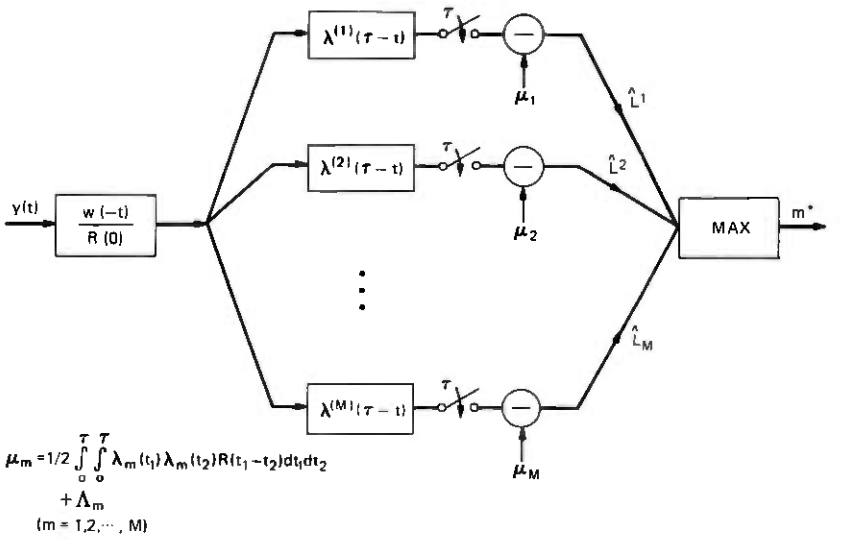


Fig. 5—Correlator filter for M-ary signaling.

a realization of the receiver is obtained by first passing the incoming signal, $y(t)$, through a filter with impulse response $w(-t)/R(0)$ to produce $P(t)$. This signal is then passed through a bank of M filters with impulse responses $\lambda_m(\mathcal{T} - t)$, $m = 1, 2, \cdots, M$ and sampled at $t = \mathcal{T}$. This is the first term in (49). The other two terms are precomputable biases. The detector then chooses the index m^*, which achieves the $\max_m \bar{L}_m(y)$, and the corresponding $\lambda_{m^*}(t)$ is declared to be the transmitted intensity.

There is a pleasing interpretation of this receiver which is reminiscent of the "linear" model discussed in Section II. If one were to consider the detection problem when the signal $I(t)$, given by (1), is replaced by its average $E[I(t)] = \bar{I}(t)$, given by (6), then the optimum detector in gaussian noise would base its decision on the likelihood function

$$\mathfrak{L} = \int_0^{\mathcal{T}} y(t)\bar{I}(t)dt - \frac{1}{2}\int_0^{\mathcal{T}} [\bar{I}(t)]^2 dt. \tag{50}$$

Substituting (6) into (50) gives

$$\begin{aligned}\mathfrak{L} &= \int_0^{\mathcal{T}} dt\, y(t) \int_0^{\mathcal{T}} w(\tau - t)\lambda(\tau)d\tau \\ &\quad - \frac{1}{2}\int_0^{\mathcal{T}} \left\{ \int_0^{\mathcal{T}} \int_0^{\mathcal{T}} w(t - t_1)\lambda(t_1)w(t - t_2)\lambda(t_2)dt_1dt_2 \right\} dt \\ &= \int_0^{\mathcal{T}} \lambda(\tau)P(\tau)d\tau - \frac{1}{2}\int_0^{\mathcal{T}} \int_0^{\mathcal{T}} \lambda(t_1)\lambda(t_2)R(t_1 - t_2)dt_1dt_2. \end{aligned} \tag{51}$$

Note that (51) differs from (49) only by the bias term Λ_m, the probability that no photons have arrived at the photodetector. We conclude, therefore, that the optimum detector structure in the case of small α^2 is thus "matched" to the average signal.

5.2 Optimum detection of PAM signals via the Viterbi algorithm

We will now develop the optimum receiver structure (still for small α^2) when the intensity is a pulse-amplitude modulated (PAM) signal†

$$\lambda(t) = \sum_{n=0}^{k} a_n f(t - nT), \qquad 0 \leqq t \leqq \mathcal{T}, \tag{52}$$

where each a_n can assume the binary values 0 or 1, $f(t)$ is a positive-valued pulse that incorporates the distortion of the optical medium, $1/T$ is the symbol rate, and $\mathcal{T} > kT$. Note that in writing (52) we have dropped the subscript m which we have used to identify the transmitted signal (intensity), since for PAM signaling it is generally more convenient to think of the receiver as finding that sequence $\{a_n\}$ which maximizes the likelihood. Substituting (52) in (49) and emphasizing that the likelihood function is now to be regarded as a function of a particular data sequence (which uniquely corresponds to a specific intensity) gives

$$\bar{L}(a_1, a_2, \cdots, a_k) = \sum_{n=1}^{k} a_n z_n - \frac{1}{2} \sum_{n,m=1}^{k} a_n a_m \mathcal{K}_{n-m}, \tag{53}$$

where

$$z_n = \int_0^{\mathcal{T}} [P(t) - \tfrac{1}{2}] f(t - nT) dt \tag{54}$$

is the response at time nT of a filter matched to $f(t)$ when the input is $P(t) - \frac{1}{2}$, and the correlation-type function $\mathcal{K}$ is defined by

$$\begin{aligned} \mathcal{K}_{n-m} &= \int_0^{\mathcal{T}} d\tau \left(\int_0^{\mathcal{T}} dt f(t - nT) w(t - \tau) \right) \\ &\qquad \times \left(\int_0^{\mathcal{T}} dt' f(t' - mT) w(t' - \tau) \right) \\ &= \int_0 U(\tau - nT) U(\tau - mT) dt \\ &= \int_0^{\mathcal{T}} U(\tau) U[\tau - (n - m)T] dt, \end{aligned} \tag{55}$$

† Note that we have neglected the dark current λ_0. This obviously does not alter the final results. Also, the results can, in a straightforward manner, be extended to the multilevel case.

with

$$U(\tau - nT) = \int_0^{\mathcal{T}} f(t_1 - nT)w(\tau - t_1)dt_1,$$

and the observation time $\mathcal{T}$ is taken to be extremely large ($\mathcal{T} \gg T$).

The receiver structure indicated by (53) to (55) is similar to the maximum likelihood (ML) receiver for detecting a PAM signal distorted by a noisy linear channel.[17] The received signal is first passed through the matched filter $w(-t)$, and then (minus the bias term $\frac{1}{2}$) matched to $f(-t)$. The result is sampled at the synchronous instants nT. This produces the set of sufficient statistics $\{z_n\}$, from which the hypothesis-insensitive bias term $\frac{1}{2}\sum_{n,m=1}^{k}\sum a_n a_m \mathcal{K}_{n-m}$ is subtracted to produce the likelihood function.

The method by which the likelihood (53) is sequentially maximized in real time has become known as the Viterbi algorithm (VA), as a result of its application to the analogous problem of ML detection of linearly distorted PAM data signals.

The VA is a dynamic programming algorithm that uses the "finite memory" of $\mathcal{K}_n$, i.e., the fact that there will always be a $\bar{k}$ such that, for all practical purposes,

$$\mathcal{K}_n = 0, \qquad |n| > \bar{k}. \tag{56}$$

Because of (56), it is easy to see that the likelihood, (53), can be written in the recursive form

$$L(a_1, a_2, \cdots, a_k) = L(a_1, a_2, \cdots, a_{k-1}) + a_k z_k - \tfrac{1}{2}a_k \sum_{m=k-\bar{k}}^{k} \mathcal{K}_{k-m}. \tag{57}$$

By introducing the sequence of state vectors $\{\sigma_n\}$, where

$$\sigma_n = (a_{n-(\bar{k}-1)}, \cdots, a_n), \qquad n = 1, 2, \cdots, k, \tag{58}$$

the likelihood can be written in the form

$$L(\sigma_1, \cdots, \sigma_k) = L(\sigma_1, \cdots, \sigma_{k-1}) + h(z_k; \sigma_k). \tag{59}$$

As is well known, the maximization of the function $L(\sigma_1, \cdots, \sigma_k)$ with respect to its arguments is amenable to solution via dynamic programming since (59) is satisfied. Since this is the case, the optimum receiver now assumes the structure shown in Fig. 6.

In summary, it has been shown that the ML receiver for the limiting case of small s/n has a structure that is asymptotically approximated by the receiver designed to detect a known signal in gaussian noise (with the inclusion of certain precomputed bias terms). We remark at this point that the application of the Viterbi algorithm is, of course, only productive when intersymbol interference is the dominant im-

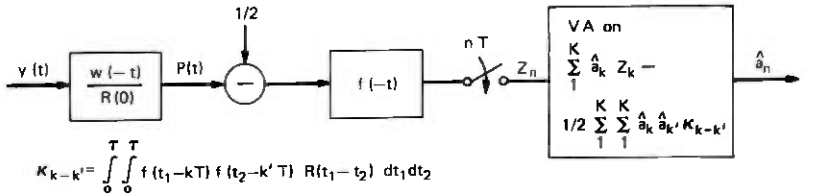


Fig. 6—Optimum detector (large noise) for PAM signaling.

pairment. In the context of the above discussion, this will be manifested in the values of $\mathcal{K}_n$ for $n \neq 0$. These values depend on the data rate relative to the channel dispersion. As in data transmission over voice-band channels, other methods of processing such as linear and decision feedback equalization should provide good results so long as the inter-symbol interference is not inordinately large. It is clear from (53) that when the distortion is small enough so that the quadratic term can be neglected, the optimization of the likelihood with respect to the data symbols can be carried out on a term-by-term or bit-by-bit basis. In other words, passing z_n through a slicer provides optimum detection. As the distortion becomes more severe, the quadratic term appearing in (53) must be retained. The linear receivers reported by Personick[5–7] and Messerschmitt[8,9] can be obtained from (53) by differentiating this expression with respect to the data symbols and then quantizing the result to the legitimate transmitted data levels. As the distortion increases still further, it becomes necessary to maximize (53), as it stands, via the Viterbi algorithm. Selecting one of these detectors in any given situation requires an evaluation of the error probability to quantify the effect of distortion on the system performance.

VI. PERFORMANCE ANALYSIS OF THE OPTIMUM DETECTOR FOR BINARY ONE-SHOT SIGNALING

6.1 An upper bound on the error rate (a simple example)

Having a description of the optimum detector structure for $\alpha^2 \to 0$, it is interesting to inquire how well it performs in certain signaling situations. Unfortunately, the M-ary mode of operation is extremely difficult to analyze, and even the general binary case poses insurmountable mathematical difficulties. We have, however, been able to analyze several special cases of interest that provide insight as to the effect of various system parameters on performance.

In the binary signaling case, information is conveyed by sending either intensity $\lambda_1(t)$ or $\lambda_2(t)$ with equal probability. From (51), the ML detector has the realization shown in Fig. 7. The detector, in this

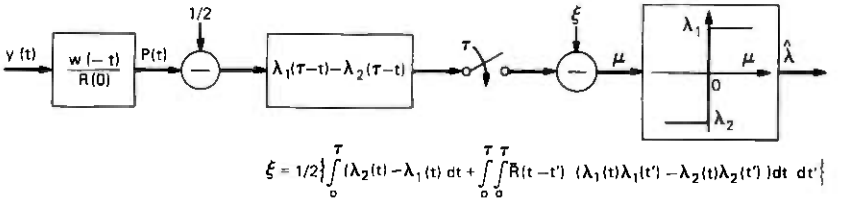


Fig. 7—Optimum detector for binary signaling ($\alpha^2 \to 0$).

situation, computes the statistic

$$\mu = \int_0^{\mathcal{T}} [\lambda_1(t) - \lambda_2(t)]P(t)dt - \frac{1}{2}\int_0^{\mathcal{T}} [\lambda_1(t) - \lambda_2(t)]dt - \frac{1}{2}\int_0^{\mathcal{T}}\int \bar{R}(t-\tau)[\lambda_1(t)\lambda_1(\tau) - \lambda_2(t)\lambda_2(\tau)]dtd\tau, \quad (60)$$

and μ is then compared to zero. When $\mu > 0$, it is decided that $\lambda_1(t)$ was sent, and when $\mu \leqq 0$, $\lambda_2(t)$ is chosen. In (60), the indicated quantities are normalized such that

$$P(t) = \frac{1}{R(0)}\int_{-\infty}^{\infty} y(\tau)w(t-\tau)d\tau$$

and $\bar{R} = R/R(0)$.

The probability of error is

$$P_e = \tfrac{1}{2}\Pr[\mu \geqq 0 | y(t) = I_1(t) + n(t), 0 \leqq t \leqq \mathcal{T}] + \tfrac{1}{2}\Pr[\mu < 0 | y(t) = I_2(t) + n(t), 0 \leqq t \leqq \mathcal{T}], \quad (61)$$

where

$$I_1(t) = \sum_1^{\nu} w(t - t_n) \qquad \text{with} \qquad E[\nu] = \int_0^t \lambda_1(\xi)d\xi,$$

and where

$$I_2(t) = \sum_1^{\nu} w(t - t_n) \qquad \text{with} \qquad E[\nu] = \int_0^t \lambda_2(\xi)d\xi.$$

It turns out that the evaluation of (61) is not mathematically tractable when λ_1 and λ_2 are arbitrary positive time functions. Even reasonable bounds on (61) are difficult to calculate in general. However, for constant intensities, exponentially tight upper bounds can be obtained. While the restriction to constant intensities might appear severe, it is shown in the appendix that in the absence of both dark current and gaussian noise the optimum choice of signals will have one intensity equal to zero while the other is arbitrary and need only satisfy a power constraint. Here we wish to illustrate a bounding approach for one special case where the upper bound can be obtained in closed form. We analyze the error rate for a system slightly modified from that depicted

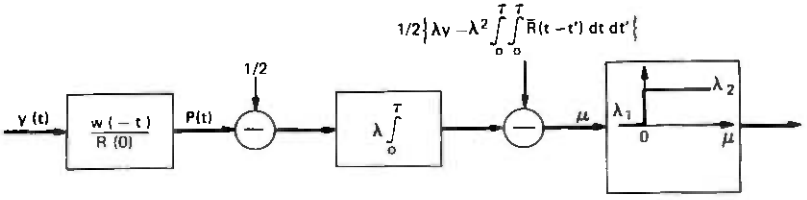


Fig. 8—Optimal detector for α^2 small and $\lambda_1 = 0$ and $\lambda_2 = \lambda$.

in Fig. 8 for $\lambda_1 = 0$ and $\lambda_2 = \lambda$. The modification will involve adjusting the threshold† so that our upper estimate of the probability of error when λ_1 is sent is equal to the estimate of probability of an error in the complementary situation.

In the binary system under consideration, the information symbol 1 is encoded into the intensity function $\lambda_1(t) = \lambda$, $0 \leqq t \leqq \mathcal{T}$ and the information symbol 0 into the intensity $\lambda_2(t) = 0$, $0 \leqq t \leqq \mathcal{T}$. Notice that the dark current is assumed to be zero. The detector structure we wish to analyze is depicted in Fig. 9. Here, the information-bearing Poisson process is passed through a matched filter $w(-t)/R(0)$, then integrated, and the result compared with a threshold set at F. If μ (refer to the block diagram) exceeds F, the symbol 1 is chosen and if $\mu \leqq F$, the symbol 0 is chosen. Our chief interest in this example is to exhibit the interplay between the various parameters in this extremely simple but informative situation.

As seen in the diagram,

$$\mu = \nu \int_0^{\mathcal{T}} R(t)dt + \int_0^{\mathcal{T}} \int_0^{\mathcal{T}} n(\tau)w(t - \tau)dtd\tau \tag{62}$$

or, equivalently, the test statistic may be written as

$$\mu_0 = \nu + n_0, \tag{63}$$

which is compared to a threshold. Note that μ_0 is just a scaled version of μ, and n_0 is a zero-mean gaussian random variable with

$$E\{n_0^2\} = N_0 \frac{\int_0^{\mathcal{T}} \int_0^{\mathcal{T}} R(t - \tau)dtd\tau}{\left[\int_0^{\mathcal{T}} R(t)dt\right]^2} \triangleq \sigma^2. \tag{64}$$

Observe that, in this situation, the receiver is just a counter since the test statistic represents the total number of photon counts observed in the entire observation interval plus an added gaussian random variable.

† By the threshold, we mean the bias terms appearing in (60), i.e., the last two terms.

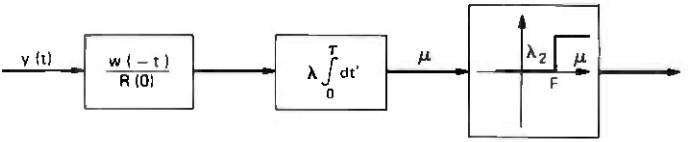


Fig. 9—Detector for α^2 small and $\lambda_1 = 0$, $\lambda_2 = \lambda$ with threshold modified so error probabilities are equal.

The integer random variable ν is Poisson-distributed with

$$E[\nu] = \lambda T \qquad \text{when} \qquad 1 \text{ is sent } (H_1)$$

and

$$E[\nu] = 0 \qquad \text{when} \qquad 0 \text{ is sent } (H_0), \tag{65}$$

where H_0 and H_1 are symbols distinguishing the two situations. The probability of error is then explicitly given by

$$P_e = \tfrac{1}{2} \Pr [\mu > F | H_0] + \tfrac{1}{2} \Pr [\mu \leqq F | H_1], \tag{66}$$

where we have made the assumption that 0s and 1s are transmitted with equal probability.

Since (66) cannot be expressed in closed form, we seek an exponentially tight upper bound. Applying the Chernoff bounding technique, we notice that the error rate under the null hypothesis, H_0, can be upper bounded immediately since under this hypothesis $\nu = 0$. Applying the bound yields

$$P_0 = \Pr [\mu > F | H_0] \leqq \exp \left\{ - \frac{F^2}{2\sigma^2} \right\}. \tag{67}$$

The second term in (66) can likewise be upper bounded since the moment generating function of ν under H_1 is known. This procedure gives

$$P_1 = \Pr [\mu \leqq F | H_1] \leqq \exp \left\{ \theta F + \frac{\theta^2 \sigma^2}{2} \right\} M_{\nu | H_1}(-\theta), \qquad \theta \geqq 0, \tag{68}$$

where

$$M_{\nu | H_1}(\theta) = E\{e^{\nu\theta} | H_1\} = \exp [\lambda T (e^{\theta} - 1)]$$

for $\theta \geqq 0$. The bound (68) then becomes

$$P_1 \leqq \exp \left\{ \theta F + \frac{\sigma^2 \theta^2}{2} + \lambda T (e^{-\theta} - 1) \right\}, \qquad \theta > 0, \tag{69}$$

and it is optimized by finding a θ^* such that

$$E(\theta^*, F) = \min_{\theta > 0} \left\{ \theta F + \frac{\sigma^2 \theta^2}{2} + \lambda T (e^{-\theta} - 1) \right\}. \tag{70}$$

To make the upper bounds on P_1 and P_2 equal, we select an $F = F_0$ such that

$$E(\theta^*, F_0) = \frac{F_0^2}{2\sigma^2}.$$

This, then, yields the final upper bound on the error rate

$$P_e \leqq \exp\,(-F_0^2/2\sigma^2). \tag{71}$$

By differentiating (70), we see that for a positive solution to exist it is required that $0 < F < \lambda\mathcal{T}$. Unfortunately, such a solution cannot be obtained in closed form. However, lower bounding $1 - e^{-\theta}$ by $\theta - \theta^2/2$, which in turn upper bounds (69), we find that

$$\theta^* = \frac{\lambda\mathcal{T} - F}{\sigma^2 + \lambda\mathcal{T}} > 0, \tag{72}$$

and consequently

$$P_1 \leqq \exp\left\{-\theta^*(\lambda\mathcal{T} - F) + \frac{\theta^{*2}}{2}(\lambda\mathcal{T} + \sigma^2)\right\}, \tag{73}$$

where θ^* has been chosen to provide the tightest bound.

Having θ^*, the threshold F_0 is obtained from

$$\frac{F_0^2}{\sigma^2} = \frac{(\lambda\mathcal{T} - F_0)^2}{\lambda\mathcal{T} + \sigma^2}.$$

Solving this quadratic equation and selecting the only reasonable root for F_0 give

$$F_0 = -\sigma^2 + \sqrt{\sigma^4 + \sigma^2\lambda\mathcal{T}}. \tag{74}$$

Substituting (74) into (72), the bound on the error rate finally becomes

$$P_e < \exp\left[-\frac{K}{2}\{\sqrt{1 + C} - \sqrt{C}\}^2\right], \tag{75}$$

where $K = \lambda\mathcal{T}$ and $C = \sigma^2/K = \dfrac{\text{Average Noise Power}}{\text{Average Shot Noise Power}}$.

It is instructive to express the bound (75) in the following alternative form

$$P_e \leqq e^{-\rho f(c)}, \tag{76}$$

where

$$\rho = \frac{1}{2}\frac{K^2}{K + \sigma^2} = \frac{\text{Average (signal)}^2}{\text{Average Total Noise Power}}$$

and where $f(c) = [1 + c - \sqrt{c^2 + c}]^2$.

As can be checked, $f(c)$ is a monotonically decreasing function of c,

and has the properties

$$\lim_{c \to 0} f(c) = 1$$

$$\lim_{c \to \infty} f(c) = \tfrac{1}{4}.$$

Thus, $P_e \leqq e^{-\rho f(c)} \to e^{-\rho}$, as $c \to 0$.

This is the situation that prevails when the shot noise dominates. On the other hand,

$$P_e \leqq e^{-\rho/4}, \qquad \text{as} \qquad c \to \infty,$$

which is the situation when the gaussian noise dominates.

6.2 Implications of the error bound

The first observation concerning (75) is that, as $C \to 0$, $P_e \leqq \exp\{-K/2\}$. This can be achieved by making $\sigma^2 \to 0$. This implies that either the gaussian noise is zero or that the number of counts is very large. However, in the absence of gaussian noise (as well as dark current), it is clear that the only way to make an error is when there are not any counts ($\nu = 0$) under H_1. The chance that $\nu = 0$ under H_1 is just $\exp\{-K\}$. In the absence of gaussian noise, this is clearly the very best performance one can hope for. Notice that the upper bound predicts an outcome which is 3-dB poorer than this ideal. The factor of 2 in the exponent of (75) is attributed to our bounding technique. What, in fact, happens as $\sigma^2 \to 0$ is that θ^* increases, and that the lower bound $\theta - \theta^2/2$ becomes loose, the upshot being the factor of 2 in the exponent. To see that this factor of 2 is indeed a quirk of the parabolic approximation to the exponential, consider the exponent in (69) as $\sigma^2 \to 0$. It is clear that the optimum threshold and θ are, respectively, zero and infinity, which when substituted in (69) does indeed give $e^{-\lambda \mathcal{T}}$ ($K = \lambda \mathcal{T}$).

Another aspect of the bound, however, is that ideal performance can be achieved with this detector structure (which is optimum for $\sigma^2 \to \infty$, the large gaussian noise situation) when the noise vanishes ($\sigma^2 \to 0$). This suggests that for the case of constant intensities the linear threshold detector is robust, i.e., it performs well over the entire range of σ^2 (or α^2).

We now use the error bound to determine the number of counts required, for reasonable operating physical parameters, to achieve a desired error rate. Note that, from (64), after a simplifying calculation on the double integral, we obtain

$$\sigma^2 = \frac{N_0}{A^2} 2 \left(\mathcal{T} A - \int_0^{\mathcal{T}} tR(t)dt \right) = \frac{2N_0}{A} \left[\mathcal{T} - \frac{\int_0^{\mathcal{T}} tR(t)dt}{\int_0^{\mathcal{T}} R(t)dt} \right], \quad (77)$$

where

$$A = \int_0^{\mathcal{T}} R(t)dt.$$

Introducing the pulse stretch factor,

$$S = \int_0^{\mathcal{T}} tR(t)dt \Big/ \int_0^{\mathcal{T}} R(t)dt < \mathcal{T}, \tag{78}$$

into (78) and recalling that $\alpha^2 = R(0)/N_0$ yields explicitly

$$\sigma^2 = \frac{2}{\alpha^2} \frac{1-r}{r} \cdot \frac{R(0)S}{A}, \tag{79}$$

where $0 < r = S/T < 1$. What, then, can be said about the choice of the parameter r? Can it be selected at will? Within a good approximation, $SR(0)/A \sim 1$. Clearly, the best choice of r appears to be unity, since $r = 1$ reduces the noise variance to zero. Recall, however, that, when the mathematical model was initially introduced, it was tacitly assumed that the observation interval was much larger than the width of the pulses emanating from the photodetector so that edge effects could be neglected. This alone would restrict the range of r to be no more than, say, 0.1, which would indicate that r does not appear to be an independent parameter. With $r = 0.1$, we may conclude from (79) that the effective gaussian variance of the scaled system is roughly

$$\sigma^2 = 20/\alpha^2. \tag{80}$$

Returning now to (75), we see that ideal performance is achieved when

$$C = \frac{\sigma^2}{K} \ll 1,$$

and when (80) is substituted in the above, we arrive at the condition that

$$\frac{20}{K\alpha^2} \ll 1 \rightarrow K\alpha^2 \gg 20. \tag{81}$$

As an example, let $\alpha^2 = 1/400$, which, according to S. Personick,† is a reasonable number for this parameter. This implies that $K \gg 8000$ is required to achieve ideal performance (i.e., the error rate in this range approaches zero like e^{-K}). On the other hand, suppose it is desired that $P_e \leqq 10^{-9}$. This would imply that

$$\frac{K}{2} \{\sqrt{(20/K\alpha^2) + 1} - \sqrt{20/K\alpha^2}\}^2 \sim 20.$$

† Private communication.

For instance, $\alpha^2 \sim 1/400$ implies that K is on the order of 1200. The above discussion quantifies the facts that to achieve good performance the total number of counts must be large or, if the background gaussian noise is small then fewer counts are needed to provide satisfactory performance.

6.3 Some conclusions concerning optimum detection for constant intensities

Note that the linear receiver, which is optimum when $\alpha^2 \to 0$, seems to be robust—at least for binary systems signaling with constant intensities. The optimum detector in the small s/n case ($\alpha^2 \to 0$) yields a decision variable based on the total number of observed counts as evidenced from (63). Of course, for the error probability bound to be tight, the average number of counts, K, must be large enough that $\sigma^2/K \ll 1$. On the other hand, we saw that the optimum detector structure in the case of large s/n ($\alpha^2 \to \infty$) combined with narrow pulses† ($r \ll 1$) is also a counter. The only difference is that the counts in the $\alpha^2 \to 0$ detector are linearly corrupted by gaussian noise, while the counts in the $\alpha^2 \to \infty$ detector are determined by quantizing the incoming signal to the nearest integer in the presence of the added gaussian noise. The latter operation is, of course, nonlinear. Nevertheless, when the added noise is small ($\alpha^2 \to \infty$), the two operations are approximately equivalent, thus explaining the robustness of the linear receiver and the results of our theory.

VII. LARGE SIGNAL-TO-NOISE RATIO ($\alpha^2 \to \infty$) AND NARROW PULSES

When a photomultiplier or avalanche photodiode is used to provide direct detection, the parameter α^2 is much larger than unity. In this application, the response of the photodetector to a single electron or hole is much larger than the background gaussian noise. In this situation, intuition dictates that the detector make use of the "estimated" arrival times of the individual photons. Here we discuss a special case that will bring out the essential structure of the optimum detector. The situation examined is when there is no avalanche gain and the individual pulses $w(t)$ are time-limited to an interval much smaller than the observation interval. The more general situation is treated in Section VIII.

The approach taken in this section is to use the gaussian process formulation (45) and attempt to approximate the indicated expectation with respect to the $x(t)$ process. For this approach to be productive, we must assume that $R(t)$ has effective duration Δ. We may then

† This was demonstrated in the examples of Section III and is reestablished in Section VII.

approximate the integral appearing in (45) by a discrete sum, i.e.,

$$\int_0^T dt \exp [\alpha P(t) + i\alpha x(t)]\lambda_m(t) \to \Delta \sum_{j=1}^{J} \exp (\alpha^2 P_j + i\alpha x_j)\lambda_m(j\Delta), \quad (82)$$

where $P_j = P(j\Delta)$ and $x_j = x(j\Delta)$.

The implication of (82) can be viewed in several ways. Of course, as $\Delta \to 0$ and $J \to \infty$, irrespective of the correlation function $R(t)$, the discrete sum is an excellent approximation to the integral. But sampling the integrand at the rate $1/\Delta$ does not necessarily guarantee that the sum is a good approximation to the integral. Yet to derive any utility from representation (45), we must sample at a rate $1/\Delta$ so that the sequence of random variables $\{x_j\}$ can be regarded as identically and independently distributed. Unfortunately, this is the only case for which we can compute the indicated averages in a useful form. What then do we mean by (82)? To make sense of this representation, we must reinterpret the distribution of the arrival times, $\{t_n\}$. Evidently, the reason we have an integral representation instead of a sum is because we have assumed that the arrival times obey a continuous distribution. However, if we assume at the outset that the arrival times $\{t_n\}$ can occur only at a set of discrete points $\{t_n = n\Delta\}$, then (45) will contain a sum instead of an integral. This procedure is equivalent to that used to obtain (41) as the discrete version of (40). Hence, a rigorous interpretation of (45) is that the Poisson arrival times can only occur at the discrete instants of time $\{j\Delta\}$, $j = 0, 1, 2, \cdots$. If we now assume that the quantization of the arrival times to units of Δ is such that $R(\Delta) \sim 0$, then the set of random variables $\{x_j\}_{j=1}^{J}$ are mutually independent. Exploring this line of reasoning, (45) can be written as

$$e^{\Delta}L(y) = \prod_{j=1}^{J} E_x[\exp \{\Delta\lambda_j \exp (\alpha^2 P_j + i\alpha x_j)\}], \quad (83)$$

where $\lambda_j = \lambda(j\Delta)$, and we have suppressed the index m denoting the particular hypothesis being tested.

Expanding (83) in a power series and carrying out the indicated expectation give

$$e^{\Delta}L(y) = \prod_{j=1}^{J} \left(\sum_{n=0}^{\infty} \frac{\Delta^n(\lambda_j)^n}{n!} \exp \left\{\alpha^2 \left(nP_j - \frac{n^2}{2}\right)\right\}\right). \quad (84)$$

We are now in a position to exploit the assumed large value of α^2. In other words, we are interested in determining the behavior of (84) as $\alpha^2 \to \infty$. Towards this goal, consider the sum

$$S_j = \sum_{n=0}^{\infty} \frac{\Delta^n(\lambda_j)^n}{n!} \exp \{\alpha^2[nP_j - n^2/2]\}. \quad (85)$$

This series is in the form

$$\sum_{0}^{\infty} \beta_n \exp(\alpha^2 \gamma_n), \tag{86}$$

where β_n and γ_n have the obvious identifications.

Let $\bar{\gamma}_j$ be the largest of the γ_n and β_j be the corresponding value of β_n. Then (86) becomes

$$\bar{\beta}_j \exp(\alpha^2 \bar{\gamma}_j) \left(\sum_{n=0}^{\infty} \frac{\beta_n}{\bar{\beta}_j} \exp[\alpha^2(\gamma_n - \bar{\gamma}_j)] \right), \tag{87}$$

where each $\gamma_n - \bar{\gamma}_j$ is negative. Since (86) converges absolutely, the infinite sum can be rearranged in such a way that the exponents are decreasing; thus, the rearranged sum is recognized to be a Dirichlet series[18] in the parameter α^2. From the elementary properties of such series, we deduce that, except for the $n = j$ term, the summation portion of (87) converges to zero as $\alpha^2 \to \infty$. So, as $\alpha^2 \to \infty$, (87) behaves like

$$S_j \sim \bar{\beta}_j \exp(\alpha^2 \bar{\gamma}_j). \tag{88}$$

Now returning to the series in (85), we let n_j denote the strictly nonnegative integer attaining

$$\max_{n \in \{0,1,2,\cdots\}} \left\{ \alpha^2 n P_j - \frac{n^2}{2} \alpha^2 \right\}, \tag{89}$$

i.e., $n_j = [P]$, where $[P]$ denotes the nonnegative integer nearest to P. The corresponding coefficient becomes

$$\bar{\beta}_j = \frac{\Delta^{n_j}(\lambda_j)^{n_j}}{n_j!}. \tag{90}$$

Thus, as $\alpha^2 \to \infty$,

$$\bar{L}(y) = \log L(y) = -\Lambda + \log \times \left\{ \sum_{j=1}^{J} \left(\frac{\Delta^{n_j}(\lambda_j)^{n_j}}{n_j!} \exp[\alpha^2(n_j P_j - n_j^2/2)] \right) \right\}. \tag{91}$$

Discarding the hypothesis-insensitive terms, (91) can be rewritten in the form

$$\bar{L}_m(y) \sim -\Lambda_m + \sum_{j=1}^{J} n_j \log \lambda_j^{(m)}, \tag{92}$$

where we recall that $n_j = 0$ whenever $P_j < \frac{1}{2}$. Note that (92) is similar to the detector described by (56); however, the different statistical model (Bernoulli as opposed to Poisson occurrences) accounts for the bias term $-\Lambda_m$ appearing in (92).

The detector structure exhibited in (92) has a simple interpretation and is similar to that depicted in Fig. 3. As shown in Fig. 10, the in-

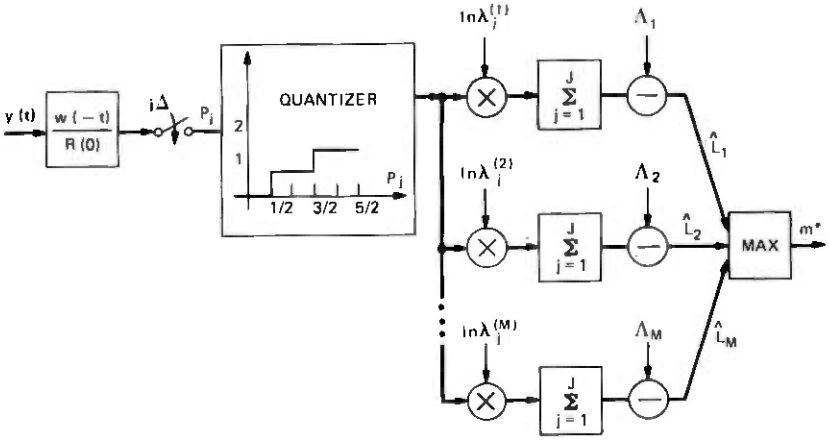


Fig. 10—Weighted counter with quantizer.

coming signal $y(t)$, having been filtered by $w(-t)/R(0)$, is sampled every Δ seconds. This is followed by quantizing the samples, P_j, to the nearest positive integer (including zero whenever $P_j < \frac{1}{2}$). The quantized samples are multiplied by the locally stored numbers $\log \lambda_j^{(m)}$ and the results summed. The sum is added to Λ_m to form the decision statistic. Since the added gaussian noise is assumed to be small and the pulse $w(t)$ is assumed to be narrow, most of the time the nearest integer at any time $t_j = j\Delta$ will be either 1 or 0, depending on whether $P_j > \frac{1}{2}$ or $P_j \leq \frac{1}{2}$, i.e., whether the receiver determines a pulse is present or absent. Consequently, the optimum detector structure may be viewed as a weighted counter, where the decision as to which intensity was transmitted is based on selecting the largest of the weighted pulse counts.

We recognize that from an implementation point of view even this seemingly simple structure may pose practical difficulties. The indicated sampling may be difficult to carry out at this high frequency. While this is indicated mathematically, in practice the peaks of the signal at the photodetector output could be used to approximate the photon arrival times and, hence, the interrogation times.

VIII. MAXIMUM LIKELIHOOD RECEIVER FOR LARGE SIGNAL-TO-NOISE RATIO ($\alpha^2 \to \infty$)

This section extends the results of the last section by indicating a general approach to the extremely complex problem of performing optimum detection when the pulses $w(t)$ are not restricted in width or shape and when avalanche gain is provided. In the presence of avalanche gain, the average signal-to-noise ratio, α^2, is large. This implies

that the photon arrival times can be accurately estimated, and these estimates can then be used to aid the detector in making accurate decisions. One objective of this section is to indicate how the optimum detector estimates the arrival times. Heuristically, the receiver attempts to "whiten" or peak up the pulse $w(t)$. The presence of gaussian noise, however small, prevents pulse whitening via linear filtering. The nonlinear manner in which the receiver estimates the arrival times is of independent interest and will be presented in the sequel.

We begin with the most general form of the likelihood function (40). While the infinite functional series appearing in (40) is quite intimidating, it has already been shown to reduce to physically interpretable receivers in the following special cases: (*i*) small signal-to-noise ratio ($\alpha^2 \to 0$) and (*ii*) large signal-to-noise ratio ($\alpha^2 \to \infty$) combined with an extremely small decorrelation time† for $R(t)$.

Since large α^2 is a practical operating condition (photomultiplier and the avalanche photodiode), we are motivated to examine the salient features of the optimal processor under these circumstances. We also specialize our development to the PAM-Poisson intensity, or data signal,

$$\lambda_m(t) = \lambda_0 + \sum_{n=0}^{N} a_n^{(m)} f(t - nT), \qquad 0 < t \leqq \mathcal{T},$$

where $f(t)$ is a known pulse shape determined by the distortion (intersymbol interference) in the optical fiber and λ_0 is again the ambient or "dark" current. Here, the optimum receiver maximizes the likelihood function with respect to the data sequence $\{a_n^{(m)}\}_{n=0}^{N}$. As it stands, the likelihood (40) is similar‡ in form to the Volterra kernel description of a general time-varying nonlinear functional. However, such generality seems to preclude any practical value, and furthermore reveals little of the receiver's essence. To obtain a good approximation to the structure of the receiver when $\alpha^2 \to \infty$, it will again be necessary to discretize the photon arrival times.

8.1 The asymptotic ($\alpha \to \infty$) likelihood function

In this section, the basic idea is to asymptotically evaluate the multidimensional sums or integrals. Note that, when $\alpha^2 \to \infty$, the $2n$-fold integrals appearing in the likelihood become increasingly sensitive to the value of the exponent, and in the limit the integral is com-

† Note that, as $R(t) \to \delta(t)$, the gaussian noise becomes transparent to the receiver (since the integrated received signal would be discontinuous whenever an impulse arrived). The receiver then assumes the form of a counter.

‡ The difference is that, in our application, the input $P(t)$ is exponentiated rather than appearing directly.

pletely determined by the coordinates that maximize the exponent. This statement is made precise by the multidimensional version of Laplace's theorem[19] which, apart from certain hypothesis-insensitive terms, gives for each n

$$\lim_{\alpha\to\infty} \int_0^{\mathcal{T}} dt_n \sum_{\mathbf{g}_n} \prod_{j=1}^{n} \rho(g_i) \prod_{j=1}^{n} \lambda(t_j) \times \exp\left\{\alpha^2\left[\sum_{m=1}^{n} g_m P(t_m) - \frac{1}{2}\sum_{m,k=1}^{n} g_m g_k R(t_m - t_k)\right]\right\} \sim \prod_{j=1}^{n} \rho(g_j^*)\lambda(t_j^*) \times \exp\left\{\alpha^2\left[\sum_{m=1}^{n} g_m^* P(t_m^*) - \frac{1}{2}\sum_{m,k=1}^{n} g_m^* g_k^* R(t_m^* - t_k^*)\right]\right\}, \quad (93)^\dagger$$

where $\{t_1^*, t_2^*, \cdots, t_n^*\}$ and $\{g_1^*, g_2^*, \cdots, g_n^*\}$ maximize the exponent,

$$\sum_{m=1}^{n} g_m P(t_m) - \frac{1}{2}\sum_{m,k=1}^{n} g_m g_k R(t_m - t_k),$$

under the constraint that $0 \leqq t_i < \mathcal{T}$, $i = 1, 2, \cdots, n$. The determination of the extremizing sets appears very difficult. For example, without avalanche gain (i.e., $g_m = 1$) and $n = 1$, it is clear that t_1^* is taken at the point where the observable $P(t)$ is a maximum. For example, when $n = 2$ the exponent becomes

$$P(t_1) + P(t_2) - R(t_1 - t_2),$$

and the choice of t_1 and t_2 is not apparent. The values of t_1^* and t_2^* tend to be near the peaks of $P(t)$, but this is not always the case.[‡] The best choice of t_1 and t_2 will, of course, depend on the interaction of the random process $P(t)$ and the correlation function $R(t)$. The problem of finding the set of points $\{t_1^*\}$ is in some sense equivalent to whitening or peaking up the pulse $w(t)$ in a nonlinear manner to minimize the noise enhancement concomitant with such an operation. Putting aside for the moment the difficulty of determining $\{t_1^*, \cdots, t_n^*\}$ and $\{g_1^*, \cdots, g_n^*\}$, we can use these values to rewrite the right-hand side of (93) as

$$\left[\prod_{j=1}^{n} \rho(g_j^*)\lambda(t_j^*)\right] \exp\left[\alpha^2\left\{\sum_{m=1}^{n} g_m^* P(t_m^*) - \frac{1}{2}\sum_{m,k=1}^{n} g_m^* g_k^* R(t_m^* - t_k^*)\right\}\right] \triangleq \gamma_n(\mathcal{T}) \exp\left[\alpha^2 \beta_n(\mathcal{T})\right], \quad (94)$$

---

† It has been assumed that there is only one set of variables $\mathbf{t} = (t_1, \cdots, t_n)$ and $\mathbf{g} = (g_1, \cdots, g_n)$, which maximize the exponent. If there are several such $\mathbf{t}^*$ and $\mathbf{g}^*$, then the right-hand side of (93) would consist of a sum of these terms. We do not pursue this approach, since the resulting structure is hopelessly complicated and appears to be impractical.

‡ This would be the case whenever $P(t)$ has equivalued maxima spaced at least a decorrelation time apart.

where we have indicated the dependence of both the coefficient and the exponent on the observation interval $\mathcal{T}$. Using (94) in (46) gives the Dirichlet series[21]

$$e^{\Lambda}L \sim \sum_{n=0}^{\infty} \gamma_n(\mathcal{T}) \cdot \exp\left[\alpha^2 \beta_n(\mathcal{T})\right]. \tag{95}$$

As $\alpha \to \infty$, it is well known that the Dirichlet series is dominated by the term with the largest exponent, i.e.,

$$\lim_{\alpha \to \infty} e^{\Lambda}L \sim \gamma_{n^*}(\mathcal{T}) \exp\left\{\alpha^2 \beta_{n^*}(\mathcal{T})\right\}, \tag{96}$$

where β_{n^*} is the largest exponent.† It is evident that n^* is an estimate of the number of (Poisson) events occurring in the interval $\mathcal{T}$ and that $t_1^*, \cdots, t_n^*$ are estimates of these occurrence times, while $g_1^*, \cdots, g_n^*$ are estimates of the avalanche gains. This is not surprising since, as $\alpha^2 \to \infty$, the vanishingly small noise implies that these estimates will be quite accurate. Hence, the receiver is intimately related to the situation considered by Bar-David,[3] where the Poisson events can be observed directly. The distinction is that estimated arrival times and avalanche gains are used rather than their true values. It is important to realize that specific estimators have been obtained for the random parameters. As we show in the sequel, the simultaneous estimation and detection described above can be recursively implemented via dynamic programming.

Since neither the exponent in (94) nor the $\prod_{j=1}^{n^*} \rho(g_i^*)$ term is hypothesis-sensitive, the relevant portion of the likelihood function is

$$L \sim e^{-\Lambda}\gamma_{n^*}(N) = e^{-\Lambda} \prod_{j=1}^{n^*} \lambda(t_j^*), \tag{97}$$

where n^* is the number of time points that maximize the exponent of (92) (which, of course, depends on $\mathcal{T}$) and $\{t_i^*\}_{i=1}^{n}$ are the values of these time points. Note that, once the exponent is jointly optimized with respect to $\mathbf{t}_n$ and $\mathbf{g}_n$, the estimate of the avalanche gain is not utilized further. This is so because the avalanche gain is a property of the photodetector and conveys no information concerning the intensity function. The asymptotic ($\alpha \to \infty$) likelihood given by (97) is exactly Bar-David's[3] likelihood formula, with the true arrival

† If the signal-to-noise ratio is not large enough so that this is not an accurate approximation, then one could designate n_* as the second largest exponent, thereby developing the more accurate series

$$L \sim \exp(-\Lambda)\gamma_{n^*} \exp(\beta_{n^*}) \left(1 + \frac{\gamma_{n_*} \exp(\beta_{n_*})}{\gamma_{n^*} \exp(\beta_{n^*})}\right).$$

times replaced by estimated arrival times. Note that the log-likelihood

$$\bar{L}_m = -\Lambda_m(\mathcal{T}) + \sum_{i=1}^{n^*} \log \lambda^{(m)}(t_i^*) \tag{98}$$

is again a weighted counter, and is similar to (98) derived in Section VII [where the pulses $w(t)$ were assumed to be narrow].

Two shortcomings are associated with the above approach, one is computational and the other involves a question of mathematical rigor. The first point is that implicit in the expression for the likelihood (97) is the ability to solve the formidable mathematical problem,

$$\max_{\substack{g_m, t_m, \text{ and } n \\ 0 \leq t_i \leq \mathcal{T},\ i=1,2,\cdots,n}} \left\{ \sum_{m=1}^{n} g_m P(t_m) - \frac{1}{2} \sum_{m,k=1}^{n} g_m g_k R(t_m - t_k) \right\}, \tag{99}$$

in real time. We are not aware of optimization techniques capable of this accomplishment. The second point involves the invocation of the large α^2 assumption in a *sequence* of operations. Recall that this assumption was used to derive (93) and then used again to obtain (96). While the validity of the preceding operations can perhaps be demonstrated (under suitable conditions), the intractable nature of (99) forces us to slightly reformulate our problem.

8.2 The optimum detector when the photon arrival times are discrete

To proceed further and obtain a physically realizable, as well as meaningful, detector, we discretize the photon arrival times. Adopting this approach, the photon arrival times are now constrained to occur at the discrete instants $j\Delta$, $(j = 0, 1, 2, \cdots, J$, where $J = \mathcal{T}/\Delta)$. This gives rise to the discrete likelihood function (41), and eq. (98) then involves only sums rather than integrals. This modified expression contains a $2n + 1$ dimensional sum, which is recognized as a *bona fide* Dirichlet series. Thus, we have avoided the mathematical question concerning the validity of an asymptotic expansion by introducing a mild relaxation of the physical set-up.

Applying the asymptotic condition to the $2n + 1$ variable summation again produces the expressions (94) to (99) where it is recognized that the variables $\{t_i\}$ are now constrained to lie on the lattice, i.e., $t_i = j_i\Delta$, where $j_i = 1, 2, \cdots, J$. We now show that, using this discrete framework, the exponent appearing in (94) can be rewritten in a form readily amenable to maximization. Note that the variables $t_1^*, t_2^*, \cdots, t_n^*$ may be thought of either as specifying a single point in n-dimensional space or as specifying n points on the interval $(0, \mathcal{T})$. This latter viewpoint turns out to be more useful.

We choose the time quantization Δ so that the probability of more than one photon arrival occurring in a time interval of size Δ is vanishingly small† under each hypothesis $\lambda_m(t)$. In this framework, the set of time points $\{t_k^*\}$ specifies n points in the interval $(0, T)$, and the exponent can be rewritten as

$$\sum_{m=1}^{n} g_m P(t_m^*) - \frac{1}{2}\sum_{m,k=1}^{n} g_m g_k R(t_m - t_k)$$

$$= \sum_{m=1}^{J} g_m q_m P(m\Delta) - \frac{1}{2}\sum_{m,k=1}^{J} g_m g_k q_m q_k R(m\Delta - k\Delta), \quad (100)$$

where $0 \leq t_m^* \leq J\Delta$ and where q_m is 0 or 1. A value of $q_m = 1$ implies that the time point $m\Delta$ is "active" in the sums appearing in (100), while $q_m = 0$ implies that it is not. If one chooses Δ to provide a coarser quantization of the time axis, as might be required by practical restrictions on the sampling rate, then it is necessary to allow q_m to assume more (integer) values than 0 and 1. To see why this must be the case, recall the physical meaning‡ of the time points $\{t_i^*\}$. It is then realistic to expect that more than one photon will have arrived in a Δ interval and consequently some $t_i^* = t_j^*$ (for $i \neq j$). The increased range of q_m is necessary to accommodate this situation. Realizing that no restriction is implied, for reasons of simplicity we assume in the sequel that Δ is chosen small enough so that $q_m = 0$ or 1. At this point, it is clear that the product $g_m q_m$ is inseparable in the optimization of (100). Note that, once the optimum values of q_m and g_m are determined, only the value of q_m plays a further role in the detection procedure. With this in mind, we let $\beta_m = q_m g_m$, where β_m will range over the allowable values of g_m as well as zero. For convenience, we call this discrete set B. In the context of this new notation, the optimization problem posed in (99) becomes

$$\max_{\substack{\beta_1,\cdots,\beta_J \\ \beta \in B}} \sum_{m=1}^{J} \beta_m P(m\Delta) - \frac{1}{2}\sum_{m,k=1}^{J} \beta_m \beta_k R(m\Delta - k\Delta), \quad (101)$$

where it is important to realize that the maximization with respect to n, appearing in (99), has been removed in (101) by eliminating the restriction that only a predetermined number of q_n's be nonzero. It is also apparent that the exponent is of the required recursive form so that the exponent can be maximized via the Viterbi algorithm. With

† This probability is $1 - e^{-\lambda} - \lambda e^{-\lambda} \approx \lambda^2$.
‡ The $\{t_j^*\}$ are estimates of the pulse arrival times.

this in mind, the likelihood function can now be written as

$$L \sim e^{-\Lambda} \prod_{j=1}^{J} [\lambda(j\Delta)]^{q_j}, \tag{102}$$

and the log-likelihood again assumes the weighted-counter form

$$\bar{L} = -\int_0^T \lambda(t)dt + \sum_{j=1}^{J} q_j \log [\lambda(j\Delta)], \tag{103}$$

which is similar to the detector described by (92) but without the restriction on the correlation function $R(t)$, i.e., $R(t)$ need not be confined to an interval Δ. The result embodied in (92) for nonoverlapping pulses can be easily derived from (101) by setting $R(m\Delta - k\Delta) = \delta_{k-m}$. The exponent then becomes $\sum_{k=1}^{J}[\beta_k P(k\Delta) - \frac{1}{2}\beta_k^2]$, which is optimized, over the integer values of β_k, by choosing β_k to be the quantized version of $P(k\Delta)$.

The structure of the optimum detector (103) is shown in Fig. 11, and is of the estimator-detector type. The arrival time indicators $\{q\}_{i=1}^{J}$ (as well as the avalanche gains) are determined by applying the Viterbi algorithm to the exponent. Once these values are available, the likelihood is computed for each hypothesis $\lambda^{(m)}(t)$ and the maximum is selected.

8.3 Optimum detection of PAM intensities

The above methodology is now applied to the optimum detection of a digital (PAM) data signal. The 2^{N+1} intensity functions in this situation are given by

$$\lambda(t) = \lambda_0 + \sum_{n=0}^{N} a_n f(t - nT), \qquad 0 \leqq t \leqq T,$$

where the effect of optical channel distortion (intersymbol interference) is included in $f(t)$.

To optimally detect these signals, it is convenient to rewrite the original likelihood expression so that time is directly expressed in units of Δ. Bringing out this dependence, the likelihood function then becomes

$$L_J \sim \exp\left\{-\int_0^{J\Delta} \lambda(t)dt\right\} \prod_{j=1}^{J} [\lambda(j\Delta)]^{q_j}, \tag{104}$$

where the index J designates time in units of Δ. Note that the exponent (101) is already expressed in this form.

It is important to emphasize that a simultaneous or *two-tier* real-time sequential optimization procedure is required to extract the ML estimate of the data sequence, $\{a_n\}_{n=0}^{N}$. The exponent is first maximized

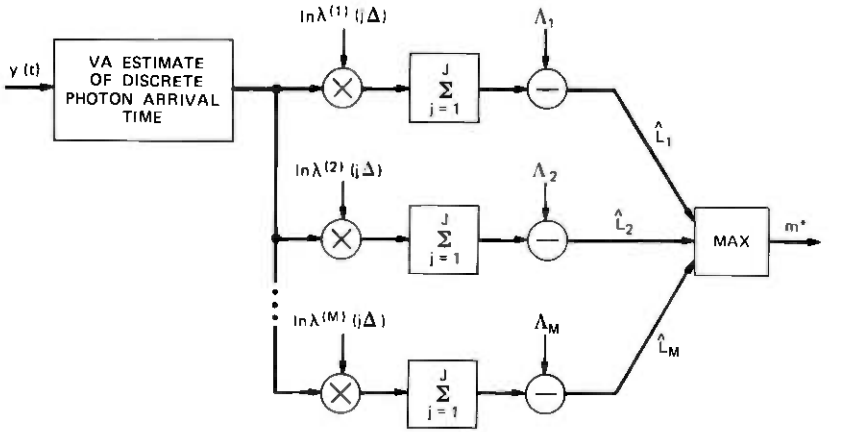


Fig. 11—Estimator detector type of weighted counter.

with respect to the $\{\beta_i\}_{j=0}^{J}$, and the corresponding q_i values are then used to maximize (104) with respect to the data symbols. The optimization of the exponent is identical to that occurring in ML data sequence estimation in the presence of intersymbol interference.[19] The maximization of the exponent will, at random intervals,† produce optimum values of $\{q_j\}$, say, $\hat{q}_0, \hat{q}_1, \cdots, \hat{q}_k$. At this instant, the optimization of the likelihood L_k can then proceed using this new information. At some later instant, $\hat{q}_{k+1}, \hat{q}_{k+2}, \cdots, \hat{q}_{k+n}$ will become available and attention again shifts to maximizing the likelihood L_{k+n}. As we shall show, the dynamic programming algorithm which maximizes the coefficient (103) is quite different from the conventional Viterbi algorithm. In fact, the application of dynamic programming to the iterative‡ maximization of this function illustrates the more general principle that dynamic programming is applicable to the iterative real-time ML sequence estimation of digital data that has undergone a wide variety of nonlinear distortion. The only requirements are that (*i*) the likelihood possesses the mathematical property of additivity and (*ii*) the nonlinearity is of finite memory so that the notion of a "state" is meaningful. In this application, both these requirements are satisfied.

To apply dynamic programming to the optimization problem exhibited in (103), we need only show that the likelihood satisfy a par-

† Owing to the merge aspect of the Viterbi algorithm.

‡ The two main virtues of dynamic programming are that (*i*) it is essentially a real-time processing scheme (although there is random signal-processing delay) and (*ii*) the number of computations is linearly proportional to time (n), as opposed to a straightforward evaluation that requires an exponentially growing number of computations.

ticular recursive form. To put the likelihood in this recursive form, we define the state vector

$$\mathbf{S}_j = (a_{j-1-\eta}, \cdots, a_j) \qquad j = \bar{f}, \bar{f}+1, \cdots, J, \tag{105}$$

where $\bar{f}$ is the memory (in units of Δ) of the dispersed pulse $f(t)$, i.e.,

$$f(n\Delta) = 0, \qquad n > \bar{f}, \tag{106}$$

and where η is the closest integer to $\bar{f}\Delta/T$.

As the optimum $\{\hat{q}_j\}$ time instants emerge from the Viterbi algorithm in a random manner (owing to the merge mechanism), they are classified according to which time segment $(0, NT)$ they belong. Once optimum time instants begin appearing that are active in the $(N+1)T$ time segment, those optimum q_n's which are in the NT time segment are available to maximize the coefficient or, equivalently, the likelihood.

By substituting the PAM signal into (104), the log-likelihood has the form

$$L_J = -\sum_{n=0}^{N} a_n F_n + \sum_{j=0}^{J} q_j \log\left(\lambda_0 + \sum_{m=0}^{N} a_m f(j\Delta - mT)\right), \tag{107}$$

where J is now interpreted as the index of the latest† merge in the Viterbi algorithm associated with the time interval $(0, NT)$ and

$$F_n = \int_0^{J\Delta} f(t - nT)dt. \tag{108}$$

It is important to keep in mind the fact that, once the decisions $(\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_J)$ are available, the iterative procedure for maximizing the likelihood proceeds in units of T. The log-likelihood can be put in the required form by letting $D = T/\Delta$ and writing the likelihood as

$$L_N = -\sum_{n=0}^{N-1} a_n F_n + \sum_{j=0}^{ND-D} q_j \log\left(\lambda_0 + \sum_{m=N-j-\bar{f}/D}^{N-1} a_m f(j\Delta - mD\Delta)\right)$$
$$+ a_N F_N + \sum_{j=ND-D+1}^{ND} q_j \log\left(\lambda_0 + \sum_{m=N-j-\bar{f}/D}^{N} a_m f(j\Delta - mD\Delta)\right). \tag{109}$$

It is crucial to realize that the last term in (109) only involves $a_{j-1-\eta}$, $a_{j-1-\eta+1}, \cdots, a_j$; therefore, with the state vector defined by (105), (109) can be written as

$$L_N = L_{N-1} + h(\mathbf{q}_N; \mathbf{S}_N), \tag{110}$$

where

$$\mathbf{q}_N = (q_{ND-D+1}, \cdots, q_{ND}). \tag{111}$$

† In other words, the next segment of optimum q_n's will penetrate beyond the time instant NT.

Fig. 12—Two-tier dynamic programming algorithm.

It is well known that, through the use of the recursion (110), dynamic programming may be applied to the maximization of L_N.

The resulting receiver is depicted in Fig. 12, and is a two-tier dynamic programming algorithm that simultaneously iterates the exponent and the coefficient to obtain a sequential (or real-time) maximum likelihood sequence estimate of the transmitted sequence $\{a_n\}$. While the above detector requires sampling at a rate that could preclude practical implementation, we remark that, in the large α^2 environment, a peak detector could be used to estimate the photon arrival times. These estimated arrival times would then be used in a dynamic programming algorithm to mitigate the effect of intersymbol interference.

IX. DISCUSSION

The communication-theoretic model for the fiber-optic communication system has proven to be quite useful. Using this model, the optimum (maximum-likelihood) receiver was exhibited under a wide variety of physical circumstances for M-ary and digital PAM signaling. Whether or not the energy in the response of the photodetector to an individual photon is large or small compared to the background gaussian noise, the detector structure turned out to be a weighted counter. The details of how the weighting is carried out have been shown to be complex in some cases. Further investigation into system performance is needed before assessing whether or not such complexity is warranted in any particular application. For values of pulse energy-to-noise ratio (α^2) much less than unity, the structure of the optimum detector can be simply instrumented in terms of analog operations on the photodetector output. On the other hand, when $\alpha^2 \gg 1$, and with or without avalanche gain, we have been unable to realize the optimum detector without first sampling the photo-detector output many times per symbol interval. This procedure may impose practical limitations on the implementation. Since the digital operations are required solely to estimate the photon arrival times, it has been pointed out that certain suboptimum operations (such as peak detection) may be used to estimate these instants. The power of maximum likelihood processing can still be used to mitigate the effect of intersymbol interference.

From a communications and information theoretic point of view, there remain many important and, as yet untouched, problems asso-

ciated with the fiber-optic channel. Sharp bounds on the performance of the various detectors are extremely difficult to obtain, and very little can be said at this time. Also, questions concerned with capacity, reliability, and complexity need be addressed.

X. ACKNOWLEDGMENT

The authors are grateful to S. D. Personick for patiently explaining to them the important aspects of the communication-theoretic model of the fiber-optic channel. We would also like to express our appreciation to D. G. Messerschmidt for his helpful participation in several informative discussions on this subject.

APPENDIX

Optimum Binary Intensities in the Absence of Gaussian Noise

In this appendix, we determine the optimum binary intensities $\lambda_1(t)$ and $\lambda_2(t)$ in the absence of gaussian noise. We proceed initially by neglecting the dark current. Of course, the optimum intensities must satisfy an energy constraint[†]

$$\int_0^{\mathcal{T}} [\lambda_1(t) + \lambda_2(t)]dt = P. \tag{112}$$

Consider the performance of a system that uses the equiprobable intensities

$$\begin{aligned} \lambda_1(t) &= \frac{P}{\mathcal{T}} \\ \lambda_2(t) &= 0 \end{aligned} \qquad 0 \leqq t \leqq \mathcal{T}. \tag{113}$$

The only way an error can be made under (113) is when $\lambda_1(t)$ is transmitted and no photons arrive; the probability of this event is

$$P_I = \tfrac{1}{2}e^{-P}. \tag{114}$$

Consider now the performance of a system that uses the arbitrary and equiprobable intensities $\lambda_1(t)$ and $\lambda_2(t)$. The probability of error for this system is

$$P_{II} = \tfrac{1}{2}P_1 + \tfrac{1}{2}P_2, \tag{115}$$

where P_1 and P_2 denote the conditional error probabilities given that $\lambda_1(t)$ and $\lambda_2(t)$ are active. Let

$$\Lambda_i = \int_0^{\mathcal{T}} \lambda_i(t)dt, \qquad i = 1, 2, \tag{116}$$

[†] Since the intensity is proportional to the transmitted optical energy, the constraint is on the average energy.

and let Λ_1 be greater than Λ_2. It is clear that, when Λ_1 is transmitted, the optimum detector must make an error when there are no photon arrivals. These observations provide the following sequence of lower bounds

$$P_{II} \geqq \tfrac{1}{2}P_1 \geqq \tfrac{1}{2}e^{-\Lambda_1}, \tag{117}$$

and since $\Lambda_1 + \Lambda_2 = P$ we have

$$P_{II} \geqq \tfrac{1}{2}e^{-\Lambda_1} \geqq \tfrac{1}{2}e^{-P} = P_I. \tag{118}$$

It is thus established that the intensities described by (113) minimize the probability of error and therefore are optimum. It is also clear that any system that has one of the intensities equal to zero, and the other arbitrary (and satisfying the power constraint), will perform equally as well as (113).

The effect of dark current on the probability of error can be made arbitrarily small by choosing $\lambda_2(t) = 0$ and picking $\lambda_1(t)$ so that the set of points where $\lambda_1(t)$ is nonzero is sufficiently small.

REFERENCES

1. Special Issue on Optical Communication, Proc. IEEE, *58*, No. 10 (October 1970).
2. W. K. Pratt, *Laser Communication Systems*, New York: John Wiley, 1969.
3. I. Bar-David, "Communication Under the Poisson Regime," IEEE Trans. on Information Theory, *IT-15*, No. 1 (January 1969), pp. 31–37.
4. R. M. Gagliardi and S. Karp, "M-ary Poisson Detection and Optical Communications," *IEEE Trans. on Communications Technology, COM-17*, No. 2 (April 1969), pp. 208–216.
5. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, Part I," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 843–874.
6. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, Part II," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 875–886.
7. S. D. Personick, "Baseband Linearity and Equalization in Fiber Optic Digital Communication Systems," B.S.T.J., *52*, No. 7 (September 1973), pp. 1175–1194.
8. D. G. Messerschmitt, "Performance of Several Equalizers in a Digital Fiber Optic Receiver," unpublished work.
9. D. G. Messerschmitt, "Optimum Mean-Square Equalization for Digital Fiber Optic Systems," ICC Conference Record, 1975.
10. S. D. Personick, "New Results on Avalanche Multiplication Statistics with Applications to Optical Detection," B.S.T.J., *50*, No. 1 (January 1971), pp. 167–189.
11. H. Melchior, M. B. Fisher, and F. R. Arams, "Photodetectors for Optical Communication Systems," Proc. IEEE, *58*, No. 10 (October 1970), pp. 1466–1486.
12. T. Kailath, "A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. on Information Theory, *IT-15*, No. 3 (May 1969), pp. 350–361.
13. T. T. Kadota, "Nonsingular Detection and Likelihood Ratio for Random Signals in White Gaussian Noise," IEEE Trans. on Information Theory, *IT-16*, No. 3 (May 1970), pp. 291–298.
14. T. Kailath, "A Further Note on a General Likelihood Formula for Random Signals in Gaussian Noise," IEEE Trans. on Information Theory, *IT-16*, No. 4 (July 1970), pp. 393–396.

15. T. T. Kadota and L. A. Shepp, "Conditions for Absolute Continuity Between a Certain Pair of Probability Measures," *Z. Wahrscheilichkeitstheorie*, Feb. 16, 1970, pp. 250–260.
16. E. V. Hoversten, D. L. Snyder, R. O. Harger, and K. Kurimoto, "Direct-Detection Optical Communication Receivers," IEEE Trans. on Communications Technology, *COM-22*, No. 1 (January 1974), pp. 17–27.
17. G. D. Forney, Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. on Information Theory, *IT-18*, No. 3 (May 1972), pp. 363–378.
18. D. V. Widder, *An Introduction to Transform Theory*, New York: Academic Press, 1971, Chapter 2.
19. D. S. Jones, "Asymptotic Behavior of Integrals," SIAM Review, *14*, No. 2 (April 1972), pp. 286–317.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

Transverse Coupling in Fiber Optics Part IV: Crosstalk

By J. A. ARNAUD

(Manuscript received March 18, 1975)

We evaluate the crosstalk between adjacent cores in an optical fiber that results from electromagnetic coupling. Means of reducing it are discussed. We find that a 0.5-μm-thick layer of silver can, in principle, reduce the crosstalk from −20 to −130 dB without significant increase of the loss. These theoretical results are obtained for two identical single-mode dielectric slabs. In reality, the slabs are not rigorously identical. Longitudinal fluctuations of slab thickness reduce the crosstalk by at least 40 dB. The slab spacing can accordingly be reduced from, typically, 11 to 6 μm for a constant crosstalk. If the slabs are made dissimilar with a relative difference in thickness of 10 percent, the spacing can be reduced further, to approximately 1.5 times the slab thickness. For example, a 15-μm spacing is required between single-mode dissimilar slabs if the nominal slab thickness is 10 μm, provided scattering can be neglected.

I. INTRODUCTION

In multichannel communication systems, crosstalk between channels is a problem that must be considered. Typically, the crosstalk should be less than −20 dB. This means that, if an optical power of 1 mW is fed into one optical guide of a cable, no more than 10 μW should be transferred into the other guides. Let us assume a typical link length of 10 km. The crosstalk measured over a 1-km-long fiber should be less than −40 dB if the power transfer is proportional to the square of the fiber length, less than −30 dB if the power transfer is proportional to the fiber length, and less than −20 dB if the power transfer is independent of the fiber length. As we shall see, the first power law is applicable to identical uniform fibers, the second to nominally identical irregular fibers, and the third to uniform dissimilar fibers.

In optical fibers, the field decays exponentially in the cladding. Therefore, a modest increase in spacing between adjacent fibers is usually sufficient to reduce the optical coupling to tolerable values.

Yet, in some cases, one needs to minimize the cross section of the cable and the spacing between adjacent fibers. Let us briefly discuss a few relevant applications. The need for minimizing the distance between *single-mode* cores in a fiber does not arise in communication systems presently envisioned for the following reasons: The fiber diameter is required to be large (e.g., larger than about 50 μm) so that the fiber is able to sustain mechanical tensions. Thus, quite a few cores can be accommodated within the fiber diameter with sufficient spacing. Furthermore, the capacity of single-mode fibers is so large there is little incentive to introduce more than one core in the same cladding. The problem of coupling between single-mode fibers (or between fibers carrying few modes) does arise, however, when one tries to increase the image-transmission capacity of a fiber bundle up to the diffraction limit, each core carrying one bit of image information. Crosstalk (image blurring) is minimized if adjacent cores are made dissimilar. However, geometrical irregularities may restore a large coupling between closely spaced cores. (This, incidentally, raises the possibility that measurement of the coupling between dissimilar, closely spaced cores gives useful information on the spectral density of the core irregularities.) The problem of coupling between single-mode dielectric waveguides also arises in integrated optics and in biology in the study of the optical behavior of the retina. The results that we present are general. They are therefore applicable, in principle, to multimode, as well as to single-mode, fibers. However, in practical multimode fibers, slow longitudinal variations of the core dimensions make the propagation constants of the modes of one core sweep randomly through the propagation constants of the modes of the other core. Thus, an averaging takes place that cannot be ignored. The problem of coupling between highly multimoded cores will be only briefly discussed.

The shielding method discussed in this paper consists of the introduction of a layer of metal, typically silver, between the adjacent optical waveguides. A reservation is in order: In some communication systems, metallic layers may be undesirable because they detract from the all-dielectric-cable properties. Shielding between adjacent fibers can be provided alternatively by low-refractive-index plastics such as Teflon® FEP ($n \approx 1.32$) that cause the optical field to decay faster than in the cladding material. The reduction in coupling, however, is much smaller than that provided by metals. Plastic materials can be made very lossy by impregnating them with dyes. High losses, however, are much less effective than small refractive indices in reducing evanescent wave coupling. Therefore, we shall consider mainly metallic layers. The practicality of metallic shields remains an open question.

In the first part of this article series,[1] a general and simple expression of the coupling between two lossy open waveguides was derived. Our formulation requires that only the normalized fields of the individual waveguides along a contour be known. In the present paper, we evaluate in detail the crosstalk between two parallel slabs caused by the electromagnetic coupling and means of reducing it. The crosstalk between two optical slabs has been evaluated by Marcuse,[2] although, in Marcuse's work, the slabs are assumed identical. In reality, unavoidable fluctuations in the slab dimensions reduce the crosstalk, as we shall see, by more than 40 dB. Marcuse has also evaluated the reduction of crosstalk provided by a layer of absorbing material located between the slabs. He found that the waveguide loss increases to intolerably high values before any significant reduction in coupling can be obtained. We find that, if the intermediate layer is metallic, the coupling can be drastically reduced without any significant increase of the waveguide loss. This discrepancy results from the fact that, for metallic layers, the permittivity is negative. For very dissimilar media, the first-order perturbation used by Marcuse is not applicable. In the present paper, we assume that the perturbation caused by the intermediate layer on the propagation is small, but we do not assume that the field in that intermediate layer is close to the field that would exist in the absence of the layer.

In Section II, we evaluate the crosstalk between optical waveguides when the axial wave numbers (or propagation constants) of the isolated guides fluctuate along the system axis. In Section III, we evaluate the spacing between slabs corresponding to a given crosstalk. In Section IV, the transmission is evaluated of a metallic layer under evanescent wave excitation and the crosstalk reduction. In Section V, we evaluate the loss that results from the introduction of a metallic layer near a slab waveguide. In Section VI, a simple approximate formula is given for the coupling between oversized round fibers. It is compared to exact results. Finally, brief comments are made in Section VI concerning the applicability of quasi-ray optics techniques in evaluating the coupling between irregular oversized fibers and the effect of bending. A few general results that do not seem available in convenient form in the literature are derived in the appendices.

II. FAST COUPLING

Solution of the coupled-mode equations when the axial wave numbers of the isolated guides are constant, or vary linearly with z, is recalled in Appendix A. In the present section, only the results are given.

Let us first assume that the coupling c between the two guides and the axial wave numbers k_1, k_2, of the isolated guides is constant (independent of z). Let a power unity be fed into guide 1 at $z = 0$ and the other guide, guide 2, be unexcited. The power in guide 2 grows, at first, according to the law (see Appendix A)

$$P_2(z) = (cz)^2. \tag{1}$$

This result is valid only as long as $\Delta z \ll 1$, where we have defined

$$\Delta \equiv \{[(k_1 - k_2)^2/4] + c^2\}^{\frac{1}{2}}. \tag{2}$$

For example, a -20-dB crosstalk ($P_2 = 0.01$) over a 1-km length of cable is obtained, according to (1), if $c = 10^{-4}$ m^{-1}. Condition $\Delta z \ll 1$ is, for identical guides, $z \ll 10$ km. However, if $(k_1 - k_2)/(k_1 + k_2) = 10^{-4}$, law (1) is applicable only if $z \ll 1$ mm, a drastically different condition. In Section III, the distance between the guides that corresponds to this particular coupling is evaluated.

Now let $k_1 - k_2$ vary linearly with z. The coupling c remains a constant. We write

$$k_1(z) = k_0 + \alpha z, \qquad k_2(z) = k_0 - \alpha z, \tag{3}$$

where k_0 and α denote constants. At large $|z|$, the coupling is insignificant because of the large value of $k_1 - k_2$. The coupling becomes important only near the origin, $z = 0$, where near-synchronism is achieved. Let a power unity be fed into guide 1, at large negative z. The power transferred to guide 2 at large positive z is exactly (see Appendix A)

$$P_2 = 1 - \exp(-\pi c^2/\alpha). \tag{4}$$

We are interested in the case where the k's are crossing very rapidly. Thus, let us assume that α is large and that, consequently, $\pi c^2/\alpha$ is small. In that approximation,

$$P_2 = \pi c^2/\alpha \ll 1. \tag{5}$$

In most practical systems, $k_1 - k_2$ oscillates as a function of z. A significant amount of coupling between two guides takes place only near the crossing points. To develop an understanding of the effects of longitudinal variations of the difference of the axial wave numbers $k_1(z)$ and $k_2(z)$, we model the difference in wave numbers as a simple sinusoid, i.e.,

$$\tfrac{1}{2}(k_1 - k_2) = \delta \sin(\Omega z), \tag{6}$$

where δ denotes the peak deviation of $(k_1 - k_2)/2$ and $2\pi/\Omega$ the period of oscillation. It seems reasonable to assume that the phases of the signals picked up by fiber 2 at the successive crossing points are un-

correlated and that, consequently, the powers add up. This incoherency is a consequence of the fluctuations of the phase of the optical field between successive crossing points. According to (6), the slope α introduced in (3) is

$$\alpha = \delta\Omega. \tag{7}$$

The number of crossing points over a length z is $\Omega z/\pi$. Thus, the power collected by guide 2 over length z is

$$P_2 = (\Omega z/\pi)(\pi c^2)/(\delta\Omega) = c^2 z/\delta. \tag{8}$$

Note that P_2 is independent of Ω. P_2 is proportional to c^2, as was the case in the absence of fluctuations, but it varies linearly with z rather than being proportional to z^2. Let us compare P_2 in (8) and P_2 in (1). The ratio of these two collected powers is

$$\frac{P_2 \text{ (uniform fibers)}}{P_2 \text{ (nonuniform fibers)}} = \delta z. \tag{9}$$

It seems reasonable to assume that, over a length of 1 km ($z = 10^9$ μm), the relative variations of the axial wave number are larger than 10^{-4}: $\delta/k > 10^{-4}$. For the single-mode slab considered in the next section, this number corresponds to a fluctuation of the slab thickness of 0.01 μm. Because k is of the order of 2π μm^{-1}, the reduction in coupling owing to the lack of identity between the two slabs is, in that case, of the order of 50 dB. The results obtained are therefore much too conservative if we assume that the optical guides are identical in evaluating the crosstalk.

III. EVALUATION OF COUPLING BETWEEN TWO SLABS

Let us consider two identical dielectric slabs having thickness $2d$ and material free wave number k. The free wave number in the medium between the slabs (cladding) is denoted k_s, and the spacing between the slabs is denoted $2D$. (See Fig. 1. The intermediate layer is to be ignored for the moment.) The expression for the coupling c between the fundamental H waves is well known (see, for example, Ref. 1):

$$c = \kappa R \exp(-2\kappa D), \tag{10}$$

where

$$\kappa \equiv (k_z^2 - k_s^2)^{\frac{1}{2}} \tag{11a}$$

$$R = (k_z d)^{-1}[1 + (1/\kappa d)]^{-1}[1 - (\kappa^2 d^2/F^2)] \tag{11b}$$

$$F^2 \equiv (k^2 - k_s^2)d^2. \tag{11c}$$

k_z denotes the axial wave number of the isolated slabs (previously denoted k_1 and k_2 for the two waveguides). If we require that only one H mode propagate (for simplicity, we shall ignore the E waves), the

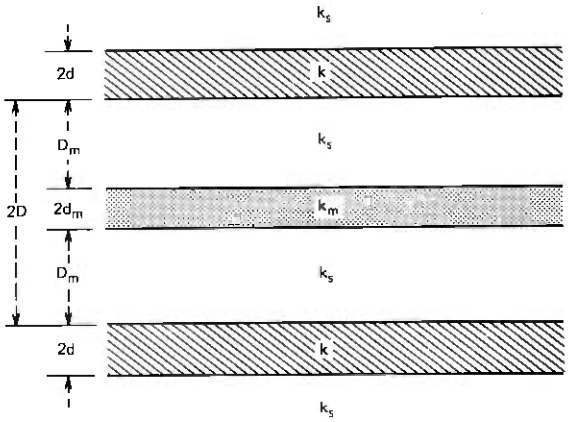


Fig. 1—Coupled dielectric slabs with thickness $2d$ and free wave number k. The cladding medium has free wave number k_s. Crosstalk can be reduced by introducing a metallic layer with free wave number k_m (almost purely imaginary) and thickness $2d_m$.

maximum value of F is $\pi/2$. The theory of dielectric slabs shows that, for that value of F, $\kappa d = 1.28$. Thus, the coupling is

$$c = (0.24/kd^2) \exp(-2.56D/d), \tag{12}$$

where we have made the approximation $k_z \approx k$ in the first term. Thus, for a constant relative spacing D/d, the coupling between two single-mode slabs varies as the inverse of the square of their thickness.

Let us evaluate c for the numerical values

$$2d = 1.32\ \mu\text{m}, \quad k = 2\pi \times 1.45\ \mu\text{m}^{-1}, \quad k_s = 2\pi \times 1.4\ \mu\text{m}^{-1}. \tag{13}$$

Thus,

$$F = \pi/2, \quad k_z d \approx kd = 8.88. \tag{14}$$

If we substitute these results in (11b) and (12), we obtain

$$R = 0.021, \quad c(\text{in m}^{-1}) = 4 \times 10^4 \exp(-3.88D), \tag{15}$$

where D is in μm. If the slabs are identical, -20-dB crosstalk in 1 km is obtained, as we have seen in Section II, when $c = 10^{-4}$ m^{-1}. This corresponds, according to (15), to a spacing

$$2D = 11\ \mu\text{m}. \tag{16}$$

If the slabs have some irregularities, with $\delta/k = 10^{-4}$ (corresponding to a variation of slab thickness of 0.01 μm), -20-dB crosstalk is obtained when $c = 0.25$ m^{-1}. This coupling corresponds to a smaller spacing: $2D = 6.2\ \mu$m. If the slab thickness is chosen equal to 10.5 μm, keeping $F = \pi/2$ ($\Delta n/n = 5 \times 10^{-4}$), the spacing required for

identical slabs and −20-dB crosstalk over a 1-km length is $2D = 66$ μm, a rather large spacing.

If the two slabs are uniform but are made deliberately dissimilar, a lower crosstalk is obtained. The relative difference δ/k in axial wave numbers is approximately $0.5\ (\Delta d/d)/(kd)^2$, where $\Delta d/d$ is the relative difference in thickness of the two slabs ($F = \pi/2$). For example, if one slab has a thickness $2d$ equal to 1.32 μm and the other has a thickness equal to 1.2 μm, the relative difference in k_z is: $\delta/k = 0.65 \times 10^{-3}$. The maximum relative power that can be transferred from one slab to the other is, according to eq. (39), equal to $(c/\Delta)^2$, where $\Delta \approx \delta$. Thus, a −20-dB crosstalk corresponds, for the above value of δ, to a coupling $c = \delta/10 = 580\ \text{m}^{-1}$. The slab spacing $2D$ corresponding to that coupling is given by (15). We obtain $2D = 2.2$ μm. More generally, we find that $D \approx 1.5d$ for any d, if F is kept equal to $\pi/2$ and $\Delta d/d = 0.1$. Thus, a considerable reduction in spacing is tolerable, in principle, if the slabs are made dissimilar. However, fast fluctuations along the z-axis of the slab dimensions with a period of the order of $\pi/\delta \approx 100$ μm would reestablish synchronism between the two slabs. Fluctuations that are too small in amplitude to deteriorate the propagation under normal conditions (e.g., no significant coupling to the radiation modes) may nevertheless introduce a large crosstalk when the slabs are very close to each other. Thus, the result obtained above, that the spacing between slabs can be reduced to $1.5 \times (2d)$ if the fibers are made dissimilar, may not hold true in practical conditions.

IV. TRANSMISSION THROUGH A METALLIC LAYER UNDER EVANESCENT WAVE EXCITATION

The results in Section II show that the crosstalk power P_2 is proportional to the square of the coupling c. We have shown in Ref. 1 that, for identical slabs and a symmetrical configuration, the coupling c is proportional to the square of the normalized field halfway between the two slabs. Thus, the crosstalk is proportional to the fourth power of the normalized field halfway between the two slabs. If we introduce a metallic layer of thickness $2d_m$, symmetrically centered between the two slabs as shown in Fig. 1, the crosstalk is reduced in proportion to the fourth power of the field in the middle of the metallic layer. This field reduction, denoted t (for transmission), is evaluated in the present section.

Let us consider an evanescent wave with axial wave number $k_z > k_s$, where k_s denotes the free wave number in the medium. This wave decays in the x direction according to

$$E(x) = E_0 \exp(-\kappa x), \tag{17}$$

$$\kappa \approx (k_z^2 - k_s^2)^{\frac{1}{2}}. \tag{18}$$

Let us now introduce a metallic layer with complex wave number $k_m \equiv k_{mr} + ik_{mi}$ and thickness $2d_m$. The ratio t of the field in the middle of the layer to the field at the same point in the absence of the layer is derived in Appendix B. Provided the layer is sufficiently thick or, more precisely, that

$$\text{Real } (\kappa_m d_m) \gg 1, \tag{19}$$

where

$$\kappa_m \equiv (k_z^2 - k_m^2)^{\frac{1}{2}}, \tag{20}$$

we have

$$t = [4\kappa\kappa_m/(\kappa + \kappa_m)^2] \exp [(\kappa - \kappa_m)d_m]. \tag{21}$$

At a free-space wavelength $\lambda_0 = 1\ \mu\text{m}$, $k_0 = 2\pi\ \mu\text{m}^{-1}$, the wave number of silver is almost purely imaginary,[3]

$$k_m^2 \equiv (k_{mr} + ik_{mi})^2 = (0.2k_0 + i5k_0)^2 = -985 + 79i \text{ (in } \mu\text{m}^{-2}), \tag{22}$$

and, for a typical glass, assumed lossless ($n_s = 1.4$),

$$k_s^2 = n_s^2 k_0^2 = (1.4k_0)^2 = 77.4\ \mu\text{m}^{-2}. \tag{23}$$

With the value of $\kappa^2 \equiv k_z^2 - k_s^2 = 3.76\ \mu\text{m}^{-2}$ in (14), and k_m^2, k_s^2 in (22) and (23), we obtain $\kappa_m = 32 - 1.3i$, and, from (21), a power transmission

$$T \equiv tt^* = 0.062 \exp (-60d_m), \tag{24}$$

where d_m is in μm, provided

$$d_m \gg 0.03\ \mu\text{m}. \tag{25}$$

Because the crosstalk power P_2 is proportional to the square of the power transmission T, the introduction of a layer of silver of thickness $2d_m$ between the two slabs reduces the crosstalk in dB by

$$20 \log_{10} (T) = 520d_m, \tag{26}$$

where d_m is in μm. For example, if the layer thickness is $2d_m = 0.5\ \mu$m, the crosstalk is reduced by 130 dB. This reduction is independent of the initial value of the crosstalk, within the approximations made. Thus, a 0.5-μm-thick layer of silver is sufficient to ensure a complete isolation of adjacent fibers, at a wavelength $\lambda_0 = 1\ \mu$m.

Surface polaritons can be guided near the dielectric ($k_s^2 > 0$) and metallic ($k_m^2 < 0$) interface. However, the losses of such modes are extremely high over a distance of 1 km. The cladding modes are also strongly attenuated, and it seems that they can be safely ignored. For comparison, let us consider, in place of the metallic layer, a low-index plastic material of the Teflon type, with a refractive index $n = 1.32$. We now have $k_m^2 = 69\ \mu\text{m}^{-2}$ and $\kappa_m = 3.47\ \mu\text{m}^{-1}$. We obtain a cross-

talk reduction equal to $26d_m$ in dB, where d_m is in μm. Thus, a 50-dB reduction in crosstalk requires a 4-μm-thick layer of low-index plastic material.

V. LOSS INTRODUCED BY A METALLIC LAYER

We are now concerned with the fact that, because the refractive index of a metal is not purely imaginary, the presence of the metallic layer may increase significantly the loss of the modes guided by the fibers. This loss depends critically on the distance between the metallic layer and the fibers and, therefore, on the distance between the two fibers. The loss suffered by the fiber is influenced by the complex reflection of the metallic layer for evanescent waves. This reflection, strictly speaking, depends on the thickness of the metallic layer. Exact expressions are given in Appendix B. However, in all our numerical examples, the thickness of the metallic layer is so large that it can be assumed infinite. In that case, the reflection r reduces to

$$r = (\kappa - \kappa_m)/(\kappa + \kappa_m), \tag{27}$$

where κ and κ_m are defined in (18) and (20), respectively. Because the imaginary part κ_{mi} of κ_m is much smaller than the real part κ_{mr}, the imaginary part r_i of r is approximately

$$r_i \approx 2\kappa\kappa_{mi}/\kappa_{mr}^2. \tag{28}$$

If we use for k_z, k_s, and k_m the numerical values in (14), (23), and (22), respectively, we find $r_i = 0.005$.

To obtain the loss suffered by the slab, we use the perturbation formula derived in Appendix C. The variation of k_z is assumed to be small. The variation of the field near the perturbing object, however, is not assumed small. In the present case, k_z is real before perturbation. The introduction of the metallic layer causes k_z to acquire a small imaginary part, k_{zi}. The imaginary part k_{zi} of k_z is the fiber loss, in neper/unit length. There is also a small variation of the real part of k_z. This variation, however, is of no interest to us. We have (see Appendix C)

$$k_{zi} = r_i\kappa R \exp(-2\kappa D_m), \tag{29}$$

where R is the slab parameter defined in (12a) and D_m the distance between the slab and the metallic layer. The imaginary part r_i of the metallic layer reflectivity is given in (28).

For the numerical values used earlier in (14) and (15), we obtain from (29)

$$\mathcal{L}_{dB/km} = 8.7 \times 10^3 k_{zi} = 2.6 \times 10^6 \times \exp(-3.88D_m), \tag{30}$$

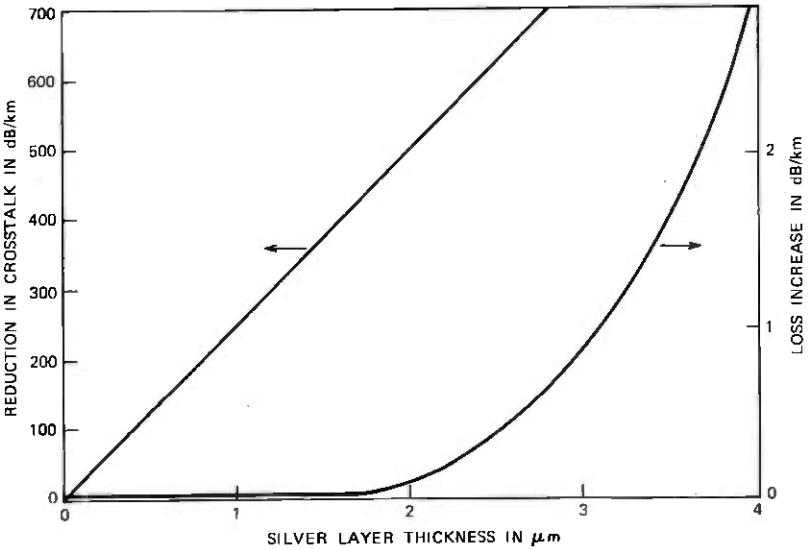


Fig. 2—Reduction in crosstalk and increase in fiber loss resulting from the introduction of a silver layer of thickness $2d_m$ (free-space wavelength = 1 μm). The dielectric slabs have a normalized frequency $F \equiv (k^2 - k_s^2)^{\frac{1}{2}}d = \pi/2$. Their spacing is kept equal to 11 μm. The loss varies with d_m only because of the change in the slab-layer spacing. In the absence of metallic layer, crosstalk is −20 dB/km.

where D_m is in μm. For $D_m = D - d_m = 5.25$ μm, the loss introduced by the metallic layer, given in (30), is only

$$\mathfrak{L} = 0.017 \text{ dB/km}. \tag{31}$$

This loss is quite negligible compared with the other losses (absorption because of impurity or scattering losses) suffered by the wave. However, because $\mathfrak{L}$ depends critically on D_m, this loss may not be negligible in all practical cases. The reduction of the crosstalk and the increase of loss caused by a silver layer of thickness $2d_m$ are shown in Fig. 2 for the dielectric slabs considered earlier, as functions of $2d_m$. Note that, if we assume for simplicity that the thickness of the metallic layer is negligible compared with the slab spacing ($2d_m \ll 2D$), the (dimensionless) ratio of k_{zi} (loss) and c is, approximately,

$$k_{zi}/c = 2(k_{mr}/k_{mi}^2)\kappa^2 d. \tag{32}$$

Thus, the best metal, from the point of view of propagation, is the one whose k_{mr}/k_{mi}^2 is the smallest.

VI. ROUND FIBERS

The general coupling formula in Ref. 1 is applicable, in principle, to round fibers. Round fibers are more often encountered in practice

than are slabs. The geometry is shown in Fig. 3. The fibers are assumed identical, with radius a and spacing $2D$. The results are given only for the scalar fundamental field $\psi \approx \mathrm{HE}_{11}$ of oversized fibers $[F \equiv (k^2 - k_z^2)^{\frac{1}{2}}a \gg 1]$. In that approximation, the normalized field is easily found to be (see Part II of Ref. 1)

$$\psi(y) = u_0(\pi^{\frac{1}{2}}k^{\frac{1}{2}}aF)^{-1} \exp(-Fy^2/2a^2), \tag{33}$$

where $u_0 \approx 2.4 \cdots$ is the first zero of the Bessel function of order zero. The y axis is tangent to the rod considered, as shown in Fig. 3. The Fourier transform of $\psi(y)$ is

$$\begin{aligned}\hat{\psi}(k_y) &= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \psi(y) \exp(-ik_y)dy \\ &= \pi^{-\frac{1}{4}}u_0k^{-\frac{1}{2}}F^{-\frac{3}{4}} \exp(-k_y^2a^2/2F).\end{aligned} \tag{34}$$

Because the spectral component $\hat{\psi}(k_y)$ varies approximately as $\exp(-sx)$ as a function of x, where $s \equiv (k_z^2 - k_z^2)^{\frac{1}{2}} \approx F/a$, the

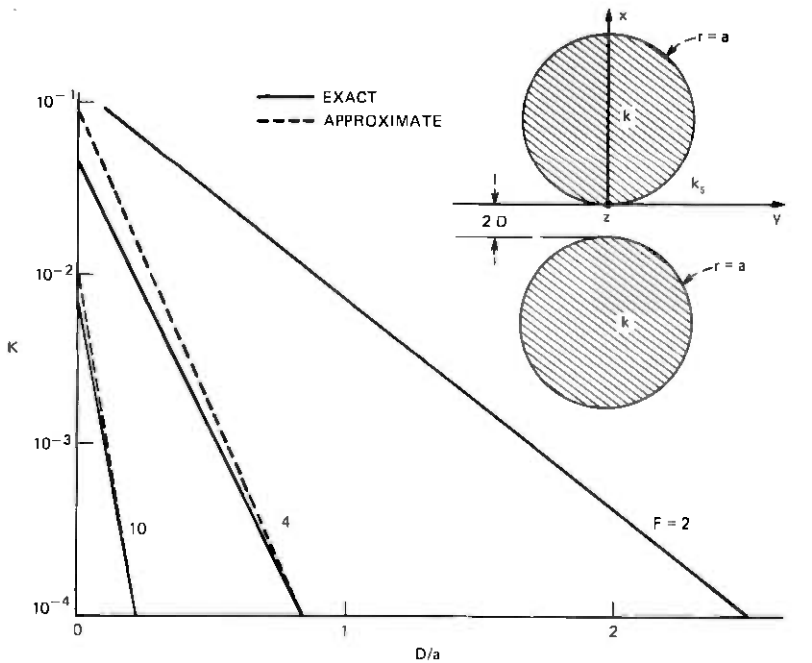


Fig. 3—Variation of the coupling between two dielectric rods of radii a as a function of their spacing $(2D)$. The dimensions and free wave numbers are shown. The parameter K is defined as $ca(1 - k_z^2/k^2)^{-\frac{1}{4}}$, and c is the coupling. The plain lines are from Snyder exact theory,[7] and the dashed lines from the theory in Ref. 1, applied to large normalized frequencies F.

coupling is

$$c = \int_{-\infty}^{+\infty} s(k_y)\hat{\psi}_2^*(k_y)\hat{\psi}_1(k_y)dk_y$$

$$= (u_0^2/\pi^{\frac{1}{2}})k^{-1}F^{-\frac{1}{2}}a^{-2}\exp(-2FD/a). \quad (35a)$$

In place of c, we can use a normalized coupling K defined by

$$K = ca[1 - (k_z^2/k^2)]^{-\frac{1}{4}}. \quad (35b)$$

In the general expression for c in (34), $\hat{\psi}_2$ and $\hat{\psi}_1$ represent the spectral components of the field of the two fibers along the y-axis at $x = 0$. The normalized coupling K is plotted in Fig. 3 (dashed lines) as a function of the ratio D/a of the fiber spacing ($2D$) to fiber diameter ($2a$). In that figure, the parameter is the normalized frequency F. For comparison, an exact result obtained by Snyder[4] is shown as a plain line. The agreement is very good for $F \gtrsim 4$.

The advantage of the method used in this section is that it is applicable when the two fibers are separated by a metallic layer. In that case, one need only introduce inside the integral sign in the first expression in (34) a term $T(k_y)$, where T denotes the power transmission of the metallic layer, defined in (24). T now depends slightly on k_y because, in the expressions given earlier for T, the axial wave number k_z should be replaced by $(k_z^2 + k_y^2)^{\frac{1}{2}}$. The effect of the dependence of T on k_y is small, however, and the value obtained earlier for T for slabs is approximately applicable to round fibers as well.

VII. MULTIMODED IRREGULAR FIBERS

We shall make only qualitative comments. In the preceding calculations, we have considered the coupling between one mode of one core and one mode of another adjacent core. If the cores can carry many modes and have dimensions that fluctuate as a function of z, with such an amplitude that the variations in axial wave numbers exceed the spacing (in axial wave numbers) between adjacent modes, some averaging takes place. The situation becomes comparable, at least over some distance, to that of a slab radiating power into a semi-infinite dielectric, a situation discussed in detail in Part II of this series of papers.[1]

Let us picture the field in slab 1 (excited at $z = 0$) as made up of two plane waves. The plane wave moving toward slab 2 tunnels into slab 2. Because of the fluctuations in axial wave numbers, the power transferred from slab 1 to slab 2 is essentially the power carried by that tunnelling wave; we can ignore the fact that this wave, after tunnelling, is reflected back and forth inside slab 2 and may tunnel back to slab 1. The power transferred from slab 1 to slab 2, then, is

proportional to z, for small z, rather than to the square of z, as is the case in the absence of irregularities. This picture is consistent with that used by Cherin,[5] who adds the powers transmitted by tunnelling rays. Let us emphasize that the validity of this quasi-ray optics approach rests on the presence of large slow fluctuations of the core dimensions. A simple calculation shows that the relative fluctuations of the slab thickness must exceed the reciprocal of the mode number. This condition is never met for the low-order modes, but it may be met by the higher-order modes. Thus, the situation is rather complicated and requires a deeper analysis. This quasi-ray technique should not be confused with that of Kapany and Burke,[6] where the slabs are assumed identical and the *fields* of the tunnelling rays, rather than their powers, are added. In the preceding discussion, we have assumed that the fiber cable is essentially straight. The coupling increases significantly if the cable is bent.[7] This effect makes it even more important to provide shields between adjacent fibers.

VIII. CONCLUSION

We have shown that a drastic reduction of crosstalk between parallel dielectric slabs can be obtained by introducing a layer of silver (thickness $\approx 0.5\ \mu$m) between adjacent slabs. The reduction, in decibels, is proportional to the imaginary part of the refractive index of the metallic layer and to the layer thickness. In many cases of practical importance, the loss introduced by this metallic layer is negligible. We have also shown that, because of unavoidable irregularities in the fiber dimensions, the crosstalk is at least 40 dB below that expected for identical fibers.

IX. ACKNOWLEDGMENTS

The author expresses his thanks to E. A. J. Marcatili for many helpful comments and stimulating discussions.

APPENDIX A

Fast and Adiabatic Coupling

Let ψ denote the field of a guide, such that $\psi\psi^*$ is the power. When two guides are weakly coupled, their respective fields ψ_1, ψ_2 approximately satisfy the well-known equations[8]

$$\begin{aligned} -id\psi_1/dz &= k_1(z)\psi_1 + c\psi_2 \\ -id\psi_2/dz &= k_2(z)\psi_2 + c\psi_1. \end{aligned} \tag{36}$$

For simplicity, we assume that the axial wave numbers k_1, k_2 of the isolated guides are real and that the coupling c is a real constant. The solution when k_1, k_2 are constant is well known. For the convenience

of the reader, this solution is derived below. The general solution of (36) is a superposition of normal modes

$$\begin{aligned}\psi_1(z) &= \psi_1^+ \exp(ik^+z) + \psi_1^- \exp(ik^-z)\\ \psi_2(z) &= \psi_2^+ \exp(ik^+z) + \psi_2^- \exp(ik^-z),\end{aligned} \tag{37}$$

where

$$k^\pm = (k_1 - k_2)/2 \pm \Delta \tag{38a}$$

$$\Delta = \{[(k_1 - k_2)^2/4] + c^2\}^{\frac{1}{2}}. \tag{38b}$$

If the initial conditions are $\psi_1(0) = 1$, $\psi_2(0) = 0$, that is, if only guide 1 is excited at the origin ($z = 0$), the field in the unexcited guide, 2, is

$$\psi_2(z) = (ic/\Delta) \exp[i(k_1 + k_2)z/2] \sin(\Delta z). \tag{39}$$

Thus, for small z, the power in guide 2 increases as

$$P_2(z) = (cz)^2, \qquad \Delta z \ll 1. \tag{40}$$

This result is independent of $k_1 - k_2$. See Fig. 4.

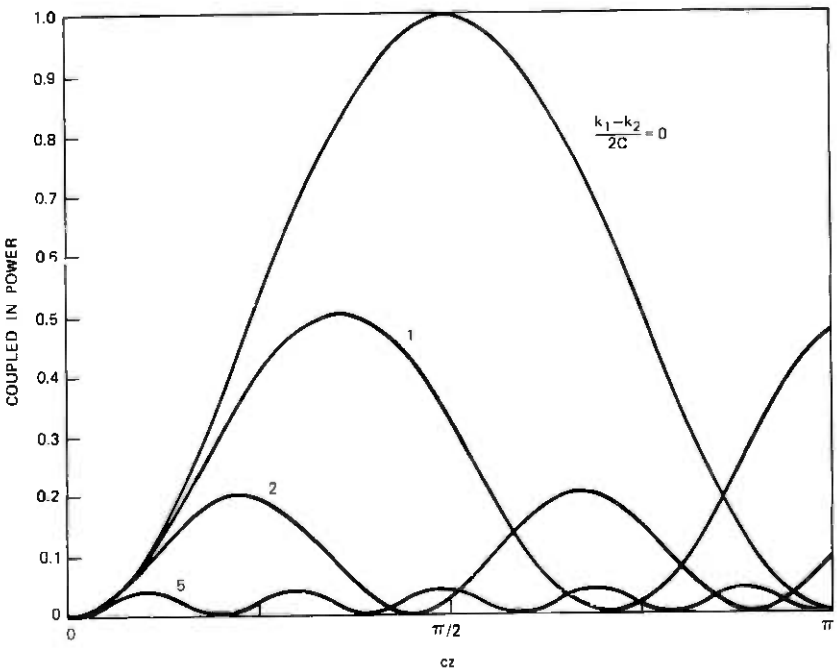
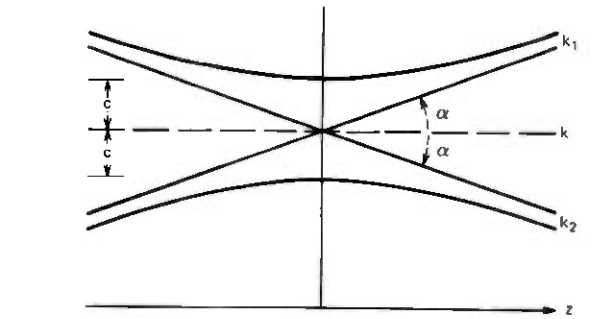
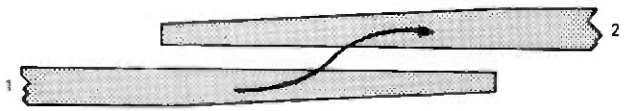


Fig. 4—Variation of the optical power picked up by fiber 2, where only fiber 1 is excited at $z = 0$, as a function of the normalized axial distance. The axial wave numbers of the isolated fibers are assumed to be constant but different [parameter $(k_1 - k_2)/2c$]. Note that the behavior for small cz is independent of $k_1 - k_2$.



(a)



(b)

Fig. 5—(a) Linear variation of the axial wave number of the isolated waveguides as a function of the axial coordinate z. The hyperbolas represent the normal mode wave numbers. (b) Adiabatic coupling in fiber optics. All the power from one fiber is transferred to the other fiber if the k's vary sufficiently slowly. This principle is applicable to multimode fibers.

Let now the axial-wave numbers k_1, k_2 of the isolated guides vary linearly with z

$$k_1(z) = k_0 + \alpha z, \qquad k_2(z) = k_0 - \alpha z, \tag{41}$$

where k_0 and α denote constants. Synchronism takes place only near the origin, $z = 0$. Let us set

$$\psi_{1,2}(z) = A_{1,2}(z) \exp(ik_0 z) \tag{42}$$

in (36). After differentiation and substitution, we obtain an equation for A_1,

$$(d^2A_1/dz^2) + (\alpha^2 z^2 + c^2 - i\alpha)A_1 = 0. \tag{43a}$$

A similar equation holds for A_2 that we need not write down. Equation (43a) is the equation for parabolic cylinder functions. The asymptotic form of the solution, valid for $-\pi/2 \leqq \arg(z) \leqq \pi$ is, for a power unity at $z = -\infty$ (see Ref. 9),

$$A_1(z) = \exp\left[i(\alpha/2)z^2 + i(c^2/2\alpha)\log(-z)\right], \qquad z \ll c/\alpha \tag{43b}$$

$$A_1(z) = \exp\left[i(\alpha/2)z^2 + i(c^2/2\alpha)\log(z) - \pi c^2/2\alpha\right], \qquad z \gg c/\alpha, \tag{43c}$$

as we easily verify by substituting (43b) in (43c) and neglecting terms of order z^{-2}. To go from (43b) to (43c), note that $\log(-z)$

$= i\pi + \log(z)$. Note also that a change in the unit with which z is measured affects only the amplitude of A_1, which is arbitrary.

The power in guide 2 after the interaction has taken place, that is, for large positive z, is, according to (43c),

$$P_2 = 1 - A_1 A_1^* = 1 - \exp(-\pi c^2/\alpha). \tag{44}$$

Let us first assume that $\pi c^2/\alpha$ is very small compared with unity, that is, the k's are crossing very rapidly. In that case, guide 1 transfers only a small amount of power to guide 2, equal to $\pi c^2/\alpha$. This is the result used in the text.

When $\pi c^2/\alpha$ is very large compared with unity, that is, when the variation of $k_1 - k_2$ is very slow, almost all the power from guide 1 is coupled to guide 2. This is the principle of the Cook adiabatic coupler.[10] This mechanism is applicable also to multimode dielectric waveguides. It may be used to couple two optical fibers because the dimensions are not critical. Only slowness is required.[11] (See Fig. 5.)

APPENDIX B

Transmission and Reflection at a Metallic Layer Under Evanescent Wave Excitation

Let the metallic layer have a complex free wave number $k_m \equiv k_{mr} + ik_{mi}$ and a thickness d_m. The surrounding medium is assumed to have a real free wave number k_s. The field has the general form (see Fig. 6)

$$E(x) = \begin{cases} E_0[\exp(-\kappa x) + r\exp(\kappa x)], & x \leqq 0 \\ E^- \exp(-\kappa_m x) + E^+ \exp(\kappa_m x) & 0 \leqq x \leqq d_m \\ E_0 t \exp(-\kappa x) & x \geqq d_m, \end{cases} \tag{45}$$

where

$$\kappa \equiv (k_z^2 - k_s^2)^{\frac{1}{2}} \tag{46}$$

is real, and

$$\kappa_m \equiv (k_z^2 - k_m^2)^{\frac{1}{2}}. \tag{47}$$

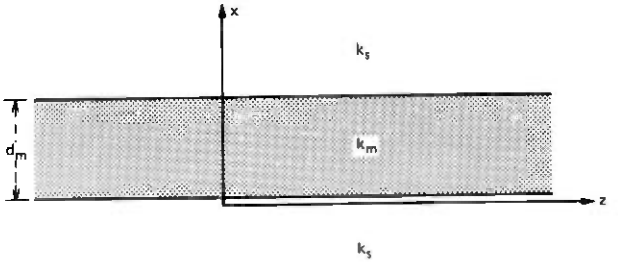


Fig. 6—Transmission of a metallic layer with thickness d_m and free wave number k_m under evanescent wave excitation (axial wave number $k_z > k_s$). At large negative x, the field is assumed unperturbed by the layer.

The axial wave number k_z is assumed to be real and larger than k_s. By specifying that E and dE/dx are continuous at the boundaries, $x = 0$, $x = d_m$, we obtain the reflection r and the transmission t:

$$r = [(\kappa/\kappa_m) - (\kappa_m/\kappa)][2 \coth (\kappa_m d_m) + (\kappa_m/\kappa) + (\kappa/\kappa_m)]^{-1} \quad (48)$$

$$t = \exp (\kappa d_m)\{\cosh (\kappa_m d_m) + \tfrac{1}{2}[(\kappa_m/\kappa) + (\kappa/\kappa_m)] \sinh (\kappa_m d_m)\}^{-1}. \quad (49)$$

We shall now assume that the metallic layer is thick in the sense that Real $(\kappa_m d_m) \gg 1$. These conditions are well satisfied for the metallic layers that we consider in the main text. In that case, (48) and (49) reduce to

$$r = (\kappa - \kappa_m)/(\kappa + \kappa_m) \quad (50)$$

$$t = [4\kappa\kappa_m/(\kappa + \kappa_m)^2] \exp [(\kappa - \kappa_m)d_m], \quad (51)$$

respectively. Equations (50) and (51) are the results used in the text.

APPENDIX C

Loss Introduced by a Metallic Layer

Let us consider a uniform reciprocal waveguide and let a uniform rod be introduced that perturbs the propagation of the waveguide (Fig. 7a). We assume that the perturbing rod does not support trapped modes or, if it does, that the axial wave numbers of these trapped modes are sufficiently far away from that, k_{z0}, of the waveguide. No resonant coupling is assumed to take place.

We shall first recall a very general result. Let $\mathbf{E}^+$, $\mathbf{H}^+$ and $\mathbf{E}_p$, $\mathbf{H}_p$ denote two time-harmonic fields at the same frequency in the same medium. If we assume that the medium is reciprocal (that is, that the tensor permittivity is symmetrical), it readily follows from the Maxwell equations that the divergence of the vector

$$\mathbf{J} = \mathbf{E}^+ \times \mathbf{H}_p + \mathbf{H}^+ \times \mathbf{E}_p \quad (52)$$

is equal to zero. Thus, the flux of $\mathbf{J}$ through any closed surface is equal to zero. In what follows, an $\exp(-i\omega t)$ term is omitted.

Now let $\mathbf{E}^+$, $\mathbf{H}^+$ be the field propagating in the $-z$ direction along an open waveguide. The dependence of $\mathbf{E}^+$ and $\mathbf{H}^+$ on z is denoted: $\exp(-ik_{z0}z)$. Let $\mathbf{E}_p$, $\mathbf{H}_p$ be the field propagating in the $+z$ direction in the presence of the perturbing rod with an $\exp(ik_z z)$ dependence on z. The closed surface S is taken as the surface shown in Fig. 7a bounded by the planes $z = 0$ and $z = dz$, the volume of the perturbing rod being excluded. For that choice, the medium enclosed by S is the same for both fields. We can therefore use the result stated earlier that the flux of $\mathbf{J}$ through S is zero. Let us consider the various contributions to that flux. The flux of $\mathbf{J}$ through the plane $z = dz$ differs from the

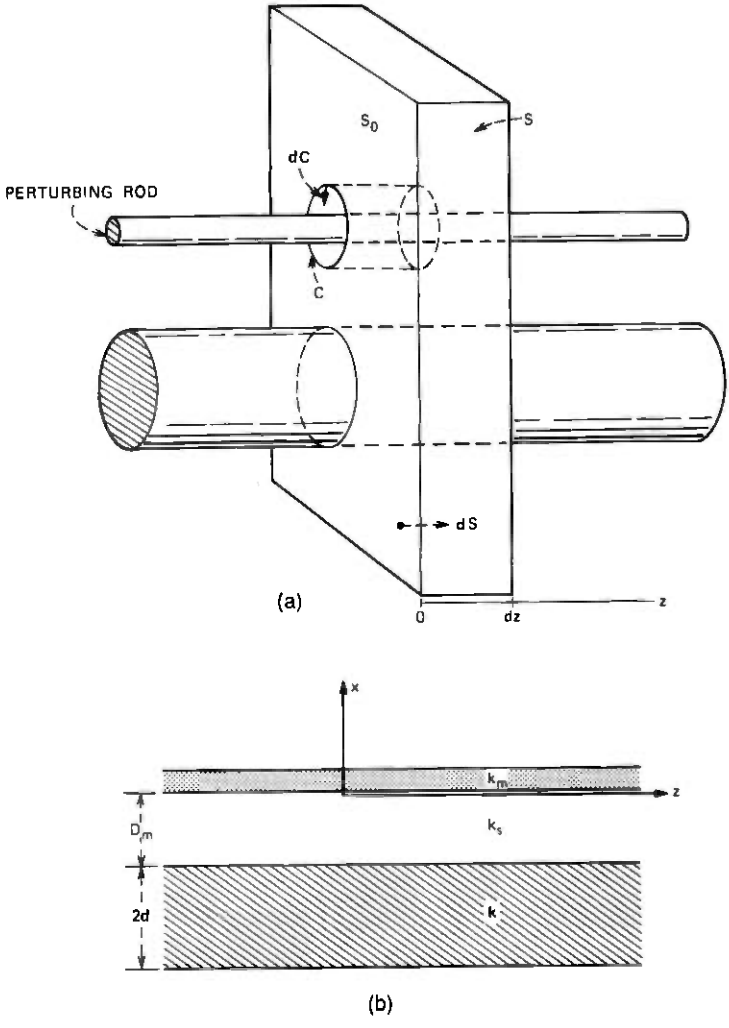


Fig. 7—(a) Schematic for the derivation of the general perturbation formula. The dielectric waveguide is perturbed by a small lossy rod. The closed surface S extends to infinity where the fields considered are assumed to vanish. (b) Application to the perturbation of H waves guided by a dielectric slab (k) by a lossy slab (k_m).

flux of J through the plane $z = 0$ only by a factor $-\exp[i(k_z - k_{z0})dz]$. The difference between these fluxes is, therefore, $i(k_z - k_{z0})dz$ for small dz. Because we are considering only trapped modes, the flux at infinity is zero. The flux through the surface surrounding the perturbing rod is dz times the line integral of $\mathbf{J} \cdot d\mathbf{C}$, with $d\mathbf{C}$ a vector perpendicular to the contour surrounding the rod, pointing inward, whose

length is the elementary arc length. Thus, we have exactly

$$i\Delta k_z = \int_C \mathbf{J}\cdot\mathrm{d}\mathbf{C} \Big/ \int_{S_0} \mathbf{J}\cdot\mathrm{d}\mathbf{S}, \tag{53}$$

where

$$\Delta k_z \equiv k_z - k_{z0}. \tag{54}$$

S_0 denotes the transverse plane, $z = 0$ minus the area enclosed by C, and $\mathrm{d}\mathbf{S}$ denotes a vector directed along the z axis whose length is the elementary area. The derivation given above is almost identical to that in Ref. 1 for coupled waveguides. We now assume that the perturbation is small. Thus, we can replace $\mathbf{E}_p$, $\mathbf{H}_p$ by the unperturbed field $\mathbf{E}$, $\mathbf{H}$ propagating in the $+z$ direction in the integral over S_0 in (53). This is not permissible, however, for the integral over C, in general.

Let (52) be specialized to the H waves guided by a dielectric slab shown in Fig. 7b. In that case, $\mathbf{E}$ has only one component: $E_y \equiv E(x)$, $H_z = (1/i\omega\mu_0)\partial E/\partial x$, and $H_x = -(k_z/\omega\mu_0)E$. Taking into account $E_y^+ = E_y$ and $H_z^+ = H_z$ (see Ref. 1), we obtain

$$\Delta k_z = \left[(E\partial E_p/\partial x) - (E_p\partial E/\partial x)\right] \Big/ \left(2k_z \int_{-\infty}^{+\infty} E^2 dx\right), \tag{55}$$

where we have assumed that E_p differs significantly from E only near the perturbing slab. The unperturbed field is, for $-D_m < x < 0$,

$$E = \exp(-\kappa x), \tag{56}$$

and the perturbed field is that given in (45)

$$E_p = \exp(-\kappa x) + r\exp(\kappa x), \tag{57}$$

where

$$\kappa \equiv (k_z^2 - k_s^2)^{\frac{1}{2}}.$$

The amplitudes in (56) and (57) are so chosen that $E_p \approx E$ for large negative x, e.g., $x = -D_m$.

We first evaluate

$$(E\partial E_p/\partial x) - (E_p\partial E/\partial x) = 2r\kappa, \tag{58}$$

where we have used (56) and (57). Note that the result (58) is independent of x (for $-D_m < x < 0$). Substituting (58) in (55), the imaginary part of k_z is found

$$k_{zi} = r_i\kappa R\exp(-2\kappa D_m), \tag{59}$$

where r_i denotes the imaginary part of r, evaluated in Appendix B. We have introduced in (59) the field strength parameter

$$R = \left(k_z \int_{-\infty}^{+\infty} E^2 dx\right)^{-1}. \tag{60}$$

In the above definition of R, the field is assumed to be unity at the guide-cladding boundary. For a dielectric slab, the value of R is given in (12). Equation (59) is the result used in the main text.

REFERENCES

1. J. A. Arnaud, "Transverse Coupling in Fiber Optics—Part I: Coupling Between Trapped Modes," B.S.T.J., *53*, No. 2 (February 1974), pp. 217–224; "Transverse Coupling in Fiber Optics—Part II: Coupling to Mode Sinks," B.S.T.J., *53*, No. 4 (April 1974), pp. 675–696.
2. D. Marcuse, "The Coupling of Degenerate Modes in Two Parallel Dielectric Waveguides," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1791–1816.
3. M. Born and E. Wolf, *Principles of Optics*, New York: Pergamon Press, 1965, p. 623.
4. A. W. Snyder, "Leaky-Ray Theory of Optical Waveguides of Circular Cross Section," Appl. Phys., *4*, No. 4 (September 1974), pp. 273–298.
5. A. H. Cherin and E. J. Murphy, "Quasi-Ray Analysis of Crosstalk Multimode Optical Fibers," B.S.T.J., *54*, No. 1 (January 1975), pp. 17–45.
6. N. S. Kapany and J. J. Burke, *Optical Waveguides*, New York: Academic Press, 1972.
7. E. A. J. Marcatili, "Bends in Optical Dielectric Guides," B.S.T.J., *48*, No. 5 (September 1960), pp. 2103–2132.
8. S. E. Miller, "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 4 (July 1955), pp. 807–822.
9. G. H. Wannier, "Probability of Violation of the Ehrenfest Principle in Fast Passage," Physics, *1*, No. 5 (May-June 1965), pp. 251–253.
10. J. S. Cook, "Tapered Velocity Couplers," B.S.T.J., *34*, No. 4 (July 1955), pp. 807–822.
11. M. G. F. Wilson and G. A. Teh, "Tapered Optical Directional Coupler," IEEE J. of Microwave Theory and Technique, *MTT-23*, No. 1 (January 1975), pp. 85–92.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

Faster-Than-Nyquist Signaling

By J. E. MAZO

(Manuscript received March 27, 1975)

The degradation suffered when pulses satisfying the Nyquist criterion are used to transmit binary data in noise at supraconventional rates is studied. Optimum processing of the received waveforms is assumed, and attention is focused on the minimum distance between signal points as a performance criterion. An upper bound on this distance is given as a function of signaling speed. In particular, the pulse energy seems to be the minimum distance up to rates of transmission 25 percent faster than the Nyquist rate, but not beyond.

Some mathematical aspects related to the above problem are also considered. In particular, the minimum distance is rigorously shown to be nonzero for all transmission rates. This is tantamount to showing that, in the singular case of linear prediction, perfect prediction cannot be approached with bounded prediction coefficients.

I. INTRODUCTION

The use of Nyquist pulses

$$g(t) = \frac{\sin(\pi t/T)}{(\pi t/T)}$$

to send binary (or multilevel) data without intersymbol interference over a channel of bandwidth $W = (1/2T)Hz$ is classic. If we assume that one receives the pulse train

$$u(t) = \sum_{n=N_1}^{N_2} a_n g(t - nT), \quad a_n = \pm 1, \text{ independently,} \tag{1}$$

in additive white gaussian noise of two-sided spectral density $N_0/2$, then the optimum detector has a bit-error rate P_e given by

$$P_e = Q\left(\frac{2\sqrt{E}}{\sqrt{2N_0}}\right), \tag{2}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy \equiv \tfrac{1}{2} \operatorname{erfc} \frac{x}{\sqrt{2}}, \tag{3}$$

erfc $(\cdot)$ denoting the co-error function, and E being the energy in the pulse $g(t)$. In our case, $E = T$. Asymptotically, for large signal-to-noise ratios, (2) becomes

$$P_e \sim \frac{1}{2}\sqrt{\frac{N_0}{\pi E}} \exp\left(-\frac{E}{N_0}\right). \tag{4}$$

We now address the following question: Suppose that in transmitting (1) we obtain a performance from (2) that is more than satisfactory. Thus, we may have a P_e of 10^{-6} or 10^{-7} when 10^{-5} would be adequate. To what extent can we trade this "excess performance" for speed by replacing T by $T' < T$ in (1), while keeping transmitted power constant? In other words, we still use pulses

$$g(t) = B \frac{\sin(\pi t/T)}{(\pi t/T)}, \tag{5}$$

but send them at intervals $T' < T$. We call this faster-than-Nyquist transmission and shall characterize T' by writing $T' = \rho T$, $0 < \rho < 1$. A particular motivation for this problem is to mathematically model, in a simple way, what would happen if voice-band telephone channels are "pushed" to their limits with more rapid transmission of pulses than has been conventional.

While simple detectors that match filter and sample can still be used for faster-than-Nyquist transmission, their performance is suboptimum.[1] We are concerned here with optimum detectors. Since exact analysis of nonlinear detectors is not presently feasible, we choose to give our detectors the benefit of the doubt and work rather with lower bounds to P_e. Nevertheless, interesting results can be obtained regarding the trade-off considered here. To see why degradation in error rate is inevitable, note that (2) is the well-known matched filter bound for antipodal pulses, each of energy E, which must bound performance for bit detection with a sequence of (perhaps interfering) pulses. On the other hand, as T' decreases, pulses are sent faster and the energy E in each pulse must be decreased in direct proportion so that the power E/T' is kept constant. This is an immediate unavoidable element in performance degradation, and may be regarded as a "fair" trade-off. Another cause of degradation is the degree to which the optimum detector can cope with the interference among pulses, i.e., the fact that the performance will drop below that of (2). Here, bounds other than (2) are useful, and in fact are the first item taken up in the next section.

II. DISCUSSION OF LOWER BOUND FOR ERROR RATE

Assuming (1) is received in white noise and an optimum detector is used for detecting the kth bit, a lower bound on the chance of making an error on this kth bit will now be derived. Since the data a_n are independent, this bound also serves for any sequence (1) starting at $n = N_1' \leqq N_1$, and ending at $n = N_2' \geqq N_2$. We begin with the fact that, for a binary hypothesis problem with equal *a priori* probabilities and having $p_+(x)$ or $p_-(x)$ as the two probability densities of the received signal x under the two respective hypotheses, one way[2] to write the probability of error is

$$P_e = \frac{1}{2} \int \min \left[p_+(x), p_-(x) \right] dx. \tag{6}$$

If we let $u_i^{\pm}(t)$ be a particular one of the equiprobable 2^N signals in (1), $N = N_2 - N_1$, which have ± 1 in the kth position, then formally

$$p_{\pm}(x) = \frac{1}{2^N} \sum_{i=1}^{2^N} p_{\pm}^i(x), \tag{7}$$

where $p_{\pm}^i(x)$ is the density of the observations conditioned on the entire sequence. Thus,

$$P_e = \frac{1}{2} \cdot \frac{1}{2^N} \cdot \int \min \left(\sum_{i=1}^{2^N} p_+^i(x), \sum_{j=1}^{2^N} p_-^j(x) \right) dx$$

$$\geqq \frac{1}{2} \frac{1}{2^N} \sum_{i=1}^{2^N} \int \min \left[p_+^i(x), p_-^{j(i)}(x) \right] dx. \tag{8}$$

In writing (8), we have made use of the fact that the minimum of two sums with an equal number of terms is at least as large as the sum of the minimum of the two ith terms of each series. Of course, each series can be arranged in any permuted order before the pair-wise minimum is taken and, thus, the pairings i with $j(i)$ are indicated in (8) to allow for this permutation. Now

$$\frac{1}{2} \int \min \left[p_+^i(x), p_-^{j(i)}(x) \right] dx \tag{9}$$

is the probability of error with two fixed signals and has the well-known evaluation

$$Q\left(\frac{d[i, j(i)]}{\sqrt{2N_0}} \right), \tag{10}$$

where

$$d^2(i, j) = \int_{-\infty}^{\infty} \left[u_+^i(x) - u_-^j(t) \right]^2 dt \tag{11}$$

is the "distance" between two sequences (1) which differ in the kth

position. Equation (8) then reads

$$P_e \geqq \frac{1}{2^N} \sum_{i=1}^{2^N} Q\left(\frac{d[i, j(i)]}{\sqrt{2N_0}}\right) \tag{12}$$

for any set of pairings $[i, j(i)]$. The bound (12) is intimately related to Forney's lower bound,[3] although our derivation is quite different. Forney's bound in the present situation reads

$$P_e \geqq p_m Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right), \tag{13}$$

where $d_{\min}$ is the minimum distance between signals (1) which differ in the kth position, and p_m is the probability that a sequence chosen at random has a sequence with opposite polarity in the kth position at distance $d_{\min}$. Equation (12) can be made to yield something like (13). Thus, in (12) discard all terms except for those pairings $[i, j(i)]$ such that $d[i, j(i)] = d_0$. Then (12) implies

$$P_e \geqq \frac{\text{no. of pairings}}{2^N} Q\left(\frac{d_0}{\sqrt{2N_0}}\right). \tag{14}$$

The coefficient in front of the Q function corresponds to the probability coefficient in (13). Choosing $d_0 = d_{\min}$ yields (13), but when we will not be able to find $d_{\min}$, eq. (14) will serve our purpose.

III. ESTIMATING THE MINIMUM DISTANCE

Clearly, in (14) we should like to find the smallest d_0 to maximize the lower bound, provided the coefficient is not too small. In our problem, $d_{\min}^2$ is given by

$$\frac{d_{\min}^2}{4E} = \inf_{N;\ \{a_l = \pm 1,\ 0\}} \frac{1}{2\pi\rho} \int_{-\rho\pi}^{\rho\pi} \left|1 - \sum_{l=1}^{N} a_l e^{il\theta}\right|^2 d\theta, \tag{15}$$

where we have normalized by dividing by the pulse energy E. The expression (15) comes from taking the Fourier transform of (11) and manipulating the resulting expression slightly. We note particularly that in (15) only positive values of l need be considered, since

$$\left|e^{iK\theta}\left(1 - \sum_{\substack{l=-K \\ l\neq 0}}^{M} a_l e^{il\theta}\right)\right|^2 = \left|1 - \sum_{l=1}^{M+K} b_l e^{il\theta}\right|^2$$

if $a_{-K} \neq 0$. We have set $b_l = -a_{-K}a_{s-K}$ if $l \neq K$ and $b_l = -a_{-K}$ if $l = K$.

We cannot claim to have found the minimum value of (15). However, a simple numerical effort has yielded the results for $d_0^2/4E$ shown in Fig. 1, where d_0 refers to the smallest distance we have found. We note in particular that d_0 is the pulse energy for ρ decreasing from 1

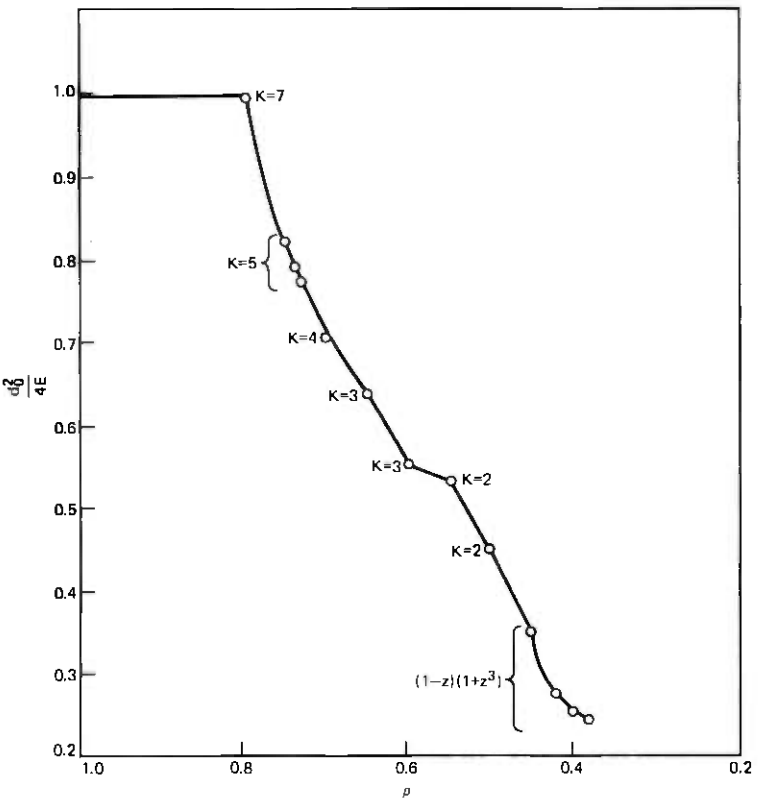


Fig. 1—The smallest distances between signal sequences that we have found are shown here for different values of signaling rate. Labeling a point by K indicates that the polynomial is

$$p(z) = 1 + \sum_{j=1}^{K} (-1)^j z^j.$$

to 0.8, or, in other words, for rates exceeding the Nyquist rate by 25 percent [percentage of excess = $100(1/\rho - 1)$]. Thus, $d^2_{\min}/4$ cannot be the pulse energy for $\rho < 0.8$ for this problem. By the time ρ has decreased to 0.5, $d_0^2/4E$ has dropped to 0.465. (G. J. Foschini has informed the author that the use of the polynomial $p(z) = 1 - z + z^3 - z^4 + z^6 - z^7$, $z = \exp(i\theta)$, results in the value 0.410 for $d_0^2/4E$ at $\rho = 0.5$.) Except for some points in the neighborhood of $\rho = 0.4$, the values for d_0^2 have been obtained by considering numerically the best value of K which minimizes, for not too large K,

$$\frac{1}{2\pi\rho} \int_{-\rho\pi}^{\rho\pi} \left| 1 + \sum_{l=1}^{K} (-1)^l e^{il\theta} \right|^2 d\theta. \tag{16}$$

These points are labeled with the appropriate value of K in Fig. 1.

Somewhat surprisingly, the larger values of K are responsible for decreasing d_0 initially ($K = 7$ at $\rho = 0.8$), and then K gradually becomes smaller ($K = 2$ at $\rho = 0.5$). The value obtained with $K = 1$ always was suboptimum, as was the limiting value of (16) when $K \to \infty$, which is easily shown to be

$$\frac{1}{\pi\rho} \tan \frac{\rho\pi}{2}. \tag{17}$$

Why were the sequences given in (16) deemed to be of interest in the first place? The most interesting reason stems from the following argument. If one considers the Fourier transform of a doubly infinite pulse sequence like (1) when pulses are being sent faster than Nyquist and when the special case of the alternating sequence $a_n = (-1)^n$ is being sent, one finds that the Fourier transform consists of delta functions spaced at all odd multiples of π/T', that is, the Fourier transform is out-of-band, which suggests zero received energy. Actually, the doubly infinite model and its δ-function Fourier transforms are idealizations representing limiting behavior for signals consisting of pulses extending from $(-N, N)$ and N becoming large. We are really concerned with limiting behavior of the energy contained in the frequency interval $(-\pi/T, \pi/T)$, with $T > T'$, and evidently for the present case, if $S_N(\omega)$ is the Fourier transform of the truncated pulse sequence,

$$\lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} |S_N(\omega)|^2 d\omega \neq \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} |\lim S_N(\omega)|^2 d\omega = 0. \tag{18}$$

In spite of the above subtlety, however, sequences which are alternating at least over part of their range are interesting and one might expect difficulty distinguishing between one such sequence and its negative.

In addition to the normalized distances given in Fig. 1, Fig. 2 plots the numerical values of lower bounds computed from expression (14), as well as the matched filter bound. These curves all assume constant power. Curves with initial ($\rho = 1$) error rates with 10^{-5} and 10^{-7} are chosen as examples in Fig. 2. In both cases, an order of magnitude of degradation in error rate is seen for a 25-percent increase in bit rate ($\rho = 0.8$) using only the matched filter bound. Decreasing ρ further on the 10^{-7} curve illustrates further degradations using (14) with an appropriate value of K. These bounds do not show a departure from the matched filter bound for as small a value of ρ as Fig. 1 would suggest, because the coefficient $1/2^K$ to be used in (14) swamps the effect of the decreasing "minimum" distance. For the 10^{-5} curve, this effect extends to even smaller ρ and no lower bound other than the matched filter one is shown for that case.

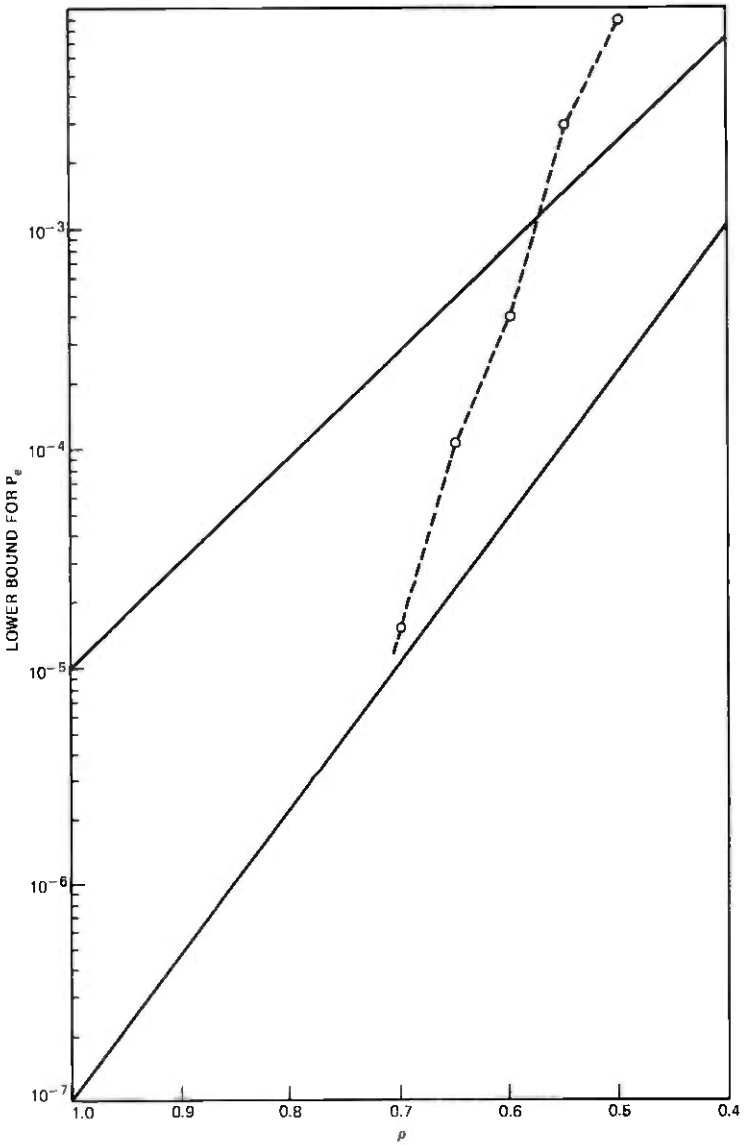


Fig. 2—Lower bounds on error rate vs signaling speed for two initial ($\rho = 1$) cases. The solid curves are both matched filter bounds. The dashed curve is based on minimum-distance considerations and applies to the 10^{-7} case. All curves are drawn for constant power.

IV. TWO MATHEMATICAL QUESTIONS

As we have already emphasized, the infimum of the right member of (15) over all the indicated trigonometric polynomials with ± 1, 0 coefficients is not displayed in Fig. 1. Figure 1 simply shows the

smallest values we have found. Next, we want rigorously to establish here that $d^2_{\min} \neq 0$ if $\rho \neq 0$. Note that this would not be the case if the coefficients a_l in (15) were allowed to be any real numbers. In fact, for any nonnegative function $f(\theta)$ with $\ln f(\theta) \in L_1(-\pi, \pi)$, we have the Szegö theorem[4] which states

$$\inf_{N;\, a_l \text{ real}} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) \left| 1 - \sum_1^N a_l e^{il\theta} \right|^2 d\theta = \exp \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\theta) d\theta. \quad (19)$$

Expressions such as (19) occur, in particular, in linear prediction theory.

In our case, $f(\theta) = 0$ if $|\theta| > \rho\pi$ and $\ln f(\theta)$ is not L_1; but the appropriate limit of (19) indicates zero to be the infimum, which is the correct answer.[4] Thus, there is some cause to wonder if $d^2_{\min}$ as defined in (15) is zero as well. We shall in fact show it is slightly more.

Theorem 1: *Let β be any positive (finite) real number and require $|a_l| \leqq \beta$, $l = 1, 2, \cdots$. Then*

$$\inf_{N;\, |a_l|} \frac{1}{2\pi} \int_{-\rho\pi}^{\rho\pi} \left| 1 - \sum_1^N a_l e^{il\theta} \right|^2 d\theta > 0, \quad \rho \neq 0. \quad (20)$$

Proof: We first note that if there exists a sequence $\{p_n(\theta)\}_{n=1}^{\infty}$ of trigonometric polynomials of the form

$$p_n(\theta) = \sum_{l=1}^{n} a_l(n) e^{il\theta}, \quad |a_l| \leqq \beta < \infty \quad (21)$$

such that

$$\frac{1}{2\pi} \int_{-\rho\pi}^{\rho\pi} |1 - p_n(\theta)|^2 d\theta \to 0, \quad (22)$$

then, for any $G(\theta) \in L_2(-\rho\pi, \rho\pi)$,†

$$\int_{-\rho\pi}^{\rho\pi} G(\theta) p_n(\theta) d\theta \to \int_{-\rho\pi}^{\rho\pi} G(\theta) d\theta. \quad (23)$$

This is simply a statement of the fact that if $p_n(\theta)$ converges strongly to unity, it also converges weakly to unity. Now it is easy to see from (23) and the form of $p_n(\theta)$ that

$$\beta \sum_1^{\infty} \left| \int_{-\rho\pi}^{\rho\pi} d\theta\, e^{in\theta} G(\theta) \right| \geqq \left| \int_{-\rho\pi}^{\rho\pi} G(\theta) d\theta \right|. \quad (24)$$

Or, in other words, if

$$\beta < \sup_{G(\theta) \in L_2(-\rho\pi, \rho\pi)} \frac{\left| \int_{-\rho\pi}^{\rho\pi} G(\theta) d\theta \right|}{\sum_1^{\infty} \left| \int_{-\rho\pi}^{\rho\pi} e^{in\theta} G(\theta) d\theta \right|}, \quad (25)$$

† In addition to $G(\theta) \in L_2(-\rho\pi, \rho\pi)$ it will sometimes be convenient to regard $G(\theta) \in L_2(-\pi, \pi)$ but having support confined to $(-\rho\pi, \rho\pi)$.

then (22) cannot be true. In particular, if (25) holds with $\beta \geq 1$, then $d^2_{\min}$ is strictly positive. Regarding $G(\theta) \in L_2(-\pi, \pi)$ but supported on $[-\rho\pi, \rho\pi]$, and calling

$$g(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\theta t} G(\theta) d\theta, \tag{26}$$

$$g_n \equiv g(n),$$

the right member of (25) contains the quantity

$$\frac{|g_0|}{\sum_1^\infty |g_n|}. \tag{27}$$

Clearly, we have a question concerning the sample values g_n at the nonnegative integers of a function whose bandwidth is strictly less than π. Normalizing (27) with $g_0 = 1$, (25) prompts the question: How small can $\sum_1^\infty |g_n|$ be? If it can be zero, then (25) would be true for any finite β. In fact, by Carlson's lemma,[5] which states that a band-limited function having a bandwidth less than π is uniquely determined by its sample values taken at integers along a half line, it follows that if $g_0 = 1$, then $\sum_1^\infty |g_n| \neq 0$. But Carlson's lemma does not say that $\sum_1^\infty |g_n|$ cannot be made arbitrarily small under these conditions. Lemma 1 (see below) shows that $\sum_1^\infty |g_n|$ can be arbitrarily small. Thus, the right member of (25) is infinity, implying the truth of Theorem 1.

An immediate corollary of Theorem 1 is that for the singular case of Szego's theorem [$f(\theta)$ vanishing on an interval] the infimum value of zero cannot be approached without using *unbounded* coefficients.

Lemma 1: *Let* $g(t)$ [*not identically zero and* $\in L_2(-\infty, \infty)$] *have Fourier transform* $G(\theta)$ *supported on* $(-\rho\pi, \rho\pi)$ *for some fixed* ρ, $0 < \rho < 1$. *Denote the samples of* $g(t)$ *at the integers by* g_n [*as in eq.* (26)], *and fix the normalization of* $g(t)$ *by setting* $|g_0| = 1$. *Then*

$$\inf \sum_1^\infty |g_n| = 0, \tag{28}$$

where the infimum is taken over all $g(t)$ *having the indicated properties.*

Proof: We begin with the simple, but crucial, remark that it is sufficient that there be, for any ρ, a function $h(t; \rho) \in L^2(-\infty, \infty)$ whose Fourier transform is supported on $(-\rho\pi, \rho\pi)$, such that $h(0, \rho) = 1$ and such that $\sum_1^\infty |h_n(\rho)|^2$ can be arbitrarily small.† This is sufficient,

† We are grateful to H. J. Landau for pointing this out. Landau has also supplied an independent proof of the above refinement to Carlson's lemma, which we give in the appendix.

because to make (27) large (for some fixed value of ρ) we would just need to take

$$g(t) = h^2\left(t, \frac{\rho}{2}\right) \tag{29}$$

for an appropriate $h(t, \rho/2)$. Clearly, $g(t)$ is band-limited to ρ and is $L^2(-\infty, \infty)$ because $h(t, \rho/2)$ is bounded:

$$h\left(t, \frac{\rho}{2}\right) = \frac{1}{2\pi} \int_{-\rho\pi/2}^{\rho\pi/2} H(\theta) d\theta \leqq \frac{1}{2\pi}\left(\rho\pi \cdot \int_{-\rho\pi/2}^{\rho\pi/2} |H(\theta)|^2 dt\right). \tag{30}$$

But can we really find an appropriate $h(t)$ such that

$$h_0 = 1, \quad \sum_1^\infty |h_n|^2 < \epsilon, \tag{31}$$

or, equivalently, can we find a real $h(t)$, band-limited to $(-\rho\pi, \rho\pi)$, such that

$$(h_0 - 1)^2 + \sum_1^\infty h_n^2 < \epsilon? \tag{32}$$

Indeed we can, and in fact the answer may be extracted from an article by Salz[6] which discusses mean-square decision feedback equalization. Salz, in Section V of his paper, considered the equalization problem for faster-than-Nyquist signaling. His minimization problem was of the form in (32) plus an added term for the noise variance; $h(t)$ corresponds to the output of the equalizer when one pulse of the form $\sin \rho\pi t/\rho\pi t$ is the input. He remarks, in the last sentence on page 1354 of his paper, that the quantity that corresponds to (32) plus added output noise variance goes to zero as the input noise variance decreases. Hence, if we choose $h(t)$ to be the output pulse of a decision-feedback equalizer whose taps have been optimized for the case of sufficiently small input noise, then (32) will be sufficiently small. Thus, Lemma 1 is proven.

The second question we discuss in this section is the rapidity with which the minimum distance decreases as ρ approaches zero. We develop this in Theorem 2.

Theorem 2:

$$\lim_{\rho \to 0} \frac{d_{\min}^2(\rho)}{\rho^k} = 0 \quad \textit{for any } k > 0. \tag{33}$$

Proof: The proof is a simple construction. Consider the polynomials

$$P_L(z) = \prod_{i=0}^{L} (1 - z^{2^i}). \tag{34}$$

Clearly, $P_L(z)$ has a zero of order $(L + 1)$ at $z = +1$, and has ± 1 coefficients, with $P_L(0) = +1$. Now, for small ρ, the $(L + 1)$st order

zero at $z = 1$ implies

$$\frac{1}{2\pi}\int_{-\rho\pi}^{\rho\pi} |P_L(e^{i\theta})|^2 d\theta = 0(\rho^{2L+3}) \tag{35}$$

for all integer L. Equation (33) follows immediately.

Short of finding $d^2_{\min}$ exactly, there are a few mathematical questions that suggest themselves and that may be less difficult than the full problem. Thus, Fig. 1 prompts one to ask if there is a neighborhood of $\rho = 1$, where $d^2_{\min}/4$ is the pulse energy? Another question has to do with pulse design. Given that $G(\theta)$ is symmetric, positive, L_2, and supported on $(-\rho\pi, \rho\pi)$, is $G(\theta) =$ constant the best choice to maximize the minimum distance (subject to fixed pulse energy)?

V. ACKNOWLEDGMENTS

It is a pleasure to acknowledge many helpful discussions with M. A. Kaplan, H. J. Landau, B. F. Logan, and H. O. Pollak during this work. We acknowledge the contribution of B. F. Logan who supplied an early proof that $d^2_{\min} \neq 0$ if $\rho > \frac{1}{2}$, approximately.

APPENDIX

Landau's Proof

In Section IV we present another proof that

$$\sup \frac{g_0^2}{\sum_{n=1}^{\infty} |g_n|^2} = \infty, \tag{36}$$

where the sup is taken over all $g(t) \in L^2(-\infty, \infty)$, which are band-limited to $(-\rho\pi, \rho\pi)$. Our proof in the text relied on the published results of work by Salz.[6] Here we give a self-contained, but more mathematical, proof of (36) which was developed by H. J. Landau.

Suppose (36) is not true, i.e., suppose that

$$\sum_1^{\infty} |g_n|^2 \geqq \frac{g_0^2}{k} > 0 \quad \text{for all } g(t) \text{ of BW} = \rho\pi. \tag{37}$$

Then,

$$|g_0|^2 \leqq k \sum_1^{\infty} |g_n|^2. \tag{38}$$

From Carlson's lemma, g_0 is a linear functional on the l_2 sequence $\{g_1, g_2, \cdots, g_k, \cdots\}$ and, from (38), this linear functional is bounded. Therefore, by the standard Riesz representation† for bounded linear

† Not all l_2 sequences $\{g_i\}$ give rise to an appropriate $g(t)$, and hence, the linear functional g_0 is not defined on all of l_2. Therefore, before using the Riesz theorem, the Hahn-Banach theorem should be invoked to extend g_0 to a bounded linear functional on all of l_2.

functionals, we may write

$$g_0 = \sum_1^\infty b_\nu g_\nu, \quad \sum_1^\infty b_\nu^2 < \infty, \tag{39}$$

where the b_ν do not depend on $g(t)$. We now consider the function

$$p(z) = 1 - \sum_1^\infty b_n z^n, \tag{40}$$

which is analytic for $|z| < 1$. For any $G(\theta) \in L_2(-\rho\pi, \rho\pi)$, we may write, using (39),

$$\begin{aligned} \int_{-\rho\pi}^{\rho\pi} G(\theta)d\theta &= \sum_1^\infty b_n \int_{-\rho\pi}^{\rho\pi} e^{in\theta}G(\theta)d\theta \\ &= \int_{-\rho\pi}^{\rho\pi} \left(\sum_1^\infty b_n e^{in\theta}\right)G(\theta)d\theta. \end{aligned} \tag{41}$$

Therefore,

$$\lim_{|z|\to 1} \int_{-\rho\pi}^{\rho\pi} \left(1 - \sum_1^\infty b_\nu z^\nu\right)G(\theta)d\theta = 0 \tag{42}$$

for all $G(\theta) \in L_2(-\rho\pi, \rho\pi)$.† By the completeness of L_2, we must have $1 - \sum_1^\infty b_n e^{in\theta} = 0$ a.e. on $(-\rho\pi, \rho\pi)$. Since the radial limit of the H_2 function $p(z)$ vanishes on a set of positive measure, $p(z)$ itself must vanish for $|z| < 1$. (See Ref. 7, p. 373, Theorem 17.18.) However, $p(0) = 1$, and, hence, we have a contradiction, denying the validity of (37).

REFERENCES

1. B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," IEEE Trans. Inform. Theory, *IT-14*, No. 4 (July 1968), pp. 563–568.
2. T. T. Kadota and L. A. Shepp, "On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, *IT-13*, No. 2 (April 1967), Appendix A.
3. G. D. Forney, "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," IEEE Trans. Com., *COM-20*, No. 1 (February 1972), pp. 76–77.
4. Ulf Grenander and Gabor Szego, *Toeplitz Forms and Their Applications*, Berkeley: University of California Press, 1958, Chapter 3.
5. E. C. Titchmarsh, *Theory of Functions*, 2nd ed., London: Oxford University Press, 1952, p. 186, Section 5.81.
6. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–1373.
7. W. Rudin, *Real and Complex Analysis*, 2nd ed., New York: McGraw-Hill, 1974.

† This is a simple application of Ref. 7, page 366, Theorem 17.10 supported by the fact that strong convergence in L_2 implies weak convergence.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

Single-Integration, Adaptive Delta Modulation

By P. CUMMISKEY

(Manuscript received March 11, 1975)

An estimate of optimum performance is derived for a single-integration, adaptive delta modulator. Several simulations of adaptive delta modulators with single integrators have all produced signal-to-noise ratios near or below the estimate.

The derivations presented here indicate that the performance of a single-integration delta modulator is dependent on the correlation between adjacent samples of the input signal and on the probability density function of its derivative. The relationship between the probability density of the derivative of the input signal and optimum performance, in turn, explains why signal-to-noise ratios taken on sine waves are greater than those recorded while processing speech signals.

I. INTRODUCTION

In this paper, an equation is derived for the optimum signal-to-noise ratio (s/n) of a single-integration, adaptive delta modulator. Mean-square quantizing noise is a mathematically tractable quantity which appears to be a reasonably good measure of overall performance. It was felt that an understanding of the relationships between this quantity and the character of the input signal would be useful. The derivations and data presented here all contribute to this end. Other practical considerations, such as subjective evaluation,[1] transmission errors,[2] and tandem encoding,[3] have been discussed elsewhere.

Several simulations[4] of single-integration, adaptive delta modulators on a variety of speech signals have produced s/n's near or below the performance estimate suggested in this paper. It is further suggested that this estimate is very close to the upper bound on the performance of such coders. The s/n formula also provides an explanation of the disparities between s/n's taken on sine waves and those obtained while coding speech signals.

A block diagram of a single-integration, adaptive delta modulator is shown in Fig. 1. At the encoder, the difference between an input

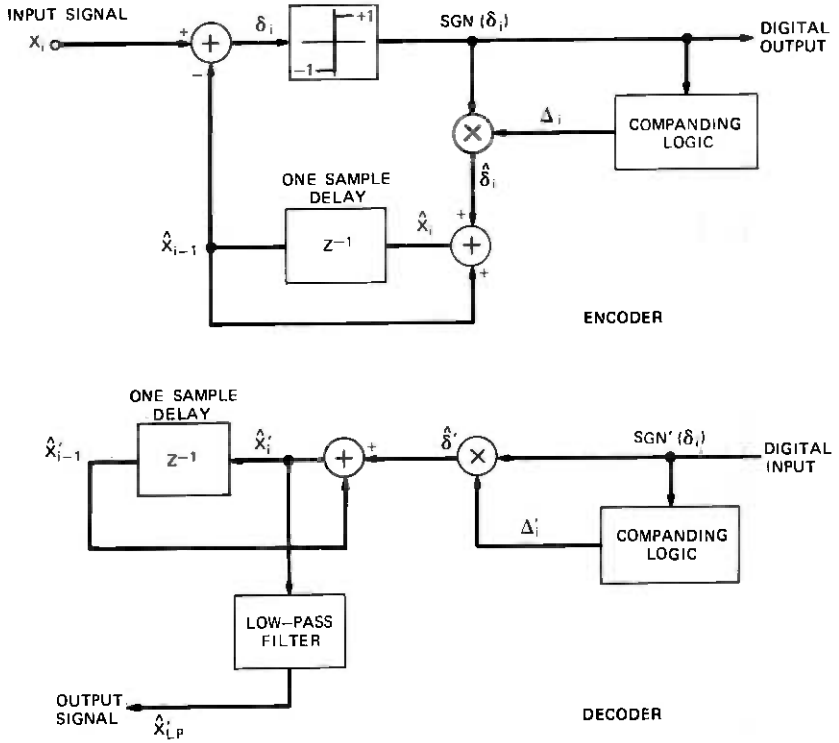
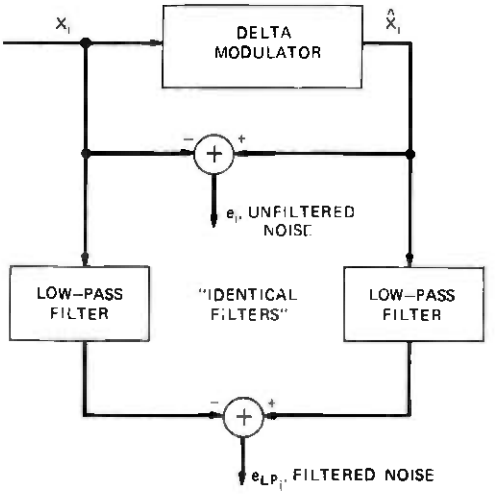


Fig. 1—Single-integration delta modulation.

sample, x_i, and the previous output sample, $\hat{x}_{i-1}$, is quantized to one of two levels and coded. The code symbols, sgn (δ_i) through sgn (δ_{i-N}) (where N may be any positive integer) are then interrogated by the companding logic, and the step size, Δ_{i-1}, is altered before the ith sample is encoded. The quantized approximation to the difference, $\hat{\delta}_i = \Delta_i$ sgn (δ_i), is added to the previous output to obtain the present output sample.

The decoder operates in the same manner as the encoder except that the circuit is driven from the transmission channel rather than from a local comparator. The quantized signal at the decoder, $\hat{x}'_i$, is low-pass filtered to eliminate noise components outside the band of x_i (i.e., frequencies greater than f_{LP}), and a replica of the input signal is thus regenerated at the desampling filter.

The signal-to-noise ratios referred to in this paper were taken in the following manner. First the noise was obtained as shown in Fig. 2 and then the ratio of input signal power to noise power was taken. The technique used by DeJager for sine wave s/n's is described in Ref. 5.



$$s/n = \frac{\sum_{i=1}^{N} X_i^2}{\sum_{i=1}^{N} e_i^2}$$

$$s/n_{LP} = \frac{\sum_{i=1}^{N} X_i^2}{\sum_{i=1}^{N} e_{LP_i}^2}$$

N = TOTAL NUMBER OF SPEECH SAMPLES OR 1500 TO 3000 SINE WAVE SAMPLES.

Fig. 2—Quantizing noise measurement.

II. EXACT S/N FORMULAS

The following equations are derived from the diagram in Fig. 1.

$$\delta_i = x_i - \hat{x}_{i-1} \tag{1}$$

$$\hat{x}_i = \hat{x}_{i-1} + \hat{\delta}_i. \tag{2}$$

If the quantizing error is defined as

$$e_i \equiv \hat{x}_i - x_i, \tag{3}$$

then, from (1) and (2), the following relationship holds:

$$e_i = \hat{\delta}_i - \delta_i. \tag{4}$$

From (3), it can be concluded that

$$\hat{x}_i = x_i + e_i, \tag{5}$$

and likewise that

$$\hat{x}_{i-1} = x_{i-1} + e_{i-1}. \tag{6}$$

Therefore, (1) may be rewritten as

$$\delta_i = x_i - x_{i-1} - e_{i-1}. \tag{7}$$

The average power in the prediction error is therefore

$$E(\delta_i^2) = E(x_i^2) + E(x_{i-1}^2) + E(e_{i-1}^2) - 2E(x_i x_{i-1}) - 2E(x_i e_{i-1}) + 2E(x_{i-1} e_{i-1}), \tag{8}$$

where the E functions are expected or average values. It is now noted

that, for quasi-stationary signals,

$$E(x_i^2) = E(x_{i-1}^2),^* \tag{9}$$

and that

$$E(e_{i-1}^2) = E(e_i^2).^* \tag{10}$$

Therefore, eq. (8) may be reduced to

$$\frac{E(\delta_i^2)}{E(x_i^2)} = 2\left[1 - \frac{E(x_i x_{i-1})}{E(x_i^2)} - \frac{E(x_i e_{i-1})}{E(x_i^2)} + \frac{E(x_{i-1} e_{i-1})}{E(x_i^2)}\right] + \frac{E(e_i^2)}{E(x_i^2)}. \tag{11}$$

The s/n at the quantizer is given as

$$\text{s/n}_Q = \frac{E(\delta_i^2)}{E[(\hat{\delta}_i - \delta_i)^2]} = \frac{E(\delta_i^2)}{E(e_i^2)}. \tag{12}$$

The s/n before filtering is defined as

$$\text{s/n} = \frac{E(x_i^2)}{E(e_i^2)}. \tag{13}$$

Note that (11) is equal to (12) divided by (13) or that

$$\frac{E(\delta_i^2)}{E(x_i^2)} = \frac{\text{s/n}_Q}{\text{s/n}}. \tag{14}$$

Hence, by substituting into (11) and transposing terms, an equation for the unfiltered s/n is obtained.

$$\text{s/n} = \frac{\text{s/n}_Q - 1}{2[1 - [E(x_i x_{i-1})]/E(x_i^2) - [E(x_i e_{i-1})]/E(x_i^2) + [E(x_{i-1} e_{i-1})]/E(x_i^2)]}. \tag{15}$$

III. ASSUMPTIONS AND APPROXIMATE FORMULAS

The variance of the prediction error is unknown because δ_i contains quantizing noise [see (7)]. Therefore, δ_i cannot be optimally quantized.

No meaningful information can be obtained directly from eqs. (1) through (15) without making some approximations or assumptions about the unknown terms [s/n_Q, $E(x_i e_{i-1})$ and $E(x_{i-1} e_{i-1})$]. Several measurements and simulations taken by the author and others before him support the following assumptions.

(*i*) The optimum step size will yield the same signal-to-noise ratio at the quantizer that can be achieved by quantizing the noise-free part of δ_i (i.e., the derivative of the input signal, $x_i - x_{i-1}$).

---

* To the extent that (9) and (10) are equations, (15) may be called an equation. Some awkward anomalies exist with regard to eq. (15); however, none of these is relevant to the problem.

(*ii*) The quantizing noise is the same as that generated by optimum quantization of $(x_i - x_{i-1})$, and therefore

$$E[(x_i - x_{i-1})e_{i-1}] = 0. \tag{16}$$

Hence,

$$E(x_i e_{i-1}) - E(x_{i-1}e_{i-1}) = 0. \tag{17}$$

Given the above assumptions, (15) reduces to

$$\text{s/n} = \frac{(\text{s/n}_{qop} - 1)}{2\{1 - [E(x_i x_{i-1})]/E(x_i^2)\}}, \tag{18}$$

where s/n_{qop} is the s/n achieved when $x_i - x_{i-1}$ is optimally quantized.

(*iii*) Finally, in an optimum modulator the quantizing noise spectrum is flat. Then the ratio of overall noise to the inband noise is equal to the ratio of half the sampling frequency to the bandwidth of the input signal.

Hence, the s/n taken on the filtered signal, $\hat{x}'_{LP}$, is equal to the unfiltered s/n multiplied by the ratio of half the sampling frequency to the cutoff frequency of the filter.

$$\text{s/n}_{LP} = \frac{[\text{s/n}_{qop} - 1]\left[\dfrac{f_s}{2f_{LP}}\right]}{2\left[1 - \dfrac{E(x_i x_{i-1})}{E(x_i^2)}\right]}, \tag{19}$$

where f_s is the sampling rate and f_{LP} is the cutoff frequency of the desampling filter or the bandwidth of the input signal.

Equation (19) is identical to Nitadori's signal-to-noise equation[6] for differential PCM. Nitadori cautions against its use in cases where the quantization is coarse, however. In this paper, eq. (19) is derived using somewhat different assumptions which, in fact, do appear to hold for delta modulation.

The validity of the three assumptions given above is the main point of this paper. When these assumptions hold, an important relationship between the amplitude distribution of the derivative of the input signal and s/n performance can be drawn.

IV. RELATIONSHIP BETWEEN S/N_{qop} AND PROBABILITY DENSITY FUNCTION OF $x_i - x_{i-1}$

Paez and Glisson,[7] among others, have shown that the amplitude probability distribution of speech and its derivatives is closely approximated by the gamma distribution. Figure 3 shows that this dis-

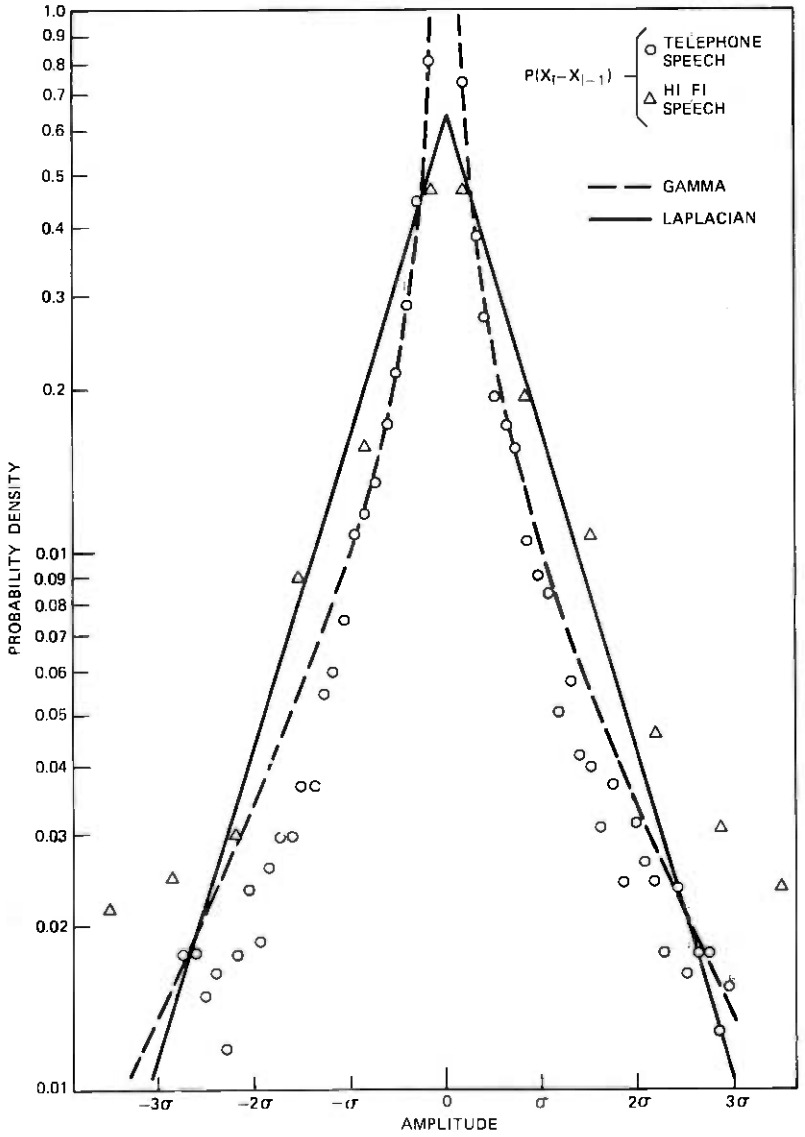


Fig. 3—The amplitude probability density function of $(x_i - x_{i-1})$ as compared with the Laplacian and gamma distributions.

tribution closely approximates the probability distribution of $(x_i - x_{i-1})$ for telephone signals used in my simulations. The distribution of $(x_i - x_{i-1})$, taken on the speech used in Jayant's simulations, lies closer to a Laplacian distribution, however.

Table I — Signal-to-noise ratios of two-level quantizer output

Probability Density Function		s/n$_{qap}$
Gamma	$P(y) = \frac{3^{\frac{1}{4}} \exp(-\sqrt{3}\lvert y\rvert/2\sigma)}{\sqrt{8\pi\sigma\lvert y\rvert}}$	1.50
Laplacian	$P(y) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}\lvert y\rvert}{\sigma}\right)$	2.00
Gaussian	$P(y) = \frac{\exp(-y^2/2\sigma^2)}{\sqrt{2\pi}\sigma}$	2.75
Rectangular	$P(y) = \frac{1}{A} - \frac{A}{2} < y < \frac{A}{2}$ $P(y) = 0 - \frac{A}{2} > y > \frac{A}{2}$	4.00
Sinusoidal $y = \cos\theta$ or $y = \sin\theta$	$P(y) = \frac{1}{\pi\sqrt{1-y^2}} - 1 \leqq y \leqq 1$	5.28

Given a distribution that is symmetrical about the origin, the quantization step is optimum when

$$\int_0^\infty (y - \Delta)P(y)dy = 0, \tag{20}$$

where y relates to $(x_i - x_{i-1})$. With the step set at the optimum size, Paez and Glisson, Max,[8] and others have calculated the noise power at the output of a two-level quantizer,

$$E[(\hat{y} - y)^2] = 2\int_0^\infty (y - \Delta)^2 P(y)dy, \tag{21}$$

and achieved the s/n's shown in Table I.

V. COMPARISONS WITH SIMULATIONS

The correlations between adjacent samples was taken on speech obtained using a carbon-button, telephone transducer. Similar data were obtained by N. S. Jayant on speech recorded from a high-fidelity transducer. Both signals were processed by Jayant's adaptive delta modulator with a one-bit memory, where the step size is multiplied by 1.5 if the present and previous code words, sgn (δ_i) and sgn (δ_{i-1}), are alike, or by 0.66 if they differ. In all the simulations, the sampling and desampling filter cutoff frequencies are set at 3.3 kHz, except for the telephone speech recorded at 24 kHz. In this case, the cutoff was

reduced to 3 kHz. The telephone signals sampled at 48 kHz were also encoded by a single-integration delta modulator designed by D. E. Blahut.[9] Blahut's encoder also performs close to the estimate [eq. (19)], when processing telephone speech. Among the numerous coders tested, no single-integration delta modulator was found that performs significantly better than Blahut's or Jayant's.

In Table II, performance estimates based on eq. (19) are compared with the s/n's obtained using Blahut's and Jayant's delta modulators. To account for difference in probability density functions (see Fig. 3), the estimates were made with s/n_{qop} equal to 1.5 for telephone signals, and 2.0 for high-fidelity signals.

The performance estimate given by eq. (19) is within 3.3 dB of the s/n's obtained in simulations with Jayant's delta modulator. The s/n's taken on Jayant's and Blahut's coders, while processing telephone speech at 48 kHz and 24 kHz, are essentially equal to the estimate. In these cases, the signal level was carefully adjusted until optimum performance was obtained, then further data were taken to verify eq. (19). (See Table III.)

The results shown in Table III lend great support to the approximations made in deriving eq. (19). The noise terms do effectively cancel, leaving a residue that is at least an order of magnitude smaller than the noise-free terms in the denominator of (15) (see lines 5 and 6 in Table III). The estimates for noise rejection at the desampling filter and for quantizer performance (s/n_Q) are within 0.8 dB of the figures obtained in the simulations.

Both coders were simulated with a 60-dB range of step sizes, and both were started with the step size equal to the minimum and the

Table II — Performance estimates

s/n_{qop}	Sampling Rate (kHz)	$\frac{E(x_i x_{i-1})}{E(x_i^2)}$	Estimate $10 \log_{10}$ (s/n_{LP}) (dB)	Delta Modulator Performance (dB)	
				Jayant's	Blahut's
2.0	20	0.989	21.3	18.0	—
*1.5	24	0.957	13.7†	14.5	—
2.0	40	0.997	30.0	28.0	—
*1.5	48	0.9897	22.6	22.9	22.7
2.0	60	0.999	36.5	34.0	—

* Telephone speech: The acoustic-to-electronic response of the new 500-type, stations sets[10] indicates that signal components in the 100-Hz to 3.3-kHz band are differentiated, and that components below 100 Hz are severely attenuated. Hence, correlation between adjacent samples is lower for telephone speech than for high-fidelity speech.

† At 24-kHz sampling, $f_{LP} = 3$ kHz.

Table III — Verification of eq. (19)

	Sampling Frequency (kHz)	Estimates	Coder Performance	
			Jayant's	Blahut's
s/n_{LP}	24	13.7 dB	14.5 dB	—
	48	22.6 dB	22.9 dB	22.7 dB
s/n	24	5.92 (i.e., 7.72 dB)	5.95 (i.e., 7.75 dB)	—
	48	24.3 (i.e., 13.9 dB)	24.3 (i.e., 13.9 dB)	26.1 (i.e., 14.2 dB)
Noise rejection at the desampling filter $\approx 10 \log_{10}(f_s/2f_{LP})$	24	6.0 dB	6.8 dB	—
	48	8.7 dB	9.0 dB	8.5 dB
s/n_Q	24	1.5	1.531	—
	48	1.5	1.555	1.569
$\left[1 - \frac{E(x_i x_{i-1})}{E(x_i^2)}\right]$	24	—	0.0422	—
	48	—	0.0103	0.0103
$\left[\frac{E(x_i e_{i-1})}{E(x_i^2)} - \frac{E(x_{i-1} e_{i-1})}{E(x_i^2)}\right]$	24	0	−0.00238	—
	48	0	−0.00112	−0.00060

predictor voltage equal to zero. As the average input signal level was varied over a 40-dB range, it was found that the s/n varied by 3 dB. In either coder, it was found that when performance fell significantly below the estimate (19), the following phenomena were observed:

(*i*) Quantizer performance and unfiltered s/n changed *slightly* (in some cases, these parameters increased in value).
(*ii*) The noise terms no longer effectively canceled.
(*iii*) There was a dramatic reduction in noise rejection at the desampling filter. It appears that when the correlation between the difference signal, $(x_i - x_{i-1})$, and the noise (16) becomes significant, more noise *must* shift into the passband of the desampling filter.

Hence, the approximations used in deriving eq. (19) do appear to describe the optimum condition.

These results have been obtained using both an HP2100A minicomputer and an IBM 370, and therefore are repeatable. Moreover, further validation by others using other encoders is desirable.

VI. SINE WAVE PERFORMANCE

Another interesting check on the theory is the fact that it explains why researchers everywhere achieve much higher s/n with sine wave inputs than with speech signals. DeJager's formula [see eq. (22)] indicates that the s/n taken on a sine wave at any frequency below 3 kHz is greater than the s/n that we predict or obtain for telephone speech.

$$\mathrm{s/n_{DeJager}} = (0.04)\,\frac{f_s^3}{f^2 \cdot f_{LP}}, \tag{22}$$

where f is the frequency of the input sine wave.

The amplitude probability distribution and $\mathrm{s/n}_{qop}$ for a sine wave were given in Table I. Substitution of the value in Table I into eq. (19) yields an estimate for sine wave s/n's.

$$\mathrm{s/n_{sine\ wave}} = \frac{4.28\left(\frac{f_s}{2f_{LP}}\right)}{2\left[1 - \frac{E(x_i x_{i-1})}{E(x_i^2)}\right]}. \tag{23}$$

Equation (23), in turn, is approximately equivalent to DeJager's formula. This relationship can be shown as follows. Let $x = \sin(2\pi ft)$; then

$$\frac{E(x_i x_{i-1})}{E(x_i^2)} = \frac{\int_0^{1/f} [\sin(2\pi ft)] \cdot \sin(2\pi ft + 2\pi f/f_s)dt}{\int_0^{1/f} \sin^2(2\pi ft)dt}, \tag{24}$$

or

$$\frac{E(x_i x_{i-1})}{E(x_i^2)} = \cos\left(\frac{2\pi f}{f_s}\right). \tag{25}$$

If the delay angle, $(2\pi f/f_s)$, is sufficiently small, then

$$2\left[1 - \cos\left(\frac{2\pi f}{f_s}\right)\right] \approx 1 - \cos^2\left(\frac{2\pi f}{f_s}\right) \approx \frac{4\pi^2 f^2}{f_s^2}. \tag{26}$$

When (26) is substituted into (23), we obtain something very close to DeJager's formula:

$$\mathrm{s/n_{sine\ wave}} = (0.054)\,\frac{f_s^3}{f^2 f_{LP}}. \tag{27}$$

For $f = 800$ Hz, $f_{LP} = 3.3$ kHz, and $f_s = 48$ kHz, estimates of 33.5 and 34.7 dB are obtained using (22) and (27). Under these same conditions, signal-to-noise readings of 26 to 27 dB were obtained in simula-

tions of Jayant's delta modulator. Under similar conditions, DeJager[5] obtained a maximum s/n of about 30 dB on a linear delta modulator.

VII. CONCLUSIONS

The optimum performance of Blahut's and Jayant's delta modulators is very close to the estimate, (19), when processing speech signals. Further experimentation with step-size compandors, without a change in the prediction technique, will not produce significantly higher signal-to-noise ratios. Equation (19) applies to a delta modulator with a single, ideal integrator; therefore, it does not preclude improvements through the use of fixed, higher-order networks.

In addition, it has been shown that delta modulator performance is dependent on the amplitude probability distribution of the derivative of the input signal. This dependence should be tested on a variety of signals and probability density functions. The theory also implies that a relationship exists between the amplitude distributions of differential waves at the input and optimum s/n, when higher-order networks are used.

Finally, I wish to call attention to the fact that the s/n performance of a delta modulator is significantly less for telephone signals than for low-pass filtered, high-fidelity signals, or for sine waves.

VIII. ACKNOWLEDGMENTS

I would like to thank D. E. Blahut and J. A. Miller for supplying me with working programs. Acknowledgments must also be extended to D. J. Hunsberger for instructing me on the use of the Hewlett Packard 2100A computer. H. J. Fletcher, D. E. Blahut, and N. S. Jayant have engaged in many useful discussions from which I have drawn insights. N. S. Jayant must also be thanked for supplying data related to his simulations. W. R. Daumer's efforts in collecting data for subjective tests have also provided useful information. I should also like to thank N. S. Jayant and C. E. Nahabedian for their help with the manuscript.

REFERENCES

1. N. S. Jayant and A. E. Rosenburg, "The Preference of Slope-Overload to Granularity in Delta Modulation of Speech," B.S.T.J., *50*, No. 10 (December 1971), pp. 3117–3125.
2. P. T. Nielsen and J. L. Flanagan, "Effects of Channel Errors and Adaptive Delta Modulation Systems," unpublished work.
3. N. S. Jayant and K. Shipley, "Multiple Delta Modulation of a Speech Signal," Proc. IEEE (lett.), *59* (September 1971), p. 1382.
4. N. S. Jayant, "An Adaptive Delta Modulator with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.

5. F. DeJager, "Delta Modulation, A Method of PCM Transmission Using a One-Unit Code," Phillips Research Report, 7 (1952), pp. 442–446.
6. K. Nitadori, "Statistical Analysis of ΔPCM," Electronics and Communications in Japan, *48*, No. 2 (February 1965), pp. 17–26.
7. M. D. Paez and T. H. Glisson, "Minimum Mean Square Quantization in Speech, PCM and DPCM Systems," IEEE Trans. on Comm., *20*, No. 2 (April 1972), pp. 225–230.
8. J. Max, "Quantization for Minimum Distortion," IRE Trans. Information Theory, *IT-6* (March 1960), pp. 7–12.
9. D. E. Blahut, U.S. Patent 3,784,922.
10. A. H. Iglis and W. L. Tuffnell, "An Improved Telephone Set," B.S.T.J., *30*, No. 1 (January 1951), pp. 239–270.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

A Low-Bit-Rate Interframe Coder for Videotelephone

By B. G. HASKELL and R. L. SCHMIDT

(Manuscript received January 7, 1975)

It has been suggested that customers for videotelephone service may be more interested in graphical information and in views of stationary objects than in head-and-shoulder views of people engaged in conversation. For this reason, an interframe coder simulation was constructed of a system that transmits graphics with full seven-bit PCM resolution, but displays scenes containing much movement with visible smearing in the moving areas.

With the coder operating at 200 kb/s (0.1 bit per pel for a 1-MHz signal), a very usable (somewhat reduced-resolution) graphics picture can be transmitted in about one-half second, which is about as fast as the human eye can assimilate the information. A full-resolution picture is built up after 3 to 5 seconds but, except for high-detail scenes, it is very difficult to tell the difference between the half-second picture and the 5-second picture.

Head-and-shoulder views of people engaged in low-key conversations are transmitted with quite adequate picture quality. Moving lips appear somewhat smeared, but it may not be enough to be objectionable if the audio is suitably delayed. However, large area movement is very visibly smeared—even to the point of being unrecognizable at moderate speeds. Whether or not this feature makes the coder unusable depends upon the value the user places on high-quality animated face-to-face conversation.

Briefly, the coder works as follows: First, the signal is temporally prefiltered. Then moving-area pels are sent as line-to-line differences of frame-to-frame differences. As the buffer fills, field-to-field, pel-to-pel, and frame-to-frame subsampling as well as adaptive quantization are brought in as needed to reduce the data rate.

I. INTRODUCTION AND SUMMARY

The use of videotelephone for graphical information and for views of stationary objects has profound implications in long-distance transmission of video signals via frame-to-frame coding, where the required

data rate is directly dependent on the amount of movement to be accommodated in the scene.

If it can be shown that visible degradation of moving areas in a television picture is not detrimental to the effectiveness of visual communication, then a significant saving in transmission costs is possible with frame-to-frame coding.[1] With these techniques, stationary areas of pictures would be transmitted with full resolution, while moving areas would be sent with visibly reduced resolution.

Transmission of graphics or still pictures can be accomplished in a particularly pleasing way, subjectively speaking. A reduced-resolution, but quite recognizable, picture appears very quickly at the receiver. Full resolution is then built up over a period of time that depends on the transmission channel data rate. However, for the majority of pictures, it is difficult for an observer to tell the difference between the full resolution picture and the earlier-appearing reduced-resolution picture. In this regard, such a system would be much more usable for interactive visual communication than would a facsimile or slow-scan system operating at the same data rate where a complete picture would not be visible for a relatively long time. Also, interframe coding can handle small amounts of movement such as adding a few lines to a sketch or using a pointer with stationary graphics, whereas a slow-scan system would be very unsatisfactory.

Scenes of people engaged in conversation do not fare as well as scenes in which there is little or no movement. Moving areas such as a person's lips and eyes are visibly smeared and, depending on the data rate, large-area movement is jerky because of coder overload. Even so, a well-behaved subject can present a very decent picture to the receiver if he or she is aware of the limitations of the medium. However, it is this aspect of low-bit-rate interframe coding that raises questions in most people's minds. Whether or not this feature makes such techniques unusable depends upon the value the user places on high-fidelity, animated, face-to-face conversation.

To move closer to the answers to some of these questions, an interframe coder simulation was constructed for 1-MHz videotelephone signals that was designed to operate in the hundreds of kilobits per second range (below $\frac{1}{4}$ bit per picture element). Many techniques are used to adaptively reduce the moving-area resolution (both spatial and temporal) in proportion to the amount of motion, and to restore full resolution to the display as quickly as possible after motion ceases.

With such a system operating at 200 kb/s (0.1 bits/pel), a recognizable, somewhat reduced-resolution graphics picture is displayed at the receiver in about one-half second. Full resolution requires 3 to 5 seconds. With head-and-shoulder views of people engaged in low-key

conversation, moving lips appear somewhat smeared, but this may not be enough to be objectionable if the audio is suitably delayed. However, large-area movement is very visibly smeared and jerky, even to the point of being unrecognizable at moderate speeds.

With the system operating at 50 kb/s, a reduced-resolution graphics picture requires about 2 seconds for transmission, while full resolution takes 10 to 15 seconds. At 50 kb/s, face-to-face conversation loses much of its naturalness. Lip motion can be followed only if the subject remains otherwise absolutely still, and large area motion is portrayed as a series of snapshots occurring at a rate of about 1 per second. It is interesting to note, however, that, even at 50 kb/s, useful interactive visual communication is still possible using interframe coding whereas, with slow-scan operating at the same data rate and requiring about 10 seconds per frame for transmission, interactive communication is severely hampered.

In the following sections, the technical aspects of the coder and the simulation are discussed.

II. MULTIMODE CONDITIONAL REPLENISHMENT

It is well known that, in a television signal, successive frames are very much alike. The frame-to-frame differences are negligibly small except in areas of the picture that contain moving objects. Thus, if frame memories are provided at the transmitter and receiver of a video communication system, it is necessary only to transmit those areas of each frame where the frame differences are significant. The remaining picture elements (pels) can be repeated from the previous frame. This technique is called conditional replenishment.[2] Conditional replenishment requires addressing the pels which are transmitted (the changed pels or "moving-area" pels) and buffers at the transmitter and receiver.

For example, in Ref. 3 a conditional replenishment coder for eight-bit PCM videotelephone signals* is described which operates at 2 Mb/s (one bit per pel on the average) and uses a number of techniques to reduce the bit rate required for transmission. The pels to be transmitted are addressed along the line in clusters, and their amplitudes are sent as frame-to-frame differences. When the transmitter buffer starts to fill, indicating active motion, only every other changed pel is transmitted,[3,4] with the unsampled pels being replaced by the average of their neighbors. When the buffer fills completely, replenishment is stopped for one frame period, allowing the buffer to empty before resuming transmission.

* 30-Hz frame rate, 271 lines, 2:1 interlace, 3 dB down at 1 MHz, 2-MHz sampling rate, 8-bits/sample, 210 visible samples/line.

Other multimode conditional replenishment coders are described in Refs. 5, 6, and 7. A variety of techniques control the rate of data generation to prevent buffer overflow.

Other functions of conditional replenishment coders, such as the sending of synchronizing information and the accommodation of transmission errors, are also discussed in Refs. 1 to 7.

III. LINEAR PREDICTIVE CODING

A linear predictive coder forms a prediction of each pel to be sent by computing a linear combination of previously transmitted pels. The difference between the actual value and the prediction is then quantized, coded, and transmitted. The inverse process takes place at the receiver. The better the prediction, the smaller the entropy of the differential signal and the bit rate required for transmission. Figure 1 shows two successive frames with interlacing assumed (two interlaced fields per frame). Suppose Z is a moving-area pel we wish to transmit. Pels A, B, C, G, and H are in the field presently being scanned; pels D, E, F, R, S, and T are in the previous field; and the remaining pels are one frame period back from the present field. Pel M is the previous frame value of Z, and if it is used as a prediction of Z, then Z − M, the differential signal which is transmitted, is the frame difference as discussed above.

In Refs. 8 and 9 it was found that using M + (B − J) as a prediction of Z resulted in a relatively-low-entropy, differential signal compared with other nonadaptive predictive coders. In this case, the transmitted differential signal is the line-to-line difference of the frame-difference signal (Z − M) − (B − J).

Transmitting line differences of frame differences has several other advantages as well. Since it does not use pels along the present line

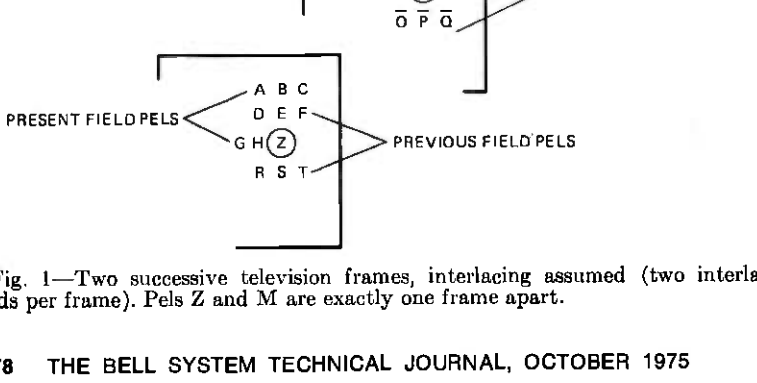


Fig. 1—Two successive television frames, interlacing assumed (two interlaced fields per frame). Pels Z and M are exactly one frame apart.

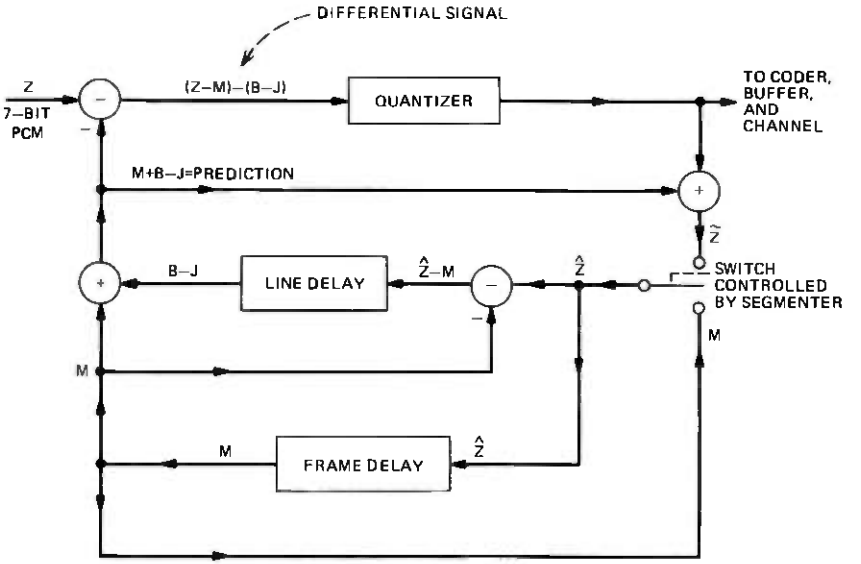


Fig. 2—Predictive coder which transmits only the moving-area pels. The differential signal (Z-M)—(B-J) is the line-to-line difference of the frame-to-frame difference. The segmenter (not shown) determines whether or not Z is a moving-area pel. If it is, the switch is put in the up position and $\tilde{Z}$, a new quantized value, enters the frame memory. Otherwise, the switch is put in the down position and the previous frame value M is recirculated. In any event, $\hat{Z}$ is the value displayed at the receiver in the absence of transmission errors.

or pels in the previous field for its prediction, pel subsampling and field subsampling can be employed without affecting the performance of the predictor. Also, it has been found that relatively few quantization levels are required to produce a good quality picture. Starting with seven-bit PCM, 11-level quantization* of the line difference of frame difference is sufficient for most pictures and most speeds of movement, whereas 30- to 40-level quantization is required for the frame-difference signal.

Figure 2 shows a single-mode conditional replenishment coder which transmits quantized line differences of frame differences in the "moving area." This is the predictive technique used in the coder described in this paper. As with all conditional replenishment coders, a "segmenter" is required to divide the picture into moving parts and stationary parts,[6,7] logic must be provided for sending addressing and synchronizing information, and a buffer is needed to smooth the data rate prior to transmission. If Z is a moving-area pel, the switch is in the up position to allow the quantized representation $\tilde{Z}$ to pass through

* On a scale of $0 \cdots 127$, the quantization levels are 0, ± 1, ± 3, ± 10, ± 23, ± 48.

to the frame memory. If Z is a stationary-area pel, the switch is in the down position, and the previous frame pel M is repeated. At the receiver, the inverse process takes place, and the value $\hat{Z}$ is displayed.

To take advantage of the low entropy of the line difference of frame difference, a variable word-length coder should be used to code the quantized moving-area differential signal. A suitable code for 11-level quantization is given in Table I. The four-bit code word 0000 is reserved for signaling the end of a cluster of significant changes.[3] In a later section of this paper, nine-level quantization is discussed. The first nine code words of Table I are suitable for nine-level quantization.

IV. TEMPORAL FILTERING

A simple method of reducing the data rate in an interframe coder for television pictures is to subsample in the temporal direction and transmit only every other frame (odd field followed by even field) which enters the coder, i.e., send frames at a rate of only 15 Hz. At the receiver in place of each missing frame, one would display either the previous frame or an interpolation of the previous frame and the upcoming frame. However, when using this technique jerkiness is visible in the displayed picture for all except the very slowest movement.

The jerkiness is due to aliasing in the temporal-axis frequency domain, i.e., the input signal has significant power above the half-sampling frequency (here, 7.5 Hz). Aliasing can be reduced by filtering the input signal to reduce as much as possible the power above 7.5 Hz in the temporal frequency domain. Instead of jerkiness, the displayed signal then exhibits blurring in the moving area in proportion to the speed of movement. Many viewers find this type of distortion prefer-

Table I — Variable word-length code suitable for 11-level quantization with code word 0000 reserved for indicating the end of a cluster of significant changes

L_0	1
$+L_1$	01
$-L_1$	001
$+L_2$	0001000
$-L_2$	0001001
$+L_3$	0001010
$-L_3$	0001011
$+L_4$	0001100
$-L_4$	0001101
$+L_5$	0001110
$-L_5$	0001111

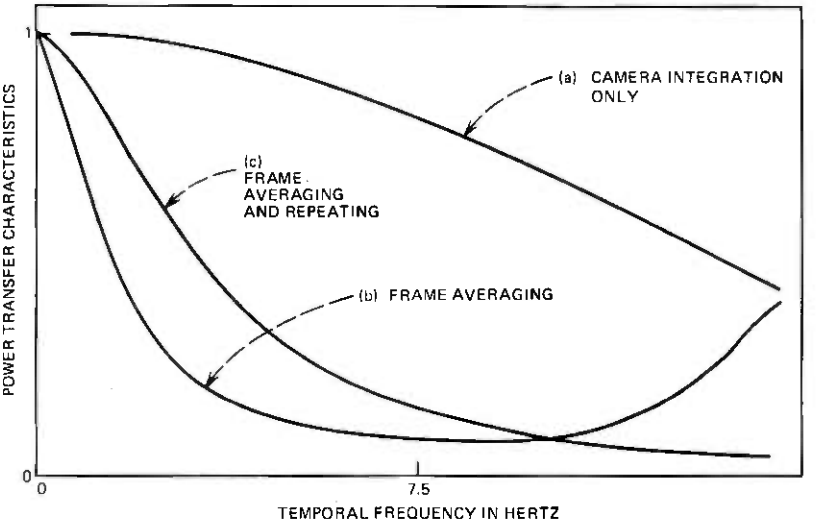


Fig. 3—Power transfer characteristics versus temporal frequency. (a) Light integration by the camera alone. (b) Simple frame averaging. (c) Averaging plus frame repeating as shown in Fig. 4.

able to jerkiness, since it is already present to some degree in all television pictures.

Ideal low-pass filtering using a $(\sin x)/x$ impulse response filter would require several frame memories. In this paper, we use a method of temporal filtering employing only one frame memory, namely, the one normally present in the interframe coder.

Some temporal filtering already takes place in a normal television camera because of its integrating action. Figure 3a shows the power transfer characteristic (derived in the appendix) owing to integration of the light falling on the camera target.

Additional temporal filtering using a frame memory can be carried out by a simple averaging of the incoming frame and the previous frame. The power transfer characteristic of this type of filtering (derived in the appendix) is shown in Fig. 3b. It is down by about 8 dB at 7.5 Hz.

Figure 4 shows the implementation of frame repeating plus temporal averaging. The switch is held in the down position during alternate input frames. Otherwise, it performs conditional replenishment under control of the segmenter as in Fig. 2. In this case, the "previous frame" coming out of the frame memory during conditional replenishment is not the previous frame at all, but, as a result of the frame repeating, it is actually the frame that was coded two frames ago. Because of this fact, increased temporal filtering occurs. Figure 3c shows the

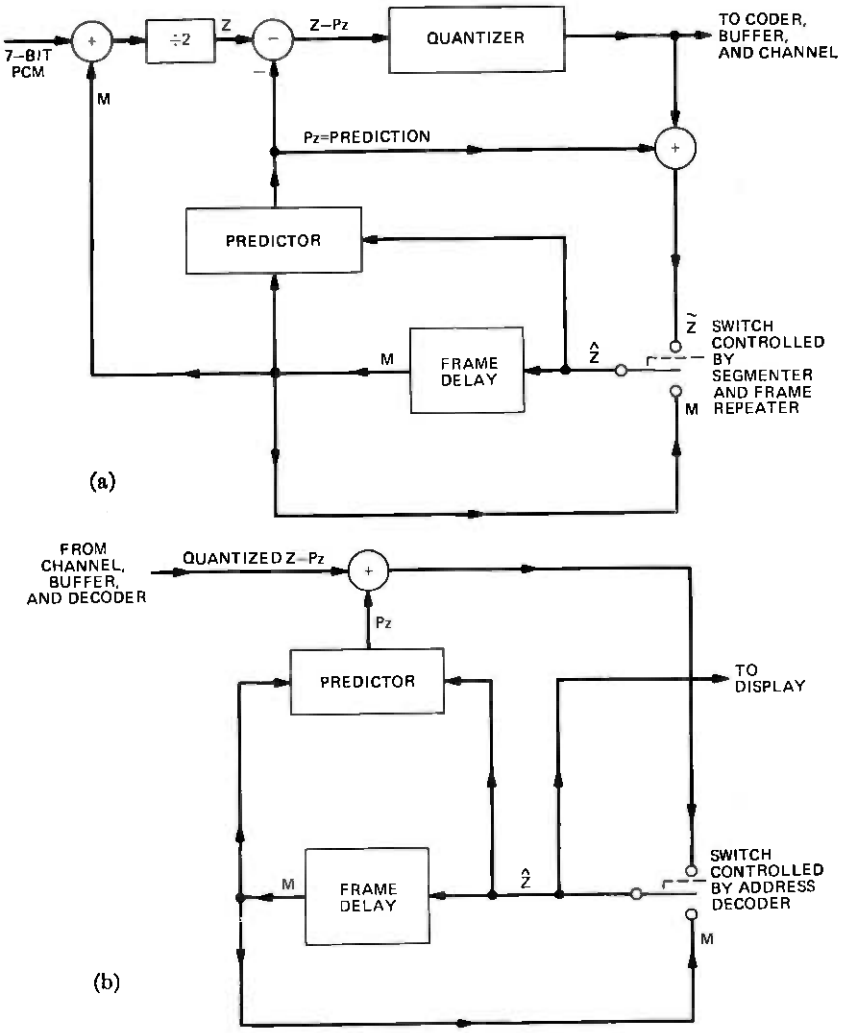


Fig. 4—Implementation of predictive coding with temporal filtering and frame repeating. (a) Transmitter. (b) Receiver. During alternate incoming frames, the switches are held in the down position, thus recirculating the contents of the frame memory, and no data are fed to the transmitter buffer.

power transfer characteristic of frame repeating plus temporal averaging (also derived in the appendix). It falls off much faster than curve b, and is down by about 10 dB at 7.5 Hz. However, unlike curve b it rises again at higher frequencies.

Temporal averaging and frame repeating as shown in Fig. 4 has been implemented, and jerkiness is difficult to detect. However,

blurring is quite visible when the subject moves. Lip motion is also blurred somewhat.

Temporal filtering helps to reduce the data rate in two ways. First, as already mentioned, only every other television frame need be transmitted. Simple frame repeating at the receiver is sufficient to display a picture in which jerkiness is difficult to detect. Second, the blurring of the moving area makes the signal more amenable to predictive coding. With the blurred picture, the differential signal is smaller on the average, thus reducing its entropy and the bit rate required for transmission.

V. SEGMENTING, ADDRESSING, AND SYNCHRONIZING

The coder uses simple, well-known techniques for segmenting the picture into moving and stationary areas.[3] Ordinary seven-bit PCM requires updating of pels which have changed by 2 or more on a scale of $0 \cdots 127$ to present good picture quality in slowly moving areas. However, temporal filtering as shown in Fig. 4 amounts to halving all the frame differences. Thus, in the coder described here, frame differences larger in magnitude than 1 on a scale of $0 \cdots 127$ are detected and labeled as significant changes. As is described later, the frame-difference threshold is raised to 2 to reduce the data rate when buffer overflow threatens.

Significant changes because of camera noise are dealt with as in Ref. 3. That is, a change is ignored if the two pels on the left and the two pels on the right have not changed significantly.

Positioning information for the transmitted pels is also sent as in Ref. 3. The start of a cluster of significant changes is signaled by an eight-bit address indicating its position. The end of a cluster is specified by sending a four-bit code word which is distinguishable from the quantizer output code words (see Table I).

Small gaps between clusters are more efficiently handled by transmitting the pels therein than by ending one cluster and starting a new one.[3] This technique is called gap-bridging.

In Ref. 8, it was found that the entropy of the quantized line difference of frame-difference signal was somewhat above two bits per moving-area pel. Since each new cluster requires twelve bits for addressing, the coder bridges gaps of six pels or less prior to conditional replenishment.

Synchronizing is handled as follows. Since there are less than 256 visible pels along a line, frame sync, field sync, error detection words, and other events which occur relatively rarely can be signaled conveniently using eight-bit code words that are distinguishable from the eight-bit cluster addresses. However, line-to-line sync is not handled as easily.

If line sync were signaled with an eight-bit word, then, with the $\approx$8-kHz line rate used here, 64 kb/s would be devoted to line sync. For a coder operating at a few hundred kb/s, this is much too high a proportion of the total bit rate.

The method of line sync proposed for the coder requires slightly more than one bit per line. With frame repeating, this amounts to about 4 kb/s being used for line sync. The method relies on the fact that the first pel in the first cluster of a line is usually located to the left of the last pel of the last cluster of the previous line. In this case, no additional information need be transmitted to tell the receiver that a new line has begun. However, the receiver must be told if the above situation does not apply, and it must also be told which lines in the picture contain no clusters.

oooooooooooooooooooooo
ABCDooooooooooooEFGH
oooooooooooooooooooooo
oooooooooooooooooooooo
ooIJKLooooMNPoooooooo
oooooooooooooooooooooo
oooooooooooooooooooQRS

Consider the field of pels shown above. Pels labeled A, B, C, $\cdots$, R, S have changed significantly and must be transmitted along with their cluster addresses. Pels labeled o will not be transmitted. Since the cluster ABCD is the first one in the field, the receiver need not be told that a new line is starting. It only needs to be told the number of lines at the beginning of the field that contains no clusters. A string of zeros equal in number to this amount followed by a one suffices to convey this information to the receiver. For implementation reasons which become apparent later, this string of bits is transmitted after the address word of cluster ABCD and before the pels A, B, C, D and the end-cluster message are sent. Cluster EFGH is sent in the normal manner, i.e., address, pels, and end-cluster.

Since pel I is to the left of pel H, the receiver can tell from the address of cluster IJKL that a new line has begun. Following the address word of cluster IJKL, the bits 001 are transmitted, indicating that two intervening lines contained no clusters. Cluster MNP is sent in the normal manner.

Since pel Q occurs to the right of pel P, the receiver cannot tell from the address of cluster QRS that a new line has begun. A special reserved address word must be transmitted to indicate a new line. Following this, the address of cluster QRS and the bits 01 are transmitted as usual. If small gaps between clusters are bridged, then the

above procedure should be modified somewhat. In this case, the special reserved address word need be transmitted only if

$$\text{Q-address} > \text{P-address} + \text{minimum gap size}.$$

A system using these ideas would operate sequentially as follows:

(*i*) At the start of each field
  (*a*) A field sync word is transmitted.
  (*b*) An address register is set to maximum value.
  (*c*) A counter is reset to zero.
(*ii*) The counter is incremented by 1 at the end of each line which contains no clusters to be transmitted.
(*iii*) When the first cluster of a line is encountered
  (*a*) A check is made to see if the address of the first pel exceeds that in the address register. If it does, a special reserved eight-bit word is transmitted which is distinguishable from all the normal cluster address words. This should not occur very often when movement is significant.
  (*b*) The address of the cluster is transmitted.
  (*c*) A string of zeros is transmitted equal in number to the value stored in the counter. None are sent if the counter equals zero.
  (*d*) A one is transmitted, and the counter is reset to zero.
(*iv*) Normal conditional replenishment then resumes and continues until the end of the line.
(*v*) The address of the last pel of the last cluster of the line is added to the minimum gap size, and the result is stored in the address register.
(*vi*) Operation continues with Step (*ii*).

This technique was tested, and with scenes containing slow, moderate, or rapid movement the number of special words that had to be transmitted rarely exceeded two per field (≈ 0.5 kb/s when frame repeating is employed). With no movement, the clusters of significant changes resulting from noise occurred randomly, and the number of special words was higher. But in this case the overall data rate is very small, and thus the special words do not overload the coder.

VI. MODE CONTROL

For a given transmission bit rate, a higher overall picture quality can be obtained if the coding is adapted to the amount of movement in the scene. For an interframe coder, the fullness of the transmitter buffer is the simplest and most useful measure of the amount of movement.[1–3] Imminent buffer overflow is a direct indication that

the data rate being generated is too high and that the displayed moving area resolution should be reduced.

The basic operating mode of the low bit-rate coder is shown in Fig. 4, i.e., temporal filtering, frame repeating, and transmission of line differences of frame differences in the moving area. As with previous coders operating at higher bit rates,[2,3,5–7] the moving-area resolution is reduced by switching to a lower resolution mode if the buffer queue length exceeds some fixed threshold. Thus, as shown in Fig. 5, if the buffer queue length exceeds T_1, then coding mode 1 is invoked; if it exceeds T_2, then coding mode 2 is invoked; etc. Mode 4 is frame repeating, i.e., the switch in Fig. 4 is held in the down position, and no data are generated except synchronizing information. In this way, buffer overflow is prevented.

When motion in the scene ceases and the size of the moving area decreases, the buffer begins to empty, and a higher-resolution coding mode should be used. To prevent oscillations between coding modes, a higher-resolution mode is not invoked until the end of a field, and then only if the buffer queue length is below T_1 for modes 1 and 2 and T_2 for modes 3 and 4. Thus, for example, a change from mode 1 to mode 2 is possible any time the buffer queue length exceeds T_2, but a change from mode 2 to mode 1 can occur only at the end of a field in which the buffer queue length falls below T_1.

VII. MODES USED IN THE CODER

Mode 0 is the previously mentioned basic operating mode shown in Fig. 4. An odd field and an even field are coded as shown in Fig. 6a. Then the next two fields are skipped; at the receiver, the frame is repeated by displaying the stored signal. Mode 0 is the highest resolution mode of the coder.

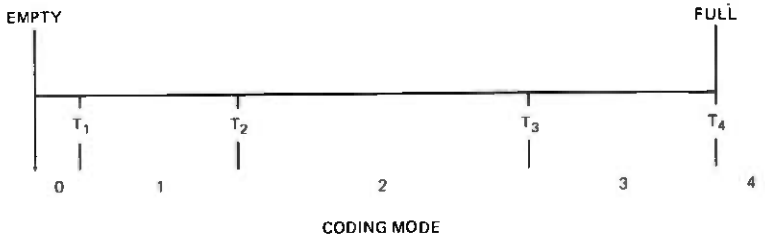


Fig. 5—Switching between coding modes under control of the transmitter buffer causes the moving area resolution to be reduced as the amount of motion in the scene increases. Mode 0 codes with the highest resolution, mode 4 with the lowest. When the buffer queue length exceeds T_i, mode i is invoked (except for mode 3 which is invoked at the end of the field). At the end of the field in which the buffer queue length falls below T_1 for modes 1 and 2 or T_2 for modes 3 and 4, mode i is revoked and mode $i-1$ is invoked. With this strategy, oscillations between coding modes are prevented.

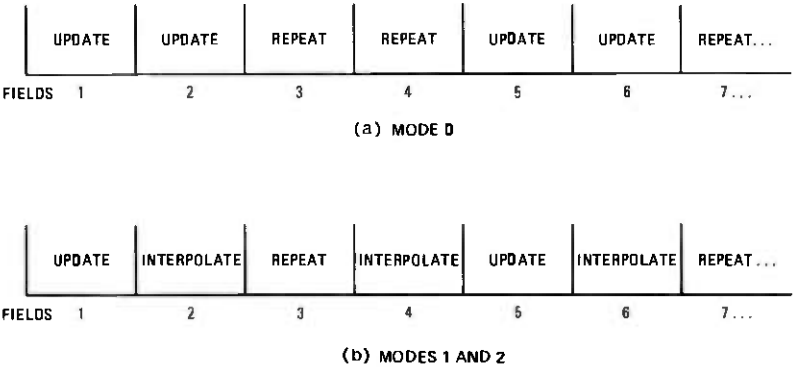


Fig. 6—(a) Simple frame repeating is used in mode 0. Two fields are updated, then two fields are repeated. (b) Frame repeating and field interpolation are used in modes 1 and 2. Only one out of four fields is updated. No data are generated for the remaining three fields.

Mode 1 is interpolation of even fields. In this mode, the data rate is halved by not transmitting even-numbered fields as shown in Fig. 6b. Instead, an interpolation between the previous odd field and the upcoming odd field[10] is displayed, thus reducing the vertical resolution in the picture by a factor of two.

Field interpolation is implemented as shown in Fig. 7. If, during input of an even field, mode 1 is invoked, then the conditional re-

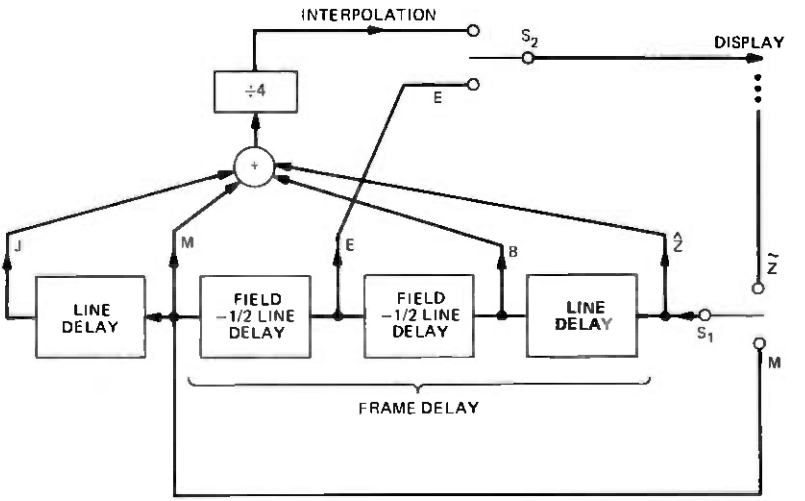


Fig. 7—Implementation of field interpolation. S_1 is held in the down position during input of repeated fields and interpolated fields. No data are generated for them. One field period later, S_2 is put in the up position to display interpolated fields and in the down position to display updated and repeated fields (see Fig. 1).

plenishment switch S_1 is held in the down position for the remainder of the field, and no updating occurs. During input of the next two successive odd fields, switch 2 is held in the up position to display interpolated values for the even fields. Otherwise, it is in the down position. Display of two interpolated fields is necessary because of the aforementioned frame repeating which would otherwise display the invalid contents of the frame memory.

Mode 2 consists of the field interpolation of mode 1 plus 2:1 horizontal subsampling,[3,4] i.e., only every other moving-area pel along a line is transmitted. The untransmitted pels are obtained from their neighbors by interpolation. Subsampling reduces the data rate by a factor of almost 2 over mode 1.

Mode 2 also employs coarser quantization of the line difference of frame-difference signal and an increase of the frame-difference threshold used by the segmenter. When mode 2 is invoked, the frame-difference threshold is raised from 1 to 2 on a scale of $0 \cdots 127$, and the two smallest nonzero levels of the quantizer are switched out of operation, reducing the number of levels to nine. The outputs of the nine-level quantizer are coded using the first nine code words of Table I. Coarser quantization reduces the entropy of the differential signal, and raising the frame-difference threshold reduces the number of pels that must be transmitted. Together they reduce the data rate by a factor of about 1.5, but this figure depends very much on the picture material and on the amount of movement in the scene.

Mode 3 is frame repeating at the end of a field. When mode 3 is invoked, all conditional replenishment is halted. The contents of the frame memory are displayed for odd-numbered fields, and interpolated values are displayed for even-numbered fields. But unlike the other modes, it is invoked only at the end of a field. The purpose of this is to avoid the picture breakup associated with the stopping of conditional replenishment in the middle of a field. As the amount of motion in the scene increases, mode 3 causes the coder to progressively operate in 4:1 frame repeating, 6:1 frame repeating, or as much as is necessary

Table II — Modes of the coder with mode 0 the highest resolution mode

Mode	
0	Temporal filtering and frame repeating.
1	Mode 0 plus interpolation of even fields.
2	Mode 1 plus 2:1 horizontal subsampling, increased frame-difference threshold, and coarser quantization.
3	Frame repeat at end of field.
4	Instantaneous frame repeat.

to accommodate the rate of data generation. Mode 3 is revoked at the end of the field during which the buffer queue length falls below T_2 (not T_1, as with modes 1 and 2).

Mode 4 is instantaneous frame repeating. It is rarely used and is invoked only to prevent data from being lost in an uncontrolled manner because of buffer overflow. It is revoked at the end of the field, and normal frame repeating under mode 3 then resumes. The modes of the coder are summarized in Table II.

VIII. CHOICE OF BUFFER QUEUE LENGTH THRESHOLDS

The objective of the coder is to operate in the mode that best matches the data generation rate with the channel transmission rate. Also, oscillation between modes must be avoided since it adversely affects picture quality in some cases. Correct choice of the buffer queue length thresholds is very important in accomplishing these objectives. As an example, the following illustrates how the thresholds might be chosen for a 200-kb/s channel rate.

Mode 0 is used only when there is little or no motion in the scene. Its most important function occurs just after motion in the scene has ceased and mode 1 (interpolation of even fields) has been revoked. The objective is to update the even field and restore full vertical resolution as quickly as possible. Shortly after even field update has begun, the buffer queue length will exceed T_1 and updating will cease. Little or no data will be produced for the remainder of the field and for the next three field periods. If during this time the buffer empties, then transmission time will have been wasted. Thus, T_1 should be chosen large enough so that the buffer cannot empty in four field periods (1/15 second). For 200 kb/s, $T_1 > 13333$ bits.*

During mode 1, data are produced in only one field out of four (see Fig. 6b). If the overall data generation rate happens to equal the channel transmission rate, then the coder should not produce any data if it should switch to mode 0, and it should not switch to mode 2. In Fig. 6 at the end of field 1 coded in mode 1, the buffer queue length will exceed T_1, and thus field 2 will be interpolated, field 3 will be repeated, and field 4 will be interpolated even if the coder drops into mode 0. To prevent mode 2 from being switched in during a mode 1 odd-field update, T_2 must be large enough to accommodate the accumulated difference between the data generation rate and the channel transmission rate. Somewhat more than three field periods of channel data may have to be buffered. Thus, for 200 kb/s, $T_2 > T_1 + 10000$ bits.

* This figure can be halved if, in mode 0, fields 4, 8, ··· (see Fig. 6) are updated instead of repeated.

With mode 2, the same sort of argument applies. Switching to adjacent modes when the data generation rate and the channel transmission are well matched can be prevented by separating the thresholds by more than three field periods of channel data. Thus, T_2 and T_3 should be more than 10000 bits larger than the next lower threshold.

Mode 3 (4:1, 6:1, $\cdots$frame repeating) is invoked at the *end* of a field in which the buffer queue length exceeds T_3. It is revoked at the end of a field in which the buffer queue length falls below T_2. If the buffer queue length exceeds T_3 at the end of field 1 in Fig. 6b, then 4:1 frame repeating will occur if the buffer queue length is still above T_2 at the end of field 4. This can be guaranteed by choosing $T_3 - T_2$ larger than three field periods of channel data. Thus, for 200 kb/s, $T_3 > T_2 + 10000$ bits. If $T_4 - T_3$ exceeds four field periods of channel data, then normal 6:1 frame repeating with no picture breakup can occur; if it exceeds eight field periods of channel data, then normal 8:1 frame repeating can occur; etc.

T_4 also determines the transmission delay owing to buffering. If 300 ms is the maximum tolerable one-way delay,[1] then T_4 must not exceed 0.3 $\times$ channel rate. For 200 kb/s, $T_4 < 60000$ bits.

Table III gives suitable buffer queue length threshold values for 200-kb/s and 50-kb/s operation. The distances $(T_2 - T_1)$ and $(T_4 - T_3)$ can be reduced somewhat without seriously affecting coder operation, and other more complex mode control strategies can probably be devised that do not require as much buffering. But for purposes of assessing the possible trade-offs between picture quality and channel bit rate, these settings are a valid compromise.

IX. DIGITAL TRANSMISSION ERRORS

It seems to be a general rule that the more the redundancy in a stream of information is reduced, the more vital the remaining information becomes. This is especially true for a low bit-rate interframe coder for television signals. Errors in the data which arrive at the receiver will usually cause discrepancies in its frame memory of which the transmitter is unaware. Thus, if no means are provided to ac-

Table III — Buffer queue length thresholds for 200 and 50 kb/s

Threshold	200-kb/s Operation	50-kb/s Operation
T_1	13400	3350
T_2	24900	6230
T_3	45000	11250
T_4	60000	15000

commodate for digital transmission errors, they will cause visible picture degradations that will last forever.

Many schemes for handling errors have been suggested.[1,3] A simple method, called forced updating, is to transmit a portion of the data as PCM. After a period of time, data in the receiver frame memory will be corrected and visible errors will disappear. However, using channel bit rates which are relatively low compared with PCM, the time required for correction can be quite long. For example, if 10 percent of the channel data is devoted to PCM, then at 200 kb/s all errors in the picture can be corrected in about 22 seconds. But with one error in 10^6 bits, for example, the average time between errors is 5 seconds. Thus, with this technique, errors are always present and, with differential coding of the type discussed previously, highly visible. Other error control techniques are obviously necessary.

Randomly occurring, isolated errors and occasional bursts of errors can be dealt with fairly easily by lowering the information bit rate and using forward-acting error correction codes. However, long bursts of errors in the bit stream present much more of a problem. Cluster addressing, variable word-length coding, and DPCM all serve to increase the vulnerability of the system to digital transmission errors. A long burst of errors would, in most instances, cause picture breakup for many seconds, until some updating procedure could restore the receiver frame memory to its proper state.

Although long bursts of digital transmission errors cannot easily be corrected, they can, in most cases, be detected fairly easily. The receiver could then switch to a frame repeat mode during that portion of the picture for which the frame memory is known to be in error, thus avoiding the picture breakup associated with free running operation. With no movement in the scene, errors would not affect picture quality. With movement, however, errors could, for example, cause the lower half of the picture to freeze for several seconds until it could be updated via PCM.

Recovery from transmission errors can be speeded up considerably if the transmitter can be made aware of their existence and general location by feedback from the receiver. The transmitter could then simply zero out the offending portion of its frame memory and send a control signal telling the receiver to do the same. The erroneous portion of the receiver frame memory will then be updated automatically.

Somewhat more complicated schemes can be devised that utilize feedback of error status and retransmission of incorrectly received data blocks. Some extra buffering is needed (the required amount depends on the channel delay), but erroneous data will not enter the

receiver frame memory except on rare occasions when the error detection algorithm fails. With these techniques, periods of very noisy transmission simply cause the transmitter buffer to fill, which automatically invokes lower-resolution coding modes or frame repeating to match the rate of data generation with the currently available channel capacity.

X. SIMULATION OF THE CODER

A simulation of the coder was constructed to observe what the picture quality would be in an actual system in the absence of digital transmission errors. By and large, the simulator performed all the coding operations that would significantly affect picture quality; however, many shortcuts were taken.

Synchronization of the camera, PCM coder, and simulator was maintained through the same 2-MHz clock; thus, phase-locking and stability problems were sidestepped. Peak-signal-to-rms noise ratio of the input video signal was above 40 dB; thus, problems of analog transmission to the coder were not considered.

The field delays were obtained using a core memory configured as a tapped delay line. Most of the other circuitry was TTL or MOS. A buffer was not constructed. Instead, an up-down counter and threshold detection logic was used to implement the mode control features previously discussed. This approach also made construction of a variable word-length coder unnecessary, although presently available inexpensive solid-state ROMs make this a fairly easy task. The display was obtained by incorporating the field interpolation circuitry of Fig. 7 into the simulator. Normally, this logic is required only at the receiver.

For scenes of people engaged in conversation, it was necessary to delay the voice signal by about 100 ms to obtain a match with the moving lips. Most of this delay is due to the temporal filtering discussed previously. The remainder is due to the field delay between input and display (see Fig. 7). In fact, a completely satisfying match between voice and lips is not obtainable because of the blurring of moving areas caused by the temporal filtering.

XI. CONCLUSION

In this paper, a frame-to-frame coder for videotelephone signals is described that operates at a relatively low bit rate compared with previous coders (200 kb/s or 0.1 bit per pel for an original signal of 1 MHz). The coder was designed on the assumption that faithful

rendition of large moving areas in a scene is not essential for effective interactive visual communication to take place. Whether or not this assumption holds in the majority of situations remains to be seen, but it is conceivable that if users are made aware of the considerable economic saving involved, they will put up with a certain amount of visible distortion in the display.

Graphics and scenes containing little or no movement are portrayed without degradation. Low-key face-to-face conversations contain detectable blurring of moving areas, but for many users this may not be highly objectionable. However, large moving areas are very visibly blurred, sometimes to the point of being nonrecognizable.

The one-way transmission delay of the coder is comparable to the nominally acceptable figure of 300 ms. If the delay of the digital transmission channel is also significant compared with this, as it would be, for example, on an earth satellite circuit, then interactive communication will be severely hampered. Also, special measures must be taken to deal with digital transmission errors. The data generated by the coder are in highly sensitive form. Thus, if some of them arrive incorrectly at the receiver, precautions must be taken to ensure that they do not corrupt legitimate information which has already been received.

The techniques described here apply also to higher resolution pictures, e.g., 525-line standard broadcast rate signals. Indeed, since moving areas do not require any more resolution than with videotelephone, the channel bit rate should not be very much higher either. Graphics and scenes containing no movement would be displayed with much higher resolution. However, the coder itself would also be more expensive.

Much work remains to be done before it will be known if the techniques described here are useful in providing an acceptable compromise between slow-scan facsimile transmission and full rendition of scenes containing movement. Coding for redundancy reduction will remain practical only if costs of logic and storage fall faster than costs of transmission. Also, the requirements of future visual communication systems may change drastically after users begin to learn how to use them effectively in their day-to-day lives.

XII. ACKNOWLEDGMENTS

We are grateful for the many suggestions and criticisms as well as for the technical support we have received from D. J. Connor, J. C. Candy, K. A. Walsh, and our colleagues in the Visual Communication Research Department.

APPENDIX

Here, analytical expressions are given for the power transfer P versus temporal frequency characteristics of Fig. 3. Let $z(t)$ be the light intensity falling on a point of the television camera target, $x(t)$ the output signal as that point is read out of the camera, and $y(t)$ the temporally filtered signal. T is the time between normal frames, i.e., 1/30 second.

Fig. 3a— Camera integration only.

$$x(t) = \frac{1}{T}\int_{t-T}^{t} z(s)ds \tag{1}$$

$$= \frac{1}{T}\int_{-\infty}^{t} z(s)ds - \frac{1}{T}\int_{-\infty}^{t-T} z(s)ds. \tag{2}$$

Taking Fourier transforms,

$$X(\omega) = \frac{1}{j\omega T}Z(\omega) - \frac{1}{j\omega T}Z(\omega)e^{-j\omega T} \tag{3}$$

$$P_a = \left|\frac{X(\omega)}{Z(\omega)}\right|^2 = \frac{2}{\omega^2 T^2}(1 - \cos\omega T). \tag{4}$$

Fig. 3b—Temporal averaging.

$$y(t) = \tfrac{1}{2}x(t) + \tfrac{1}{2}y(t - T). \tag{5}$$

Taking Fourier transforms,

$$Y(\omega) = \tfrac{1}{2}X(\omega) + \tfrac{1}{2}Y(\omega)e^{-j\omega T} \tag{6}$$

$$\left|\frac{Y(\omega)}{X(\omega)}\right|^2 = \frac{1}{(5 - 4\cos\omega T)} \tag{7}$$

$$P_b = \left|\frac{Y(\omega)}{Z(\omega)}\right|^2 = \frac{2(1 - \cos\omega T)}{\omega^2 T^2(5 - 4\cos\omega T)}. \tag{8}$$

Fig. 3c—Temporal averaging and frame repeating.

$$y(t) = \tfrac{1}{2}x(t) + \tfrac{1}{2}y(t - 2T). \tag{9}$$

From (7),

$$\left|\frac{Y(\omega)}{X(\omega)}\right|^2 = \frac{1}{(5 - 4\cos 2\omega T)} \tag{10}$$

$$P_c = \left|\frac{Y(\omega)}{Z(\omega)}\right|^2 = \frac{2(1 - \cos\omega T)}{\omega^2 T^2(5 - 4\cos 2\omega T)}. \tag{11}$$

REFERENCES

1. B. G. Haskell, F. W. Mounts, and J. C. Candy, "Interframe Coding of Videotelephone Pictures," Proc. IEEE, *60*, No. 7 (July 1972), pp. 792–800.
2. F. W. Mounts, "A Video Encoding System Using Conditional Picture-Element Replenishment," B.S.T.J., *48*, No. 7 (September 1969), pp. 2545–2554.
3. J. C. Candy, M. A. Franke, B. G. Haskell, and F. W. Mounts, "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1889–1917.
4. R. F. W. Pease and J. O. Limb, "Exchange of Spatial and Temporal Resolution in Television Coding," B.S.T.J., *50*, No. 1 (January 1971), pp. 191–200.
5. J. B. Millard, Y. C. Ching, and D. M. Henderson, private communication.
6. D. J. Connor, B. G. Haskell, and F. W. Mounts, "A Frame-to-Frame *Picturephone*® Coder for Signals Containing Differential Quantizing Noise," B.S.T.J., *52*, No. 1 (January 1973), pp. 35–51.
7. J. O. Limb, R. F. W. Pease, and K. A. Walsh, "Combining Intraframe and Frame-to-Frame Coding for Television," B.S.T.J., *53*, No. 6 (July–August 1974), pp. 1137–1173.
8. B. G. Haskell, "Entropy Measurements for Nonadaptive and Adaptive Frame-to-Frame Linear Predictive Coding of Videotelephone Signals," unpublished work.
9. D. J. Connor, private communication.
10. J. O. Limb and R. F. W. Pease, "A Simple Interframe Coder for Video Telephony," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1877–1888.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 8, October 1975
Printed in U.S.A.

A Novel Implementation of Digital Phase Shifters

By R. E. CROCHIERE, L. R. RABINER, and R. R. SHIVELY

(Manuscript received March 24, 1975)

A novel technique is presented for implementing a variable digital phase shifter which is capable of realizing noninteger delays. The theory behind the technique is based on the idea of first interpolating the signal to a high sampling rate, then using an integer delay, and finally decimating the signal back to the original sampling rate. Efficient methods for performing these processes are discussed in this paper. In particular, it is shown that the digital phase shifter can be implemented by means of a simple convolution at the sampling rate of the original signal.

I. INTRODUCTION

In digital systems, linear phase shift or delay of a signal waveform by an integer multiple of the sampling period is a simple process that can be implemented as a cascade of unit delays in the network. If, however, it is desired to delay the signal waveform by an amount not equal to an integer multiple of the sampling period, then the process is considerably more difficult. In this case, the signal must be interpolated to obtain new samples of its waveform at noninteger sample times.

In this paper, we propose a novel implementation for achieving such noninteger delays. The theory is based on the application of the concepts of decimation and interpolation proposed by Schafer and Rabiner[1] and Crochiere and Rabiner.[2] It is shown that the actual implementation of the phase shifter or interpolator can be achieved by means of a simple convolution.

Applications in which such noninteger delays in the signal waveform are required often occur when digital systems must interface with analog systems. For example, in the cancellation of echoes, digital systems are often used to generate artificial echoes by means of a simulation of an echo model. These artificial echoes are then subtracted from the original analog signal to cancel its echo. For best cancellation, the digital simulated echo may have to be delayed by a noninteger multiple of the sampling period.

A second potential application occurs when multiple signals must be processed together such as in a phased-array antenna system (e.g., for seismic processing). In this case, the signal waveforms from the various elements must be shifted by noninteger multiples of the sampling period relative to each other.

A third application of noninteger delays is in pitch, synchronous synthesis of speech.[3] In this case, a parametric representation of speech is generated at a fixed sampling rate (usually 100 Hz); however, the synthesis parameters are required at time instances between the sampling intervals to avoid producing transients in the synthesized signal. Using the variable phase shifter proposed in this paper, the synthesis parameters can be readily interpolated to any point between sampling intervals.

II. BASIC CONCEPTS OF THE PHASE SHIFTER

Figure 1 illustrates the basic operation of the phase shifter. To implement a delay of l/D samples, where l and D are any integers, the sampling rate, f_r, of the input signal $x(n)$ is first increased by an integer factor D [by inserting $D - 1$ zero-valued samples between each sample of $x(n)$]. The resulting signal $v(n)$ is then filtered by a low-pass filter $h(n)$ (generally a linear-phase FIR filter is used here) to remove its periodic frequency components, which are centered about integer multiples of the original sampling frequency.[1,2] The output of the filter $u(n)$ is an interpolated version of the input signal $x(n)$. The signal $u(n)$ is then delayed by l samples at the high sampling rate to produce the signal $w(n) = u(n - l)$. It will be assumed that l satisfies the condition.

$$0 \leqq l \leqq D - 1. \tag{1}$$

Finally, the output $y(n)$ is obtained by desampling or decimating $w(n)$, i.e., by choosing every Dth sample of $w(n)$. The net effect is to delay the original signal $x(n)$ by a noninteger delay of $(l/D)T$ where $T = 1/f_r$ is the sampling period at the low rate. In addition, an integer delay is introduced in the signal due to the delay of the low-pass filter $h(n)$.

The structure in Fig. 1 can be analyzed in a straightforward manner. Let $X(e^{j\omega})$, $V(e^{j\omega})$, $W(e^{j\omega})$, $Y(e^{j\omega})$, and $H(e^{j\omega})$ be the Fourier transforms of $x(n)$, $v(n)$, $w(n)$, $y(n)$, and $h(n)$ respectively. Then, the relationships

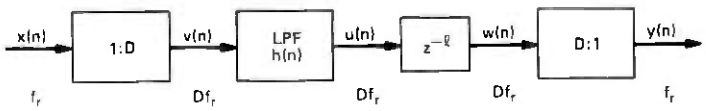


Fig. 1—Block diagram of the phase shifter.

between them can be given as in Refs. 1 and 2:

$$V(e^{j\omega}) = X(e^{j\omega D}), \tag{2}$$

$$W(e^{j\omega}) = H(e^{j\omega})e^{-j\omega l}V(e^{j\omega}), \tag{3}$$

and

$$Y(e^{j\omega}) = \frac{1}{D}\sum_{m=0}^{D-1} W(e^{-j2\pi m/D}e^{j\omega/D}). \tag{4}$$

In eq. (4), the terms in the summation for $m = 1, 2, \cdots, D-1$ correspond to high-frequency components of $W(e^{j\omega})$, which are aliased into the low-frequency band from 0 to $f_r/2$ due to the desampling process. We assume that the low-pass filter $H(e^{j\omega})$ attenuates these high-frequency components to a point where such aliasing can be considered negligible. That is, it has a stop-band cutoff frequency of $1/2D$ (normalized to the high sampling rate Df_r) and a stop-band ripple δ_s that is sufficiently small to prevent aliasing. With these assumptions, (4) becomes

$$Y(e^{j\omega}) \cong \frac{1}{D} W(e^{j\omega/D}) \tag{5}$$

and with the aid of (2) and (3) it can be written as

$$\begin{aligned} Y(e^{j\omega}) &\cong \frac{1}{D} H(e^{j\omega/D})e^{-j\omega l/D}V(e^{j\omega/D}) \\ &\cong \frac{1}{D} H(e^{j\omega/D})e^{-j\omega l/D}X(e^{j\omega}). \end{aligned} \tag{6}$$

We now assume that $H(e^{j\omega})$ is a FIR filter with exactly linear phase and has a unit sample response duration of N samples. Then, its delay will be $(N-1)/2$ samples at the high sampling rate. If it is desired that this delay be an integer delay at the low sampling rate, then N must be chosen such that $(N-1)/2$ is an integer multiple of D. That is,

$$\frac{N-1}{2} = ID, \tag{7}$$

where I is a positive integer and

$$N = 2ID + 1. \tag{8}$$

If the particular application does not require that the delay of the filter appear as an integer delay at the low sampling rate, then condition (8) is entirely optional and need not be used.

We now impose the constraint that the passband response of $H(e^{j\omega/D})$ have a gain of D and be essentially flat (i.e., have very small passband ripples). Then the filter response of $H(e^{j\omega/D})$ over the pass-

band is approximately [assuming (8) applies]

$$H(e^{j\omega/D}) \cong De^{-(j\omega/D)[(N-1)/2]} \cong De^{-j\omega I}. \tag{9}$$

Substituting (9) into (6) gives the final desired result:

$$\frac{Y(e^{j\omega})}{X(e^{j\omega})} \cong e^{-j\omega I}e^{-j\omega l/D} \tag{10}$$

or in terms of z-transforms

$$\frac{Y(z)}{X(z)} \cong z^{-I}z^{-l/D}. \tag{11}$$

Thus, the structure in Fig. 1 is essentially an all-pass network [over the passband of $H(e^{j\omega/D})$] with a fixed integer delay of I samples due to the processing delay of the low-pass filter $h(n)$ and a variable noninteger delay of l/D samples. If N does not satisfy condition (8), then I in eqs. (10) and (11) will not be an integer. In either case, the output, $y(n)$, in Fig. 1 is an approximation to $x(n - l/D - I)$.

III. IMPLEMENTATION OF THE PHASE SHIFTER

The design of the phase shifter in Fig. 1 suggests a structure which involves two different sampling rates. In this section, we show that the actual implementation of the phase shifter can be achieved considerably easier as a straightforward convolution at the low sampling rate.

Since the duration of $h(n)$ is N samples and $D - 1$ out of every D samples of $v(n)$ are zero valued, the filter $h(n)$ spans approximately N/D nonzero samples of $v(n)$. More precisely, because of the constraint imposed on N in (8), $h(n)$ spans Q nonzero samples of $v(n)$ for the computation of some output points and $Q - 1$ nonzero samples of $v(n)$ for the computation of other output points [Q is defined in eq. (13)]. To avoid this implementation difficulty, it is convenient to consider instead a new filter $h'(n)$ whose length N' is

$$N' = QD \geqq N, \tag{12}$$

where $h'(n)$ is obtained by extending $h(n)$ with $N' - N$ zero-valued coefficients. Obviously, the filter $h'(n)$ has the same exact frequency response and delay as $h(n)$, but it spans *exactly* Q nonzero samples of $v(n)$ [although one nonzero sample of $v(n)$ may be multiplied by a zero valued coefficient of $h'(n)$]. Since we wish to keep N' as small as possible, consistent with (12) we can choose Q to be

$$Q = \left\lceil \frac{N}{D} \right\rceil, \tag{13}$$

where the brackets indicate that the number is rounded to the next largest integer.

With these assumptions, we can now relate the output $y(n)$ in Fig. 1 to $x(n)$ and $h'(n)$ by the expression[2]

$$y(n) = \sum_{k=0}^{Q-1} h'[kD + (-l) \oplus D]x(n-k), \tag{14}$$

where $\oplus$ corresponds to modulo addition. By letting

$$g_l(k) = h'[kD + (-l) \oplus D] \quad k = 0, 1, \cdots, Q-1, \tag{15}$$

(14) then becomes

$$y(n) = \sum_{k=0}^{Q-1} g_l(k)x(n-k), \tag{16}$$

which is the form of a simple convolution. Therefore, the phase shifter can be implemented by a Q point convolution of $x(n)$ with $g_l(n)$, where $g_l(n)$ is an appropriate subset of the coefficients of $h'(n)$. To obtain a zero incremental phase shift, we use the coefficients $\{g_0(0) = h'(0), g_0(1) = h'(D), \cdots, g_0(Q-1) = h'[(Q-1)D]\}$. To obtain a delay of $(l/D)T$ (or a phase shift of $\omega l/D$), we use the coefficients $\{g_l(0) = h'[(-l) \oplus D], g_l(1) = h'[D + (-l) \oplus D], \cdots, g_l(Q-1) = h'[(Q-1)D + (-l) \oplus D]\}$. If we want a variable phase shifter, we can store all D sets of coefficients and use the appropriate set as suggested in Fig. 2.

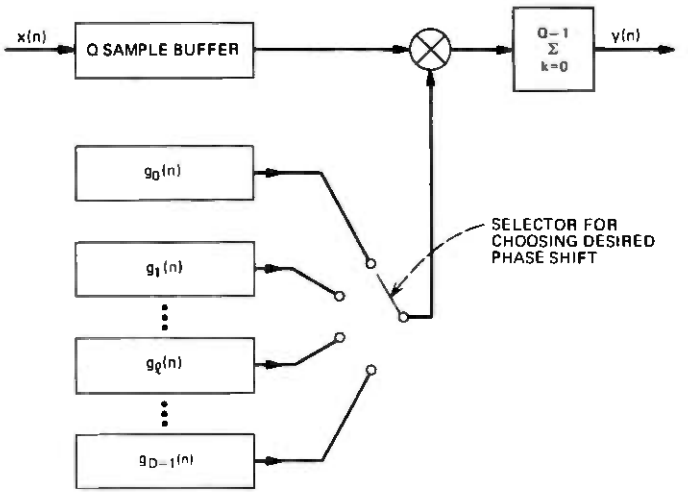


Fig. 2—A practical implementation of a variable phase shifter.

V. CONCLUSIONS

We have presented a method for designing an incremental digital phase shifter that can shift the phase of a waveform by a noninteger number of samples. Conceptually, the process can be thought of as a sample rate increase, a delay, and a sample rate decrease as indicated in Fig. 1. Practically, it can be implemented as a straightforward convolution as shown in Fig. 2. From the discussion of the theory, it is also clear that the design trade-offs of the phase shifter are directly related to the characteristics of the low-pass FIR filter. That is, the passband ripples of $H(e^{j\omega})$ determine how close the phase shifter is to an ideal all-pass network (over the passband), and the stop-band ripples determine the amount of distortion due to aliasing. Finally, the cutoff frequency of the filter determines the usable frequency range of the phase shifter.

REFERENCES

1. R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," Proc. IEEE, *61*, No. 6 (June 1973), pp. 692–702.
2. R. E. Crochiere and L. R. Rabiner, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation, and Narrow Band Filtering," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 5 (October 1975), pp. 444–456.
3. L. R. Rabiner, "Digital Formant Synthesizer for Speech Synthesis Studies," J. Acoust. Soc. Amer., *43*, No. 4 (April 1968), pp. 822–828.

Contributors to This Issue

Jacques A. Arnaud, Dipl. Ing., 1953, Ecole Supérieure d'Electricité, Paris, France; Docteur Ing., 1963, University of Paris; Docteur es Science, 1972, University of Paris; Assistant at E.S.E., 1953–1955; CSF, Centre de Recherche de Corbeville, Orsay, France, 1955–1966; Warnecke Elec. Tubes, Des Plaines, Illinois, 1966–1967; Bell Laboratories, 1967—. At CSF, Mr. Arnaud was engaged in research on high-power traveling-wave tubes and noise generators. He is presently studying the propagation of optical waves. Senior Member, IEEE; Member, Optical Society of America.

Ronald E. Crochiere, B.S. (E.E.), 1967, Milwaukee School of Engineering; M.S. (E.E.), 1968, and Ph.D. (E.E.), 1974, Massachusetts Institute of Technology; Raytheon, 1968–1970; M.I.T. Research Laboratory of Electronics, 1970–1974; Bell Laboratories, 1974—. Mr. Crochiere has worked on the design of microwave phase shifters, on digital network theory, and digital filter structures. He is presently engaged in research in speech communications and digital signal processing. Member, Sigma Xi and IEEE G-ASSP Technical Committee on Digital Signal Processing.

Peter Cummiskey, B.S. (E.E.), 1963, Fairleigh Dickinson University; M.S. (E.E.), 1969, and Dr. Eng. Sci., 1973, Newark College of Engineering; Bell Laboratories, 1962—. Mr. Cummiskey has been engaged in research in the areas of speech synthesis, an adaptive time waveform coding (DPCM and delta modulation). He has also worked on minicomputer and automatic voice response systems.

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 mainly been engaged in analytical work concerning the transmission of signals. Currently, he is working in the area of data communication theory. Member, IEEE, Sigma Xi, Mathematical Association of America, American Men of Science, New York Academy of Sciences.

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 has worked on problems in data transmission and digital signal processing. He is a member of the Communication Theory and the Technology Forecasting and Assessment committees of the IEEE Communications Society, and is an Associate Editor of the IEEE Transactions on Communications. Member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

Barry G. Haskell, B.S., 1964, M.S., 1965, and Ph.D. (Electrical Engineering), 1968, University of California, Berkeley; Research Assistant, University of California, 1965–1968; Bell Laboratories, 1968—. Mr. Haskell is a member of the Visual Communication Research Department. His primary interest is the efficient coding of pictures for transmission at reduced bit rate. Member, Phi Beta Kappa, Sigma Xi, IEEE.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematical Research Center. Member, American Physical Society, IEEE.

Lawrence R. Rabiner, S.B. and S.M. (E.E.), 1964, Ph.D. (E.E.), 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. Coauthor, *Theory and Application of Digital Signal Processing*. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; President, IEEE G-ASSP Ad Com; member, G-ASSP Technical Committee on Digital Signal Processing, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Technical Committee on Speech Communication of the Acoustical Society; former Associate Editor of the G-ASSP Transactions.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked

on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems, and is currently a supervisor in the Advanced Data Communications Department. During the academic year 1967–68 he was on leave as Professor of Electrical Engineering at the University of Florida. Member, Sigma Xi.

Robert L. Schmidt, Brookdale Community College, 1972—; Bell Laboratories, 1972—. Mr. Schmidt is a member of the Visual Communications Research Department specializing in digital coding systems intended to reduce bandwidth of television signals. He is presently involved with a software-controlled system for use with broadcast television signals.

R. R. Shively, B.S. (E.E.), 1956, M.S. (E.E.), 1957, and Ph.D. (E.E.), 1963, University of Illinois; IBM, 1957–1963; Bell Laboratories, 1963—. Mr. Shively has worked on the design of computer systems. Since joining Bell Laboratories, he has worked on the design of the SAFEGUARD computing system, implementation of the first fast-Fourier-transform processor, and analysis of large-scale computer system performance. Currently, he is involved in the study of programmable signal analysis systems.

Aaron D. Wyner, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information and communication theory and related mathematical problems. He is presently Head of the Communications Analysis Research Department. He spent the year 1969–1970 visiting the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, the Technion, Haifa, Israel on a Guggenheim Foundation Fellowship. He has also been a full- and part-time faculty member at Columbia University and the Polytechnic Institute of Brooklyn. He has been chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, has served as an associate editor of the Group's *Transactions*, and has served as cochairman of two international symposia. He is presently first vice-president of the IEEE Information Theory Group. Fellow, IEEE, member, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.