THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 November 1978 Number 9

Calculation of Steady-State Probabilities for Content of Buffer with Correlated Inputs

By W. A. MASSEY and J. A. MORRISON

(Manuscript received March 2, 1978)

In a previous paper, a model for the behavior of a switching node that receives data from many terminals over low-speed access lines was considered. In this paper, we give the details of an alternate procedure for calculating the steady-state probabilities for the buffer content. It is shown that a finite system of linear equations may be obtained for calculating the steady-state probability that the buffer content is i. In a particular case of interest, explicit formulas are derived for the number of equations which arise in this procedure, for each value of i. Some detailed calculations are given for one example.

I. INTRODUCTION

Mathematical models for the behavior of a switching node that receives data from a (large) number of terminals over low-speed access lines have been considered by Gopinath and Morrison,[1,2] and some particular examples have been investigated by Fraser, Gopinath, and Morrison.[3] In this paper, we consider one of the models and give the details of an alternate procedure, which was alluded to by Gopinath and Morrison,[1] for calculating certain steady-state probabilities.

We first describe the model which we will consider. It is assumed that the data are received at the switching node in the form of packets of fixed size. As the packets arrive, they are placed in a buffer, which is a first-in-first-out queue. The buffer processes packets at a uniform rate, provided that it is not empty. In an actual computer network, the buffer capacity is finite, and a packet is lost if the buffer is full when it attempts to enter it. In our mathematical model, it is assumed that the buffer has

infinite capacity, so that no overflow is possible, and we are interested in calculating the steady-state probability that the buffer content (i.e., the number of packets in the buffer) exceeds the proposed capacity of the buffer.

We let the time that it takes for the buffer to process a packet through the node be our unit of time. We suppose that ξ_n is the number of packets which enter the buffer in the time interval $(n,n+1]$. If b_n denotes the buffer content at time n, then the buffer content at time $n+1$ is given by the equation

$$b_{n+1} = (b_n - 1)^+ + \xi_n, \tag{1}$$

where $a^+ = \max(a,0)$. The quantity ξ_n is a random variable, and hence so is b_n.

Consider the case in which each message from a terminal consists of exactly two packets which are separated by k units of time, where k is an integer. The packets are spread apart since the speed of the access lines is slower than the buffer processing rate. If x_n denotes the number of first packets entering the buffer in the interval $(n,n+1]$, then $\xi_n = x_n + x_{n-k}$, since x_{n-k} is the number of second packets entering in this interval which belong to messages whose first packets entered k intervals earlier. It was shown,[1,3] under suitable conditions, that if the number of terminals is large, then it is a reasonable approximation to assume that the random variables x_i are independently and identically distributed (i.i.d.).

A generalization of the above model was considered,[1] in which the number of packets entering the buffer in the interval $(n,n+1]$ is

$$\xi_n = \sum_{j=0}^{k} \alpha_j x_{n-j}, \tag{2}$$

where the nonnegative integer valued random variables x_i are i.i.d. and the constant coefficients α_i are nonnegative integers. It is assumed, without loss of generality, that $\alpha_0 \neq 0 \neq \alpha_k$. This is the model which we consider in this paper. It corresponds to a fixed pattern for each message. A more general model was considered[2] which allows for randomness in the message pattern, e.g., a random number of packets in a message. It would be of interest to obtain results for the more general model, analogous to those derived in this paper for the model corresponding to (2). This could be the topic of a future paper.

The results are stated and proved in a series of propositions, lemmas, theorems, and corollaries. In Section II, an explicit expression is first given for the steady-state probability that the buffer is empty, under the assumption that the mean arrival rate is less than unity. The steady-state probability that the buffer content is i is expressed in terms of the steady-state probabilities corresponding to a certain $(k+1)$-dimensional Markov process. Criteria for the proper states of this process

are obtained, and it is shown that, for fixed k, a finite number of these states correspond to a prescribed buffer content. The fundamental relation satisfied by the steady-state probabilities corresponding to the $(k + 1)$-dimensional process is derived.

In Section III, it is shown how this fundamental relation may be iterated, so as to obtain a finite system of linear equations for calculating the steady-state probabilities corresponding to a prescribed buffer content. Numerous subsidiary quantities are defined to establish the required reduction formulas. The use of the reduction formulas to obtain the desired steady-state probabilities is described in Section IV.

In Section V, attention is turned to the particular case $\xi_n = x_n + x_{n-k}$, so that $\alpha_0 = 1 = \alpha_k$, and $\alpha_j = 0$ otherwise, in (2). Explicit formulas are derived for the number of equations which occur in the calculation of the steady-state probabilities corresponding to a prescribed buffer content. In Section VI, the steady-state probabilities corresponding to an empty buffer are calculated in the case $k = 4$.

II. THE FUNDAMENTAL RELATION

We assume that the mean arrival rate at the buffer is less than unity, and we are interested in determining the quantities

$$\kappa_i = \lim_{n \to \infty} \Pr(b_n = i), \tag{3}$$

where b_n satisfies (1) subject to (2). Hence, κ_i is the steady-state probability that the buffer content is i. We will see that the determination of these quantities involves the determination of certain other steady-state probabilities, as discussed by Gopinath and Morrison.[1] It was proved[2] that all these steady-state probabilities exist. We proceed to state, and prove, the results in a series of propositions, lemmas, theorems, and corollaries. We first give an explicit expression[1] for κ_0, the steady-state probability that the buffer is empty.

Proposition 1: $\kappa_0 = 1 - \mu_k E(x)$ *where* $\mu_k = \sum_{i=0}^{k} \alpha_i$ *and* $E(x)$ *is the expectation of any* x_n.

(This result may be derived by solving (77) in Ref. 2 for the marginal generating function $\phi_k(s)$, and letting $s \to 1$. This was the method of proof used in Ref. 1.)

We remark that, from (2), $E(\xi_n) = \mu_k E(x)$. Note that our assumption that the mean arrival rate is less than unity implies that $\kappa_0 > 0$.

To determine the other κ_i's, it will be convenient to use the following quantities:

$$\theta_n^{(r)} = \sum_{i=r}^{k} \alpha_i x_{n+r-i-1} \qquad \text{for } r = 1, \ldots, k. \tag{4}$$

Since we are using the first packets of a message to count the number

of intermediate packets at some later time, the $\theta_n^{(r)}$ correspond in this sense to the packet contribution prior to time n to ξ_{n+r-1}. We will determine the κ_i's by exploiting the recursive relations between b_n, ξ_n, and $\theta_n^{(r)}$.

Let Z^l be the direct sum of a countable number of copies of Z, the set of integers. For l, a nonnegative integer, we define the following collection of subsets:

$$N^l = \{(n_0, \ldots, n_l, 0, \ldots) | n_0, \ldots, n_l \geq 0\}.$$

Clearly, we have $N^0 \subset N^1 \subset \ldots \subset Z^l$. We now define a random $k+1$-tuple variable

$$\mathbf{B}_n = (b_n, \theta_n^{(1)}, \ldots, \theta_n^{(k)}, 0, \ldots). \tag{5}$$

Using $\mathbf{B}_n$, we can define a map U that sends Z^l into [0,1]. Given $\mathbf{m} \in Z^l$, with $\mathbf{m} = (m_0, m_1, \ldots)$, we define

$$U(\mathbf{m}) = \lim_{n \to \infty} Pr(\mathbf{B}_n = \mathbf{m}). \tag{6}$$

We can then recover any κ_i from the $U(\mathbf{m})$'s via the relation

$$\kappa_i = \sum_{m_0 = i} U(\mathbf{m}). \tag{7}$$

This summation looks unwieldy, but we will show that this is not the case.

We first establish

Proposition 2: $(b_{n-l} - 1)^+ + \sum_{i=1}^{l} \xi_{n-i} \leq b_n + l - 1$ *for* $l \geq 1$.

Proof: Use induction on l.

$(l = 1)$ $\quad (b_{n-1} - 1)^+ + \xi_{n-1} = b_n$, from (1).

$(l \to l + 1)$ Note that $(b_{n-l} - 1) \leq (b_{n-l} - 1)^+$, and hence, using (1),

$$(b_{n-l-1} - 1)^+ + \sum_{i=1}^{l+1} \xi_{n-i} = b_{n-l} + \sum_{i=1}^{l} \xi_{n-i}$$

$$\leq (b_{n-l} - 1)^+ + 1 + \sum_{i=1}^{l} \xi_{n-i} \leq b_n + l \; **$$

(The double asterisk is used throughout the paper to denote the end of a proof.)

We now prove

Theorem 3: $U(\mathbf{m}) \neq 0$ *implies that* $\mathbf{m} \in N^k$, α_k *divides* m_k *and, for* $l = 1, \ldots, k$,

$$\sum_{i=1}^{l} m_{k-i+1} \leq \sum_{j=1}^{l} \alpha_0^{-1} \alpha_{k-l+j} (m_0 + j - 1).$$

Proof: To have a nonzero probability that $\mathbf{B}_n = \mathbf{m}$, it is immediate that

$\mathbf{m} \in N^k$. Also, $\theta_n^{(k)} = \alpha_k x_{n-1} = m_k$, and for a nonzero probability, x_{n-1} must take on an integer value.

Using Proposition 2, we have for $l = 1, \ldots, k$,

$$b_n + l - 1 \geq \sum_{i=1}^{l} \xi_{n-i} = \sum_{i=1}^{l} \sum_{j=0}^{k} \alpha_j x_{n-i-j} \geq \alpha_0 \sum_{i=1}^{l} x_{n-i}.$$

Therefore,

$$\sum_{i=1}^{l} x_{n-i} \leq \alpha_0^{-1}(b_n + l - 1). \tag{8}$$

Now

$$\sum_{i=1}^{l} \theta_n^{(k-i+1)} = \sum_{i=1}^{l} \sum_{j=k-i+1}^{k} \alpha_j x_{n+k-i-j}$$

$$= \sum_{j=k-l+1}^{k} \sum_{i=k+1-j}^{l} \alpha_j x_{n+k-i-j}.$$

Hence, if we make the substitutions $j = \tau + k - l$ and $i = \sigma + l - \tau$, we obtain

$$\sum_{i=1}^{l} \theta_n^{(k-i+1)} = \sum_{\tau=1}^{l} \alpha_{\tau+k-l} \left(\sum_{\sigma=1}^{\tau} x_{n-\sigma} \right)$$

$$\leq \sum_{\tau=1}^{l} \alpha_0^{-1} \alpha_{\tau+k-l}(b_n + \tau - 1),$$

using (8).

So, if $\theta_n^{(r)} = m_r$ and $b_n = m_0$, the m_r's must satisfy these conditions. **

Corollary 4: For fixed m_0 and k, there can only be a finite number of $\mathbf{m}$ *such that* $U(\mathbf{m}) \neq 0$.

Such $\mathbf{m}$ that satisfy the criteria of Theorem 3 will be called proper states. From (7), each κ_i then is the sum over only a finite number of these.

To derive the fundamental relation satisfied by $U(\mathbf{m})$ we need

Proposition 5: $\theta_{n+1}^{(r)} = \alpha_r x_n + \theta_n^{(r+1)}$ *for* $r = 1, \ldots, k-1$ *and* $\theta_{n+1}^{(k)} = \alpha_k x_n$.

Proof: From (4), for $r = 1, \ldots, k-1$,

$$\theta_{n+1}^{(r)} = \sum_{i=r}^{k} \alpha_i x_{n+r-i}$$

$$= \alpha_r x_n + \sum_{i=r+1}^{k} \alpha_i x_{n+r+1-i-1}$$

$$= \alpha_r x_n + \theta_n^{(r+1)},$$

and $\theta_{n+1}^{(k)} = \alpha_k x_n$ by definition. **

We now define a map from Z^l into itself called T_γ, where γ is a nonnegative integer:

$$T_\gamma(\mathbf{m}) = R(\mathbf{m}) + (\gamma, -(\gamma - 1)^+, 0, \ldots),$$

where R is the right shift operator. More explicitly, we can write

$$T_\gamma(\mathbf{m}) = (\gamma, m_0 - (\gamma - 1)^+, m_1, \ldots). \tag{9}$$

Theorem 6:

$$U(\mathbf{m}) = p(\sigma) \sum_{\gamma \geq 0} U(T_\gamma(\mathbf{m} - \sigma \nu_0)),$$

where $\sigma = \alpha_k^{-1} m_k, p(\sigma) = Pr(x_n = \sigma)$ *and* $\nu_0 = (\alpha_0, \ldots, \alpha_k, 0, \ldots)$.

Proof: By Proposition 5, $\theta_{n+1}^{(k)} = m_k$ implies $x_n = \alpha_k^{-1} m_k = \sigma$. If σ is not a nonnegative integer, then $p(\sigma) = 0$ and $U(\mathbf{m}) = 0$, from (5) and (6), and the equation holds trivially.

Now we let σ be a nonnegative integer. Recall from Proposition 5 again that

$$\theta_n^{(r+1)} = \theta_{n+1}^{(r)} - \alpha_r x_n, \qquad \text{for } r = 1, \ldots, k-1.$$

Also, from (1), (2) and (4), we have

$$\theta_n^{(1)} = \xi_n - \alpha_0 x_n = b_{n+1} - (b_n - 1)^+ - \alpha_0 x_n.$$

So, if $b_n = \gamma$ and $\mathbf{B}_{n+1} = (m_0, \ldots, m_k, 0, \ldots)$, it will be necessary and sufficient, from (5), that $x_n = \sigma$ and

$$\mathbf{B}_n = (\gamma, m_0 - (\gamma - 1)^+ - \alpha_0 \sigma, m_1 - \alpha_1 \sigma, \ldots, m_{k-1} - \alpha_{k-1} \sigma, 0, \ldots).$$

In more compact notation, for $\mathbf{m} \in N^k$ we have $\mathbf{B}_{n+1} = \mathbf{m}$, $b_n = \gamma$ iff $\mathbf{B}_n = T_\gamma(\mathbf{m} - \sigma \nu_0)$, $x_n = \sigma$.

But x_n is independent of b_n and hence, from (4) and (5), of $\mathbf{B}_n$. Therefore,

$$\begin{aligned} \Pr(\mathbf{B}_{n+1} = \mathbf{m}) &= \sum_{\gamma \geq 0} \Pr(\mathbf{B}_{n+1} = \mathbf{m}, b_n = \gamma) \\ &= \Pr(x_n = \sigma) \sum_{\gamma \geq 0} \Pr(\mathbf{B}_n = T_\gamma(\mathbf{m} - \sigma \nu_0)). \end{aligned}$$

The theorem follows by letting $n \to \infty$ and using (6). ∎∎

The fundamental relation in Theorem 6 satisifed by $U(\mathbf{m})$ was stated by Gopinath and Morrison,[1] in less compact notation. They also showed[1,2] that, once the steady-state probabilities $U(\mathbf{m})$ with $m_0 = 0$, corresponding to an empty buffer, were obtained, then the steady-state generating function for the buffer content could be calculated in terms of the generating functions for some marginal distributions. In this paper, we show how the quantities $U(\mathbf{m})$ may be calculated for any value of m_0, so that the steady-state probability that the buffer content is i may be calculated with the help of (7). In fact, we show how to iterate the fundamental relation in Theorem 6 so as to obtain a finite system of linear equations for calculating $U(\mathbf{m})$ for a fixed value of m_0. This procedure was alluded to by Gopinath and Morrison[1] in the case $m_0 = 0$.

III. REDUCTION FORMULAS

We first remark that, if $m_k = 0$ then the summation in the fundamental relation for $U(\mathbf{m})$ in Theorem 6 includes the term corresponding

to $\gamma = m_0 + 1$, so that this does not give a closed system of equations for $U(\mathbf{m})$ for a given value of m_0. To carry out the desired iteration of the fundamental relation, it is convenient to define some new quantities. Accordingly, we let

$$U^{(1)}(\mathbf{m}) = \sum_{\gamma \geq 0} U(T_\gamma(\mathbf{m})), \tag{10}$$

and, for $r = 1, \ldots, k-1$, define

$$U^{(r+1)}(\mathbf{m}) = \sum_{\gamma \geq 1} U^{(r)}(T_\gamma(\mathbf{m})). \tag{11}$$

Note that the summation starts at $\gamma = 0$ in (10), but at $\gamma = 1$ in (11). In terms of the definition in (10), Theorem 6 may be restated as

Theorem 6':

$$U(\mathbf{m}) = p(\sigma)U^{(1)}(\mathbf{m} - \sigma\nu_0),$$

where $\sigma = \alpha_k^{-1}m_k$, $p(\sigma) = Pr(x_n = \sigma)$ *and* $\nu_0 = (\alpha_0, \ldots, \alpha_k, 0, \ldots)$.

The $U^{(r)}$'s are intimately related to the U's, and analogous statements can be made about them.

Theorem 7: $U^{(r)}(\mathbf{m}) \neq 0$ *implies, for* $k \geq 1$ *and* $r = 1, \ldots k$, *that* $\mathbf{m} \in N^{k-r}$ *and, for* $k \geq 2$ *and* $r = 1, \ldots, k-1$, *that* α_k *divides* m_{k-r} *and, for* $l = 1, \ldots, k-r$,

$$\sum_{i=1}^{l} m_{k-i-r+1} \leq \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(m_0 + r + j - 1).$$

Proof: Use induction on r.

$(r = 1)$ From (10), $U^{(1)}(\mathbf{m}) \neq 0$ implies that $U(T_\gamma(\mathbf{m})) \neq 0$ for some $\gamma \geq 0$. By Theorem 3, $T_\gamma(\mathbf{m}) \in N^k$ and α_k divides $(T_\gamma(\mathbf{m}))_k$, for some $\gamma \geq 0$. Hence, from (9), $\mathbf{m} \in N^{k-1}$ and, for $k \geq 2$, α_k divides m_{k-1}. Using the inequalities in Theorem 3 on $T_\gamma(\mathbf{m})$, we have

$$\sum_{i=1}^{l} (T_\gamma(m))_{k-i+1} \leq \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(\gamma + j - 1),$$

for $l = 1, \ldots, k$. This translates into

$$\sum_{i=1}^{l} m_{k-i} \leq \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(\gamma + j - 1),$$

for $l = 1, \ldots, k-1$, and

$$\sum_{i=1}^{k} m_{k-i} - (\gamma - 1)^+ \leq \sum_{j=1}^{k} \alpha_0^{-1}\alpha_j(\gamma + j - 1).$$

Since $(T_\gamma(\mathbf{m}))_1 = m_0 - (\gamma - 1)^+$, we must have $\gamma \leq m_0 + 1$ in order for $T_\gamma(\mathbf{m}) \in N^k$. It is necessary that the m_i's satisfy the above inequalities for the largest possible γ, so we let $\gamma = m_0 + 1$. Then

$$\sum_{i=1}^{l} m_{k-i} \le \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(m_0 + j),$$

for $l = 1, \ldots, k-1$, and the other inequality is redundant, being trivial for $k = 1$, and implied for $k \ge 2$ by the inequality for $l = k - 1$.

$(r \to r+1)$ We consider $r \le k-2$, for $k \ge 3$. From (11), $U^{(r+1)}(\mathbf{m}) \ne 0$ implies that $U^{(r)}(T_\gamma(\mathbf{m})) \ne 0$ for some $\gamma \ge 1$, so that $T_\gamma(\mathbf{m}) \in N^{k-r}$ and α_k divides $(T_\gamma(\mathbf{m}))_{k-r}$. Hence, from (9), $\mathbf{m} \in N^{k-r-1}$ and α_k divides m_{k-r-1}. Also, for some $\gamma \ge 1$,

$$\sum_{i=1}^{l} (T_\gamma(\mathbf{m}))_{k-i-r+1} \le \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(\gamma + r + j - 1),$$

for $l = 1, \ldots, k-r$. As before, $\gamma \le m_0 + 1$, and for $l = 1, \ldots k-r-1$ we obtain the inequalities

$$\sum_{i=1}^{l} m_{k-i-r} \le \sum_{j=1}^{l} \alpha_0^{-1}\alpha_{k-l+j}(m_0 + r + j).$$

As before, the inequality for $l = k - r$ is redundant.

It follows from the above that $U^{(k-1)}(\mathbf{m}) \ne 0$ implies that $\mathbf{m} \in N^1$, for $k \ge 2$. Hence, from (11), $U^{(k)}(\mathbf{m}) \ne 0$, for $k \ge 2$, implies that $T_\gamma(\mathbf{m}) \in N^1$ for some $\gamma \ge 1$, so that, from (9), $\mathbf{m} \in N^0$. But we have already shown for $k = 1$ that $U^{(1)}(\mathbf{m}) \ne 0$ implies that $\mathbf{m} \in N^0$. Hence, $U^{(k)}(\mathbf{m}) \ne 0$ implies that $\mathbf{m} \in N^0$ for $k \ge 1$. ••

Corollary 8: For fixed m_0, k, r, there is only a finite number of $\mathbf{m}$ such that $U^{(r)}(\mathbf{m}) \ne 0$. Moreover, each sum that defines each $U^{(r)}$ is finite.

Proof: The first assertion is clear. For the second, we use (10) and (11) and the fact that $(T_\gamma(\mathbf{m}))_1 = m_0 - (\gamma - 1).^+$ ••

Before we derive the relations for $U^{(r)}(\mathbf{m})$ corresponding to Theorem 6′, we need some more definitions. For $r = 0, \ldots, k$ we define

$$\mu_r = \sum_{i=0}^{r} \alpha_i, \tag{12}$$

and, for $r = 0, \ldots, k-1$,

$$\nu_r = (\mu_r, \alpha_{r+1}, \ldots, \alpha_k, 0, \ldots). \tag{13}$$

We will make use of

Proposition 9: T_γ has the properties:

(*i*) $T_\gamma(\mathbf{m} + \mathbf{m}') = T_\gamma(\mathbf{m}) + R(\mathbf{m}')$.

(*ii*) *For integers $\gamma \ge 1$ and $\gamma' \ge 0$,*

$$T_{\gamma+\gamma'}(\mathbf{m}) = T_\gamma(\mathbf{m}) + (\gamma', -\gamma', 0 \ldots).$$

(*iii*) *For integers $\gamma \ge 1$ and $\sigma \ge 0$, and $r = 1, \ldots, k-2$,*

$$T_{\gamma+\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r = T_\gamma(\mathbf{m} - \sigma\nu_{r+1}).$$

Proof: (i) and (ii) follow directly from (9). Also, using (ii), for integers $\gamma \geq 1$ and $\sigma \geq 0$, we have

$$\begin{aligned} T_{\gamma+\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r &= T_\gamma(\mathbf{m}) + (\sigma\mu_r, -\sigma\mu_r, 0, \ldots) - \sigma\nu_r \\ &= T_\gamma(\mathbf{m}) - \sigma(0, \mu_r + \alpha_{r+1}, \alpha_{r+2}, \ldots, \alpha_k, 0, \ldots) \\ &= T_\gamma(\mathbf{m}) - R(\sigma\nu_{r+1}) = T_\gamma(\mathbf{m} - \sigma\nu_{r+1}). \quad ** \end{aligned}$$

For $k \geq 2$, $r = 0, \ldots, k-2$, and σ a nonnegative integer, we define

$$S_r(\mathbf{m};\sigma) = T_{\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r. \tag{14}$$

Also, for $k \geq 2$, $r = 1, \ldots, k-1$ and $s = 1, \ldots, r$, and $\alpha_k^{-1} m_{k-r}$ a nonnegative integer, we define $V_r^{(s)}(\mathbf{m})$:

$$V_r^{(s)}(\mathbf{m}) = \sum_{\gamma_1, \ldots, \gamma_{r-s} \geq 1} U^{(s)}(S_{s-1}(T_{\gamma_1} \circ \ldots \circ T_{\gamma_{r-s}}(\mathbf{m}); \alpha_k^{-1} m_{k-r})),$$

if $s \neq r$, where $\circ$ denotes composition of the operators, and

$$V_r^{(r)}(\mathbf{m}) = U^{(r)}(S_{r-1}(\mathbf{m}; \alpha_k^{-1} m_{k-r})). \tag{15}$$

Lemma 10: $V_{r+1}^{(s)}(\mathbf{m}) = \sum_{\gamma \geq 1} V_r^{(s)}(T_\gamma(\mathbf{m}))$ *for* $k \geq 3, r = 1, \ldots, k-2$ *and* $s = 1, \ldots, r$, *and* $\alpha_k^{-1}\,(T_\gamma(\mathbf{m}))_{k-r}$ *a nonnegative integer.*

Proof: We will only consider the case $r \neq s$. The proof for $r = s$ requires only a slight modification. From (15),

$$\sum_{\gamma \geq 1} V_r^{(s)}(T_\gamma(\mathbf{m}))$$

$$= \sum_{\gamma \geq 1} \sum_{\gamma_1, \ldots, \gamma_{r-s} \geq 1} U^{(s)}(S_{s-1}(T_{\gamma_1} \circ \ldots \circ T_{\gamma_{r-s}} \circ T_\gamma(\mathbf{m}); \sigma)),$$

where $\sigma = \alpha_k^{-1}(T_\gamma(\mathbf{m}))_{k-r}$. However, from (9), $(T_\gamma(\mathbf{m}))_{k-r} = m_{k-r-1}$ for $k - r \geq 2$. Therefore, $\sigma = \alpha_k^{-1}(T_\gamma(\mathbf{m}))_{k-r} = \alpha_k^{-1}(\mathbf{m})_{k-r-1}$, and if we let $\gamma = \gamma_{r+1-s}$, then the above expression is equal to $V_{r+1}^{(s)}(\mathbf{m})$. $**$

Theorem 11: *Let* $k \geq 2$. *For* $r = 1, \ldots, k-1$, *we have the following formulas*:

$$U^{(r)}(\mathbf{m}) = p(\sigma)\left[U^{(r+1)}(\mathbf{m} - \sigma\nu_r) + \sum_{s=1}^{r} V_r^{(s)}(\mathbf{m})\right]$$

where $\sigma = \alpha_k^{-1}(\mathbf{m})_{k-r}$ *and* $\sigma \neq 0$. *If* $\sigma = 0$, *then*

$$U^{(r)}(\mathbf{m}) = p(0)[U^{(r+1)}(\mathbf{m}) + V_r^{(1)}(\mathbf{m})].$$

Proof: We use induction on r.

$(r = 1)$ Note here that the two cases coincide. From (10) and Theorem 6′,

$$U^{(1)}(\mathbf{m}) = \sum_{\gamma \geq 0} U(T_\gamma(\mathbf{m})) = \sum_{\gamma \geq 0} p(\sigma) U^{(1)}(T_\gamma(\mathbf{m}) - \sigma\nu_0),$$

where $\sigma = \alpha_k^{-1}(T_\gamma(\mathbf{m}))_k$. But $(T_\gamma(\mathbf{m}))_k = m_{k-1}$ for $k \geq 2$; therefore, $\sigma = \alpha_k^{-1} m_{k-1}$ and so σ is independent of γ, and

$$U^{(1)}(\mathbf{m}) = p(\sigma) \sum_{\gamma \geq 0} U^{(1)}(T_\gamma(\mathbf{m}) - \sigma\nu_0).$$

If σ is not a nonnegative integer, then $p(\sigma) = 0$ and $U^{(1)}(\mathbf{m}) = 0$, and the required result holds trivially. If σ is a nonnegative integer, then we let $q = \gamma - \sigma\mu_0$ and obtain

$$U^{(1)}(\mathbf{m}) = p(\sigma) \sum_{q \geq -\sigma\mu_0} U^{(1)}(T_{q+\sigma\mu_0}(\mathbf{m}) - \sigma\nu_0).$$

The zeroth term of $T_{q+\sigma\mu_0}(\mathbf{m}) - \sigma\nu_0$ is q, so, by Theorem 7, any terms where $q < 0$ vanish. Hence,

$$\begin{aligned} U^{(1)}(\mathbf{m}) &= p(\sigma)\left[U^{(1)}(T_{\sigma\mu_0}(\mathbf{m}) - \sigma\nu_0) + \sum_{q \geq 1} U^{(1)}(T_{q+\sigma\mu_0}(\mathbf{m}) - \sigma\nu_0)\right] \\ &= p(\sigma)\left[U^{(1)}(S_0(\mathbf{m};\sigma)) + \sum_{q \geq 1} U^{(1)}(T_q(\mathbf{m} - \sigma\nu_1))\right] \\ &= p(\sigma)[V_1^{(1)}(\mathbf{m}) + U^{(2)}(\mathbf{m} - \sigma\nu_1)]. \end{aligned}$$

The last two steps follow from (11), (14), and (15), and Proposition 9, and the fact that $\sigma = \alpha_k^{-1} m_{k-1}$, to use the definition of $V_1^{(1)}$.

$(r \to r+1)$ We consider $r \leq k-2$, for $k \geq 3$, and first assume that $\sigma = \alpha_k^{-1} m_{k-r-1} \neq 0$. But $(T_\gamma(\mathbf{m}))_{k-r} = m_{k-r-1}$, for $r \leq k-2$. Hence, $\sigma = \alpha_k^{-1}(T_\gamma(\mathbf{m}))_{k-r} \neq 0$, and we may use our inductive hypothesis on $U^{(r)}(T_\gamma(\mathbf{m}))$. From (11), since σ is independent of γ, we obtain

$$U^{(r+1)}(\mathbf{m}) = p(\sigma) \sum_{\gamma \geq 1}\left[U^{(r+1)}(T_\gamma(\mathbf{m}) - \sigma\nu_r) + \sum_{s=1}^{r} V_r^{(s)}(T_\gamma(\mathbf{m}))\right].$$

If $\sigma \neq 0$ is not a positive integer, then $p(\sigma) = 0$ and $U^{(r+1)}(\mathbf{m}) = 0$, and the required result holds trivially.

If σ is positive, then, using Lemma 10, we have

$$U^{(r+1)}(\mathbf{m}) = p(\sigma)\left[\sum_{\gamma \geq 1} U^{(r+1)}(T_\gamma(\mathbf{m}) - \sigma\nu_r) + \sum_{s=1}^{r} V_{r+1}^{(s)}(\mathbf{m})\right].$$

Also, if we let $q = \gamma - \sigma\mu_r$, then

$$\sum_{\gamma \geq 1} U^{(r+1)}(T_\gamma(\mathbf{m}) - \sigma\nu_r) = \sum_{q \geq 1 - \sigma\mu_r} U^{(r+1)}(T_{q+\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r).$$

But σ and μ_r are positive integers, so $\sigma\mu_r \geq 1$. Hence, by a similar argument to the case $r = 1$,

$$\begin{aligned} \sum_{\gamma \geq 1} U^{(r+1)}(T_\gamma(\mathbf{m}) - \sigma\nu_r) &= \sum_{q \geq 0} U^{(r+1)}(T_{q+\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r) \\ &= U^{(r+1)}(T_{\sigma\mu_r}(\mathbf{m}) - \sigma\nu_r) + \sum_{q \geq 1} U^{(r+1)}(T_q(\mathbf{m} - \sigma\nu_{r+1})) \\ &= V_{r+1}^{(r+1)}(\mathbf{m}) + U^{(r+2)}(\mathbf{m} - \sigma\nu_{r+1}), \end{aligned}$$

where we have used (11), (14), and (15), and Proposition 9. Conse-

quently,

$$U^{(r+1)}(\mathbf{m}) = p(\sigma)\left[U^{(r+2)}(\mathbf{m}-\sigma\nu_{r+1}) + \sum_{s=1}^{r+1} V_{r+1}^{(s)}(\mathbf{m})\right],$$

where $\sigma = \alpha_k^{-1}m_{k-r-1}$.

Finally, we consider the case $\sigma = 0$, so that $(T_\gamma(\mathbf{m}))_{k-r} = m_{k-r-1} = 0$. Then, using (11) and the inductive hypothesis on $U^{(r)}(T_\gamma(\mathbf{m}))$, we have

$$\begin{aligned} U^{(r+1)}(\mathbf{m}) &= p(0)\sum_{\gamma\geq 1}[U^{(r+1)}(T_\gamma(\mathbf{m})) + V_r^{(1)}(T_\gamma(\mathbf{m}))] \\ &= p(0)[U^{(r+2)}(\mathbf{m}) + V_{r+1}^{(1)}(\mathbf{m})], \end{aligned}$$

from Lemma 10. **

Having derived the reduction formulas of Theorem 11, we now comment on the quantities $V_r^{(s)}(m)$ defined in (15), under the assumption that $\alpha_k^{-1}m_{k-r}$ is a nonnegative integer. It may be verified, from the definitions in (9), (13), and (14), that, for $k \geq 2$ and $r = 1, \ldots, k-1, S_{r-1}(\mathbf{m};\alpha_k^{-1}m_{k-r}) \in N^{k-r}$ implies that $\mathbf{m} \in N^{k-r}$. Also, for $k \geq 3$, $r = 2, \ldots, k-1, s = 1, \ldots, r-1$, and positive integers $\gamma_1, \ldots, \gamma_{r-s}$, $S_{s-1}(T_{\gamma_1}\circ \ldots \circ T_{\gamma_{r-s}}(\mathbf{m});\ \alpha_k^{-1}m_{k-r}) \in N^{k-s}$ implies that $\mathbf{m} \in N^{k-r}$. It follows, from Theorem 7, that, for $k \geq 2$, $r = 1, \ldots, k-1$ and $s = 1, \ldots, r$, $V_r^{(s)}(\mathbf{m}) \neq 0$ implies that $\mathbf{m} \in N^{k-r}$. Also, for $r = 0, \ldots, k-1$, we note that $(\mathbf{m} - \sigma\nu_r) \in N^{k-r-1}$, where $\sigma = \alpha_k^{-1}m_{k-r}$ is a nonnegative integer, implies that $\mathbf{m} \in N^{k-r}$.

IV. THE STEADY-STATE PROBABILITIES

We define the sets

$$\Omega_i = \{U(\mathbf{m}) | 0 \leq m_0 \leq i, \mathbf{m} \text{ a proper state}\}, \tag{16}$$

where the proper states satisy the criteria of Theorem 3. The sets Ω_i are finite, for fixed k, by Corollary 4. We will first show how to calculate the elements of Ω_0. Then, as shown by Gopinath and Morrison,[1,2] the steady-state generating function for the buffer content can be calculated in terms of the generating functions for some marginal distributions. The marginals are finitely solvable, in the sense that a finite number of components of the marginal distributions can be solved for, from a finite number of linear equations. However, we will give an alternate method for calculating the steady-state probability that the buffer content is i, which also involves a finite number of linear equations. In fact, by induction on i, we show how to calculate the elements of Ω_i, $i = 1,2, \ldots$, and hence $\kappa_1, \ldots, \kappa_i$, from (7).

We begin by defining the sets

$$\Lambda_i = \{U^{(r)}(\mathbf{m}) | 0 \leq m_0 \leq i, 1 \leq r \leq k, \mathbf{m} \text{ a proper state}\}, \tag{17}$$

where the proper states satisfy the criteria of Theorem 7. The sets Λ_i are finite, for fixed k, by Corollary 8. We also define the sets Λ_i^*, which are

obtained from Λ_i by deleting the single element $U^{(k)}(i,0,0,\ldots)$, that is

$$\Lambda_i^* = \Lambda_i \sim \{U^{(k)}(i,0,0,\ldots)\}. \tag{18}$$

We will first show how to determine the elements of Λ_0^*, and thence the elements of Ω_0. We will make use of

Lemma 12: *For $k \geq 2$, $r = 1,\ldots,k-1$ and $s = 1,\ldots r$, the quantities $V_r^{(s)}(\mathbf{m})$ are linear combinations of the elements of Λ_0^*.*

Proof: From (9), (13), and (14), it follows that $(S_{s-1}(\mathbf{m};\sigma))_0 = 0$. The result is then a consequence of the definitions in (15), (17), and (18). **

If $k=1$, then Λ_0 contains the single element $U^{(1)}(0,0,\ldots)$, since $\mathbf{m} \in N^{k-r}$ for a proper state, and hence the set Λ_0^* is empty. If $k \geq 2$, then Λ_0^* contains at least one element, namely, $U^{(k-1)}(0,0,\ldots)$. (If $k = 2$, this might be the only element.) We now apply the reduction formula of Theorem 11 to each element of $\Lambda_0^* \sim U^{(k-1)}(0,0,\ldots)$. But for $m_0 = 0$ and σ a positive integer, $(\mathbf{m} - \sigma\nu_r)_0 < 0$, and hence $U^{(r+1)}(\mathbf{m} - \sigma\nu_r) = 0$. Hence, from Lemma 12, we obtain a system of homogeneous linear equations which contain as unknowns only the elements of Λ_0^*. Note that we have omitted the reduction formula for $U^{(k-1)}(0,0,\ldots)$. Since there is one more unknown than the number of equations, we can solve for the elements of Λ_0^* to within a multiplicative constant.

We are now in a position to determine the elements of Ω_0. If $k = 1$, then, from the inequality in Theorem 3, Ω_0 contains just the single element $U(0,0,\ldots) = \kappa_0$, from (7), and κ_0 is given by the formula in Proposition 1. If $k \geq 2$, then the elements of Ω_0 are given by Theorem 6′ in terms of elements of Λ_0^*, since $(\mathbf{m} - \sigma\nu_0)_0 \leq 0$ if $m_0 = 0$ and σ is a nonnegative integer, and $U^{(1)}(\mathbf{m} - \sigma\nu_0) = 0$ if $(\mathbf{m} - \sigma\nu_0)_0 < 0$. Hence, the elements of Ω_0 are determined to within a multiplicative constant, which is determined by (7), in terms of κ_0. The elements of Λ_0^* are now also completely determined.

We next turn our attention to the calculation of the elements of Λ_i^* and Ω_i, for $i = 1,2,\ldots$. First, however, we need

Lemma 13: *The assumption $\mu_k E(x) < 1$ implies that $p(0) > 0$.*

Proof: $E(x) = \sum_{i=1}^{\infty} ip(i) \geq \sum_{i=1}^{\infty} p(i) = 1 - p(0)$. But, from (12), since $\alpha_0 \neq 0 \neq \alpha_k$, it follows that $\mu_k \geq 2$, and hence $E(x) < \frac{1}{2}$. **

We have shown how to determine the elements of Λ_0^* and Ω_0. We will show how to determine the elements of Λ_i^* and Ω_i, for $i = 1,2,\ldots$. We first consider the special case $k=1$, and use induction on i.

Theorem 14: *For $k = 1$, if the elements of Λ_i^* and Ω_i are known, then the elements of $\Lambda_{i+1}^* \sim \Lambda_i^*$ and $\Omega_{i+1} \sim \Omega_i$ may be determined.*

Proof: From (17) and (18), since $k = 1$ and $\mathbf{m} \in N^{k-r}$ for a proper state,

$$\Lambda_{i+1}^* \sim \Lambda_i^* = \{U^{(1)}(i,0,\ldots)\}.$$

But, from Theorem 6′,

$$U(i,0,\ldots) = p(0)U^{(1)}(i,0,\ldots).$$

Since $p(0) > 0$, by Lemma 13, and $U(i,0,\ldots) \in \Omega_i$, this equation determines $U^{(1)}(i,0,\ldots)$. Also, from Theorem 6′, $U(i+1,m_1,0,\ldots)$ is determined for $m_1 \neq 0$, if there are any such elements in Ω_{i+1}, since $\sigma > 0$ and so $(\mathbf{m} - \sigma\nu_0)_0 < i+1$. The remaining element of $\Omega_{i+1} \sim \Omega_i$ is $U(i+1,0,\ldots)$, since $\mathbf{m} \in N^1$ for a proper state. But, from (9) and (10),

$$U(i+1,0,\ldots) = U^{(1)}(i,0,\ldots) - \sum_{\gamma=0}^{i} U(\gamma, i-(\gamma-1)^+,0,\ldots),$$

so that the remaining element is determined. ⁂

We now consider the general case, and establish

Theorem 15: *For $k \geq 2$, if the elements of Λ_i^* are known, then the elements of $\Lambda_{i+1}^* \sim \Lambda_i^*$ may be determined.*

Proof: From Theorem 11,

$$U^{(k-1)}(i,0,\ldots) = p(0)[U^{(k)}(i,0,\ldots) + V_{k-1}^{(1)}(i,0,\ldots)],$$

which equation was omitted for $i = 0$. This equation determines $U^{(k)}(i,0,\ldots)$, by Lemmas 12 and 13, since $U^{(k-1)}(i,0,\ldots) \in \Lambda_i^*$. Also, if $m_0 = i+1$, $1 \leq r \leq k-1$ and $\sigma = \alpha_k^{-1} m_{k-r} > 0$, then $U^{(r)}(\mathbf{m})$ is determined by Theorem 11, since $(\mathbf{m} - \sigma\nu_r)_0 < i+1$ for $\sigma > 0$. The remaining elements of $\Lambda_{i+1}^* \sim \Lambda_i^*$ are $U^{(r)}(\mathbf{m})$ with $m_0 = i+1$, $1 \leq r \leq k-1$ and $m_{k-r} = 0$.

But from (11),

$$U^{(k-1)}(i+1,0,\ldots) = U^{(k)}(i,0,\ldots) - \sum_{\gamma=1}^{i} U^{(k-1)}(\gamma, i-\gamma+1,0,\ldots),$$

where the summation is absent if $i = 0$. This determines $U^{(k-1)}(i+1,0,\ldots)$, and if $k = 2$ this is the only remaining element in $\Lambda_{i+1}^* \sim \Lambda_i^*$. If $k \geq 3$, there still remain $U^{(r)}(\mathbf{m})$ with $m_0 = i+1$, $1 \leq r \leq k-2$ and $m_{k-r} = 0$, and from Theorem 11,

$$U^{(r)}(\mathbf{m}) = p(0)[U^{(r+1)}(\mathbf{m}) + V_r^{(1)}(\mathbf{m})].$$

But we have just determined $U^{(k-1)}(i+1,0,\ldots)$, and so we know $U^{(k-1)}(i+1,m_1,\ldots)$ for $m_1 \geq 0$. Hence, from the above equation, by Lemma 12, we may determine $U^{(k-2)}(\mathbf{m})$ with $m_0 = i+1$, and $m_2 = 0$. We then know $U^{(k-2)}(\mathbf{m})$ with $m_0 = i+1$ and $m_2 \geq 0$. By iteration of the above equation, we may determine any remaining elements of $\Lambda_{i+1}^* \sim \Lambda_i^*$. ⁂

Lemma 16: *For $k \geq 2$, the elements of Ω_i are determined by elements of Λ_i^*, for $i = 1,2,\ldots$.*

Proof: The result follows from Theorem 6′. ⁂

We have shown that the elements of Λ_i^* and Ω_i, for $i = 1,2,\ldots$, may be determined explicitly, once the elements of Λ_0^*, and Ω_0, are known.

The determination of the elements of Λ_0^*, however, involves the solution of a homogeneous system of linear equations.

V. A PARTICULAR CASE

We now confine our attention to the particular case $\xi_n = x_n + x_{n-k}$, so that, from (2),

$$\alpha_0 = 1 = \alpha_k, \quad \alpha_j = 0 \text{ otherwise.} \tag{19}$$

We are interested in determining the number of proper states $\mathbf{m}$ of $U(\mathbf{m})$, and also of $U^{(r)}(\mathbf{m})$, as defined by the criteria of Theorems 3 and 7. We show in the appendix that these criteria lead to a precise count of the number of nonzero U's and $U^{(r)}$'s when (19) holds, if $p(i) > 0$, $i = 0,1,2,\ldots$.

We will make use of

Lemma 17: *For* $r = -1,0,1,\ldots$, *and* $s = 1,2,\ldots$, *the number of elements of* $\mathbf{n} \in N^{s-1}$ *which satisfy the conditions* $\sum_{i=0}^{l-1} n_i \leq r + l$ *for* $l = 1,\ldots,s$ *is*

$$P(r,s) = \binom{r+2s}{s} - \binom{r+2s}{s-2} = \frac{(r+2)(r+2s+1)!}{s!(r+s+2)!} \equiv F(r,s). \tag{20}$$

Proof: We use induction on s.

($s = 1$) The number of n_0 with $0 \leq n_0 \leq r+1$ is clearly $r + 2 = F(r,1)$.

($s \to s+1$) Now $\sum_{i=0}^{l-1} n_i \leq r + l$ for $l = 1,\ldots,s+1$ implies that $n_0 \leq r+1$ and $\sum_{i=1}^{l} n_i \leq r + l + 1 - n_0$ for $l = 1,\ldots,s$. Hence,

$$P(r,s+1) = \sum_{n_0=0}^{r+1} P(r+1-n_0,s) = \sum_{i=0}^{r+1} P(i,s) = \sum_{i=0}^{r+1} F(i,s),$$

from the inductive hypothesis. But, as may be verified,

$$F(i,s) = F(i-1,s+1) - F(i-2,s+1). \tag{21}$$

Hence,

$$\sum_{i=0}^{r+1} F(i,s) = F(r,s+1), \tag{22}$$

since $F(-2,s+1) = 0$. ∎∎

Corollary 18: *For fixed* m_0 *and* k, *the number of proper states* $\mathbf{m}$ *of* $U(\mathbf{m})$ *is* $F(m_0 - 1,k)$, *and the number of proper states* $\mathbf{m}$ *of* $U^{(r)}(\mathbf{m})$ *is* $F(m_0 + r - 1, k - r)$, *for* $r = 1,\ldots,k$.

Proof: The results follow from (19), Theorems 3 and 7, and Lemma 17. Note that, for $r = k$, the only proper state of $U^{(k)}(\mathbf{m})$ is $(m_0,0,\ldots)$, since $\mathbf{m} \in N^0$, and we have $F(m_0 + k - 1,0) = 1$. ∎∎

From (16) and Corollary 18, it follows that the number of elements of Ω_0 is

$$|\Omega_0| = F(-1,k) = \frac{(2k)!}{k!(k+1)!}. \tag{23}$$

Also, the number of elements of $\Omega_i \sim \Omega_0$ is

$$|\Omega_i| - |\Omega_0| = \sum_{m_0=1}^{i} F(m_0 - 1,k) = F(i-2,k+1)$$
$$= \frac{i(i+2k+1)!}{(k+1)!(i+k+1)!}, \tag{24}$$

from (20) and (22). From (17) and Corollary 18, the number of elements of Λ_i is

$$|\Lambda_i| = \sum_{m_0=0}^{i} \sum_{r=1}^{k} F(m_0 + r - 1, k - r). \tag{25}$$

But

$$F(-(r+4),r+s+2) = \frac{-(r+2)(r+2s+1)!}{(r+s+2)!s!} = -F(r,s). \tag{26}$$

Therefore, from (21) and (26), we have

$$F(m_0 + r - 1, k - r) = -F(-(m_0 + r + 3), m_0 + k + 1)$$
$$= -[F(-(m_0 + r + 4), m_0 + k + 2) - F(-(m_0 + r + 5), m_0 + k + 2)].$$

Hence, if we sum and use (26), we obtain

$$\sum_{r=1}^{k} F(m_0 + r - 1, k - r)$$
$$= -[F(-(m_0+5), m_0 + k + 2) - F(-(m_0 + k + 5), m_0 + k + 2)]$$
$$= F(m_0 + 1, k - 1) - F(m_0 + k + 1, -1) = F(m_0 + 1, k-1). \tag{27}$$

From (21) and (25), it follows that

$$|\Lambda_i| = F(i,k) - F(-1,k). \tag{28}$$

Note, from (23), (24), and (28), that

$$|\Omega_i| + |\Lambda_i| = F(i-2,k+1) + F(i,k) = F(i-1,k+1),$$

from (21).

Of particular interest, for $k \geq 2$, is the number of equations required to determine the elements of Λ_0^* to within a multiplicative constant, namely $|\Lambda_0^*| - 1 = |\Lambda_0| - 2$, from (18). But, from (21) and (28),

$$|\Lambda_0| = F(0,k) - F(-1,k) = F(1,k-1) = \frac{3(2k)!}{(k-1)!(k+2)!}.$$

The first few values of $|\Lambda_0| - 2$ and $|\Omega_0|$, as given by (23), are

k	2	3	4	5	6
$\|\Lambda_0\| - 2$	1	7	26	88	295
$\|\Omega_0\|$	2	5	14	42	132

We also have the asymptotic result

$$\lim_{k\to\infty} \frac{(|\Lambda_0| - 2)}{|\Omega_0|} = 3.$$

VI. AN EXPLICIT EXAMPLE

We here consider the example corresponding to $k = 4$ in (19), so that $\xi_n = x_n + x_{n-4}$. We will explicitly determine the elements of Ω_0 for this example. As discussed in Section IV, the reduction formula of Theorem 11 is applied to each element of $\Lambda_0^* \sim U^{(3)}(0,0,\ldots)$. Then the elements of Ω_0 are determined with the help of Theorem 6′ and the normalization condition (7) with $i = 0$.

From (17), (18), and Theorem 7, the elements of Λ_0^*, with an obvious change of notation, are

$$U^{(1)}_{0m_1m_2m_3}, \quad m_3 \le 1, m_2 + m_3 \le 2, m_1 + m_2 + m_3 \le 3, \tag{29}$$

$$U^{(2)}_{0m_1m_2}, \quad m_2 \le 2, m_1 + m_2 \le 3, \tag{30}$$

and

$$U^{(3)}_{0m_1}, \quad m_1 \le 3, \tag{31}$$

where m_1, m_2 and m_3 are nonnegative integers. From (16) and Theorem 3, the elements of Ω_0 are

$$U_{0m_1m_2m_30}, \quad m_3 \le 1, m_2 + m_3 \le 2, m_1 + m_2 + m_3 \le 3. \tag{32}$$

But from Theorem 6′, again with an obvious change of notation,

$$U_{0m_1m_2m_30} = p_0 U^{(1)}_{0m_1m_2m_3}. \tag{33}$$

From (7), the normalization condition is

$$\kappa_0 = \sum U_{0m_1m_2m_30} = p_0 \sum U^{(1)}_{0m_1m_2m_3}, \tag{34}$$

where the summations are over the range of subscripts satisfying the inequalities in (29) and (32).

We now apply the reduction formula of Theorem 11 to each element of $\Lambda_0^* \sim U^{(3)}_{00}$, and note, from (12) and (19), that

$$\mu_r = 1, \qquad r = 0,1,2,3. \tag{35}$$

From (9), (13), and (14), with $\mathbf{m} = (m_0,m_1,m_2,m_3,0,\ldots)$, we have

$$\mathbf{m} - m_3\nu_1 = (m_0 - m_3,m_1,m_2,0,\ldots), \tag{36}$$

and

$$S_0(\mathbf{m};m_3) = (0,m_0 - (m_3 - 1)^+,m_1,m_2,0,\ldots). \tag{37}$$

Hence, from (15) and (37),

$$V_1^{(1)}(\mathbf{m}) = U^{(1)}_{0,m_0-(m_3-1)^+,m_1,m_2}. \tag{38}$$

It follows from Theorem 11 that

$$U^{(1)}_{0m_1m_20} = p_0(U^{(2)}_{0m_1m_2} + U^{(1)}_{00m_1m_2}), \tag{39}$$

and

$$U^{(1)}_{0m_1m_21} = p_1 U^{(1)}_{00m_1m_2}, \tag{40}$$

since $U^{(2)}_{-1,m_1,m_2} = 0$.

Similarly, with $\mathbf{m} = (m_0,m_1,m_2,0,\ldots)$,

$$\mathbf{m} - m_2\nu_2 = (m_0 - m_2,m_1,0,\ldots), \tag{41}$$

$$S_1(\mathbf{m};m_2) = (0,m_0 - (m_2 - 1)^+,m_1,0,\ldots), \tag{42}$$

and

$$S_0(T_{\gamma_1}(\mathbf{m});m_2) = (0,\gamma_1 - (m_2 - 1)^+,m_0 - (\gamma_1 - 1)^+,m_1,0,\ldots). \tag{43}$$

Hence, from (15),

$$V_2^{(2)}(\mathbf{m}) = U^{(2)}_{0,m_0-(m_2-1)^+,m_1}, \tag{44}$$

and

$$V_2^{(1)}(\mathbf{m}) = \sum_{\gamma_1\geq 1} U^{(1)}_{0,\gamma_1-(m_2-1)^+,m_0-(\gamma_1-1)^+,m_1}. \tag{45}$$

It follows from Theorem 11 that

$$U^{(2)}_{0m_10} = p_0(U^{(3)}_{0m_1} + U^{(1)}_{010m_1}), \tag{46}$$

and, for $m_2 \neq 0$,

$$U^{(2)}_{0m_1m_2} = p_{m_2}(U^{(2)}_{0,-(m_2-1)^+,m_1} + U^{(1)}_{0,1-(m_2-1)^+,0,m_1}). \tag{47}$$

Hence,

$$U^{(2)}_{0m_11} = p_1(U^{(2)}_{00m_1} + U^{(1)}_{010m_1}) \tag{48}$$

and

$$U^{(2)}_{0m_12} = p_2 U^{(1)}_{000m_1}. \tag{49}$$

Next, with $\mathbf{m} = (m_0,m_1,0,\ldots)$,

$$\mathbf{m} - m_1\nu_3 = (m_0 - m_1,0,\ldots), \tag{50}$$

$$S_2(\mathbf{m};m_1) = (0,m_0 - (m_1 - 1)^+,0,\ldots), \tag{51}$$

$$S_1(T_{\gamma_1}(\mathbf{m});m_1) = (0,\gamma_1 - (m_1 - 1)^+,m_0 - (\gamma_1 - 1)^+,0,\ldots), \tag{52}$$

and

$$S_0(T_{\gamma_1}(T_{\gamma_2}(\mathbf{m}));m_1)$$
$$= (0,\gamma_1 - (m_1 - 1)^+,\gamma_2 - (\gamma_1 - 1)^+, m_0 - (\gamma_2 - 1)^+,0,\ldots). \quad (53)$$

Hence, from (15),

$$V_3^{(3)}(\mathbf{m}) = U^{(3)}_{0,m_0-(m_1-1)^+}, \quad (54)$$

$$V_3^{(2)}(\mathbf{m}) = \sum_{\gamma_1\geq 1} U^{(2)}_{0,\gamma_1-(m_1-1)^+,m_0-(\gamma_1-1)^+}, \quad (55)$$

and

$$V_3^{(1)}(\mathbf{m}) = \sum_{\gamma_1,\gamma_2\geq 1} U^{(1)}_{0,\gamma_1-(m_1-1)^+,\gamma_2-(\gamma_1-1)^+,m_0-(\gamma_2-1)^+}. \quad (56)$$

It follows from Theorem 11 that, for $m_1 \neq 0$,

$$U^{(3)}_{0m_1} = p_{m_1}\Big(U^{(3)}_{0,-(m_1-1)^+} + U^{(2)}_{0,1-(m_1-1)^+,0} + \sum_{\gamma_1\geq 1} U^{(1)}_{0,\gamma_1-(m_1-1)^+,1-(\gamma_1-1)^+,0}\Big). \quad (57)$$

We now write out in full the nontrivial equations corresponding to (39), (40), (46), (48), (49), and (57), omitting terms which are identically zero. From (39) we have

$$U^{(1)}_{0000} = p_0(U^{(2)}_{000} + U^{(1)}_{0000}), \qquad U^{(1)}_{0010} = p_0(U^{(2)}_{001} + U^{(1)}_{0001}),$$
$$U^{(1)}_{0020} = p_0U^{(2)}_{002}, \qquad U^{(1)}_{0100} = p_0(U^{(2)}_{010} + U^{(1)}_{0010}),$$
$$U^{(1)}_{0110} = p_0(U^{(2)}_{011} + U^{(1)}_{0011}), \qquad U^{(1)}_{0120} = p_0U^{(2)}_{012},$$
$$U^{(1)}_{0200} = p_0(U^{(2)}_{020} + U^{(1)}_{0020}), \qquad U^{(1)}_{0210} = p_0U^{(2)}_{021},$$
$$U^{(1)}_{0300} = p_0U^{(2)}_{030}, \quad (58)$$

and from (40) we have

$$U^{(1)}_{0001} = p_1U^{(1)}_{0000}, \qquad U^{(1)}_{0011} = p_1U^{(1)}_{0001}, \qquad U^{(1)}_{0101} = p_1U^{(1)}_{0010},$$
$$U^{(1)}_{0111} = p_1U^{(1)}_{0011}, \qquad U^{(1)}_{0201} = p_1U^{(1)}_{0020}. \quad (59)$$

Next, from (46) we have

$$U^{(2)}_{000} = p_0(U^{(3)}_{00} + U^{(1)}_{0100}), \qquad U^{(2)}_{010} = p_0(U^{(3)}_{01} + U^{(1)}_{0101}),$$
$$U^{(2)}_{020} = p_0U^{(3)}_{02}, \qquad U^{(2)}_{030} = p_0U^{(3)}_{03}, \quad (60)$$

from (48) we have

$$U^{(2)}_{001} = p_1(U^{(2)}_{000} + U^{(1)}_{0100}),$$
$$U^{(2)}_{011} = p_1(U^{(2)}_{001} + U^{(1)}_{0101}), \qquad U^{(2)}_{021} = p_1U^{(2)}_{002}, \quad (61)$$

and from (49) we have

$$U^{(2)}_{002} = p_2U^{(1)}_{0000}, \qquad U^{(2)}_{012} = p_2U^{(1)}_{0001}. \quad (62)$$

Finally, from (57) we have

$$U_{01}^{(3)} = p_1(U_{00}^{(3)} + U_{010}^{(2)} + U_{0110}^{(1)} + U_{0200}^{(1)}),$$
$$U_{02}^{(3)} = p_2(U_{000}^{(2)} + U_{0010}^{(1)} + U_{0100}^{(1)}), \qquad U_{03}^{(3)} = p_3 U_{0000}^{(1)}. \tag{63}$$

We may eliminate the 13 nonzero quantities $U_{0m_1m_2}^{(2)}$ and $U_{0m_1}^{(3)}$ from (58) to (63), and solve for the 14 nonzero quantities $U_{0m_1m_2m_3}^{(1)}$ to within a multiplicative constant. It is found that

$$U_{0000}^{(1)} = a_0, \qquad U_{0001}^{(1)} = p_1 a_0, \qquad U_{0011}^{(1)} = p_1^2 a_0,$$
$$U_{0111}^{(1)} = p_1^3 a_0, \qquad U_{0020}^{(1)} = p_0 p_2 a_0, \qquad U_{0300}^{(1)} = p_0^2 p_3 a_0,$$
$$U_{0120}^{(1)} = U_{0210}^{(1)} = U_{0201}^{(1)} = p_0 p_1 p_2 a_0, \tag{64}$$

and

$$U_{0010}^{(1)} = p_1 \Delta a_0, \qquad U_{0101}^{(1)} = p_1^2 \Delta a_0,$$
$$U_{0110}^{(1)} = p_1^2(1 + p_0 p_1)\Delta a_0, \qquad U_{0200}^{(1)} = p_0 p_2(1 + p_0 p_1)\Delta a_0,$$
$$U_{0100}^{(1)} = p_1[1 + p_0^2(p_1^2 + p_0 p_2)(1 + p_0 p_1)]\Delta a_0, \tag{65}$$

where

$$\Delta = \{1 - p_0 p_1[1 + p_0^2(p_1^2 + p_0 p_2)(1 + p_0 p_1)]\}^{-1}. \tag{66}$$

The constant a_0 is determined by the normalization condition (34). These results are consistent with those derived by a different method.[3]

APPENDIX

We show here that in the particular case corresponding to (19), the criteria of Theorems 3 and 7 lead to a precise count of the number of nonzero U's and $U^{(r)}$'s if $p(i) > 0$, $i = 0,1,2,\ldots$. We first prove

Theorem 19: *If* (19) *holds*, $p(i) > 0$, $i = 0,1,2,\ldots$, $\mathbf{m} \in N^k$ *and*

$$\sum_{i=1}^{l} m_{k-i+1} \leq m_0 + l - 1, \qquad l = 1,\ldots,k, \tag{67}$$

then $U(\mathbf{m}) \neq 0$.

Proof: From (4), (5), and (19), it follows that

$$\mathbf{B}_n = (b_n, x_{n-k}, \ldots, x_{n-1}, 0, \ldots). \tag{68}$$

It was shown[2] that the irreducible Markov chain, with state space consisting of those states which communicate with $(0,0,\ldots)$, is positive recurrent. Moreover, it was also shown that, in the present notation, the state $(i_0,0,\ldots)$ communicates with the state $(0,0,\ldots)$, where i_0 is a positive integer. Hence, with probability 1, the state $(i_0,0,\ldots)$ occurs

infinitely often. We now assume that $\mathbf{B}_{n-k}$ is given by

$$b_{n-k} = m_0 + k - \sum_{i=1}^{k} m_{k-i+1}, \qquad x_{n-2k+r-1} = 0, \qquad r = 1, \ldots, k. \tag{69}$$

Also, with positive probability, since $p(i) > 0, \quad i = 0,1,2, \ldots,$

$$x_{n-k+r-1} = m_r, \qquad r = 1, \ldots, k. \tag{70}$$

We will show that (67), (69), and (70) imply that $b_n = m_0$, and hence, from (6), that $U(\mathbf{m}) \neq 0$.

We first show, by induction, that

$$b_{n-k+r} = m_0 + k - r - \sum_{i=1}^{k-r} m_{k-i+1} \geq 1, \qquad r = 0, \ldots, k-1. \tag{71}$$

This is true for $r = 0$, from (67) and (69).

$(r \to r+1)$ We consider $r = 0, \ldots, k-2$, for $k \geq 2$. Since $\xi_n = x_n + x_{n-k}$, it follows from (1), (69), and (70) that

$$\begin{aligned} b_{n-k+r+1} &= (b_{n-k+r} - 1)^+ + m_{r+1} \\ &= b_{n-k+r} - 1 + m_{r+1} \\ &= m_0 + k - (r+1) - \sum_{i=1}^{k-r-1} m_{k-i+1} \geq 1, \end{aligned} \tag{72}$$

from (67). This completes the inductive proof of (71). Finally, with the help of (71), we obtain

$$b_n = (b_{n-1} - 1)^+ + m_k = b_{n-1} - 1 + m_k = m_0. \quad **$$

We now prove

Theorem 20: Suppose that (19) *holds and* $p(i) > 0, \quad i = 0,1,2, \ldots$. *Also suppose, for* $k \geq 1$ *and* $r = 1, \ldots, k$, *that* $\mathbf{m} \in N^{k-r}$ *and, for* $k \geq 2$ *and* $r = 1, \ldots, k-1$, *that*

$$\sum_{i=1}^{l} m_{k-i-r+1} \leq m_0 + r + l - 1, \qquad l = 1, \ldots, k-r. \tag{73}$$

Then $U^{(r)}(\mathbf{m}) \neq 0$.

Proof: Use induction on r. We note, from (6), (10), and (11), that $U^{(r)}(\mathbf{m}) \geq 0, \quad r = 1, \ldots, k$.

$(r = 1)$ Let

$$\hat{\mathbf{m}} = T_{m_0+1}(\mathbf{m}) = (m_0 + 1, 0, m_1, \ldots), \tag{74}$$

from (9). We will show that $U(\hat{\mathbf{m}}) \neq 0$, which implies that $U^{(1)}(\mathbf{m}) \neq 0$, from (10). If $k = 1$, then $\hat{\mathbf{m}} = (m_0 + 1, 0, 0, \ldots)$, hence $\hat{\mathbf{m}} \in N^1$ and $0 = \hat{m}_1 \leq \hat{m}_0 = m_0 + 1$. It follows from Theorem 19 that $U(\hat{\mathbf{m}}) \neq 0$. If $k \geq 2$, then $\hat{\mathbf{m}} \in N^k$ since $\mathbf{m} \in N^{k-1}$, and, from (73),

$$\sum_{i=1}^{l} \hat{m}_{k-i+1} \le \hat{m}_0 + l - 1, \qquad l = 1, \ldots, k-1,$$

$$\sum_{i=1}^{k} \hat{m}_{k-i+1} = \sum_{i=1}^{k-1} \hat{m}_{k-i+1} \le \hat{m}_0 + k - 2 \le \hat{m}_0 + k - 1.$$

It follows from Theorem 19 that $U(\hat{\mathbf{m}}) \neq 0$.

$(r-1 \to r)$ We consider $r = 2, \ldots, k-1$, for $k \ge 3$. Then, from (74), $\mathbf{m} \in N^{k-r}$ implies that $\hat{\mathbf{m}} \in N^{k-r+1}$. Also, from (73),

$$\sum_{i=1}^{l} \hat{m}_{k-i-r+2} \le \hat{m}_0 + r + l - 2, \qquad l = 1, \ldots, k-r,$$

$$\sum_{i=1}^{k-r+1} \hat{m}_{k-i-r+2} = \sum_{i=1}^{k-r} \hat{m}_{k-i-r+2} \le \hat{m}_0 + k - 2 \le \hat{m}_0 + k - 1.$$

It follows from the inductive hypothesis that $U^{(r-1)}(\hat{\mathbf{m}}) \neq 0$. Hence, from (11), $U^{(r)}(\mathbf{m}) \neq 0$.

(k) $\mathbf{m} \in N^0$, for $k \ge 2$, implies that $\hat{\mathbf{m}} \in N^1$ and $0 = \hat{m}_1 \le \hat{m}_0 + k - 1 = m_0 + k$. Hence, $U^{(k-1)}(\hat{\mathbf{m}}) \neq 0$ and, from (11), $U^{(k)}(\mathbf{m}) \neq 0$. ..

REFERENCES

1. B. Gopinath and J. A. Morrison, "A Discrete Queueing Problem Arising in Packet Switching," Analyse et Contrôle de Systèmes (1976), Séminaires IRIA, Rocquencourt, pp. 201–210.
2. B. Gopinath and J. A. Morrison, "Discrete-Time Single Server Queues with Correlated Inputs," B.S.T.J., *56*, No. 9 (November 1977), pp. 1743–1768.
3. A. G. Fraser, B. Gopinath, and J. A. Morrison, "Buffering of Slow Terminals," B.S.T.J., *57*, No. 8 (October 1978), pp. 2865–2885.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

A Subjective Comparison of Selected Digital Codecs for Speech

By W. R. DAUMER and J. R. CAVANAUGH

(Manuscript received March 28, 1978)

Technological advances are continually increasing the economic viability of efficient codecs in telephone networks. A subjective evaluation is described here of the μ255 Pulse Code Modulation (PCM) algorithm and three more efficient techniques, Nearly Instantaneous Companding PCM (NIC PCM), Cummiskey-Jayant-Flanagan Adaptive Differential PCM (ADPCM), and Subscriber Loop Carrier Adaptive Delta Modulation (SLC ADM). These codecs are compared under the conditions of: (i) single encodings as a function of line bit rate, input level, received volume, and error rate, (ii) tandem encodings with intermediate baseband conversion, and (iii) local, exchange, and toll network reference connections where mixed tandem encodings might be found along with typical analog impairments such as loss and random noise. The simulation of the codec algorithms on a minicomputer facility enabled the production of subjective test tapes containing speech processed under these conditions. These tapes were then evaluated in listening-type subjective tests. It is shown that (i) NIC PCM, ADPCM, and SLC ADM have approximately a 12- to 16-kb/s advantage over μ255 PCM for equivalent subjective ratings, (ii) NIC PCM, ADPCM, and SLC ADM perform comparably over the range of conditions tested; and (iii) 64-kb/s μ255 PCM can be deployed in a multiple encoding environment with very few restrictions, whereas the use of lower bit rate NIC PCM, ADPCM, and SLC ADM codecs would necessitate more stringent application rules to avoid excessive degradation in tandem encoding situations.*

* Trademark of Western Electric.

I. INTRODUCTION

In recent years, a great amount of interest has been expressed in the literature on the subject of efficient encoding of voiceband signals. Much of this interest stems from the economics of bandwidth reduction that are possible with efficient codecs. Waveform codecs such as differential PCM (DPCM), adaptive differential PCM (ADPCM), delta modulation (DM), and adaptive delta modulation (ADM) are considered for general use in telephone networks because, among other reasons, they often are a reasonable compromise between the bandwidth required of the transmission channel and the terminal complexity at the ends of the channel.

Aside from the economics, there is the issue of degradations introduced by a codec in speech and voiceband signals. In fact, a signal may undergo a number of encodings by codecs of different types as it progresses through the network. An example of this is the hypothetical network of Fig. 1. Here, a mixture of analog and digital switching and transmission facilities are represented.

The toll portion of the network incorporates analog toll switches such as the No. 4 crossbar and the digital No. 4 ESS switch. The D channel banks and VIF terminals are shown in order to point out where analog-

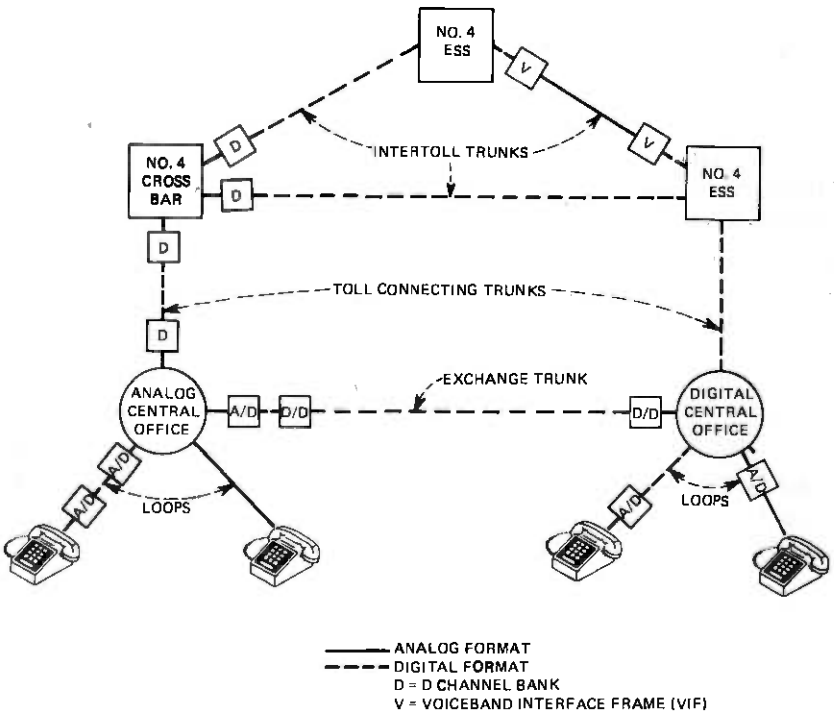


Fig. 1—Possible codec deployment in DDD network.

to-digital conversions take place. The coding law currently used in Bell System D channel banks is 64-kb/s μ255 PCM.[1]

The exchange network environment favors the deployment of efficient codecs because of the investment in the exchange and loop plants. Subscriber carrier and remote switching systems may be installed on loops and interfaced directly to a digital central office. Other carrier systems may be implemented by a digital-to-digital (D/D) conversion from one code format to another more efficient format as shown in the exchange trunk example of Fig. 1.

The point is that a telephone call can be routed over a variety of connections in this network and might be subjected to 64-kb/s μ255 PCM encodings in the toll plant in tandem with other codecs found in the local environment. The goal of the studies reported in this paper is to gain insight into the subjective effects of a multiple encoding environment on speech. Another important component of this goal is the impact of this environment on voiceband data. However, the resources and time consumed by the speech studies did not allow simultaneous work on voiceband data in the context of this study.

Prior to the evaluation of the overall subjective effect of a network environment on speech, the contributions of the various components of the network must be well understood. The analog components can be represented by transmission impairments such as loss and additive random noise.[2] The performance of a digital component of a connection is a function of the quantizing noise characteristics of the digital encoding, the line bit rate, the speech input level to the codec, and transmission impairments such as bit errors, slips, and misframes. The initial step of the study concentrated on the evaluation of some of these impairments for single encodings by selected codecs. Next, the codecs were tandemed with themselves under controlled conditions in order to establish the multiple encoding behavior without the complications of more than one coding law. Finally, selected codecs were incorporated into a set of reference network connections. These reference connections are representative of Bell System local and toll connections with characteristics (loss, noise, talker volume) that are derived from survey data.

The purpose of this paper is to report on recently completed subjective tests of digital codecs. Detailed results of the tests are presented and preliminary analyses are discussed. However, another objective of these tests is to provide a sufficiently large data base to enable the development of analytic models of subjective behavior. These models would be used to predict the performance resulting from the introduction of digital codecs in an evolving telephone network.

II. CODEC ALGORITHMS

Four basic codec algorithms were evaluated: (*i*) μ255 PCM,[3,4] (*ii*) Nearly Instantaneous Companding (NIC) PCM,[5] (*iii*) Cummiskey-Jayant-Flanagan ADPCM,[6] and (*iv*) Subscriber Loop Carrier (*SLC**) ADM.[7] These four were chosen for one or more of the following reasons: interest in the algorithm, availability of a well-defined solftware algorithm in the time frame of these tests, total number of test conditions to be generated, and the existence of a hardware implementation of the codec. These codecs are samples of four different classes of waveform codecs, but each chosen codec is quite specific and the reader is cautioned against generalizing the results for a particular codec to an entire class of codecs. Note that NIC PCM, ADPCM, and *SLC* ADM employ adaptive quantization (or companding) but none of the codecs incorporates adaptive prediction.

A brief description of some of the characteristics of each algorithm is given here which, in conjunction with the references, should aid the reader in understanding the similarities and differences among the algorithms when the results are discussed.

2.1 μ255 PCM

Three versions of the μ255 PCM algorithm were included in the tests: the continuous law compandor[3] with a mid-tread bias and the 15-segment version[4] of the μ255 compandor with both a mid-tread and a mid-riser bias. However, most of the attention is focused on the 15-segment mid-tread algorithm, since this algorithm at 64 kb/s is used in Bell System D channel banks.† Since the idle channel noise of a mid-tread algorithm implemented on a computer is essentially nonexistent, 16 dBrnC0 of random noise is introduced at the outputs of both mid-tread algorithms. This level of noise is intended to represent that which could be achieved in a hardware implementation. The overload point in all three cases is set at the peak amplitude of an inband sine wave with an rms power of +3 dBm0.

2.2 NIC

This algorithm is a block encoding scheme[5] which compresses L-bit, 15-segment μ255 PCM ($L \geq 4$) to $L - 2$ bits. In this application, there are eight samples to a block. The largest segment number in the block is found and transmitted to the far-end decoder. The eight L-bit μ255 samples are then digitally reencoded into $L - 2$ bit uniform samples with an overload point at the top of the largest segment in the block. This yields a bit rate of $[8 \times (L - 2) + 3]$ kb/s for an 8-kHz sampling rate, where the 3-kb/s component represents the transmission of the maxi-

* Trademark of Western Electric.

† Except for older D1 channel banks, where 56-kb/s μ100 PCM is used.

mum segment number in the block every millisecond. The NIC algorithm is biased mid-tread for $L \geq 7$ bits and mid-riser for $L \leq 6$ bits. As in the PCM algorithms described above, 16 dBrnC0 of noise is added to the decoder output during mid-tread operation, and the overload point is set at +3 dBm0.

2.3 ADPCM

The Cummiskey-Jayant-Flanagan ADPCM algorithm[6] was chosen using a first-order predictor in the feedback loop with a time constant of 0.43 ms. For line bit rates greater than or equal to 32 kb/s ($L \geq 4$ bits/sample at 8-kHz sampling), the minimum quantizer step size is equal to twice the interval on the first chord of the 8-bit μ255 PCM quantizer. For $L < 4$, the minimum step-size is increased by the factor $(5 - L)$ over the step-size for $L \geq 4$. For all values of L, the ratio of the maximum step-size to the minimum step-size is 128, identical to the 8-bit μ255 PCM algorithm. No amplitude overload point was set for this algorithm, since it is a differential codec, but for $L = 4$ the algorithm is driven into slope overload with a 425-Hz sine wave at +6 dBm0. A modification of this algorithm was also investigated to determine the subjective effect of line bit errors. This modification was the introduction of step-size leak into the step-size adaptation logic. Normally,

$$\Delta_i = a_i \Delta_{i-1},$$

where Δ_i is the quantizer step-size at a particular sample interval i and a_i is the corresponding adaptation coefficient. An error in the transmission of the bit stream to the decoder will cause an error in the decoder step-size adaptation. The effect of bit errors can be minimized somewhat by the addition of step-size leak, so that

$$\Delta_i = a_i(\Delta_{i-1})\gamma,$$

where γ is less than but nearly equal to unity. In our application, γ is chosen to be 31/32. This limits the effect of an error at the decoder so that the encoder and decoder step-sizes eventually track each other. However, the adaptation process remains affected by the step-size leak in the absence of errors, and discernible distortion may be introduced into the speech.

2.4 SLC ADM

The ADM algorithm chosen was the *SLC*-40 algorithm.[7] It is currently deployed at a sampling rate of 37.7 kHz in a 40-channel loop carrier system. Of the three non-PCM coding schemes described in this section, it is the only one in commercial use in the Bell System, with approximately 1800 systems installed in the field to date. This algorithm uses an adaptive step-size where the step-size is altered whenever four suc-

cessive like sign bits are transmitted on the line. The predictor in the feedback loop has a main time constant of 0.7 ms. The minimum step-size is set to achieve an idle channel noise of 15.5 dBrnC0 at 37.7 kb/s. This codec is driven into slope overload with a +6 dBm0 input sine wave at a frequency of 800 Hz.

III. SIMULATION SYSTEM

3.1 Overview of system

A minicomputer facility which was developed at Bell Laboratories in Holmdel, N.J., is briefly described in this section. A detailed description of its capabilities and operation can be found in Ref. 8.

This system has four important characteristics: (*i*) the system configuration is independent of any particular codec algorithm; (*ii*) the system is conducive to experimentation and development of a desired codec algorithm; (*iii*) the system is capable of simulating a network connection containing mixed tandem encodings of several codecs; and (*iv*) the system is capable of automatic production of audio tapes suitable for subjective evaluation. All the codec test conditions discussed in this paper were simulated on this system.

Two PDP*-11 minicomputers are interfaced with uniform 15-bit A/D and D/A converters. The system is capable of real-time operation at sampling rates up to 72 kHz. A particular codec algorithm can be implemented on this system as long as the sampling rate of the codec is less than 72 kHz; the codec distortion dominates the 15-bit A/D and D/A converter distortion; and differences between hardware and software versions of the codec are recognized and accounted for.

3.2 Hardware configuration

Referring to Fig. 2, PDP 11/40 and 11/20 minicomputers are used, each having 28K words (16 bits/word) of memory. Each minicomputer is interfaced to a dedicated Tustin series 1500 A/D and D/A converter system. The A/D and D/A converter systems were measured and adjusted to ensure that they were capable of at least theoretical 14-bit accuracy so that the waveform codecs of Section II could be simulated.

A 1.25-million-word disk system was installed on each machine with a nine-track tape drive interfaced on the 11/40 for mass storage. The tape format is standard, so that large amounts of data can be transferred to other more powerful computer facilities for processing. An Ampex AG-440G analog tape unit is also controlled by the 11/40 so that high-quality audio tapes can be automatically prepared for subjective testing.

In addition to the dedicated equipment on each processor, there is a UNIBUS* window between the two machines to allow interprocessor

* Registered trademark of Digital Equipment Corporation.

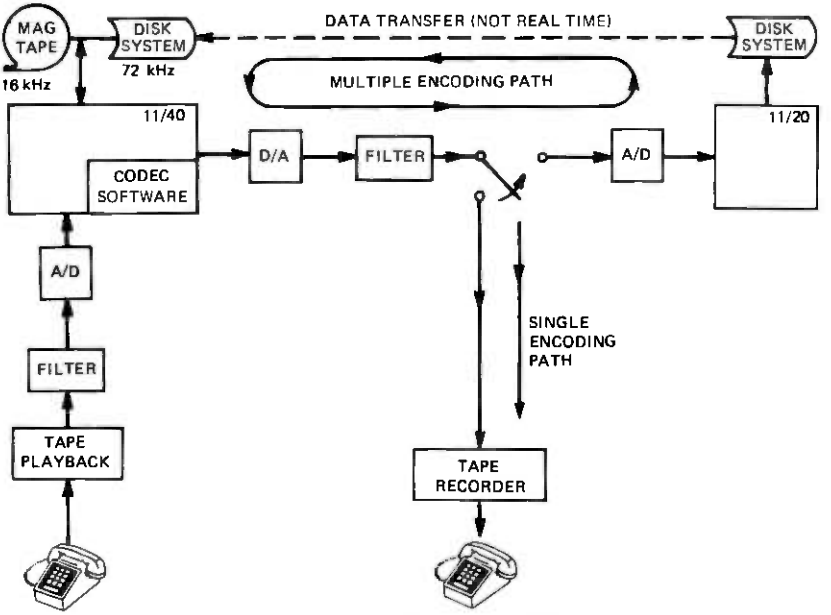


Fig. 2—Codec simulation facility.

communication and data transfer. This equipment is essential for the automatic generation of tandem encodings.

3.3 Facility operation

Referring to the lower left-hand portion of Fig. 2, a segment of an analog signal such as speech is derived from a tape recorder, telephone set, or other source. The signal is bandlimited, quantized, and stored on either tape or disk. The choice of tape or disk is determined from two considerations, the sampling rate of the A/D converter and the length of the analog segment to be digitized. The disk can store samples at rates up to 72 kHz, the magnetic tape up to 16 kHz. However, a 2400-foot tape reel can store 10 million words, the disk only 1.25 million words.

After the digits are stored on either tape or disk, they can be processed by a codec algorithm and stored back on magnetic tape or disk in preparation for playback. The output signal of the D/A converter is passed through a reconstruction filter and the codec simulation is essentially completed. The resultant analog signal can be stored on an audio tape recorder as indicated by the "single encoding" path in Fig. 2.

If a tandem encoding is to be simulated, the "switch" following the reconstruction filter is positioned to the input of the A/D converter of the 11/20. As the playback operation is proceeding on the 11/40, a simultaneous acquisition process is proceeding at the 11/20. The D/A on

the 11/40 and the A/D on the 11/20 can be timed by independent clocks. Hence, the overall D/A-A/D process can be performed asynchronously. After the acquisition at the 11/20 is complete, the 11/40 retrieves the samples from the disk system of the 11/20 via the UNIBUS window. The samples then reside on the magnetic tape unit or disk system of the 11/40. A second codec processing can take place in an identical manner as described above. A tandem encoding loop is thus defined. This loop is illustrated in the upper portion of Fig. 2. A signal can be made to traverse this loop as many times as desired, with no restrictions on sampling rate or codec type for each successive encoding. When the desired number of tandem encodings are completed, the signal can be stored on audio tape as in the single encoding case. Both the single and tandem encoding operations are performed under the automatic control of the software. Manual intervention after each loop is not necessary.

IV. SUBJECTIVE TESTING—TECHNIQUES

The subjective tests described in this paper were all conducted as listening-only tests (not conversational). The subjects listened to prerecorded speech and voted on the perceived quality. Details concerning the test facilities, selection of subjects, test circuitry, and test administration are covered in this section.

4.1 Description of tests and facilities

The subjective tests were conducted in an acoustically treated test room containing 11 cubicles permitting up to 11 subjects to be tested simultaneously. Each cubicle contains a handset over which test conditions are heard and a keyboard with five keys labeled "excellent," "good," "fair," "poor," and "unsatisfactory," which is used for registering the vote for each test condition. Associated with the keyboard are red indicator lights which are lit during the presentation of a test condition and green indicator lights which are lit to indicate the period for voting on the test condition.

The votes of each subject are recorded using a minicomputer system, a keyboard interface, and associated programs. A terminal permits monitoring of the ratings during the tests. All votes are recorded on magnetic tape for subsequent analysis.

At the start of the test session, a start signal is sent by the test administrator to the computer which then remotely actuates the tape recorder and turns on the red keyboard lights. At the end of each test condition, a tone recorded on the second track of the test tape is recognized by a tone detection circuit which signals the computer to stop the recorder and turn on the green voting lights. The computer then collects the votes. After all the votes are received from the subjects or after a 3-second timeout period (whichever occurs first), the computer extin-

guishes the green voting lights, turns on the red keyboard lights, and then starts the tape recorder for the next test condition.

4.2 Test playback system

The system used for the subjective tests is shown in Fig. 3. A dual-track tape recorder (Ampex 440G) equipped with a Dolby noise reduction unit is used to drive a standard 500-type telephone set (with a receiving rating efficiency of 21 dB) connected to 6 kft of 26-gauge, nonloaded cable and a 400-ohm, 48-Vdc feeding bridge (central office battery supply circuit). The carbon transmitter is replaced with a 90-ohm resistor to eliminate any room noise pickup. The master telephone set receiver is replaced by a 120-ohm resistor to avoid possible introduction of acoustic/inductive interference in the test conditions; a transformer-amplifier bridge on this resistor drives the 11 telephone set receivers in the cubicles as shown in Fig. 3.

The measured responses of the playback system are shown in Figs. 4 and 5. These responses reflect adjustment of the listening amplifier such that a speech level of −29 VU (volume units) at the line terminals of the telephone set (point V_2 of Fig. 3) produces an acoustic pressure of −12 dBPa* (82 dB relative to 20 μPa) for speech power averaged across the 11 receivers. The value of −12 dBPa approximates the preferred speech pressure level.

4.3 Subject selection and test administration

Subjects were selected from employees in various job classifications and age groups at Bell Laboratories in Holmdel, N.J. The sheer number of test conditions dictated that the total test program be divided into five tests. All the tests were administered independently of one another using different subjects. The number of subjects and breakdown according to sex is given in Table I. Discussions of the testing throughout the remainder of this paper will be structured under the three topics of (*i*) single encodings, (*ii*) tandem encodings, and (*iii*) the local, exchange, and toll reference connections.

For each test session, a maximum of 11 test subjects were seated in the test cubicles and supplied with test instructions shown in Fig. 6. The test administrator read the instructions and told the subjects that there would be four practice test conditions given before the start of the test. The subjects were told to vote on the four practice conditions as if they were part of the actual test. After the practice conditions were presented and voted on, the test administrator would enter the test room to answer any questions before the actual testing began.

* dBPa = dB relative to 1 Pascal which corresponds to 1 newton per square meter.

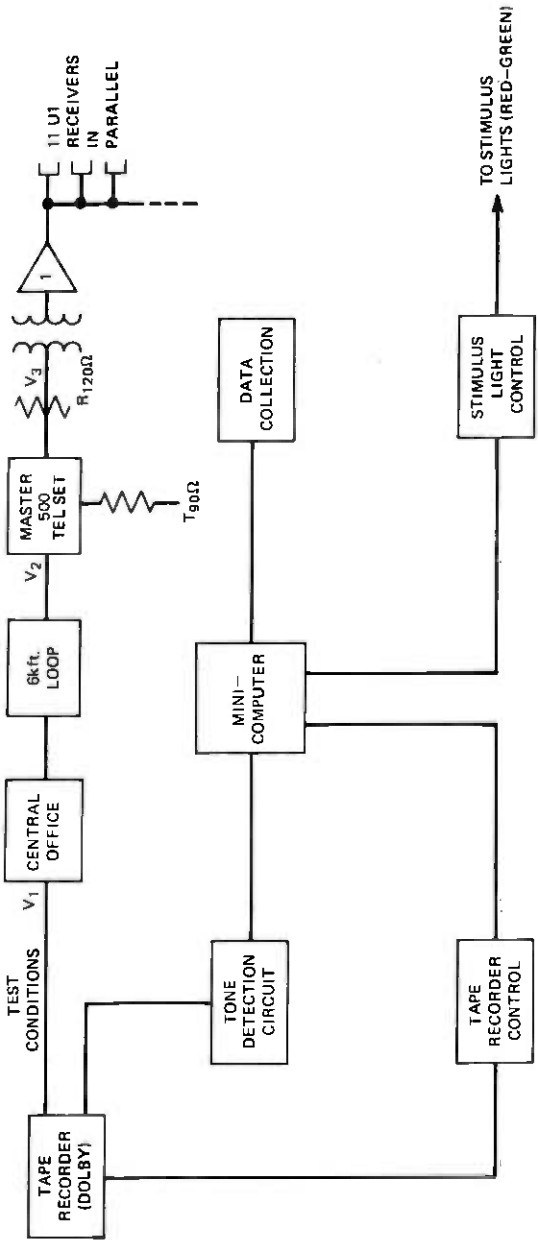


Fig. 3—Multiple listening playback system.

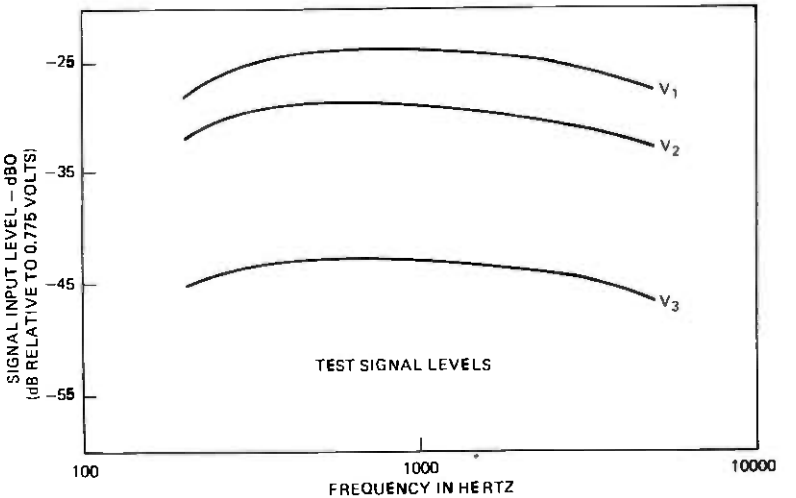


Fig. 4—Test signal levels.

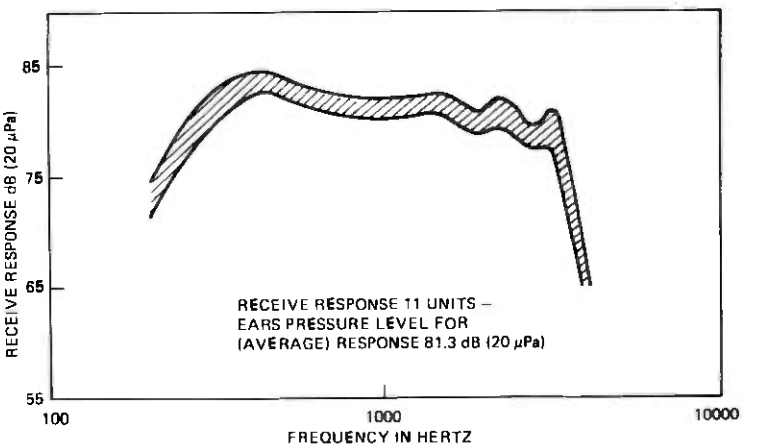


Fig. 5—Response measurements for playback system.

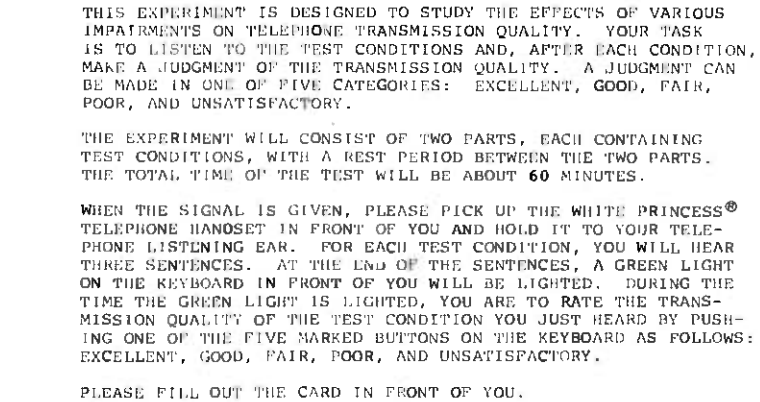
4.4 Speech sources and anchoring conditions

The source tapes were made by recording male and female speech from the output of a 500-type telephone set carbon transmitter through a 111C repeater coil and a simulated 6-kft loop using a Dolby noise reduction unit. These sources were then processed by the various codec algorithms to produce the test conditions.

Analog noise conditions were included in each test session along with codec conditions. The analog noise conditions were generated by adding white noise bandlimited from 200 Hz to 3.4 kHz to the speech so that, with the speech level set at −29 VU at the 500-type telephone set terminals, the noise is at a level of 10, 20, 30, or 40 dBrnC. The inclusion of

Table I—Number of subjects

Test	Total	Male	Female
(*i*) Single Encodings—Variation of Line Bit Rates and Levels	51	37	14
(*ii*) Single Encodings—Error Rate	52	35	17
(*iii*) Tandem Encodings plus Local and Exchange Reference Connections	53	49	4
(*iv*) Toll Reference Connections—Part 1	56	46	10
(*v*) Toll Reference Connections—Part 2	54	50	4


THIS EXPERIMENT IS DESIGNED TO STUDY THE EFFECTS OF VARIOUS IMPAIRMENTS ON TELEPHONE TRANSMISSION QUALITY. YOUR TASK IS TO LISTEN TO THE TEST CONDITIONS AND, AFTER EACH CONDITION, MAKE A JUDGMENT OF THE TRANSMISSION QUALITY. A JUDGMENT CAN BE MADE IN ONE OF FIVE CATEGORIES: EXCELLENT, GOOD, FAIR, POOR, AND UNSATISFACTORY.

THE EXPERIMENT WILL CONSIST OF TWO PARTS, EACH CONTAINING TEST CONDITIONS, WITH A REST PERIOD BETWEEN THE TWO PARTS. THE TOTAL TIME OF THE TEST WILL BE ABOUT 60 MINUTES.

WHEN THE SIGNAL IS GIVEN, PLEASE PICK UP THE WHITE PRINCESS® TELEPHONE HANDSET IN FRONT OF YOU AND HOLD IT TO YOUR TELEPHONE LISTENING EAR. FOR EACH TEST CONDITION, YOU WILL HEAR THREE SENTENCES. AT THE END OF THE SENTENCES, A GREEN LIGHT ON THE KEYBOARD IN FRONT OF YOU WILL BE LIGHTED. DURING THE TIME THE GREEN LIGHT IS LIGHTED, YOU ARE TO RATE THE TRANSMISSION QUALITY OF THE TEST CONDITION YOU JUST HEARD BY PUSHING ONE OF THE FIVE MARKED BUTTONS ON THE KEYBOARD AS FOLLOWS: EXCELLENT, GOOD, FAIR, POOR, AND UNSATISFACTORY.

PLEASE FILL OUT THE CARD IN FRONT OF YOU.

ARE THERE ANY QUESTIONS?

Fig. 6—Test instructions.

these noise conditions allowed the codec results to be referenced or anchored into the body of information that has been accumulated over the years on the effects of random noise. It is estimated that the background noise contribution from the original source tape was approximately 4 dBrnC measured at the line terminals of the telephone set.

Speech-correlated noise conditions were included because they approximate μ255 PCM quantizing noise and can be utilized in future modeling efforts. These conditions were produced by a device called the Modulated Noise Reference Unit (MNRU).[9] This introduces a noise signal to the input speech, which is directly correlated to the instantaneous amplitude of the speech. The speech-correlated noise conditions were designated $Q = 5$, $Q = 10$, etc., where Q is equal to the decibel value of the ratio of the speech power to speech correlated noise power.

In addition to the reasons given above, the analog noise and Q conditions served as control conditions by being included in every test session as a check on session-to-session differences.

4.5 Analysis of results

The votes from each test condition are combined into a vote histogram with the votes for male and female speech pooled together. These histograms contain the number of votes recorded for each of the comment categories "excellent," "good," "fair," "poor," and "unsatisfactory" as represented by the category numbers 5, 4, 3, 2, and 1, respectively. A Mean Opinion Score (MOS) is calculated for each test condition by taking the arithmetic mean of the category numbers voted. A sample standard deviation (σ) is also calculated for each vote histogram.

The male and female results are combined because a simple regression analysis of the data indicated that there was little difference between the subjective responses for male and female speakers (about 0.2 to 0.4 of a category point). The overall standard deviation of the data is on the order of 0.6 to 0.7 of a category point.

V. SUBJECTIVE TESTING—SINGLE ENCODINGS

The first round of subjective testing involves single encodings of μ255 PCM, NIC, ADPCM, and *SLC* ADM as a function of three parameters; line bit rate, input level and received volume (listening level) pairs, and line error rate.

5.1 Test design

Table IIA lists the test conditions as a function of line bit rate. Three

Table IIA—Single encoding versus line bit rate conditions

Algorithm	Bit Rate (kb/s)
Continuous Law PCM (mid-tread bias)	16,40,64
15-Segment PCM (mid-tread bias)	32,40,48,64
15-Segment PCM (mid-riser bias)	32,48
NIC PCM	19,35,43,51,59
ADPCM (with step-size leak)	16,24,32,40,48
ADPCM (without step-size leak)	16,24,32,48
SLC™ ADM	24,37.7, 48

Table IIB—Single encoding versus level variation conditions

Algorithm	Bit Rate (kb/s)
15-Segment PCM (mid-tread bias)	48
NIC PCM	51
ADPCM (with step-size leak)	48
SLC™ ADM	48
Input Level (VU)	**Received Volumes (VU)**
−0.2	−20.7,−29.0
−10.5	−20.7,−29.0,−38.3
−20.7	−29.0,−38.3
−31.0	−38.3
−41.2	−47.0

Table IIC—Single encoding versus error rate conditions

Algorithm	Bit Rate (kb/s)
15-Segment PCM (mid-tread bias)	48
NIC PCM	51
ADPCM (with step-size leak)	48
SLC™ ADM	48

versions of the μ255 PCM algorithm are included mainly for verifying that the subjective differences between continuous law, 15-segment mid-tread, and the 15-segment mid-riser algorithms are negligible. ADPCM is simulated with and without step-size leak. To economize on the total number of test conditions, various line bit rates are excluded with care taken so that subjects are exposed to a wide subjective quality range and the subjective ratings of these exclusions can be estimated by interpolation. The input speech level for these conditions is -24.8 VU (~ -23.4 dBm), the average reported by McAdoo[10] for local calls over the Bell System message network. The measured idle channel noise of each codec is given in Table III. The received volume presented to the subject is -29 VU, measured at the line terminals of the telephone set.

The conditions for single encoding as a function of input level and received volume are arrived at by estimating Bell System speech volumes and network losses for five types of network connections: (*i*) intrabuilding, (*ii*) interbuilding over a direct trunk, (*iii*) interbuilding over two tandem trunks, (*iv*) and (*v*) long and short connections over the intertoll network. For each of these five situations, the encoder and decoder of a codec are postulated to be located in the telephone set, end office, and at an intermediate point in the loop as in the example of a remote switching or pair gain system. Speech volume[10] and loss[11] distributions are used to derive the input level and received volume ranges to be tested. A condition is then defined by quantizing these ranges and assigning input level-received volume pairs to a codec where the received volume is always less than the input level. The four codecs are implemented at a single bit rate which is equal to 48 kb/s or as close to 48 kb/s as possible (51 kb/s in the case of NIC). This particular line bit rate is chosen for two reasons: (*i*) the large number of possible test conditions to be evaluated dictated the use of a single line bit rate, and (*ii*) that line bit rate should be 48 kb/s since experience has shown that the distortion introduced by a single encoding of μ255 PCM at 48 kb/s can be barely perceived and it is desirable to ascertain and compare the performance of the other codecs at a comparable line bit rate. Table IIB lists the codecs and the input level-received volume matrix used in this section of testing.

The third and final portion of the single encoding tests is concerned with the subjective effect of random transmission errors. Independent errors are introduced between the encoder and decoder at rates 10^{-1},

10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5} errors/bit. The four codecs are chosen to operate at a line bit rate equal to or nearly 48 kb/s for the same reasons given in the previous paragraph. The input level is set at −12.4 VU so that the entire dynamic range of the PCM and NIC codecs is exercised without overloading. A summary of the error rate test conditions appears in Table IIC.

5.2 Results and observations

The combined results for single encodings as a function of line bit rate, input level and received volume, and line error rate are given in the appendix, Tables V, VI, and VII, respectively. Most of these tabulated results are summarized in Figs. 7 through 10.

Figure 8 is a plot of the mean opinion score (MOS) versus line bit rate for the codecs of Table IIA. It is recalled that these results are obtained using an input speech level of −24.8 VU with a received volume of −29 VU, representing a 4.2-dB loss located after the decoder. An immediate observation that can be drawn is the subjective advantage of all the adaptive codecs (NIC, *SLC* ADM, and ADPCM) over the μ255 PCM. Below a line bit rate of 48 kb/s, this advantage is on the order of 12 to 16 kb/s for equivalent subjective ratings. At 48 kb/s and above, the leveling off of the response curves is due to two effects. The first important factor is the absolute level of idle channel noise the subjects hear. The idle

Table III—Single encoding idle channel noise measurements

Algorithm	Bit Rate (kb/s)	Idle Channel Noise (dBrnC)
True Logarithmic PCM (mid-tread)	16	16.0
True Logarithmic PCM (mid-tread)	40	16.0
True Logarithmic PCM (mid-tread)	64	16.0
15-Segment PCM (mid-riser)	32	37.7
15-Segment PCM (mid-riser)	48	26.3
15-Segment PCM (mid-tread)	32	16.0
15-Segment PCM (mid-tread)	40	16.0
15-Segment PCM (mid-tread)	48	16.0
15-Segment PCM (mid-tread)	64	16.0
NIC PCM	19	31.4
NIC PCM	35	19.6
NIC PCM	43	16.0
NIC PCM	51	16.0
NIC PCM	59	16.0
ADPCM (with step-size leak)	16	27.1
ADPCM (with step-size leak)	24	23.1
ADPCM (with step-size leak)	32	18.6
ADPCM (with step-size leak)	40	18.4
ADPCM (with step-size leak)	48	18.4
ADPCM (without step-size leak)	16	25.2
ADPCM (without step-size leak)	24	22.3
ADPCM (without step-size leak)	32	18.7
ADPCM (without step-size leak)	48	18.4
SLC ADM	24	23.3
SLC ADM	37.7	15.5
SLC ADM	48	13.3

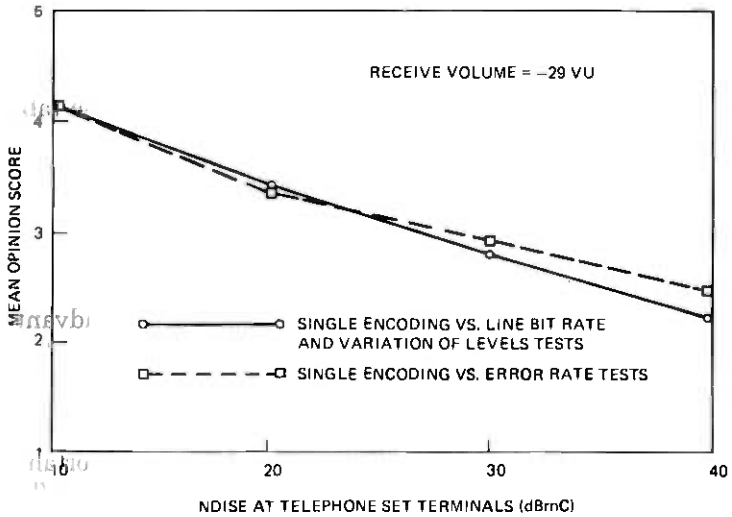


Fig. 7—Subjective results—additive random noise.

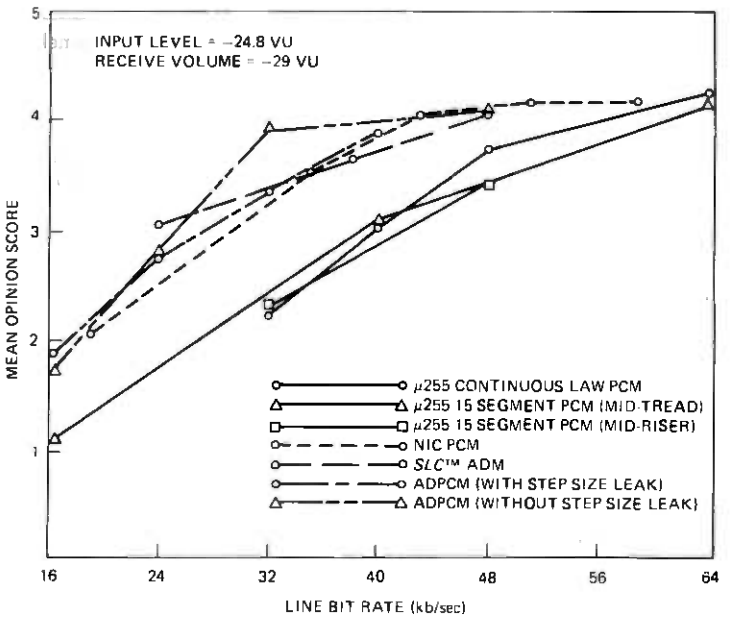


Fig. 8—Subjective results—single encodings vs line bit rate.

channel noises of the higher bit rate codec conditions listed in Table III are in the range of 13.3 to 26.3 dBrnC. The 4.2-dB loss following the decoders translates this range into 9.1 to 22.1 dBrnC at the line terminals of the telephone set. As these tests were conducted in an acoustically shielded room, the noise results of Fig. 7 show that subjects can react to

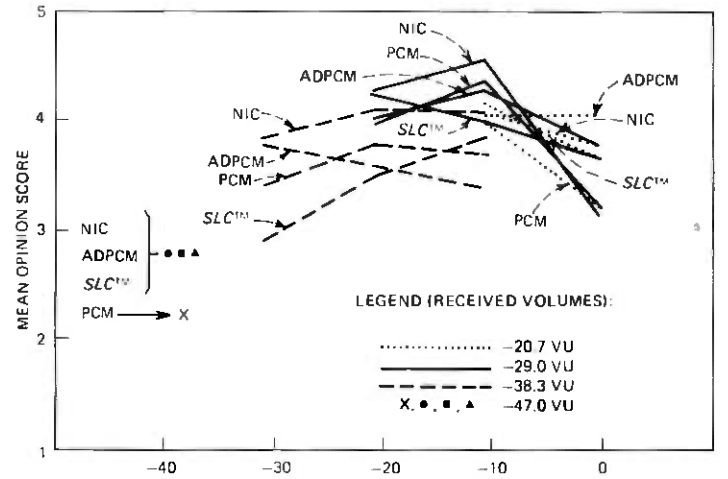


Fig. 9—Subjective results—variation of input level and receive volume.

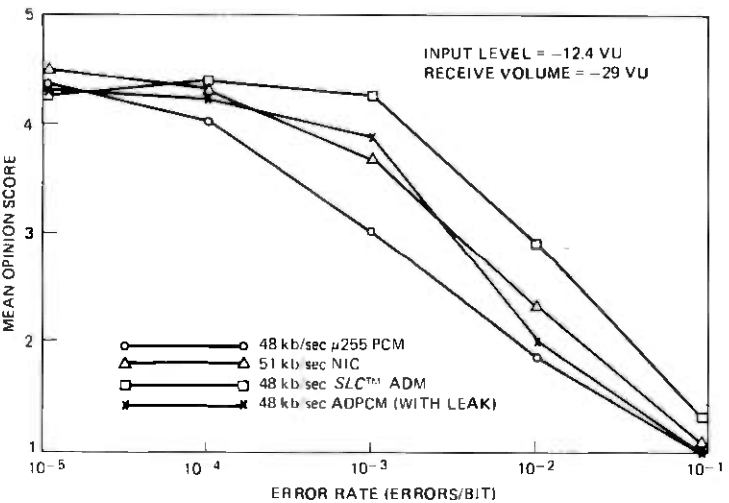


Fig. 10—Subjective results—line error rates.

noise levels on the order of 10 dBrnC, since the curves do not level off in this vicinity. Thus, the leveling-off tendency in Fig. 8 is in line with the noise response curves of Fig. 7 for the lower values of noise. A second factor may be the distortion and bandlimiting introduced by the transmitting and receiving 500-type telephone set pair.

Two other important observations can be extracted from Fig. 8. The first is that there appears to be no appreciable differences among the three versions of μ255 PCM tested, at least for bit rates greater than 32 kb/s. The other observation is that the introduction of step-size leak in the ADPCM algorithm (see Section 2.3) has a small effect on subjective

quality in the vicinity of 32 kb/s under these conditions. The intent here is not to determine the optimal amount of leak to be introduced but to demonstrate that leak can affect subjective quality.

The results for the subjective ratings as a function of input level and received volume are shown in Fig. 9. The results are grouped according to a fixed received volume with a varying input level. The solid lines represent the preferred received volume of −29 VU with the dotted and dashed lines representing received volumes of −20.7 VU and −38.3 VU, respectively. The cluster of four points on the left-hand side of Fig. 9 represents the −47.0 VU received volume where only one corresponding input level of −41.2 VU is used. Each curve is a response for a particular codec and is labeled accordingly.

The MOS results in Fig. 9 are due to a combination of effects: (*i*) quantizing distortion, (*ii*) the received volume level that the subject hears at the telephone set, (*iii*) the absolute level of idle channel noise at the set terminals which is a function of both the input level and receive volume, and (*iv*) overload distortion which is manifested as amplitude-limiting for PCM and NIC and slope overload for the ADPCM and *SLC* ADM codecs. All four of these effects must be considered collectively when interpreting these plots. With these caveats in mind, a few observations can be made here with detailed analyses left to future studies.

The NIC algorithm is rated significantly better than the PCM algorithm for nearly all input levels and received volumes. This is an expected result, since the NIC codec readjusts its dynamic range for each block of eight samples. The amplitude overload point of both the PCM and NIC codecs for the speech sources in this study occurs at an input level of approximately −12.4 VU. Thus, the subjective ratings of PCM and NIC are the greatest for the input level of −10.5 VU, where only a small number of samples are clipped and the entire dynamic range is fully exercised. At the input level of −0.2 VU, the peaks of the speech are roughly 12 dB above overload and the MOS ratings fall off for PCM and NIC at both received volumes of −20.7 and −29 VU.

The *SLC* ADM and ADPCM codecs hold up somewhat better for high input levels of speech, confirming that slope overload is "more" tolerable than amplitude overload. The ADPCM and *SLC* ADM codecs are rated comparably for the input levels of −10.5 and −20.7 VU.

Two additional general observations can be inferred from Fig. 9. First, the set of curves for the received volume of −29 VU tends to have higher MOS ratings than the sets for received volumes of −20.7 and −38.3 VU. This can be attributed to three causes: (*i*) −29 VU is the preferred received volume, (*ii*) the dynamic ranges of the PCM and NIC codecs are exercised fully with essentially no overloading, and (*iii*) the idle channel noises heard by the subjects are low because of the high input levels (0 to 4 dBrnC for the input level of −10.5 VU). These effects lead to higher MOS values than those shown for the same codecs in Fig. 8. The

second observation from Fig. 9 is that for the lowest received volume tested, −47.0 VU, the adaptive codecs are scored equivalently with PCM receiving an even lower rating due to its coarse quantization of the speech at the input level of −41.2 VU.

The final portion of the single encoding testing is concerned with the effect of line error rates on the four codecs in Table IIC. The results are tabulated in Table VII and shown in Fig. 10, where MOS is plotted versus the errors per bit on a logarithmic scale. Note that the response curves converge at low (10^{-5}) and high (10^{-1}) error rates. The 10^{-1} rate is sufficiently severe so that the four codecs are all rated "unsatisfactory" and the 10^{-5} error rate is virtually undetectable, and the four codecs are rated "good." The leveling-off of the four curves for low error rates at MOS values of 4.2 to 4.5 is slightly higher than the asymptotic value in Fig. 8. This is due to the fact that the input level of −12.4 VU and the corresponding receive value of −29 VU result in lower idle circuit noises as measured at the line terminals of the telephone set.

Significant differences among the codecs are only manifested in the area between 10^{-4} and 10^{-2} errors per bit. The *SLC* codec is the least sensitive to line errors while PCM is the most sensitive. NIC and ADPCM (with step-size leak) are rated nearly the same and fall between the *SLC* ADM and PCM extremes. Although the ADPCM codec without step-size leak was not formally tested in the presence of errors, informal listening tests have shown that it is subjectively comparable to the PCM codec.

VI. SUBJECTIVE TESTING—TANDEM ENCODINGS

6.1 Test design

This section contains a description of the tandem encoding conditions listed in Table IV. Basically, the four codec algorithms of the previous section are used here at a few selected bit rates, again because of the constraint of economizing on the total number of test conditions. Evidence from the single encoding versus line bit rate tests show that the adaptive codecs, NIC, ADPCM, and *SLC* ADM have a 12- to 16-kb/s advantage over μ255 PCM for bit rates below 48 kb/s. Thus, it was decided to compare 48-kb/s μ255 PCM against NIC, ADPCM, and *SLC* ADM in the vicinity of 32 kb/s. The 8-kHz sampling rate dictated that ADPCM and NIC operate at 32 and 35 kb/s, respectively. Since the *SLC*-40 ADM algorithm is implemented in the loop plant today, the sampling rate of that system, 37.7 kb/s, is used. Finally, there is interest in the tandem performance of D channel banks, hence 64-kb/s μ255 PCM is included as a reference. ADPCM is implemented without step-size leak since subjects did perceive some degradation for ADPCM with leak at 32 kb/s.

The input speech level to the codecs is −20 VU, approximately the speech volume averaged over Bell System local and toll connections.[10] The received volume is set at the preferred level of −29 VU. A tabulation of the tandem encoding conditions is given in Table IV.

Table IV—Tandem encoding conditions

Codec	Bit Rate (kb/s)	Number of Tandem Encodings
PCM	64	1,2,4,6,8
PCM	48	1,2,4,8
NIC PCM	35	1,2,4,8
ADPCM	32	1,2,4,8
SLC™ADM	37.7	1,2,4,8

To explain the tandem encoding process in more detail, two configurations are shown in Fig. 11 which are simply expansions of Fig. 2. The upper arrangement is applicable to the 8-kHz codecs, PCM, NIC, and ADPCM. The filtering used for bandlimiting and reconstruction is realized with transmit and receive filters with characteristics that are similar to those used in D3 channel banks.[12] Note that each time the tandem encoding loop is transversed, the filters encountered are a receive/transmit pair, analogous to a back-to-back D channel bank situation.

Tandem encodings of the 37.7-kb/s *SLC*-40 ADM are generated in a slightly different manner, as shown in the lower portion of Fig. 11. Specifically, the differences are manifested in the filtering. The *SLC* input and output filters, which are implemented in software, in conjunction with a 6.4-kHz low-pass filter provide sufficient rejection at half the 37.7-kHz sampling rate to avoid aliasing. After the desired number of tandem encodings have been simulated, a D channel bank transmit filter is inserted in the playback path prior to recording so that small bandwidth differences between the 8-kHz and 37.7-kHz *SLC* ADM conditions are eliminated.

The use of these two filtering strategies results in different in-band characteristics as illustrated in Figs. 12 and 13, where the overall responses for a single encoding and eight tandem encodings are given for the 8-kHz and the 37.7-kHz *SLC* ADM codecs, respectively.

6.2 Results and observations

The tandem encoding MOS results and the breakdown according to comment category are tabulated in Table VIII of the appendix.

These results are shown in Fig. 14, where the MOS ratings are plotted as a function of the number of tandem encodings. An immediate observation is the superiority of the 64-kb/s μ255 PCM over the other four codecs. This superiority is evidenced by the facts that the response curve for 64-kb/s PCM begins to fall off at four encodings while the responses for the other four codecs fall off at two encodings, and eight encodings of 64-kb/s PCM is rated slightly below a MOS of 4 ("good"), while eight encodings of the other codecs are rated between 3 ("fair") and 2 ("poor"). Even for eight encodings of 64-kb/s μ255 PCM, the degradation is not

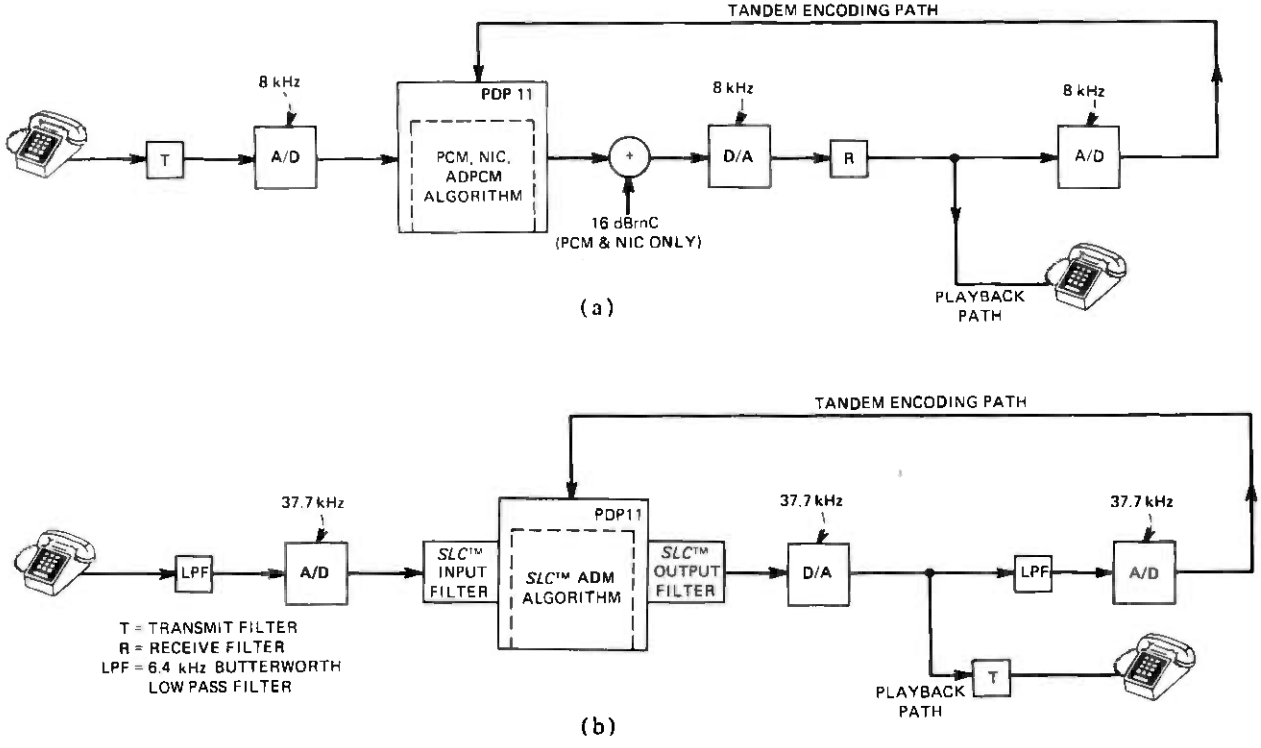


Fig. 11—Similar tandem encoding configuration. (a) PCM, NIC, and ADPCM configuration. (b) *SLC* ADM configuration.

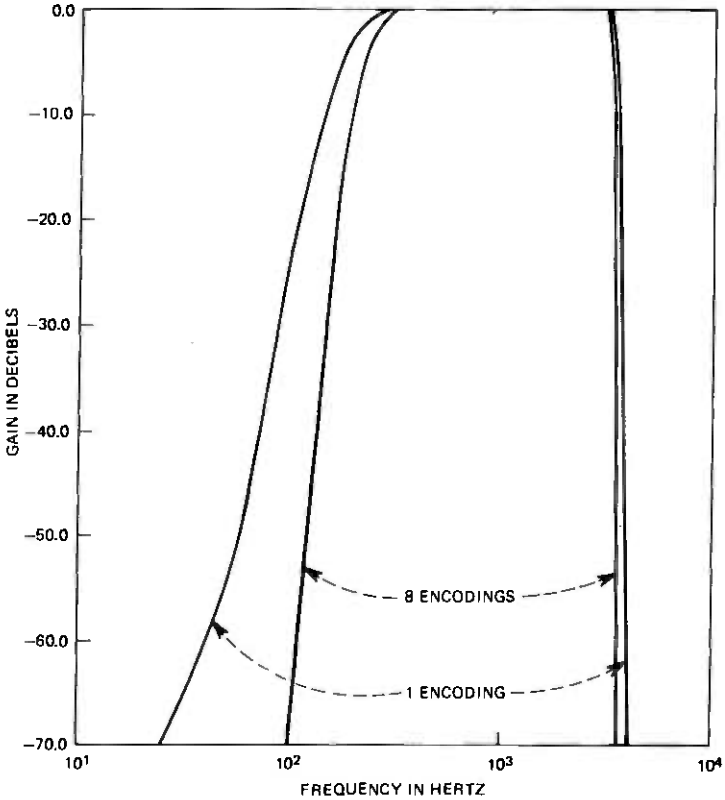


Fig. 12—Overall frequency response—8 kHz—one and eight encodings.

attributable only to quantizing noise. Accumulation of the idle channel noise and bandwidth reduction of the tandem filtering are other factors.

Another general observation is that, to a rough approximation, the three adaptive codecs (35-kb/s NIC, 32-kb/s ADPCM, and 37.7-kb/s *SLC* ADM) behave in a similar fashion to 48-kb/s PCM. This result concurs with the single encoding results of Fig. 8 where it was shown that the adaptive codecs exhibited a 12- to 16-kb/s subjective advantage over PCM.

Finally, all the curves in Fig. 14 at one encoding agree well with the single encoding results of Fig. 8 after it is recognized that the input level here is 4.8 dB greater than that of the single encoding tests. Both the single encoding versus line bit rate and tandem encoding tests were conducted at the received volume of −29 VU. Thus, the higher input level in the tandem encoding tests results in a 4.8-dB reduction in the idle channel noises heard by the subjects.

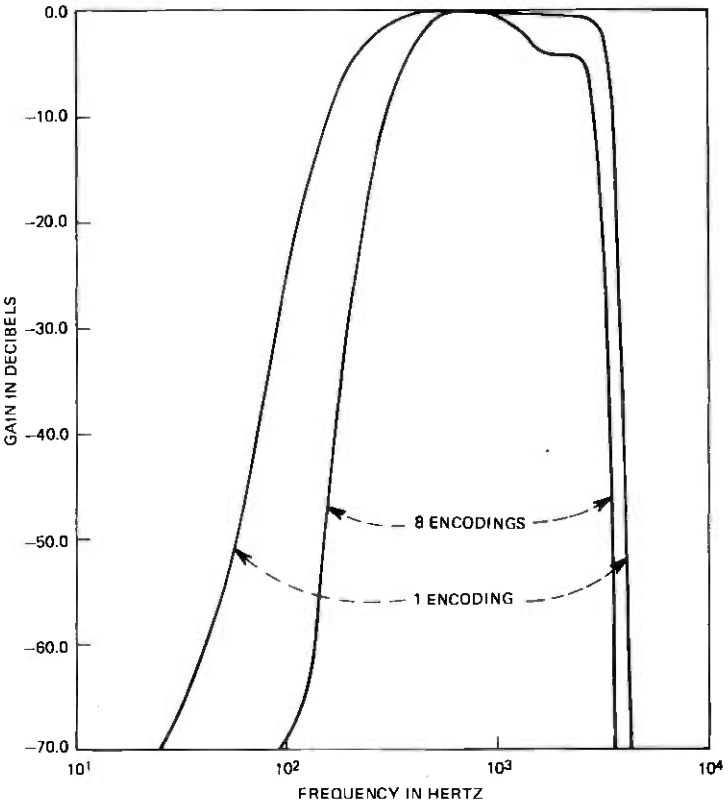


Fig. 13—Overall frequency response—37.7 kHz—one and eight encodings.

VII. SUBJECTIVE TESTING—REFERENCE CONNECTIONS

The purpose of the reference connection conditions is to evaluate the codecs of the previous section in representative network connections. Basically, these reference connections are of three types: local, exchange, and toll. The codecs are placed in connections where the environment in terms of speech volume, analog loss, and analog noise is defined from survey data. The received volumes are not held constant, but vary depending on the loss in the connection. For all connections, the codec characteristics such as overload and idle channel noise are defined in Section II. The connections are chosen in such a manner that average and worst-case situations are represented.

Many reference connections involve similar and dissimilar tandem encodings. Prior to a detailed description of the different connections, a few words on the analog interface between successive encodings are appropriate. The four codecs can be grouped into two categories determined by whether the sampling rate is 8 kHz or 37.7 kHz. This leads to four possible combinations in simulating a tandem encoding because

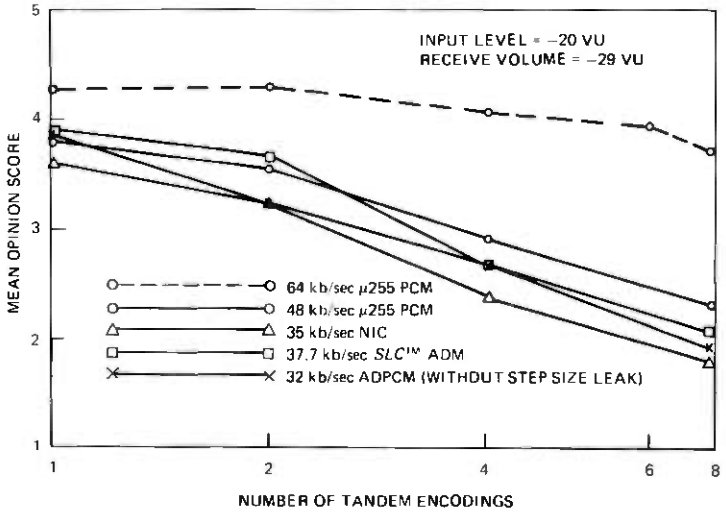


Fig. 14—Subjective results—tandem encodings.

of the filtering that is necessary. These four configurations are shown in Fig. 15. The top and bottom are identical to those shown in Fig. 11 since the sampling rates of the two codecs are the same. The two center arrangements in Fig. 15 are the cases where an 8-kHz and 37.7-kHz codec are both involved. Here the D channel bank transmit and receive filters are utilized with the *SLC* input and output filters. These four arrangements apply to the remainder of the discussions on reference connections.

7.1 Test design

7.1.1 Local reference connections

The local connection conditions are configured as shown in Fig. 16. There are four types of connections, depending on the presence of codecs in the loops: (*i*) near- and far-end loops, both analog, (*ii*) codec in near-end loop, (*iii*) codec in far-end loop, and (*iv*) codec in both loops. Note that an average loop loss of 3.7 dB[13] is assumed regardless of the presence or absence of a codec. It is recognized that this assumption may be inappropriate when a codec is present, since the loop loss might be reduced. However, this loss is fixed to avoid confounding the subjective effect of loss variation with that of introducing a codec in the loop. The input speech volume is chosen to be the average found for local calls, −24.8 VU.[10] The assumed loop loss of 3.7 dB translates this into a received volume of about −28.5 VU.

The central office is modeled as either an analog switch or a 64-kb/s μ255 PCM digital switch. When the office is analog, 16 dBrnC of random noise is added to the speech as it traverses the office. This noise value

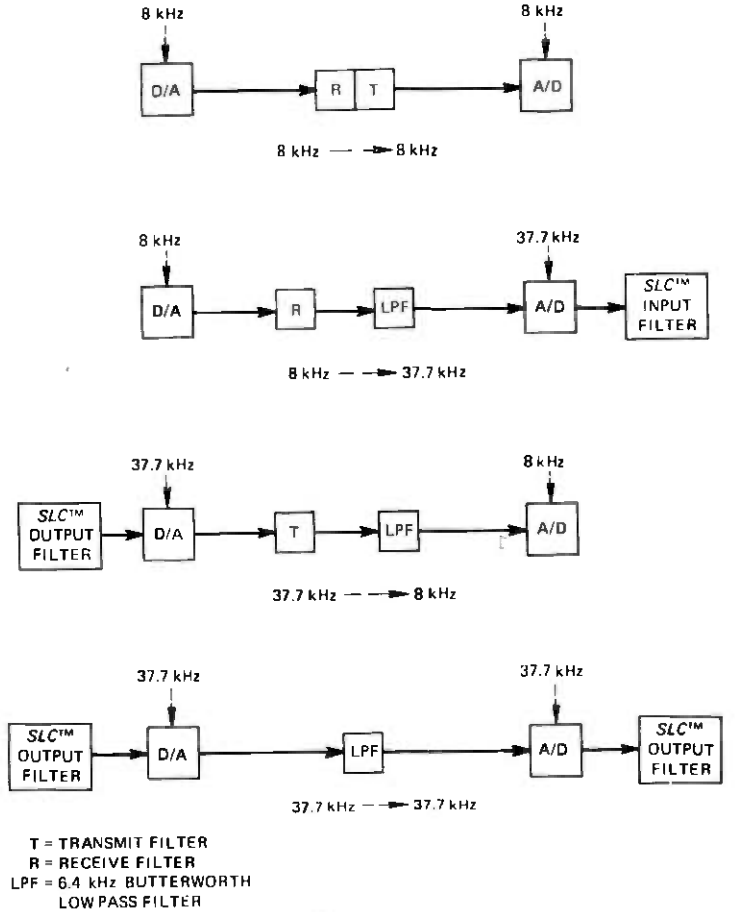


Fig. 15—Dissimilar tandem encoding configuration.

is chosen to correspond to the digital switch where 16 dBrnC is added after the PCM decoder, as described in Section 2.1. In the context of this local environment, the local reference connection conditions to be subjectively evaluated are listed in Fig. 21. In the conditions with a codec in both loops, the two codecs may be either identical or dissimilar. For the case of two dissimilar codecs, the subjective effect of ordering is investigated by including conditions where the physical locations of the codecs are interchanged.

7.1.2 Exchange reference connections

The exchange reference connection model is shown in Fig. 17. This model is a simple extension of the local connection model in that an additional central office is connected to the first office via a direct trunk. The central offices and loops, whether analog or digital, are modeled as in the local reference connections.

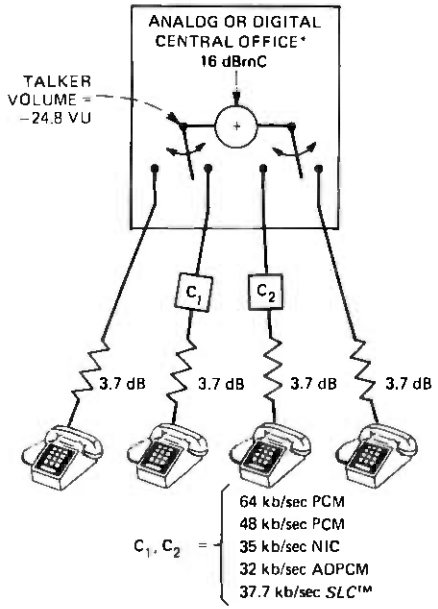
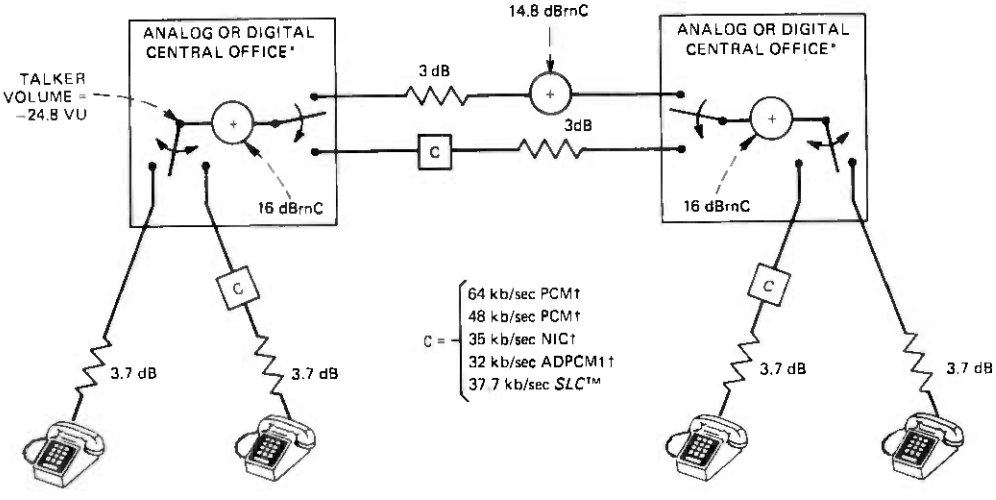


Fig. 16—Local reference connections.

The direct trunk facility between the central offices is either analog or digital. The loss used in both cases is 3 dB. In addition, 14.8 dBrnC of noise derived from survey data is introduced on the analog trunk. On the digital direct trunk, the noise introduced is the characteristic idle channel noise of the codec. The speech volume in both situations is −23.1 VU,[10] slightly higher than that found on the local reference connections.

To keep the number of conditions down to a mangeable level, codecs are introduced into the exchange connection in only one manner. If a codec is to appear in the connection, it must appear in all three components of connection simultaneously—both loops and the direct trunk. Furthermore, the three codecs must be identical; a mixture of different codecs is not allowed. Using these rules, the exchange reference connection conditions are formulated and tabulated in Fig. 22.

When both central offices are digital (64-kb/s μ255 PCM) and the codecs in the loops and direct trunk are either PCM or NIC, only a trivial code conversion is needed between the office and loop or trunk. The 3-dB direct trunk loss can be included in the far-end loop loss, and the entire connection collapses into a single encoding. For the case of the three ADPCM codecs, the analog link between the codec and digital central office can also be eliminated. All that is necessary is an intermediate D/D conversion to uniform PCM. Thus, the entire connection is digital where the 3-dB loss is a digital loss and is introduced where it is depicted in Fig. 17.



*64 kb/sec μ255 PCM DIGITAL CENTRAL OFFICE
†FOR DIGITAL CENTRAL OFFICE, CONNECTION COLLAPSES INTO A SINGLE ENCODING FOR THESE CODECS
††FOR DIGITAL CENTRAL OFFICE, ADPCM AND PCM ENCODINGS ARE SYNCHRONOUS WITH RESPECT TO EACH OTHER

Fig. 17—Exchange reference connections.

7.1.3 Toll reference connections

The toll reference connections are arrived at using the diagram in Fig. 18. Essentially, this diagram differs from Fig. 17 in that the direct trunks are replaced with the toll network wherein a connection comprises two toll-connecting trunks and one or more intertoll trunks. The toll network is assumed to be implemented on analog and digital facilities where "digital" implies only 64-kb/s μ255 PCM–No. 4 ESS switches, VIF terminals, and D channel banks.[1] The codecs described in the previous two sections will only be modeled in loops on the ends of the toll connection. Thus, a toll reference connection consists of two loops where a codec may appear and toll trunks with various mixtures of analog and digital 64-kb/s μ255 PCM facilities.

The local portion of Fig. 18 is similar to that in the local and exchange reference connections with the following exceptions. To limit the total number of test conditions, only the analog version of the central office is configured in toll connections. As before, a 16-dBrnC noise source is

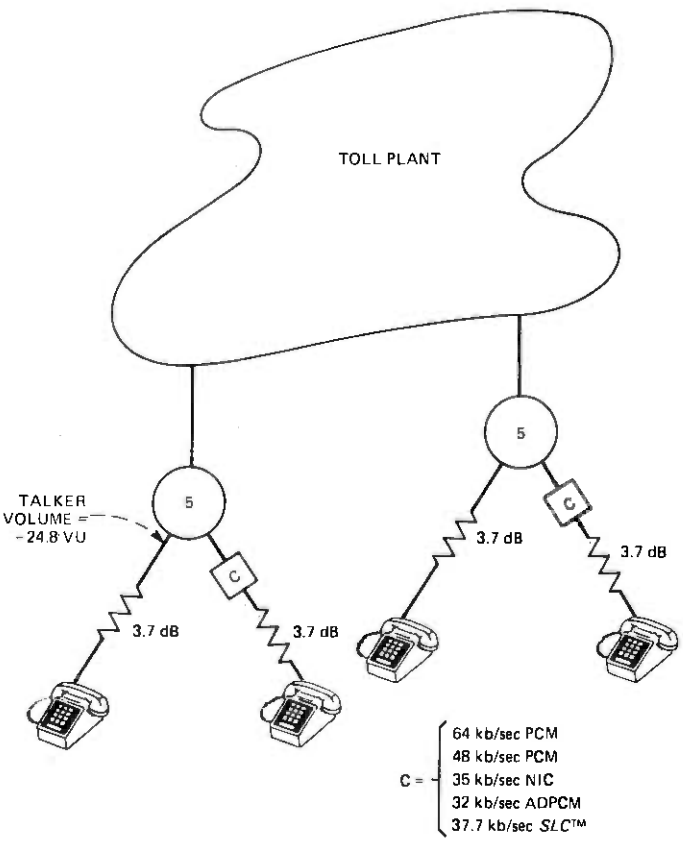


Fig. 18—Toll connection diagram.

incorporated in the central office model. The average speech volume used in the toll reference connections is significantly higher than that found in local and exchange connections and was found by McAdoo to be −16.8 VU.[10] Finally, when two codecs are simultaneously introduced in both loops on a particular connection, both codecs are always identical.

Four types of toll connections are considered here: (*i*) short, (*ii*) long, (*iii*) "worst-case" long, and (*iv*) all-digital. The short and long toll connections are further divided into three subcategories according to facility makeup: (*a*) analog switches and analog transmission facilities, (*b*) analog switches and digital transmission facilities, and (*c*) digital switches and analog transmission facilities. The "worst-case" long connection is simply derived from the long toll connection by the introduction of additional intertoll trunks. The all-digital connection consists of a single 64-kb/s μ255 PCM encoding in the toll network with the speech in digital form between the transmit and receive central offices.

The toll reference connections are now described, using Fig. 19.

(*i*) *Short Toll Connections*—The analog connection S_1 consists of two toll connecting trunks and a single short intertoll trunk. The toll connecting trunks are characterized by a VNL loss of 3 dB and an average survey noise source of 11.3 dBrnC.[14] The 60-mile intertoll trunk has a VNL design loss of 0.5 dB and a projected noise of 20.4 dBrnC.[11] The analog toll switches are assumed to add noise in an amount equivalent to 12 dBrnC. These characterizations of analog transmission and switching facilities will apply to all the toll connections discussed from this point on.

The center configuration, S_2, is a modification of S_1 where the analog toll switches are replaced with No. 4 ESS switches and VIF terminals. The VIF terminal involves a 64-kb/s μ255 PCM encoding and decoding followed by a 16-dBrnC0 noise source as described in Section 2.1. This translates into 13 dBrnC at the −3 TLP point for the speech with the codec overload set at +3 dBm0.

The bottom configuration, S_3, is another modification of S_1 with digital transmission facilities and analog switches. All three trunks are analog trunks, and VNL loss design applies. However, the noise sources on the toll-connecting and intertoll trunks have been replaced with the appropriate idle channel noises at the decoder side of the D channel banks.

(*ii*) *Long Toll Connections*—Referring to configuration L_1 in the diagram, the toll-connecting portion is identical to that of the short connection. The intertoll section consists of two intertoll trunks, a short (48 mi.) and a long (1300 mi.) one. The short trunk has loss and noise characteristics similar to the intertoll trunk in the short connection described above. Using VNL design and survey data, the long intertoll trunk has a loss of 2.0 dB and a noise of 37.8 dBrnC.

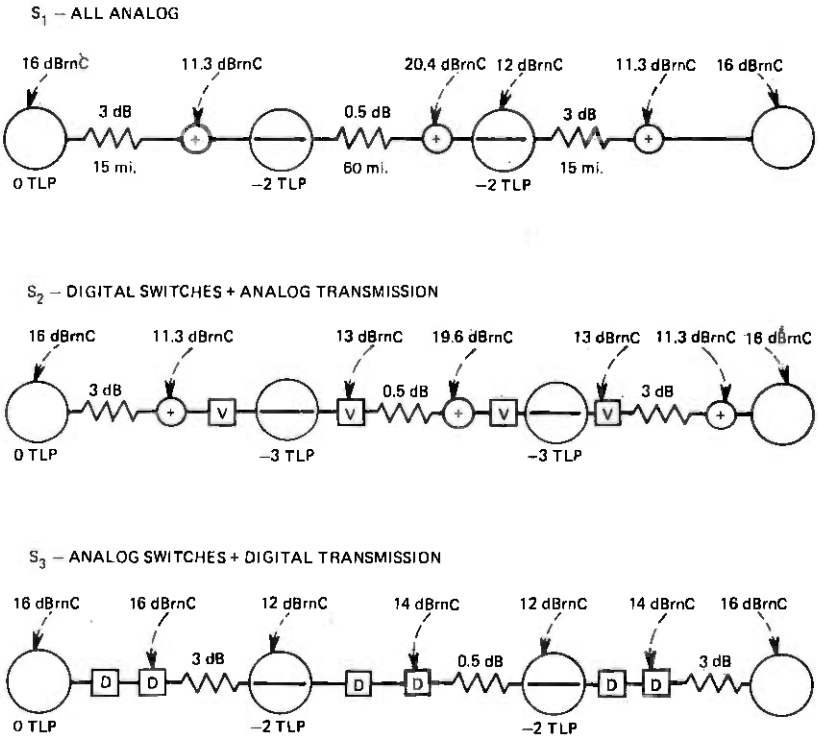


Fig. 19—Toll reference connections. (a) Short toll connections. (*Figs 19b and 19c on following pages.*)

Configuration L_1 is modified in a manner identical to that described above for the short toll connection to produce the mixed analog and digital connections, L_2, and L_3. Of significant interest here is the reduction of noise with the deployment of digital transmission facilities in configuration L_3. It is expected that this reduction in noise will manifest itself subjectively.

(*iii*) *Worst-Case Long Toll Connection*—This condition is dubbed "worst case" for two reasons. First, it is constructed from L_2 of the long toll connections by the addition of three intertoll trunks for a total of five intertoll trunks. Network statistics show that a small percentage of toll connections are made over five intertoll trunks. Second, the No. 4 ESS with analog transmission facilities type of connection contains not only all the analog impairments such as loss and noise but also the 64-kb/s μ255 PCM encodings. In this connection, there are a total of six PCM encodings in the toll portion and also codecs in the loops.

(*iv*) *All-Digital Toll Connection*—This connection is the simplest

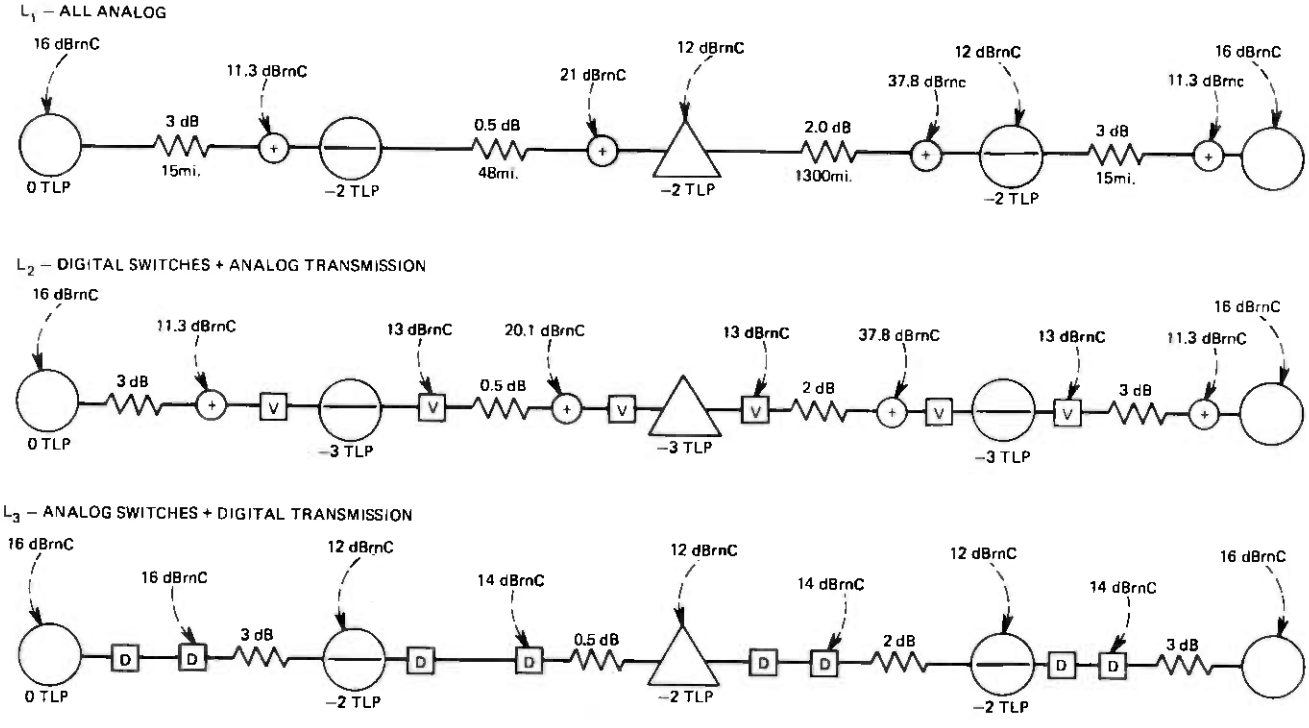


Fig. 19(b)—Long toll connections.

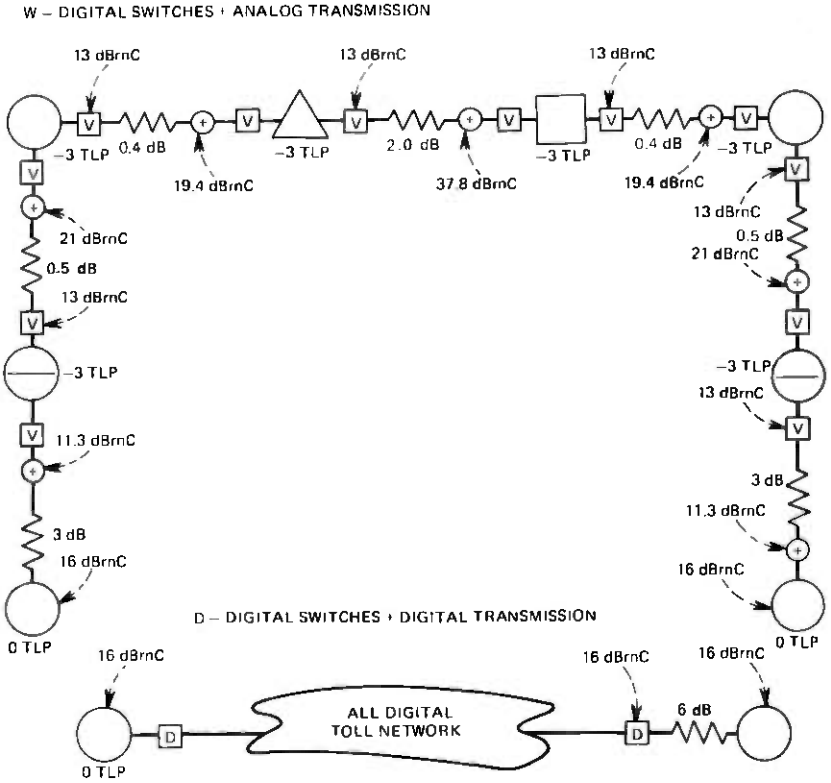


Fig. 19(c)—Worst-case and all-digit toll connections.

of the toll connections in that the toll network (switches and trunks) is purely digital. Hence, the toll network can be modeled by a single 64-kb/s μ255 PCM encoding implemented on a D channel bank on each toll-connecting trunk and a fixed 6-dB loss at the receiving central office.

The μ255 PCM at 48 and 64 kb/s, 35-kb/s NIC, 32-kb/s ADPCM, and 37.7-kb/s *SLC* ADM codecs are incorporated in the loops on the eight toll connections described above.

7.2 Results and observations

Detailed tabulations of the MOS ratings for the local, exchange, and toll reference connection conditions are given in Tables IX through XII in the appendix. A cumulative comparison of the analog noise ratings is plotted in Fig. 20, where the single encoding noise results are plotted along with the tandem encoding and reference connection results to illustrate a comparison across all the blocks of testing.

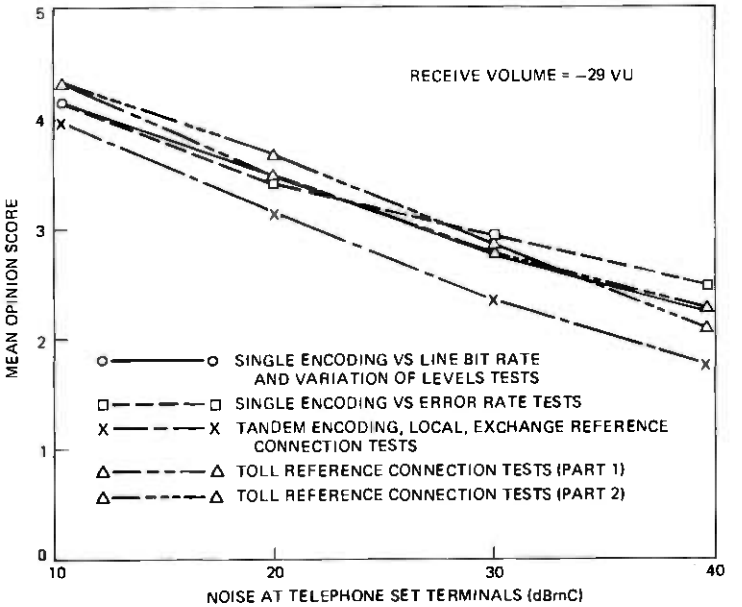


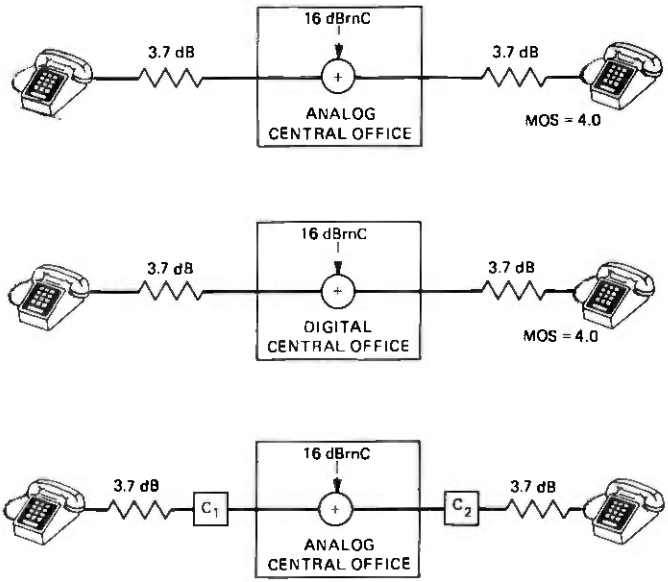
Fig. 20—Subjective results—additive random noise.

7.2.1 Local reference connections

The local reference connection results are summarized in Fig. 21. In this figure, the connections of Fig. 16 are redrawn to aid the reader in associating the connection with the results. The top configuration of Fig. 21 is the analog version of the connection where the only impairment introduced in the speech is the noise of the analog central office. This amounts to 12.3 dBrnC of noise with a speech level of −28.5 VU at the telephone set terminals. This condition is rated with a MOS of 4.0 and is in agreement with the noise results shown in Fig. 20.

The second configuration represents the digital central office case using 64-kb/s μ255 PCM. As described in Section 2.1, 16 dBrnC of noise is introduced into the speech following the decoder so that the speech level and idle channel noise at the line terminals of the telephone set are identical to those of the analog connection above it. The MOS of this connection is 4.0 and it is concluded that the 64-kb/s PCM office introduces no additional subjective distortion.

The bottom configuration in Fig. 21 is used to represent all possible combinations of codecs in the loops. The table directly beneath it gives the MOS ratings for three cases: (*i*) codec in the transmit loop with an analog receive loop, (*ii*) codec in the receive loop with an analog transmit loop, and (*iii*) codecs in both loops. The first two columns are essentially single encodings with the addition of the 16-dBrnC office noise. These results are in agreement with the single encoding results of Section 5.2.



CODEC	MEAN OPINION SCORES		
	C_1 ONLY	C_2 ONLY	C_1 AND C_2
64 kb/sec PCM	3.7	3.8	3.8
48 kb/sec PCM	3.4	3.5	3.3
35 kb/sec NIC	3.3	3.3	3.0
32 kb/sec ADPCM	3.6	3.5	3.2
37.7 kb/sec *SLC*™	3.6	3.6	3.2

C_1 \ C_2	64 kb/sec PCM	48 kb/sec PCM	35 kb/sec NIC	32 kb/sec ADPCM	37.7 kb/sec *SLC*™
64 kb/sec PCM		3.4	3.1	3.3	3.6
48 kb/sec PCM	3.5		3.0	3.0	3.2
35 kb/sec NIC	3.2	2.9		3.1	3.2
32 kb/sec ADPCM	3.4	3.2	2.9		3.3
37.7 kb/sec *SLC*™	3.5	2.8	3.1	3.1	

Fig. 21—Subjective results—local reference connections.

The third column represents two codecs in tandem with the addition of the office noise. It is observed that a second 64-kb/s μ255 PCM encoding does not introduce any additional degradation, while the MOS ratings for the other four codecs are slightly lower than those for the case of a single codec in one loop of the connection. This result is in agreement with the tandem encoding results discussed in Section 6.2.

The bottom table in Fig. 21 is a matrix of the results for mixed tandem encodings, that is, there are codecs in both loops but they are dissimilar. Reversal of the ordering of any pair of codecs is indicated by interchanging the row and column indices for any element in the matrix. A comparison of the elements in the upper and lower triangular portions

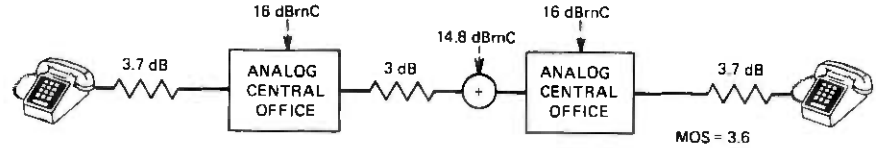
of the matrix leads to the following conclusions: (*i*) the subjective performance of a tandem encoding involving 64-kb/s PCM and one of the four lower bit rate codecs is roughly equivalent to a single encoding of the lower bit rate codec, and (*ii*) the subjective effects of ordering in dissimilar tandem encodings are small for the five codecs discussed here.

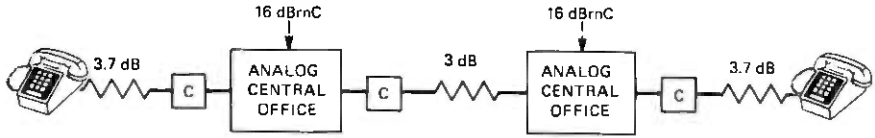
7.2.2 Exchange reference connections

The exchange reference connection results are shown in Fig. 22. In keeping with the format of the local reference connections, the top configuration in Fig. 22 represents the all-analog exchange connection. Here, 19.5 dBrnC of noise appears with the −29.8 VU speech at the telephone set terminals and is rated with a MOS of 3.6 by the subjects.

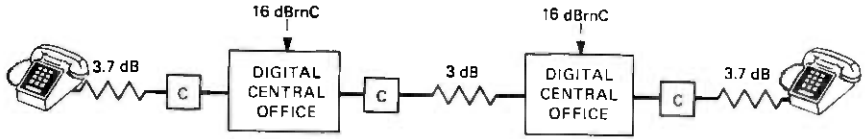
The second connection in Fig. 22 is identical to the analog connection except that three identical codecs are introduced in the exchange trunk and both loops. Note that the 14.8-dBrnC noise on the analog exchange trunk is effectively replaced by the idle channel noise of the codec. The table immediately below this connection lists the MOS results for the five codecs. The introduction of the three 64-kb/s PCM codecs does not alter the MOS rating over that of the analog connection. However, three encodings of one of the other four codecs degrades the connection somewhat.

The last configuration in Fig. 22 is a modification of the second configuration which is realized by replacing both analog central offices with digital central offices. As explained in Section 7.1, the resulting connection can be represented as either a single encoding, a series of synchronous tandem encodings with intermediate D/D conversions, or a series of asynchronous encodings with intermediate analog links, depending on code compatibility between the 64-kb/s PCM central offices and the codecs on the loops and exchange trunk. For 64-kb/s PCM, 48-kb/s PCM, and 35-kb/s NIC, the connection from the encoder of the transmit codec to the decoder of the receive codec collapses into a single encoding. The 3-dB exchange trunk loss is incorporated in the loop loss following the decoder. Thus, the first three entries in the table under this connection represent single encodings of these codecs. Consequently, the reduction of noise over the all-analog connection results in MOS ratings higher than the analog connection rating of 3.6. In the case of ADPCM, the successive encodings are performed synchronously with a corresponding reduction in noise, and the connection is rated nearly equivalent to the analog connection. The final entry for the *SLC* ADM codec is virtually identical to the analog central office case because all the noise sources are unchanged and the tandem encodings are performed asynchronously with the analog central offices replaced by 64-kb/s PCM digital central offices.





C		MOS
64 kb/sec	PCM	3.5
48 kb/sec	PCM	3.0
35 kb/sec	NIC	2.5
32 kb/sec	ADPCM	2.8
37.7 kb/sec	*SLC*™	2.8



C		MOS
64 kb/sec	PCM	4.2
48 kb/sec	PCM	3.8
35 kb/sec	NIC	3.7
32 kb/sec	ADPCM	3.7
37.7 kb/sec	*SLC*™	2.9

Fig. 22—Subjective results—exchange reference connections.

7.2.3 Toll reference connections

The final portion of the subjective testing is concerned with the effects of codecs in toll reference connections. The codecs are placed in loops on the ends of the eight toll connections of Fig. 19. This is accomplished in two steps. The first step is the placing of a codec in the receive loop only and, in the second step, identical codecs are placed in both loops.

Figure 23 displays the results for a codec in the receive loop only. The results for the longer toll connections on which 37.8-dBrnC of noise appears demonstrate that the noise is the dominant impairment and the effects of the codecs are of little importance (see connections L_1, L_2, and W). The MOS ratings of these connections are tightly grouped in the area of a MOS of 3 ("fair"), indicating that the subjects are reacting only to the noise source. Note that the noises on the intertoll trunks translate to approximately 31 dBrnC of noise at the line terminals of the telephone set for these three connections. A check with the noise results of Fig. 20 indicates agreement. It is also observed that in toll connection L_3, the 37.8-dBrnC noise is eliminated because of the deployment of digital transmission systems on the intertoll trunks. This shifts the MOS ratings upward in line with those obtained for the short (S_1, S_2, S_3) and all-digital (D) connections.

The toll connections which received the higher subjective scores (S_1, S_2, S_3, L_3, and D) show that, if the quality of the connection is good, subjects can discriminate among the five codecs and the MOS ratings are spread out over a point or so. However, it is observed that, even though the discrimination between one codec and the next may be small, on most connections the 64-kb/s PCM codec case is rated nearly equivalent to the case where both loops are analog. Also, a comparison of connections S_1 versus S_2 and L_1 versus L_2 with both loops analog serves to demonstrate that the addition of 64-kb/s PCM codecs in the analog toll portion of the connection does not significantly alter the subjective performance of the connection.

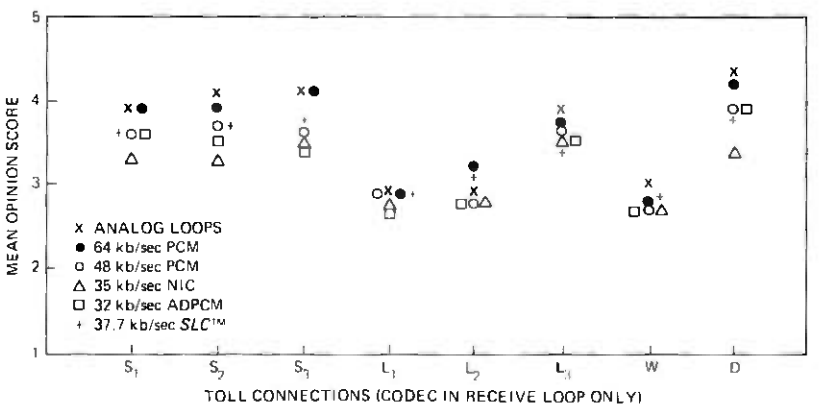


Fig. 23—Subjective results—toll reference connections.

The introduction of codecs in both loops of the toll connections results in the ratings of Fig. 24. Many of the observations for Fig. 23 apply here with minor modification. The toll connections of L_1, L_2, and W are rated the same with codecs in both loops when compared to the results for a codec in the receive loop only. In nearly all of the connections, the 64-kb/s PCM and analog loop cases again receive nearly equivalent ratings. The only differences between the results of Figs. 23 and 24 are those for connections with the higher ratings (S_1, S_2, S_3, L_3, and D). Here the spread among the MOS results for the lower bit rate codecs has increased because codecs are introduced in both loops. This indicates that subjects can perceive an additional degradation for two encodings of the lower bit rate codecs over a single encoding, a result already demonstrated in the tandem encoding results of Section 6.2.

VIII. CONCLUSIONS

Several important conclusions can be drawn from the results presented in this paper for speech:

(*i*) The 64-kb/s μ255 PCM codec can be used, with very few restrictions, in the telephone network without affecting speech performance. It can be tandemed up to eight times without introducing serious subjective degradation. It can be inserted in a variety of analog network connections with essentially no subjective penalty. However, it is shown that eight encodings of either 48-kb/s μ255 PCM, 35-kb/s NIC, 32-kb/s ADPCM, or 37.7-kb/s *SLC* ADM introduce significant subjective degradations.

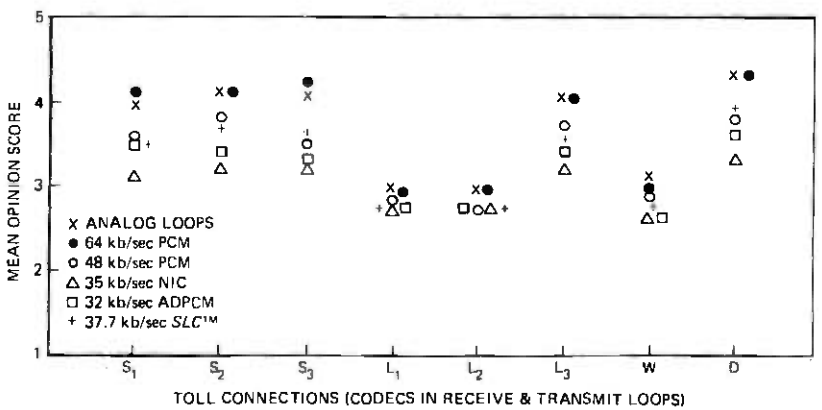


Fig. 24—Subjective results—toll reference connections.

This conclusion supports the general application of the 64-kb/s μ255 PCM in the combined analog and digital network which is likely to exist for many years to come. The use of codecs with lower bit rates would require more stringent application rules to avoid excessive degradation in tandem arrangements.

(*ii*) The 10^{-6} error rate minor alarm level for the toll switched digital network,[1] which uses 64-kb/s μ255 PCM, is more than adequate for speech since the results of this study show that the subjective degradation at a 10^{-5} error rate is negligible. In other words, if error rates could be guaranteed to never exceed 10^{-6} errors/bit, then error rate would not be a consideration for the four 48-kb/s codecs tested.

(*iii*) Connection degradation is dominated by high levels of random noise that may be found on long toll connections. However, the 64-kb/s μ255 PCM all-digital toll connection shows that the network performance will improve as the network becomes increasingly digital. As the network evolves, the performance should not be unduly limited by the choice of codec deployed in the loops.

(*iv*) For single encodings, the results show that the three adaptive coding schemes (NIC, ADPCM, and *SLC* ADM) have about a 12- to 16-kb/s advantage over μ255 PCM for equivalent subjective ratings. This advantage can also be inferred from the tandem encoding and reference connection results where 48-kb/s PCM has approximately the same rating as 35-kb/s NIC, 32-kb/s ADPCM, and 37.7-kb/s *SLC* ADM.

(*v*) At 48-kb/s, the level variation and error rate tests demonstrate that the three adaptive codecs are slightly superior to PCM under these conditions.

(*vi*) For the three adaptive coding algorithms (NIC, ADPCM, and *SLC* ADM), no single algorithm is significantly superior to the other two on an overall basis. All three perform comparably, with small variations over the range of conditions tested.

IX. ACKNOWLEDGMENTS

The authors wish to express their appreciation to a number of people at Bell Laboratories in Holmdel, N.J., who have made this study possible. W. C. Kublin, T. A. Pappas, and E. C. Stevens helped set up and conduct the subjective tests. P. C. Lopiparo contributed advice and moral support in constructing the simulation facility. Finally, discussions with J. E. Abate and J. L. Sullivan on the content of the tests are sincerely appreciated.

APPENDIX

Table V—Detailed tabulation of subjective scoring for single encoding vs line bit rate tests

Codec	Bit Rate (kb/s)	Votes	% Exc	% Good	% Fair	% Poor	% Uns.	MOS	σ
PCM	32	102	0.0	1.0	31.4	54.9	12.8	2.2	0.66
PCM	40	102	2.0	20.6	56.9	18.6	2.0	3.0	0.74
PCM	48	102	15.7	47.1	32.4	4.9	0.0	3.7	0.78
PCM	64	101	35.6	49.5	14.9	0.0	0.0	4.2	0.68
PCM_1	32	102	0.0	2.0	30.4	62.8	4.9	2.3	0.59
PCM_1	48	101	4.0	43.6	42.6	9.9	0.0	3.4	0.72
PCM_2	16	102	0.0	0.0	0.0	6.9	93.1	1.1	0.25
PCM_2	40	101	3.0	25.7	48.5	22.8	0.0	3.1	0.77
PCM_2	64	102	30.4	53.9	15.7	0.0	0.0	4.2	0.66
NIC	19	100	0.0	1.0	27.0	49.0	23.0	2.1	0.73
NIC	35	102	10.8	38.2	43.1	7.8	0.0	3.5	0.79
NIC	43	101	22.8	58.4	18.8	0.0	0.0	4.0	0.64
NIC	51	101	30.7	54.5	13.9	1.0	0.0	4.2	0.68
NIC	59	101	33.7	53.5	10.9	2.0	0.0	4.2	0.70
ADPCM	32	101	19.8	41.6	36.6	2.0	0.0	3.8	0.77
ADPCM	48	102	31.4	50.0	17.6	1.0	0.0	4.1	0.72
$ADPCM_1$	16	101	0.0	0.0	12.9	61.4	25.7	1.9	0.61
$ADPCM_1$	24	102	1.0	9.8	51.0	38.2	0.0	2.7	0.67
$ADPCM_1$	32	102	3.9	37.3	51.0	6.7	1.0	3.4	0.71
$ADPCM_1$	40	101	19.8	48.5	26.7	5.0	0.0	3.8	0.80
$ADPCM_1$	48	102	30.4	55.9	12.8	1.0	0.0	4.2	0.67
SLC™	24	102	2.0	20.6	57.8	18.6	1.0	3.0	0.71
SLC	37.7	101	11.9	46.5	36.6	5.0	0.0	3.7	0.75
SLC	48	102	29.4	51.0	15.7	3.9	0.0	4.1	0.78
Noise 10 dBrnC		203	34.0	47.3	17.2	1.5	0.0	4.1	0.74
Noise 20 dBrnC		204	9.8	30.9	51.5	7.8	0.0	3.4	0.77
Noise 30 dBrnC		203	1.5	15.3	46.8	35.0	1.5	2.8	0.76
Noise 40 dBrnC		203	1.5	3.5	26.6	54.2	14.3	2.2	0.79
Q—5 dB		203	0.0	0.0	0.0	32.0	68.0	1.3	0.47
Q—10 dB		204	0.0	0.5	18.1	54.4	27.0	1.9	0.68
Q—15 dB		204	0.0	10.3	48.5	35.8	5.4	2.6	0.74
Q—20 dB		203	2.0	32.0	53.2	12.8	0.0	3.2	0.69
Q—25 dB		203	23.2	50.3	25.1	1.5	0.0	4.0	0.73

Notation:

PCM = 15-segment PCM (mid-tread)
PCM_1 = 15-segment PCM (mid-riser)
PCM_2 = continuous law PCM (mid-tread)
NIC = NIC PCM
ADPCM = ADPCM (without step-size leak)
$ADPCM_1$ = ADPCM (with step-size leak)

Table VI—Detailed tabulation of subjective scoring for single encoding vs level variation tests

Codec	Levels	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
PCM	I_1,R_1	51	5.9	31.4	41.2	19.6	2.0	3.2	0.89
PCM	I_1,R_2	51	5.9	25.5	52.9	15.7	0.0	3.2	0.77
PCM	I_2,R_1	51	33.3	37.3	19.6	9.8	0.0	3.9	0.96
PCM	I_2,R_2	51	45.1	43.1	11.8	0.0	0.0	4.3	0.68
PCM	I_2,R_3	51	13.7	43.1	39.2	3.9	0.0	3.7	0.76
PCM	I_3,R_2	51	19.6	54.9	25.5	0.0	0.0	3.9	0.67
PCM	I_3,R_3	51	17.7	43.1	37.3	2.0	0.0	3.8	0.76
PCM	I_4,R_3	51	0.0	41.2	56.9	2.0	0.0	3.4	0.53
PCM	I_5,R_4	51	0.0	2.0	31.4	52.9	13.7	2.2	0.69
NIC	I_1,R_1	51	15.7	49.0	31.4	3.9	0.0	3.8	0.76
NIC	I_1,R_2	51	5.9	25.5	47.1	21.6	0.0	3.2	0.83
NIC	I_2,R_1	50	34.0	38.0	24.0	4.0	0.0	4.0	0.86
NIC	I_2,R_2	51	54.9	41.2	3.9	0.0	0.0	4.5	0.57
NIC	I_2,R_3	51	27.5	47.1	25.5	0.0	0.0	4.0	0.73
NIC	I_3,R_2	51	27.5	70.6	2.0	0.0	0.0	4.3	0.48
NIC	I_3,R_3	51	35.3	35.3	29.4	0.0	0.0	4.1	0.80
NIC	I_4,R_3	51	21.6	41.2	33.3	3.9	0.0	3.8	0.82
NIC	I_5,R_4	51	2.0	7.8	54.9	33.3	2.0	2.8	0.71
ADPCM	I_1,R_1	51	33.3	43.1	15.7	7.8	0.0	4.0	0.90
ADPCM	I_1,R_2	50	18.0	46.0	28.0	8.0	0.0	3.7	0.84
ADPCM	I_2,R_1	51	31.4	43.1	19.6	3.9	2.0	4.0	0.92
ADPCM	I_2,R_2	51	35.3	52.9	11.8	0.0	0.0	4.2	0.64
ADPCM	I_2,R_3	50	4.0	38.0	48.0	10.0	0.0	3.4	0.71
ADPCM	I_3,R_2	50	34.0	40.0	14.0	12.0	0.0	4.0	0.98
ADPCM	I_3,R_3	51	11.8	41.2	39.2	7.8	0.0	3.6	0.80
ADPCM	I_4,R_3	51	15.7	45.1	39.2	0.0	0.0	3.8	0.70
ADPCM	I_5,R_4	51	0.0	15.7	49.0	31.4	3.9	2.8	0.76
SLC™	I_1,R_1	51	17.7	37.3	35.3	9.8	0.0	3.6	0.88
SLC	I_1,R_2	50	16.0	36.0	42.0	6.0	0.0	3.6	0.82
SLC	I_2,R_1	51	43.1	31.4	15.7	9.8	0.0	4.1	0.99
SLC	I_2,R_2	51	21.6	54.9	23.5	0.0	0.0	4.0	0.67
SLC	I_2,R_3	51	19.6	47.1	29.4	3.9	0.0	3.8	0.78
SLC	I_3,R_2	50	34.0	52.0	14.0	0.0	0.0	4.2	0.66
SLC	I_3,R_3	51	7.8	43.1	37.3	11.8	0.0	3.5	0.80
SLC	I_4,R_3	51	3.9	7.8	60.8	27.5	0.0	2.9	0.70
SLC	I_5,R_4	50	0.0	18.0	42.0	40.0	0.0	2.8	0.73
Noise 10 dBrnC	R_1	51	23.5	39.2	27.5	7.8	2.0	3.8	0.97
Noise 10 dBrnC	R_3	51	5.9	60.8	25.5	7.8	0.0	3.7	0.71
Noise 10 dBrnC	R_4	51	3.9	9.8	43.1	43.1	0.0	2.8	0.79
Noise 20 dBrnC	R_1	51	3.9	29.4	52.9	11.8	2.0	3.2	0.77
Noise 20 dBrnC	R_3	51	17.7	54.9	27.5	0.0	0.0	3.9	0.66
Noise 20 dBrnC	R_4	51	0.0	5.9	47.1	47.1	0.0	2.6	0.60
Noise 30 dBrnC	R_1	51	3.9	11.8	43.1	37.3	3.9	2.8	0.86
Noise 30 dBrnC	R_3	51	0.0	15.7	56.9	27.5	0.0	2.9	0.65
Noise 30 dBrnC	R_4	51	2.0	2.0	25.5	64.7	5.9	2.3	0.69
Noise 40 dBrnC	R_1	51	0.0	5.9	25.5	41.2	27.5	2.1	0.87
Noise 40 dBrnC	R_3	51	0.0	0.0	23.5	58.8	17.6	2.1	0.64
Noise 40 dBrnC	R_4	51	2.0	2.0	11.8	66.7	17.6	2.0	0.74
Q—5 dB	R_1	51	0.0	0.0	5.9	19.6	74.5	1.3	0.58
Q—5 dB	R_3	51	0.0	0.0	0.0	27.5	72.5	1.3	0.45
Q—5 dB	R_4	51	0.0	0.0	0.0	15.7	84.3	1.2	0.36
Q—10 dB	R_1	51	0.0	2.0	9.8	49.0	39.2	1.8	0.71
Q—10 dB	R_3	51	0.0	0.0	11.8	62.7	25.5	1.9	0.59
Q—10 dB	R_4	51	0.0	0.0	2.0	54.9	43.1	1.6	0.53
Q—15 dB	R_1	51	9.8	7.8	51.0	29.4	2.0	2.9	0.92
Q—15 dB	R_3	51	0.0	3.9	49.0	47.1	0.0	2.6	0.57
Q—15 dB	R_4	51	0.0	5.9	7.8	70.6	15.7	2.0	0.68
Q—20 dB	R_1	51	15.7	35.3	41.2	7.8	0.0	3.6	0.84
Q—20 dB	R_3	51	3.9	23.5	54.9	17.6	0.0	3.1	0.74

(continued)

Table VI (cont)

Q—20 dB	R_4	51	0.0	3.9	37.3	52.9	5.9	2.4	0.66
Q—25 dB	R_1	50	32.0	34.0	26.0	8.0	0.0	3.9	0.94
Q—25 dB	R_3	51	7.8	39.2	39.2	13.7	0.0	3.4	0.82
Q—25 dB	R_4	51	9.8	13.7	43.1	27.5	5.9	2.9	1.02

Notation:
PCM = 48-kb/s 15-segment PCM (mid-tread)
NIC = 51-kb/s NIC PCM
ADPCM = 48-kb/s ADPCM (with step-size leak)
SLC = 48-kb/s *SLC* ADM
I_1, I_2, I_3, I_4, and I_5 = input levels of −0.2, −10.5, −20.7, −31.0, and −41.2 VU, respectively
R_1, R_2, R_3, and R_4 = received volumes of −20.7, −29.0, −38.3, and −47.0 VU, respectively

Table VII—Detailed tabulation of subjective scoring for single encoding vs error rate tests

Codec	Error Rate	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
PCM	10^{-1}	207	0.0	0.0	0.0	3.4	96.6	1.0	0.18
PCM	10^{-2}	204	0.0	2.0	8.3	63.2	26.5	1.9	0.64
PCM	10^{-3}	207	0.5	17.9	66.7	15.0	0.0	3.0	0.59
PCM	10^{-4}	207	31.4	41.1	26.6	0.5	0.5	4.0	0.80
PCM	10^{-5}	208	42.8	50.0	7.2	0.0	0.0	4.4	0.61
NIC	10^{-1}	208	0.0	0.0	0.5	7.7	91.8	1.1	0.30
NIC	10^{-2}	207	0.0	3.9	33.8	54.1	8.2	2.3	0.68
NIC	10^{-3}	207	12.1	46.4	39.1	2.4	0.0	3.7	0.71
NIC	10^{-4}	207	43.5	44.9	10.1	1.4	0.0	4.3	0.71
NIC	10^{-5}	208	52.4	41.8	5.8	0.0	0.0	4.5	0.60
ADPCM	10^{-1}	207	0.0	0.0	0.0	3.9	96.1	1.0	0.19
ADPCM	10^{-2}	206	0.0	1.5	16.5	65.0	17.0	2.0	0.63
ADPCM	10^{-3}	207	22.7	45.9	28.0	3.4	0.0	3.9	0.79
ADPCM	10^{-4}	208	35.1	52.4	12.0	0.5	0.0	4.2	0.66
ADPCM	10^{-5}	208	41.8	44.2	13.9	0.0	0.0	4.3	0.69
SLC™	10^{-1}	206	0.0	0.0	3.4	27.7	68.9	1.3	0.54
SLC	10^{-2}	206	1.0	17.5	55.3	25.2	1.0	2.9	0.71
SLC	10^{-3}	207	40.1	44.9	14.5	0.5	0.0	4.3	0.71
SLC	10^{-4}	206	46.1	44.7	9.2	0.0	0.0	4.4	0.65
SLC	10^{-5}	206	43.7	38.8	16.5	1.0	0.0	4.3	0.76
Noise 10 dBrnC		201	31.3	50.2	17.9	0.5	0.0	4.1	0.70
Noise 20 dBrnC		206	7.8	37.9	48.5	5.8	0.0	3.5	0.72
Noise 30 dBrnC		205	3.4	14.6	53.7	27.3	1.0	2.9	0.77
Noise 40 dBrnC		206	0.0	7.8	35.0	54.4	2.9	2.5	0.68
Q—10 dB		208	0.0	2.4	11.5	69.7	16.3	2.0	0.61
Q—20 dB		206	14.6	37.4	40.3	7.8	0.0	3.6	0.83

Notation:
PCM = 48-kb/s 15-segment PCM (mid-tread)
NIC = 51-kb/s NIC PCM
ADPCM = 48-kb/s ADPCM (with step-size leak)
SLC = 48-kb/s *SLC* ADM

Table VIII—Detailed tabulation of subjective scoring for number of codecs in tandem

Codec	Tandem Encodings	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
PCM	1	106	35.9	55.7	6.6	1.9	0.0	4.3	0.66
PCM	2	105	41.9	45.7	11.4	1.0	0.0	4.3	0.70
PCM	4	106	22.6	60.4	14.2	2.8	0.0	4.0	0.69
PCM	6	106	21.7	51.9	23.6	2.8	0.0	3.9	0.75
PCM	8	106	16.0	41.5	36.8	5.7	0.0	3.7	0.81

(continued)

Table VIII (cont)

PCM_1	1	106	14.2	52.8	31.1	1.9	0.0	3.8	0.70
PCM_1	2	106	12.3	34.9	46.2	6.6	0.0	3.5	0.79
PCM_1	4	106	0.0	15.1	58.5	26.4	0.0	2.9	0.63
PCM_1	8	106	0.0	2.8	34.0	53.8	9.4	2.3	0.68
NIC	1	106	7.5	51.9	33.0	7.5	0.0	3.6	0.74
NIC	2	106	2.8	34.9	42.5	19.8	0.0	3.2	0.79
NIC	4	106	0.0	7.5	34.0	46.2	12.3	2.4	0.79
NIC	8	106	0.0	0.9	13.2	48.1	37.7	1.8	0.70
ADPCM	1	106	22.6	41.5	31.1	4.7	0.0	3.8	0.83
ADPCM	2	105	1.9	34.3	49.5	14.3	0.0	3.2	0.71
ADPCM	4	106	0.9	8.5	49.1	39.6	1.9	2.7	0.70
ADPCM	8	106	0.0	0.9	17.9	53.8	27.4	1.9	0.70
SLC™	1	104	27.9	40.4	26.0	4.8	1.0	3.9	0.90
SLC	2	105	13.3	45.7	33.3	6.7	1.0	3.6	0.83
SLC	4	106	0.9	10.4	44.3	42.5	1.9	2.7	0.73
SLC	8	106	0.0	1.9	26.4	47.2	24.5	2.1	0.76
ADPCM*	1 (at 16 kb/s)	106	0.0	1.0	13.2	41.5	44.3	1.7	0.73
ADPCM*	1 (at 24 kb/s)	106	0.0	11.3	52.8	32.1	3.8	2.7	0.71
ADPCM*	1 (at 32 kb/s)	105	29.5	41.0	25.7	3.8	0.0	4.0	0.84
ADPCM*	1 (at 48 kb/s)	106	26.4	50.9	21.7	1.0	0.0	4.0	0.72

* ADPCM conditions (without step-size leak) which tie into those of Table V.

Notation:

PCM = 64-kb/s 15-segment PCM (mid-tread)
PCM_1 = 48-kb/s 15-segment PCM (mid-tread)
NIC = 35-kb/s NIC PCM
ADPCM = 32-kb/s ADPCM (without step-size leak)
SLC = 37.7-kb/s *SLC* ADM

Table IX—Detailed tabulation of subjective scoring for local reference connection tests

Connection	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
AL-ACO-AL	106	25.5	55.7	16.0	2.8	0.0	4.0	0.73
AL-DCO-AL	106	22.6	55.7	18.9	2.8	0.0	4.0	0.73
AL-ACO-P	106	9.4	57.5	29.2	3.8	0.0	3.7	0.68
AL-ACO-P_1	105	6.7	40.0	39.0	13.3	1.0	3.4	0.83
AL-ACO-N	106	3.8	34.0	50.0	12.3	0.0	3.3	0.73
AL-ACO-A	106	15.1	37.7	41.5	5.7	0.0	3.6	0.81
AL-ACO-S	106	4.7	51.9	38.7	4.7	0.0	3.6	0.66
P-ACO-AL	105	17.1	51.4	28.6	2.9	0.0	3.8	0.74
P_1-ACO-AL	106	6.6	43.4	41.5	7.5	0.9	3.5	0.77
N-ACO-AL	106	1.9	37.7	51.9	7.5	0.9	3.3	0.68
A-ACO-AL	106	3.8	47.2	40.6	8.5	0.0	3.5	0.70
S-ACO-AL	106	6.6	53.8	34.0	5.7	0.0	3.6	0.69
P-ACO-P	106	10.4	55.7	33.0	0.9	0.0	3.8	0.64
P_1-ACO-P_1	104	5.8	33.7	50.0	10.6	0.0	3.4	0.74
N-ACO-N	106	1.9	29.2	36.8	27.4	4.7	3.0	0.91
A-ACO-A	106	3.8	28.3	53.8	13.2	0.9	3.2	0.75
S-ACO-S	106	0.9	26.4	59.4	13.2	0.0	3.2	0.64
P-ACO-P_1	105	2.9	39.0	51.4	5.7	0.9	3.4	0.68
P-ACO-N	106	0.9	27.4	54.7	17.0	0.0	3.1	0.68
P-ACO-A	106	4.7	30.2	54.7	10.4	0.0	3.3	0.71
P-ACO-S	105	2.9	55.2	38.1	3.8	0.0	3.6	0.62
P_1-ACO-P	106	5.7	45.3	38.7	10.4	0.0	3.5	0.75
P_1-ACO-N	106	0.9	17.9	61.3	19.8	0.0	3.0	0.64
P_1-ACO-A	106	0.9	15.1	63.2	19.8	0.9	3.0	0.65
P_1-ACO-S	106	1.9	33.0	50.9	14.2	0.0	3.2	0.70
N-ACO-P	106	0.9	33.0	51.9	12.3	1.9	3.2	0.73
N-ACO-P_1	106	2.8	17.0	48.1	29.2	2.8	2.9	0.82
N-ACO-A	105	1.9	26.7	55.2	15.2	1.0	3.1	0.72
N-ACO-S	106	0.9	32.1	50.9	16.0	0.0	3.2	0.70
A-ACO-P	105	7.6	34.3	52.4	5.7	0.0	3.4	0.72
A-ACO-P_1	105	3.8	28.6	48.6	18.1	1.0	3.2	0.79
A-ACO-N	104	0.0	14.4	58.7	25.0	1.9	2.9	0.67

(continued)

Table IX (cont)

A-ACO-S	106	3.8	31.1	52.8	12.3	0.0	3.3	0.72
S-ACO-P	106	4.7	44.3	47.2	3.8	0.0	3.5	0.65
S-ACO-P_1	106	0.0	14.2	55.7	29.2	0.9	2.8	0.67
S-ACO-N	106	1.9	24.5	52.8	19.8	0.9	3.1	0.74
S-ACO-A	106	0.0	18.9	67.0	14.2	0.0	3.1	0.57

Notation:
P = 64-kb/s 15-segment PCM (mid-tread)
P_1 = 48-kb/s 15-segment PCM (mid-tread)
N = 35-kb/s NIC PCM
A = 32-kb/s ADPCM (without step-size leak)
S = 37.7-kb/s *SLC* ADM
AL = analog loop
ACO = analog central office
DCO = digital central office

Table X—Detailed tabulation of subjective scoring for exchange reference connection tests

Connection	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
AL-ACO-AEX-ACO-AL	106	7.5	49.1	39.6	3.8	0.0	3.6	0.68
P-ACO-P-ACO-P	106	5.7	37.7	54.7	1.9	0.0	3.5	0.63
P_1-ACO-P_1-ACO-P_1	106	3.8	23.6	40.6	27.4	4.7	2.9	0.92
N-ACO-N-ACO-N	106	0.0	7.5	44.3	42.5	5.7	2.5	0.72
A-ACO-A-ACO-A	105	0.0	12.4	61.0	21.0	5.7	2.8	0.72
S-ACO-S-ACO-S	106	0.0	7.5	61.3	30.2	0.9	2.8	0.60
P-DCO-P-DCO-P	106	36.8	50.0	13.2	0.0	0.0	4.2	0.67
P_1-DCO-P_1-DCO-P_1	106	14.2	60.4	19.8	5.7	0.0	3.8	0.73
N-DCO-N-DCO-N	106	10.4	50.9	36.8	1.9	0.0	3.7	0.68
A-DCO-A-DCO-A	106	15.1	47.2	32.1	5.7	0.0	3.7	0.79
S-DCO-S-DCO-S	105	1.9	11.4	60.0	25.7	1.0	2.9	0.69
Noise 10 dBrnC	212	20.3	56.6	23.1	0.0	0.0	4.0	0.66
Noise 20 dBrnC	211	2.8	29.9	47.9	16.6	2.8	3.1	0.82
Noise 30 dBrnC	211	0.0	8.5	30.8	46.9	13.7	2.3	0.82
Noise 40 dBrnC	212	0.0	1.9	11.8	44.8	41.5	1.7	0.74
Q—5 dB	212	0.0	0.0	0.0	10.8	89.2	1.1	0.31
Q—10 dB	211	0.0	0.9	9.0	39.8	50.2	1.6	0.69
Q—15 dB	211	0.9	5.7	37.4	46.0	10.0	2.4	0.78
Q—20 dB	211	5.7	30.3	47.4	16.1	0.5	3.3	0.81
Q—25 dB	210	32.4	48.6	15.7	3.3	0.0	4.1	0.78

Notation:
P = 64-kb/s 15-segment PCM (mid-tread)
P_1 = 48-kb/s 15-segment PCM (mid-tread)
N = 35-kb/s NIC PCM
A = 32-kb/s ADPCM (without step-size leak)
S = 37.7-kb/s *SLC* ADM
AL = analog loop
ACO = analog central office
DCO = digital central office
AEX = analog exchange trunk

Table XI—Detailed tabulation of subjective scoring for toll reference connection conditions—part 1

Connection	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
AL-S_1-AL	112	17.0	58.0	22.3	2.7	0.0	3.9	0.70
AL-S_1-P	102	19.6	50.0	30.4	0.0	0.0	3.9	0.70
AL-S_1-P_1	111	12.6	42.3	36.0	9.0	0.0	3.6	0.82
AL-S_1-N	112	2.7	35.7	47.3	13.4	0.9	3.3	0.75
AL-S_1-A	112	10.7	44.6	38.4	6.3	0.0	3.6	0.76
AL-S_1-S	111	5.4	54.1	35.1	5.4	0.0	3.6	0.68
AL-S_2-AL	111	30.6	49.5	18.0	1.8	0.0	4.1	0.74
AL-S_2-P	112	18.8	56.3	23.2	1.8	0.0	3.9	0.70
AL-S_2-P_1	111	9.9	53.2	30.6	5.4	0.9	3.7	0.77
AL-S_2-N	112	6.3	36.6	40.2	17.0	0.0	3.3	0.83
AL-S_2-A	102	8.8	40.2	41.2	9.8	0.0	3.5	0.79
AL-S_2-S	112	14.3	43.8	36.6	5.4	0.0	3.7	0.78
AL-S_3-AL	112	25.9	57.1	15.2	1.8	0.0	4.1	0.69
AL-S_3-P	112	30.4	50.0	19.6	0.0	0.0	4.1	0.70
AL-S_3-P_1	102	11.8	43.1	41.2	3.9	0.0	3.6	0.74
AL-S_3-N	102	7.8	43.1	41.2	7.8	0.0	3.5	0.75
AL-S_3-A	102	4.9	43.1	41.2	10.8	0.0	3.4	0.75
AL-S_3-S	112	15.2	49.1	30.4	5.4	0.0	3.7	0.78
AL-L_1-AL	111	2.7	15.3	51.4	27.0	3.6	2.9	0.81
AL-L_1-P	112	1.8	19.6	49.1	28.6	0.9	2.9	0.76
AL-L_1-P_1	112	0.9	16.1	55.4	27.7	0.0	2.9	0.68
AL-L_1-N	102	1.0	10.8	50.0	37.3	1.0	2.7	0.70
AL-L_1-A	112	0.9	10.7	50.0	35.7	2.7	2.7	0.72
AL-L_1-S	101	1.0	20.8	51.5	26.7	0.0	3.0	0.72
AL-L_2-AL	112	1.8	14.3	52.7	31.3	0.0	2.9	0.71
AL-L_2-P	111	2.7	28.8	46.8	21.6	0.0	3.1	0.77
AL-L_2-P_1	111	0.9	13.5	55.0	30.6	0.0	2.9	0.67
AL-L_2-N	112	0.0	13.4	51.8	33.9	0.9	2.8	0.68
AL-L_2-A	111	0.9	9.9	57.7	29.7	1.8	2.8	0.68
AL-L_2-S	111	1.8	21.6	54.1	22.5	0.0	3.0	0.72
AL-L_3-AL	112	17.9	62.5	18.8	0.9	0.0	4.0	0.63
AL-L_3-P	112	13.4	50.0	32.1	4.5	0.0	3.7	0.75
AL-L_3-P_1	112	10.7	45.5	40.2	3.6	0.0	3.6	0.72
AL-L_3-N	112	8.0	43.8	41.1	7.1	0.0	3.5	0.74
AL-L_3-A	112	8.0	42.0	42.0	8.0	0.0	3.5	0.76
AL-L_3-S	112	5.4	39.3	47.3	8.0	0.0	3.4	0.72
AL-W-AL	112	1.8	20.5	58.9	18.8	0.0	3.1	0.68
AL-W-P	112	1.8	10.7	53.6	33.9	0.0	2.8	0.69
AL-W-P_1	102	2.0	6.9	54.9	35.3	1.0	2.7	0.68
AL-W-N	101	0.0	9.9	47.5	41.6	1.0	2.7	0.66
AL-W-A	112	1.8	7.1	51.8	39.3	0.0	2.7	0.67
AL-W-S	112	0.0	10.7	55.4	33.0	0.9	2.8	0.64
AL-D-AL	111	39.6	48.6	10.8	0.9	0.0	4.3	0.68
AL-D-P	102	29.4	53.9	14.7	2.0	0.0	4.1	0.71
AL-D-P_1	112	17.9	53.6	26.8	1.8	0.0	3.9	0.71
AL-D-N	112	8.0	33.9	45.5	12.5	0.0	3.4	0.80
AL-D-A	112	18.8	57.1	20.5	3.6	0.0	3.9	0.73
AL-D-S	111	22.5	38.7	34.2	4.5	0.0	3.8	0.84
Noise 10 dBrnC	224	44.2	46.0	9.8	0.0	0.0	4.3	0.65
Noise 20 dBrnC	214	3.7	43.9	47.7	4.7	0.0	3.5	0.65
Noise 30 dBrnC	214	0.9	15.0	48.6	34.6	0.9	2.8	0.73
Noise 40 dBrnC	224	0.0	2.7	29.0	60.7	7.6	2.3	0.63
Q—5 dB	223	0.0	0.0	0.0	21.1	78.9	1.2	0.41
Q—10 dB	226	0.0	0.0	7.1	51.3	41.6	1.7	0.61
Q—15 dB	223	0.0	5.8	37.7	43.9	12.6	2.4	0.77
Q—20 dB	214	13.1	23.4	37.9	23.8	1.9	3.2	1.01
Q—25 dB	224	32.1	43.8	21.0	3.1	0.0	4.1	0.81

Notation:
P, P_1, N, A, S, AL—see Table X
S_1, S_2, S_3, L_1, L_2, L_3, W, D—see Figs. 19a, 19b, and 19c in text.

Table XII—Detailed tabulation of subjective scoring for toll reference connection conditions—part 2

Connection	Votes	% Exc.	% Good	% Fair	% Poor	% Uns.	MOS	σ
P-S_1-P	108	40.7	50.0	9.3	0.0	0.0	4.3	0.63
P_1-S_1-P_1	108	10.2	48.1	37.0	4.6	0.0	3.6	0.73
N-S_1-N	107	2.8	26.2	51.4	17.8	1.9	3.1	0.78
A-S_1-A	108	13.9	35.2	35.2	15.7	0.0	3.5	0.92
S-S_1-S	108	6.5	45.4	41.7	6.5	0.0	3.5	0.71
P-S_2-P	108	23.1	64.8	12.0	0.0	0.0	4.1	0.58
P_1-S_2-P_1	107	16.8	50.5	29.9	2.8	0.0	3.8	0.74
N-S_2-N	108	6.5	23.1	55.6	14.8	0.0	3.2	0.77
A-S_2-A	108	5.6	33.3	52.8	8.3	0.0	3.4	0.71
S-S_2-S	108	13.9	51.9	28.7	5.6	0.0	3.7	0.76
P-S_3-P	108	31.5	56.5	10.2	1.9	0.0	4.2	0.68
P_1-S_3-P_1	107	11.2	38.3	44.9	4.7	0.9	3.5	0.79
N-S_3-N	108	2.8	34.3	43.5	17.6	1.9	3.2	0.82
A-S_3-A	106	3.8	34.0	53.8	8.5	0.0	3.3	0.68
S-S_3-S	107	14.0	50.5	29.9	5.6	0.0	3.7	0.77
P-L_1-P	108	0.0	13.0	59.3	27.8	0.0	2.9	0.62
P_1-L_1-P_1	108	0.0	10.2	59.3	30.6	0.0	2.8	0.60
N-L_1-N	108	0.0	8.3	60.2	28.7	2.8	2.7	0.64
A-L_1-A	108	0.0	6.5	54.6	38.0	0.9	2.7	0.61
S-L_1-S	107	0.0	4.7	57.0	37.4	0.9	2.7	0.58
P-L_2-P	108	0.0	17.6	58.3	24.1	0.0	2.9	0.64
P_1-L_2-P_1	108	0.0	7.4	61.1	28.7	2.8	2.7	0.63
N-L_2-N	107	0.0	4.7	57.9	35.5	1.9	2.7	0.60
A-L_2-A	108	0.0	6.5	58.3	34.3	0.9	2.7	0.60
S-L_2-S	108	0.0	5.6	60.2	33.3	0.9	2.7	0.58
P-L_3-P	106	17.9	60.4	21.7	0.0	0.0	4.0	0.63
P_1-L_3-P_1	108	12.0	49.1	33.3	5.6	0.0	3.7	0.76
N-L_3-N	107	1.9	27.1	56.1	15.0	0.0	3.2	0.69
A-L_3-A	108	6.5	35.2	50.0	8.3	0.0	3.4	0.73
S-L_3-S	107	2.8	47.7	42.1	7.5	0.0	3.5	0.67
P-W-P	108	0.0	13.9	63.0	22.2	0.9	2.9	0.62
P_1-W-P_1	108	0.0	8.3	59.3	31.5	0.9	2.8	0.61
N-W-N	107	0.0	1.9	52.3	44.9	0.9	2.6	0.55
A-W-A	106	0.0	2.8	57.5	37.7	1.9	2.6	0.58
S-W-S	108	0.0	6.5	56.5	36.1	0.9	2.7	0.60
P-D-P	108	37.0	51.9	10.2	0.9	0.0	4.3	0.67
P_1-D-P_1	106	18.9	43.4	33.0	4.7	0.0	3.8	0.81
N-D-N	108	0.9	37.0	50.0	12.0	0.0	3.3	0.68
A-D-A	107	13.1	40.2	38.3	8.4	0.0	3.6	0.82
S-D-S	108	17.6	50.0	31.5	0.9	0.9	3.8	0.71
Noise 10 dBrnC	216	45.8	44.4	9.7	0.0	0.0	4.4	0.65
Noise 20 dBrnC	215	8.4	52.1	36.7	2.8	0.0	3.7	0.67
Noise 30 dBrnC	214	1.4	13.6	52.8	30.8	1.4	2.8	0.73
Noise 40 dBrnC	216	0.0	0.0	24.5	58.8	16.7	2.1	0.64
Q—5 dB	215	0.0	0.0	0.0	16.7	83.3	1.2	0.37
Q—10 dB	216	0.0	0.0	9.7	54.2	36.1	1.7	0.62
Q—15 dB	215	0.9	9.8	38.6	46.0	4.7	2.6	0.77
Q—20 dB	213	12.7	36.6	41.8	8.9	0.0	3.5	0.83
Q—25 dB	216	50.0	38.4	10.6	0.9	0.0	4.4	0.71

Notation:
See Table XI.

REFERENCES

1. J. E. Abate, L. H. Bradenburg, J. C. Lawson, and W. L. Ross, "The Switched Digital Network Plan," B.S.T.J., *56*, No. 7 (September 1977), pp. 1297–1320.
2. J. R. Cavanaugh, R. W. Hatch, and J. L. Sullivan, "Models for the Subjective Effects of Loss, Noise, and Talker Echo on Telephone Connections," B.S.T.J., *55*, No. 9 (November 1976), pp. 1319–1371.
3. B. Smith, "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 653–709.
4. C. L. Dammann, L. D. McDaniel, and C. L. Maddox, "Multiplexing and Coding," B.S.T.J., *51*, No. 8 (October 1972), pp. 1675–1700.
5. D. L. Duttweiler and D. G. Messerschmitt, "Nearly Instantaneous Companding for Nonuniformly Quantized PCM," IEEE Trans. on Comm., *COM-24* (August 1976), pp. 864–873.
6. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
7. R. J. Canniff, "Signal Processing in SLC-40, A 40 Channel Rural Subscriber Carrier," IEEE ICC 1975, June 16–18, Conference Record, Vol. 3, pp. 40-7 to 40-11.
8. W. R. Daumer, "A Digital Codec Simulation Facility," IEEE Trans. on Comm., *COM-26* (May 1978), pp. 665–669.
9. Annex 2 (Status of Noise Reference Unit Instrumentation) of Question 18/XII (Transmission Performance of Pulse-Code Modulation Systems), C.C.I.T.T. Green Book, Vol. V, published by The International Telecommunications Union, 1973.
10. K. L. McAdoo, "Speech Volumes on Bell System Message Circuits," B.S.T.J., *42*, No. 5 (September 1963), pp. 1999–2012.
11. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–1348.
12. R. A. Friedenson, R. W. Daniels, R. J. Dow, and P. H. McDonald, "RC Active Filters for the D3 Channel Bank," B.S.T.J., *54*, No. 3 (March 1975), pp. 507–530.
13. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3386.
14. J. E. Kessler, "The Transmission Performance of Bell System Toll Connecting Trunks," B.S.T.J., *50*, No. 8 (October 1971), pp. 2741–2776.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

A Loss Model for Parabolic-Profile Fiber Splices

By C. M. MILLER and S. C. METTLER

(Manuscript received April 26, 1978)

In the past, measurement results of splice loss of optical fibers have corresponded poorly to existing theory, which assumes a uniform power distribution across the cone of radiation defined by the local numerical aperture. In this paper, a model is developed in which a Gaussian power distribution across the local numerical aperture is assumed. Transmission through a splice at each point on the transmitting core is found to depend on the ratio of receiving to transmitting numerical aperture at that point. Numerical integration of these "point" transmission functions over core areas of interest yields both splice loss and the additional loss that occurs in a long fiber following the splice. This model cannot be theoretically rigorous, since it is inconsistent with boundary conditions required by the laws of light propagation. However, it has been found to predict splice loss under varying conditions with much greater accuracy than existing theory. The model has the further virtue of being able to calculate how variations in many intrinsic and extrinsic splice parameters combine to produce an overall splice loss.

I. INTRODUCTION

Calculations for the loss in an optical fiber splice, as a function of offset, tilt, diameter, or numerical aperture mismatch, have been reported by several authors.[1–3] These calculations all assumed a uniform power distribution across the cone of radiation defined by the local numerical aperture (NA). This is consistent with assuming equal mode excitation, equal mode attenuation, and no mode coupling.[4] An assumption concerning the power distribution is necessary to characterize all combinations of both intrinsic and extrinsic splice imperfections. While these assumptions allow easy calculations, correspondence with measurement data has been unacceptable.[3] Gloge[5] reported results based on a diffusion process from the uniform power distribution to the steady-state distribution calculated by Marcuse.[6] Correspondence with

measurement data is much improved for the case of small offsets; however, approximations used in the theory make it difficult to characterize other splice imperfections or splice loss for large offsets.

This paper presents a phenomenological model with an assumed power distribution selected solely due to resulting correspondence of calculated effects of splice imperfections with measurement data. The model is not intended for uses other than splice loss characterization, and since the model does not obey the laws of ray optics exactly, caution should be exercised in any other uses of it.

II. DETAILS OF THE MODEL

The power distribution across the cone of radiation defined by the numerical aperture at any point on a fiber core is assumed to be Gaussian in form. Figure 1 is a sketch of the assumed radially symmetric distribution across the cone of radiation from a given point on the fiber core. The model consists of solving for the transmission ratio in terms of transmitting and receiving numerical apertures, then integrating this ratio over the core areas of interest.

Consider the steady-state power distribution across the cone defined by the local numerical aperture to be Gaussian in form, normalized to a value of 1 at $r = 0$.

$$p(r) = e^{-r^2/2\sigma^2}, \tag{1}$$

where σ is proportional to the width of the Gaussian.

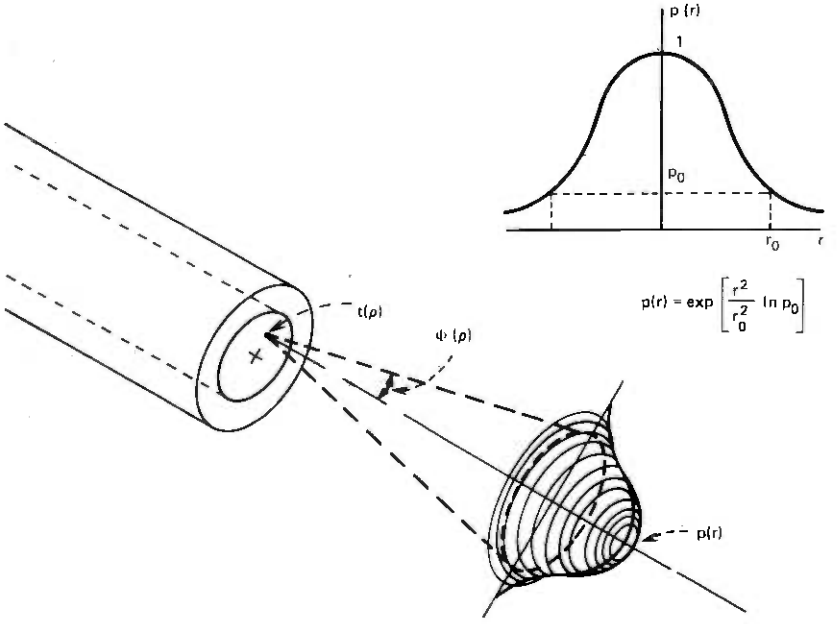


Fig. 1—Gaussian power distribution.

Numerical aperture has been defined by various investigators to be the $1/e^2$, 0.1 or 0.01 power level. This power level, p_0, determines the width of the Gaussian and will be left as a parameter. Therefore, at r_0,

$$p_0 = e^{-r_0^2/2\sigma^2} \tag{2}$$

or

$$p(r) = e^{(r^2/r_0^2)\ln p_0}. \tag{3}$$

III. NA MISMATCH

The application of this simple model can best be illustrated by considering the change in the power distribution as it propagates through a splice. The usual class of circularly symmetric index of refraction profiles[4] is used throughout this paper. In this example, transmitting and receiving fiber profiles are identical except for the value of n_{01}. Therefore, the NA as a function of transmitting fiber core radius (ρ) is

$$\mathrm{NA}_1(\rho) \approx n_{01}\sqrt{2\Delta_1}\left[1 - \left(\frac{\rho}{R}\right)^{\alpha}\right]^{1/2}, \tag{4}$$

where $\Delta_1 \approx (n_{01} - n_c)/n_{01}$ is small
n_{01} = refractive index at center of core
n_c = refractive index of cladding
α = 2 for nearly parabolic profile fibers
R = fiber core radius.

The angle $\phi_1(\rho)$ (Fig. 1) is determined by $\mathrm{NA}_1(\rho)$ to be

$$\phi_1(\rho) = \sin^{-1}\left[\mathrm{NA}_1(\rho)/n_1(\rho)\right], \tag{5}$$

where

$$n_1(\rho) = n_{01}\left[1 - 2\Delta_1\left(\frac{\rho}{R}\right)^{\alpha}\right]^{1/2}. \tag{5a}$$

The power distribution at a given value of ρ is assumed to be

$$p_1(r) = e^{(r^2/r_1^2)\ln p_0}, \tag{6}$$

where r_1 is proportional to tan ϕ_1 (Fig. 1).

For the receiving fiber, $\mathrm{NA}_2(\rho)$ will first be assumed to be less than the transmitting fiber $\mathrm{NA}_1(\rho)$, so that

$$p_2(r) = e^{(r^2/r_2^2)\ln p_0} \tag{7}$$

is the relative power distribution that corresponds to the same point on the receiving fiber. Figure 2 shows these two functions for $r_2 < r_1$ and for Gaussian distributions which are truncated at r_2 and r_1, respectively. Using these truncated distributions, we assume the power in region I is lost immediately at the splice, since power in this region lies totally outside the receiving NA. Splice transmission through this point is then

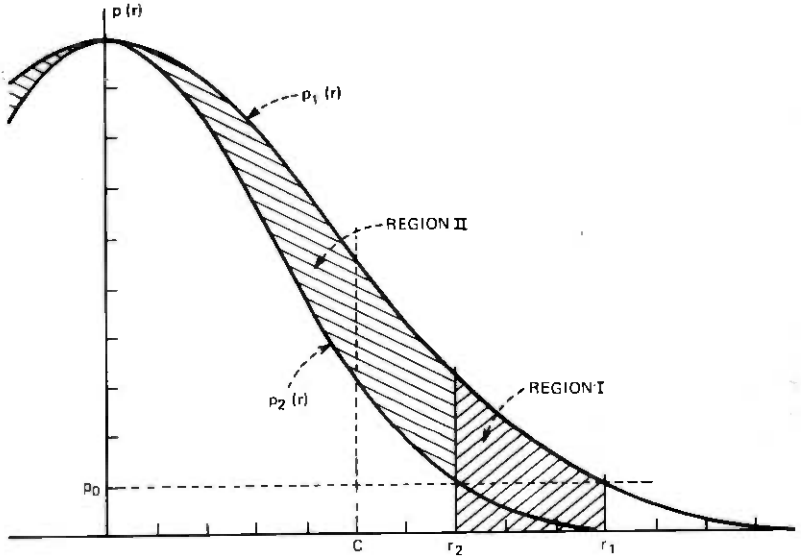


Fig. 2—Point transmission distributions.

$$t(\rho) = \frac{\int_0^{r_2} p_1(r)\, r dr}{\int_0^{r_1} p_1(r)\, r dr}, \tag{8}$$

$$t(\rho) = \frac{1}{p_0 - 1}\left[e^{(r_2/r_1)^2 \ln p_0} - 1\right] \tag{9}$$

for $r_2 < r_1$.

Since $r_1 = k \tan \phi_1$ and $r_2 = k \tan \phi_2$ where k is a constant of proportionality, then using (4) and (5),

$$\frac{r_2}{r_1} \approx \frac{\tan(\sin^{-1}\sqrt{2\Delta_2})}{\tan(\sin^{-1}\sqrt{2\Delta_1})} \approx \frac{\mathrm{NA}_2}{\mathrm{NA}_1} \tag{10}$$

for some given ρ.

For this particular example, $\mathrm{NA}_2/\mathrm{NA}_1$ is constant for every point (ρ) on the fiber core; therefore, eq. (9) gives the total transmission just after the splice for a truncated Gaussian distribution.

For the case of nontruncated Gaussian distributions, (8) and (9) become

$$t(\rho) = \frac{\int_0^{r_2} p_1(r)\, r dr + \int_{r_2}^{\infty} p_2(r)\, r dr}{\int_0^{\infty} p_1(r)\, r dr}, \tag{11}$$

$$t(\rho) = 1 + \left(\frac{r_2}{r_1}\right)^2 p_0 - e^{(r_2/r_1)^2 \ln p_0}, \tag{12}$$

for $r_2 < r_1$. For small values of p_0, (12) and (9) yield similar results. The full Gaussian is used to simplify later computations.

Since this example contains radially symmetric distributions, there is no dependence on θ. Equation (12) gives the transmission through the splice at a point for the steady-state power distribution assumed in (3). Again, for this example, $(r_2/r_1)^2$ is constant across the fiber core; therefore, (12) gives the total transmission just after the splice for a numerical aperture mismatch.

IV. COMPARISON WITH NA MISMATCH DATA

Equation (12) can be compared to measurement data for a splice with an NA mismatch ($r_2 < r_1$, receiving fiber NA < transmitting fiber NA). This transmission coefficient represents an immediate loss at the splice. Figure 3 contains measurement data along with calculations for both a uniform power distribution and a Gaussian distribution. The parameter, p_0, is set to $1/e^2$ and 0.1. As shown in Fig. 3, sensitivity to the value selected for p_0 is significant only for large NA mismatches.

Figure 3 contains data obtained using a HeNe and a GaAlAs laser source with a long input fiber (>500 m). Fibers selected for these measurements were well matched in O.D., core diameter, and α, but contained mismatches in NA. There were, however, slight differences in O.D., core diameter, and α on the order of a few percent, so that measured loss would be expected to differ somewhat from calculated values.

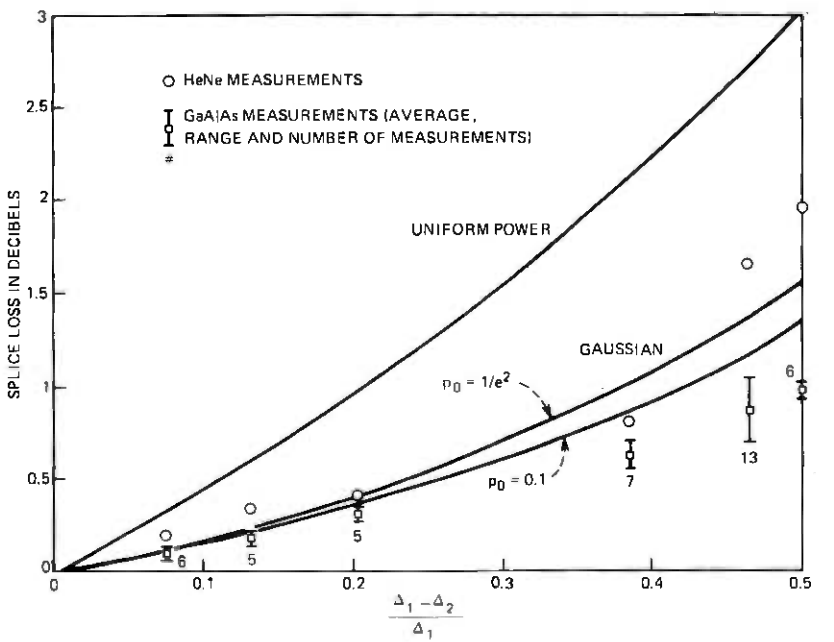


Fig. 3—Loss due to numerical aperture mismatch.

Correspondence with measured data using either source is good compared to the uniform power distribution model which predicts too much loss.

Before this model can be applied in general, the case of transmission from a smaller NA transmitting fiber to a larger NA receiving fiber must be considered. If, in Fig. 2, r_2 represents the transmitting fiber and r_1 the receiving fiber, then all the power contained in the cone of radiation defined by the transmitting NA is within the receiving NA. A unity transmission coefficient is assumed for this case.

V. NEAR-FIELD POWER DISTRIBUTION

The near-field power distribution as a function of core radius, ρ, can be calculated for the Gaussian model after a suitable weighting function is applied. Since a Gaussian distribution implies higher loss for higher order modes of propagation (the tails of the distribution), a weighting function is needed to reduce the amplitude of the Gaussian as a function of radius. This is required since only higher order modes of propagation are significant near the core-cladding interface. The weighting function assumed is the power distribution for the uniform power model. From (4),

$$P(\rho) \approx P_0 \left[1 - \left(\frac{\rho}{R}\right)^2\right] \tag{13}$$

for the case of uniform power across the cone of radiation defined by the numerical aperture where P_0 is proportional to Δ, n_0 and input power.[3] With this function used as the amplitude at $r = 0$, then

$$P(\rho) \approx P_0 \left(1 - \frac{\rho^2}{R^2}\right) \int_0^{2\pi} \int_0^{\infty} e^{(r/r_0)^2 \ln p_0} r \, dr \, d\theta. \tag{14}$$

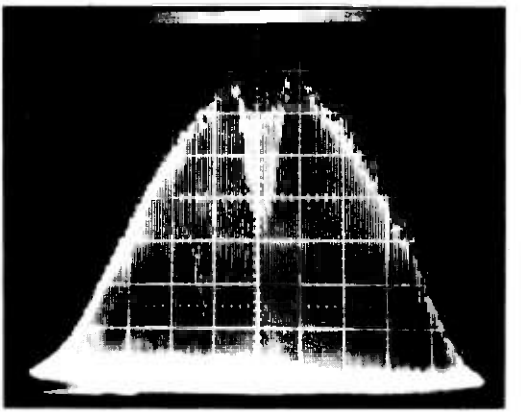
Using (4) to obtain r_0,

$$P(\rho) \approx P_0(1 - \rho^2)^2 \tag{15}$$

where all constant terms have been combined in P_0 and $R = 1$.

VI. COMPARISON WITH NEAR-FIELD POWER MEASUREMENTS

Measurements of near-field power were made with a GaAlAs laser source after propagation through a long (>1 km) fiber wrapped under tension to simulate the effects of some microbending loss (~1 dB/km). These measurements were made using a 100X objective and a TV vidicon camera. Figure 4 is a photograph of the camera output for a typical Western Electric fiber. An approximation to the curve was obtained by smoothing over the index dip and averaging power measurements for each side of the distribution. Calculations using (13) for the uniform distribution and (15) are plotted for comparison. The calculated



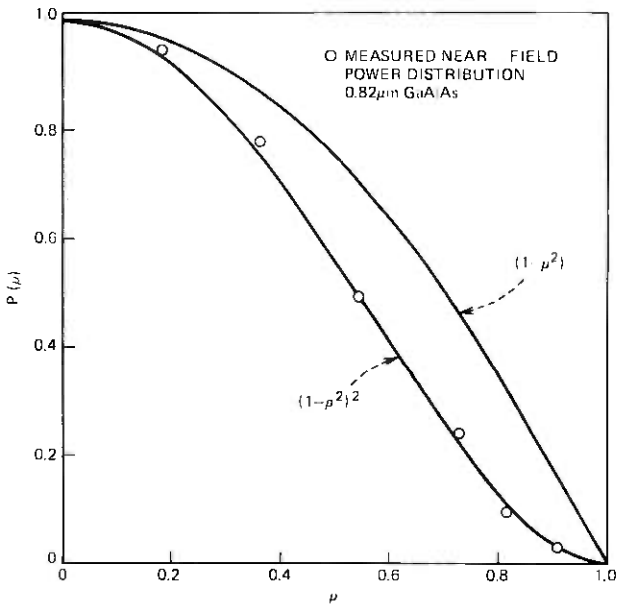


Fig. 4—Near-field power distribution.

steady-state power distribution using the Gaussian model with the $(1 - \rho^2)$ weighting function is in excellent agreement with the measured distribution.

VI. ADDITIONAL LOSS IN THE FIBER AFTER A SPLICE

A parabolic index fiber splice is known to cause additional loss in the fiber after the splice.[7] This additional loss depends on the differential mode attenuation and mode coupling characteristics of the receiving fiber and, for the fibers used in these experiments, is approximately equal to the loss at the splice for small offsets.[7] Therefore, any realistic description of an optical fiber splice must include this effect.

Referring to Fig. 2, the power in region II is seen to be within the numerical aperture of the receiving fiber; however, this power is improperly distributed. The sharp decrease in power at the edge of the receiving numerical aperture, r_2, is physically impossible; however, this effect is significantly reduced compared to the uniform power model. As this distribution propagates down the fiber length (l), we assume that a new Gaussian steady-state distribution will be generated, as shown in Fig. 5. Some portion, c, of the excess energy contained in region II of Fig. 2 will be lost in the process of reestablishing steady-state conditions. If energy couples to adjacent modes with equal probability and if all modes are equally excited and attenuated, then 1/2 of the power in region II would be lost during redistribution. Since the excess energy in region II is skewed toward higher order modes which may be lossier, c would be expected to be greater than 1/2. The loss mechanism is probably transfer of some of this excess energy to higher order propagating or leaky modes.[8]

Referring to Fig. 2, the equivalent transmission loss due to power lost from region II is

$$\Delta t(\rho) = \frac{c\left[\int_0^{r_2} p_1(r)\, r dr - \int_0^{r_2} p_2(r)\, r dr\right]}{\int_0^{\infty} p_1(r)\, r dr}. \tag{16}$$

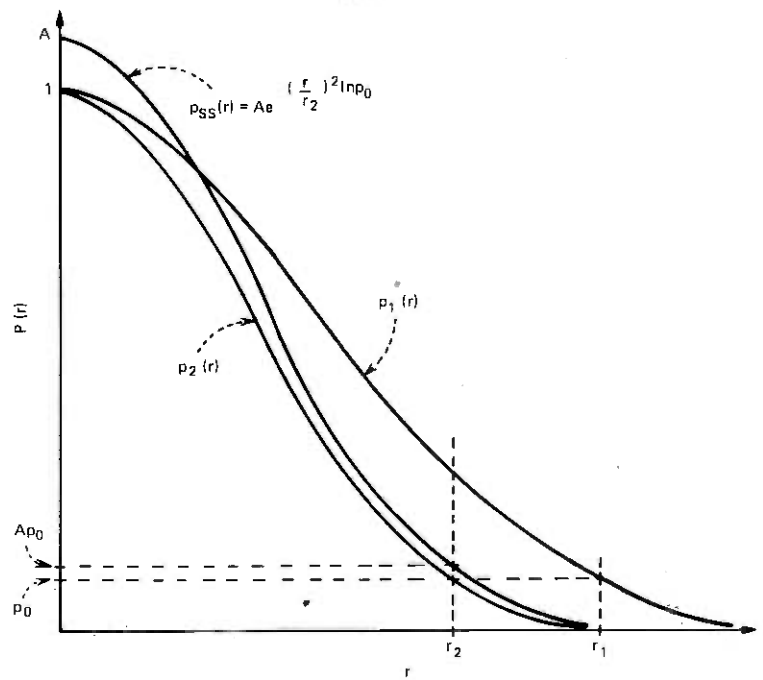


Fig. 5—Steady-state power distribution.

Evaluating this integral yields

$$\Delta t(\rho) = -ce^{(r_2/r_1)^2 \ln p_0} + c + c\left(\frac{r_2}{r_1}\right)^2 \cdot (p_0 - 1). \tag{17}$$

To approximate c, we use the centroid for the solid of revolution for region II in Fig. 2. This places one-half the excess energy in the region $0 < r < c$ and the remaining energy between $c < r < r_2$. This radius is taken as the value for c. From Fig. 2,

$$\int_0^c [p_1(r)\, rdr - p_2(r)\, rdr] = \int_c^{r_2} [p_1(r)\, rdr - p_2(r)\, rdr]. \tag{18}$$

Solving the integrals yields the following nonlinear equation:

$$e^{(c/r_1)^2 \ln p_0} - \left(\frac{r_2}{r_1}\right)^2 e^{(c/r_2)^2 \ln p_0} = \frac{1}{2}\left[1 + e^{(r_2/r_1)^2 \ln p_0} - \left(\frac{r_2}{r_1}\right)^2 (p_0 + 1)\right]. \tag{19}$$

This equation was solved for c/r_2 with $0 < r_2/r_1 < 1$ for $p_0 = 0.1$ and $p_0 = 1/e^2$, and found to be accurately approximated by a quadratic equation. For $p_0 = 0.1$,

$$c \approx 0.7994 - 0.08796 \left(\frac{r_2}{r_1}\right)^2 + 0.00846 \left(\frac{r_2}{r_1}\right)^4 \tag{20}$$

and for $p_0 = 1/e^2$,

$$c \approx 0.8041 - 0.0724 \left(\frac{r_2}{r_1}\right)^2 + 0.00636 \left(\frac{r_2}{r_1}\right)^4. \tag{21}$$

The total effect of the splice is then

$$t_{\text{tot}}(\rho) = t(\rho) - \Delta t(\rho). \tag{22}$$

$$t_{\text{tot}}(\rho) = (1 - c)\left[1 + p_0 \left(\frac{r_2}{r_1}\right)^2 - e^{(r_2/r_1)^2 \ln p_0}\right] + c\left(\frac{r_2}{r_1}\right)^2, \tag{23}$$

where $t(\rho)$ is the immediate point transmission coefficient at the splice given by (12) and $\Delta t(\rho)$ is the reduction in transmission due to power lost in reestablishing a steady-state Gaussian distribution in a long fiber after the splice given by (17). Again, we call attention to the fact that the value of c probably depends on the differential mode attenuation and mode coupling characteristics of the receiving fiber.

VIII. DIAMETER MISMATCH

The necessary parts of the model have been developed so that the effect of diameter mismatch (Fig. 6) can now be considered. Let the transmitting NA function equal

$$\text{NA}_1(\rho) = n_0 \sqrt{2\Delta}\left[1 - \left(\frac{\rho}{R_1}\right)^2\right]^{1/2} \tag{24}$$

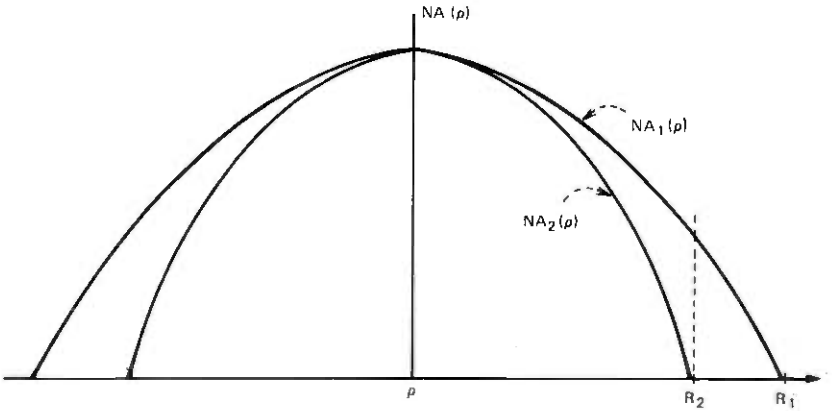


Fig. 6—Core index of refraction profiles for the case of diameter mismatch.

and the receiving NA function equal

$$\mathrm{NA}_2(\rho) = n_0 \sqrt{2\Delta} \left[1 - \left(\frac{\rho}{R_2}\right)^2\right]^{1/2}, \tag{25}$$

where R_1 and R_2 are the transmitting and receiving core radii. For $K = R_2/R_1$ and R_2 normalized to unity,

$$\left(\frac{r_2}{r_1}\right)^2 \simeq \frac{1-\rho^2}{1-K^2\rho^2}. \tag{26}$$

This ratio is not constant with ρ (except for $K = 1$), so that an integration over the receiving core is necessary. Distributions remain radially symmetric; therefore, no angle dependence is present.

$$T_{\mathrm{tot}} = \frac{\int_0^1 t(\rho)\,[1-K^2\rho^2]^2\,\rho d\rho}{\int_0^{1/K} [1-K^2\rho^2]^2\,\rho d\rho}, \tag{27}$$

where $t(\rho)$ equals eq. (12) for the effect at the splice, or by $t_{\mathrm{tot}}(\rho)$ [eq. (23)], to include the additional loss in the fiber after the splice. This integral must be solved by numerical techniques.

IX. COMPARISON WITH DIAMETER MISMATCH DATA

Figure 7 compares the calculations of short-length-diameter mismatch effects at the splice with measured results for various values of K. Fibers were drawn from the same preform to insure that α and NA mismatch were minimized. A HeNe laser source was used with approximately 1 m of fiber after the splice.

The Gaussian power distribution model shows good agreement with measurement data. Long-length-diameter mismatch data are not presently available.

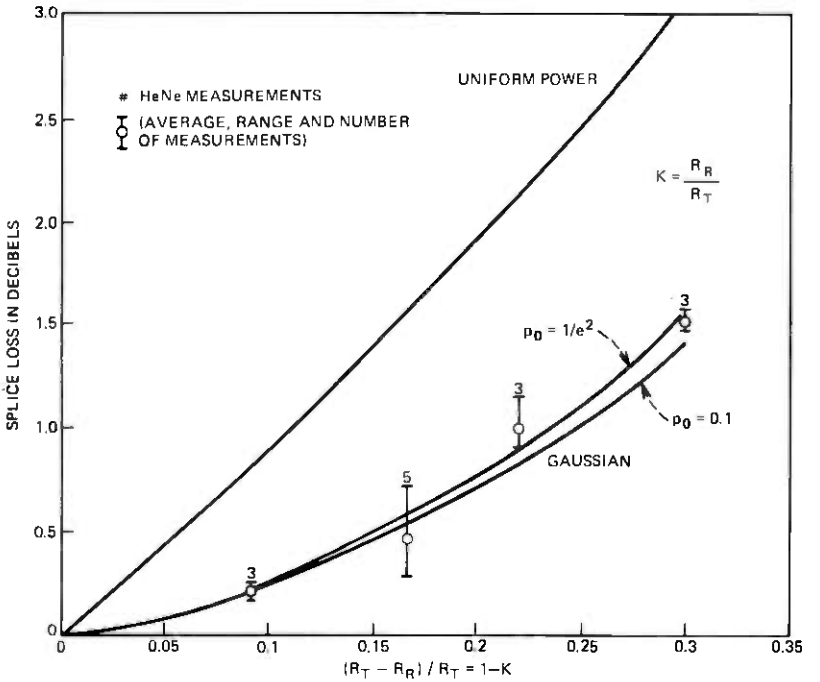


Fig. 7—Loss due to core diameter mismatch.

X. TRANSVERSE OFFSET

Perhaps the most important extrinsic splice parameter is transverse offset (axial displacement). The Gaussian model can be applied to the case of transverse offset; however, some computational difficulties are encountered. Referring to Fig. 8, the area of overlap is divided into region I, where the receiving NA is greater than the transmitting NA, and region II, where the receiving NA is less than the transmitting NA. In region I, $t(\rho)$ is unity, therefore with $R = 1$,

$$T_{\mathrm{I}} = \frac{\int_0^{\cos^{-1}d/2} \int_{d/2\cos\phi}^{1} (1-\rho^2)^2\, \rho d\rho d\phi}{\int_0^{\pi/2} \int_0^{1} (1-\rho^2)^2\, \rho d\rho d\phi}. \tag{28}$$

In region II,

$$T_{\mathrm{II}} = \frac{\int_0^{\cos^{-1}d/2} \int_{d/2\cos\phi}^{1} t(q)[1-q^2]^2\, \rho d\rho d\phi}{\int_0^{\pi/2} \int_0^{1} (1-q^2)^2\, \rho d\rho d\phi}, \tag{29}$$

where $q = \rho^2 - 2\rho d \cos\phi + d^2$.

Referring to Fig. 8, we note that the line separating regions I and II is a locus of equal NA; therefore, $t(\rho)$ is unity on this line. On the curved boundary of region II, $t(\rho)$ is zero; therefore, the derivative of $t(\rho)$ is infinite at the intersection of these lines. This point must be omitted and, since $t(\rho)$ is steep in the vicinity of this point, a large number of increments are required to evaluate eqs. (28) and (29) numerically.

$$T_{tot} = T_I + T_{II}. \tag{30}$$

If the loss at the splice is desired, (12) is used for $t(\rho)$, and if the total loss including losses required to reestablish steady state is being calculated, then (23) is used.

XI. COMPARISON WITH TRANSVERSE OFFSET DATA

Figure 9 is a comparison of transverse offset data and calculations (*i*) reported by Gloge,[5] (*ii*) using the Gaussian distribution assumption, and (*iii*) using the uniform power assumption. Measurements were made with a GaAlAs laser source and an unbroken long fiber (2 km). After a reference level was established, the fiber was broken approximately in the center and the ends spliced to obtain the reference level again. Offset was introduced and the loss measured 1 km and 2 m after the splice as a function of transverse offset. Figure 9 shows the results for small offsets. Agreement among the Gloge model, the Gaussian model, and measurement data is good for loss at the splice. The Gaussian model also agrees well for the loss occurring in the fiber after the splice.

Figure 10 shows results for large offsets. As compared with actual measurements, the Gaussian model understates the loss through a long

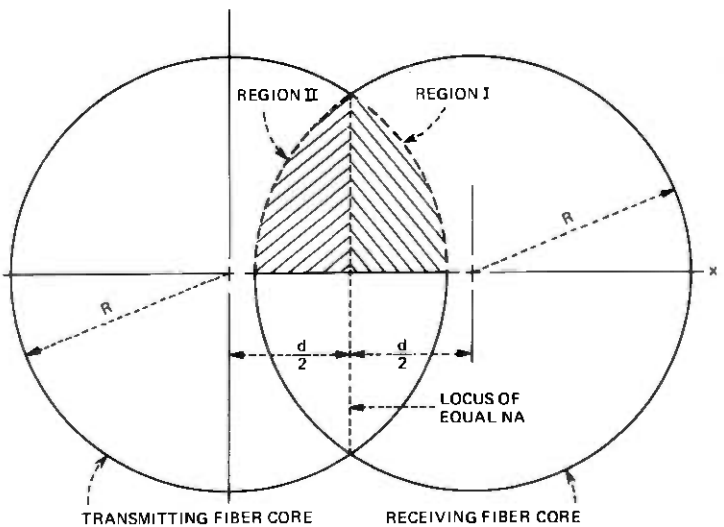


Fig. 8—Regions of overlap for offset fiber cores.

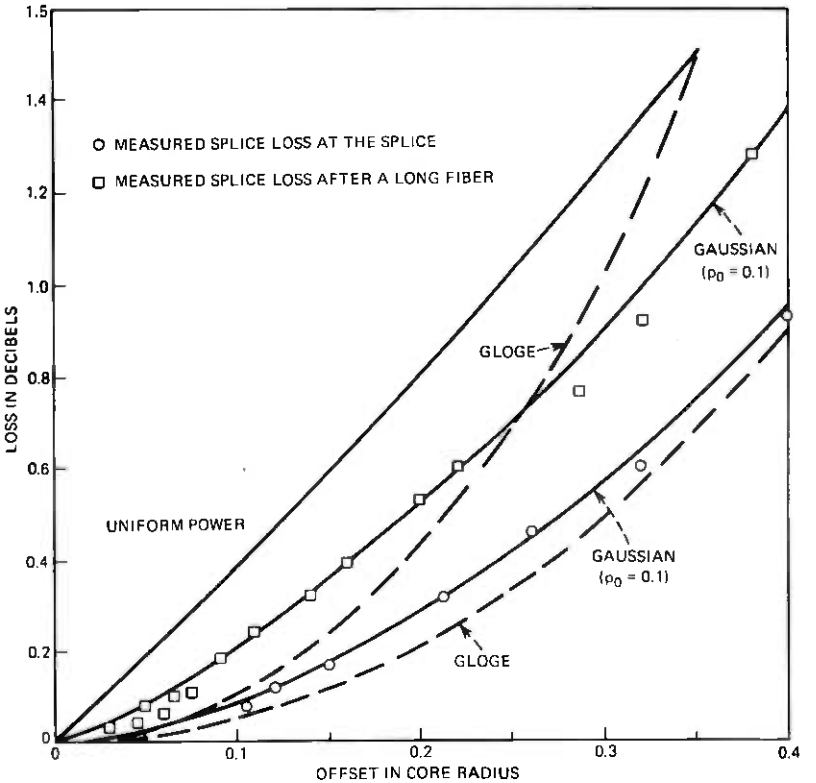


Fig. 9—Loss due to transverse offset.

fiber after the splice for large offsets. The Gaussian model cannot, at present, account for this effect for large offsets.

XII. CONCLUSIONS

A point transmission model has been used to calculate splice loss due to NA mismatch, diameter mismatch, and transverse offset in parabolic-profile fiber splices. Near-field power distributions have also been calculated. Correspondence with measurement data is much improved as compared to calculations using the uniform power distribution model.

The authors again emphasize that the Gaussian power assumption does not obey the laws of ray optics exactly and that this assumption was made primarily due to good correspondence of resulting calculations with measurement data and due to computational considerations. Future work will include the effects of α mismatch and the effects of long fibers after splices with combinations of intrinsic and extrinsic parameter mismatches.

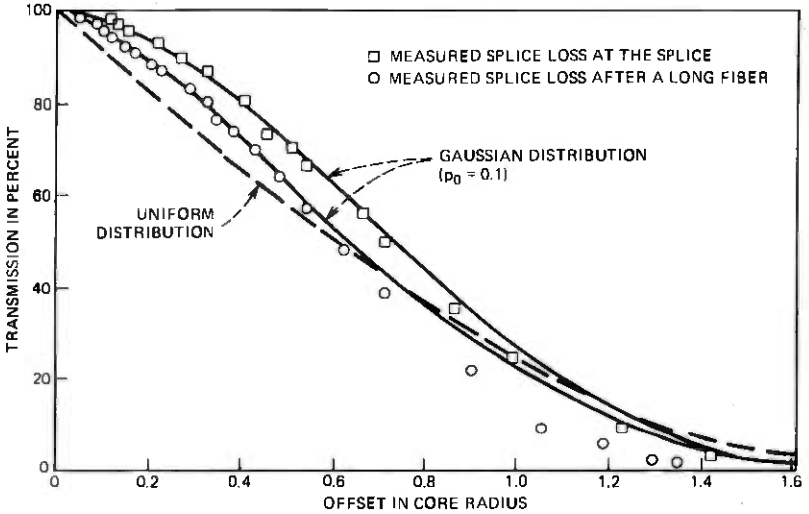


Fig. 10—Transmission vs offset for large offsets.

XIII. ACKNOWLEDGMENT

The authors appreciate the use of diameter mismatch data furnished by M. R. Gotthardt.

REFERENCES

1. F. L. Thiel, "Utilizing Optical Fibers in Communications Systems," Int. Conf. Comm., Conference Record, *II*, Session 32, June 16–18, 1975.
2. Haruhiko Tsuchiya et al., "Double Eccentric Connectors for Optical Fibers," Appl. Opt., *16*, No. 5 (May 1977), pp. 1323–1331.
3. C. M. Miller, "Transmission vs. Transverse Offset for Parabolic-Profile Fiber Splices with Unequal Core Diameters," B.S.T.J., *55*, No. 7 (September 1976), pp. 917–928.
4. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9, (November 1973), pp. 1563–1578.
5. D. Gloge, "Offset and Tilt Loss in Optical Fiber Splices," B.S.T.J., *55*, No. 7 (September 1976), pp. 905–916.
6. D. Marcuse, "Loss and Impulse Response of a Parabolic Index Fiber with Random Bends," B.S.T.J., *52*, No. 8 (October 1973), pp. 1423–1437.
7. C. M. Miller, "Realistic Splice Losses for Parabolic Index Fibers," presented at IOOC 1977, Tokyo, Japan, July 18–20, 1977.
8. M. J. Adams et al., "Splicing Tolerances in Graded-Index Fibers," Appl. Phys. Lett., *28*, No. 9 (1 May 1976), pp. 524–526.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Optimum Reception of Digital Data Signals in the Presence of Timing-Phase Hits

By D. D. FALCONER and R. D. GITLIN

(Manuscript received March 29, 1978)

Protection switching of digital radio channels results in a timing-phase discontinuity and occasional long error bursts in the demodulated data signal. Using an idealized mathematical model, we derive maximum likelihood receivers which rapidly track such delay hits, whether or not a timing-pilot tone is used. When the receiver is at a different physical location from the switch, the tracking algorithm must also sense the occurrence of a switch. A dual-mode, data-directed structure is revealed as being optimum; a narrowband tracking loop is used for steady-state operation, while a wideband tracking loop provides rapid recovery from the timing transient. An error-sensing nonlinearity, which incorporates hysteresis, inhibits erroneous noise-induced mode transitions. Oversampling of the demodulated data signal rapidly establishes a coarsely quantized, optimum sampling phase and permits the dual-mode tracking loop to operate in a data-directed manner. Data-directed operation permits greater loop bandwidths, since the data energy is not perceived as noise. Simulation of a digital data transmission system employing a dual-mode, data-directed, and coarse-quantized timing loop has demonstrated dramatic reduction in the length of error bursts following a protection switch. For example, at the data rate of 1.544 Mb/s, a conventional phase-locked loop with a 100-Hz bandwidth, when displaced a half-symbol interval by a delay hit, would typically sustain an error burst 15,000 bits in duration. When such a delay hit stresses the dual-mode timing loop, simulation has indicated error bursts on the order of 15 bits in duration.

I. INTRODUCTION

Some channels used for data transmission exhibit occasional abrupt changes in their absolute delay. For example, during severe fading, line-of-sight microwave facilities commonly switch signals from their regularly assigned channel to a protection channel which, owing to different filtering, cable lengths, etc., may impart a different delay. Such a switch, unknown to the receiver, changes the best phase with which a synchronous receiver should sample the incoming signal during each symbol interval. Until this new optimal timing phase is acquired by the receiver, data errors may proliferate if the delay change is a significant fraction of the symbol interval. The length of the succession of errors will be essentially inversely proportional to the bandwidth of the timing-recovery loop or filter.

The object of this investigation is to determine and analyze signal processing structures which rapidly respond to a sudden change in timing phase (a delay hit) while also providing accurate steady-state timing information when the protection channel is not required. Based upon an idealized channel model, we determine the maximum likelihood (optimum) receiver, and practically motivated approximations are made to provide realizable signal processors. The proposed receivers mediate the inherent conflict between using a narrowband timing recovery loop for steady-state operation, so that accurate and stable timing can be derived from the noisy received signal, and using a wideband loop to follow a timing-phase transient. As might be expected, the derived tracking loop is of the dual-mode variety; i.e., it automatically senses the state of the system (i.e., transient or steady state) and adjusts its structure accordingly. The exact form of the loop depends on the detailed manner in which the disturbances are modeled; yet it is demonstrated that the essential features of the signal processors are quite robust and have significant intuitive appeal.

In this study, our development is for an arbitrary protection-switched data communication system; however, specific simulation results and special emphasis are given to a 4-input level, Class-IV, partial response signaling format, such as used in the DUV system.[1]

In Section II, we describe the system model—principally the statistical mechanism for generating a "delay hit." The optimum receiver is described in Section III, and various suboptimum realizable structures are developed in Section IV. Digital implementation of these techniques in the partial response system is reported, via simulation, in Sections V and VI.

II. SYSTEM MODEL FOR TIMING-RECOVERY PROBLEM

In this section, we propose and develop a mathematical model for the data transmission system under consideration. Any attempt to exactly model the end-to-end data channel will be exceedingly tedious and probably fruitless, since the FM demodulator and phase-hit mechanism of most microwave facilities are highly nonlinear effects. Our approach is to isolate the major manifestations of a timing-phase discontinuity and background noise via a simple model, and then to apply maximum likelihood detection to obtain useful receivers. The validity of this approach is measured by simulation of the receiver over a real channel. We begin by writing the transmitted baseband data signal[2] as

$$s(t) = \sum_n a_n h(t - nT) + \rho_c \sin \frac{\pi}{T} t, \tag{1}$$

where $\{a_n\}$ is a sequence of independent multilevel symbols, $h(t)$ is a band-limited transmitted pulse, $1/T$ is the symbol rate, and ρ_c is the parameter which indicates the power in the pilot tone located at $1/2T$ Hz. The purpose of the pilot tone is to aid in providing the receiver timing phase and frequency. It will be assumed that the end-to-end pulse shaping used in the system is such that the desired sampling instants are $t = nT$. Whenever the pulse $h(t)$ possesses more than the minimum Nyquist bandwidth, it is convenient to rewrite (1) as

$$s(t) = \sum_n [a_n + \rho(-1)^n] h(t - nT), \tag{2}$$

where it is recognized that $\Sigma_n (-1)^n h(t - nT)$ is periodic with period $2T$, and ρ_c is the product of ρ and the energy in the pulse at $1/2T$ Hz. Since $h(t)$ is customarily band-limited to less than $1/T$ Hz, only the fundamental component of the signal $\Sigma_n (-1)^n h(t - nT)$ will be transmitted through the filter $h(t)$, thus the sinusoid may be represented by the alternating (dotting pattern) series. Indeed, in practice, a dotting pattern is frequently used to generate the tone. The transmitted signal may be rewritten as

$$s(t) = \sum_n c_n h(t - nT), \tag{3}$$

where

$$c_n = a_n + \rho(-1)^n. \tag{4}$$

Recall that partial response signals can be generated by either digital filtering of the independent data symbols, $\{a_n\}$, or by the use of special non-Nyquist pulse shapes. The receiver structures derived in the sequel will be discussed for both Nyquist and partial-response shaping. We now turn to the specific idealizations we will make to model the transmission path.

In the absence of any transmission distortion, the received baseband signal, $r(t)$, can be modeled as

$$r(t) = \mathcal{J}[s(t)] + \nu(t), \tag{5}$$

where $\mathcal{J}[s(t)]$ is the time-jittered signal and where the additive noise $\nu(t)$ will be taken as white Gaussian with spectral density N_0. For the purpose of analytical tractability, any instability in the timing phase will be modeled by representing the received signal as

$$r(t) = \sum_n c_n h(t - nT - \Delta_n) + \nu(t), \tag{6}$$

where Δ_n is a random process whose characteristics will be described below. In the above model, which is shown in Fig. 1, the phase of the pilot tone is presumed to be jittered at the discrete instants, $\{nT\}$, in the same manner as the phase of the data signal. This is accurate when the pilot tone is generated via the dotting pattern method, and the timing instabilities arise solely in the *transmitter* clock. Any timing-phase jitter, $\Delta(t)$, that occurs during transmission should properly be modeled by $r(t) = s(t - \Delta(t)) + \nu(t)$. However, this leads to analytical difficulty in characterizing the statistical nature of the random process $\Delta(t)$, as well as having to contend with jitter-induced amplitude modulation of the received signal. With this caveat in mind, we lump all sources of timing-phase jitter into the model given by (6). Of course, the utility of the above model will be measured by the performance of the derived receivers in the real-world environment.

The standard approach to the tracking of a slowly varying timing phase, and thus the timing frequency, is to use a narrowband filter centered about $1/2T$ Hz to extract the transmitted pilot tone. Of course, the bandwidth of this filter must be quite narrow to attenuate the in-band data-plus-noise energy. Extremely small effective bandwidths are

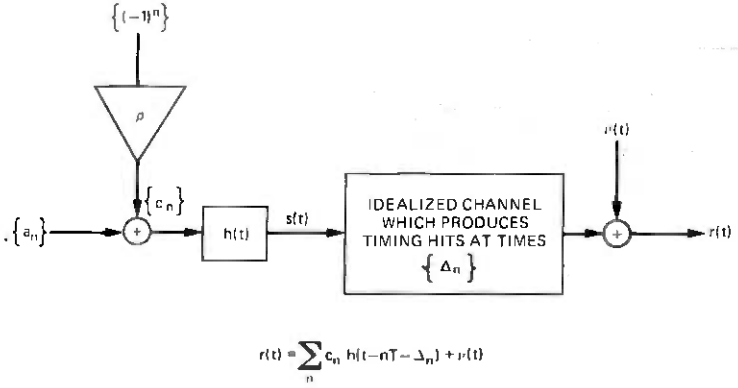


Fig. 1—Idealized model of timing hits in a digital data transmission system.

achieved in practice by following a narrowband zonal filter with a phase-locked loop (PLL) whose voltage-controlled oscillator (VCO) is tuned to $1/2T$ Hz. However, the narrow bandwidth of the PLL will preclude rapid tracking of any sudden change in the timing phase.

We digress momentarily to recall that the conventional envelope-derived timing scheme,[3,4] which does not utilize a pilot tone, will not provide timing information as the bandwidth of the system decreases to $1/2T$ Hz. It should be pointed out, however, that other nonquadratic techniques not requiring a pilot tone will provide a tone at the symbol rate for such minimum bandwidth systems; in particular, Saltzberg[5] has shown that the average of $\operatorname{sgn}[s(t)s(t-T)]$ provides a tone at $1/T$ Hz, and it is apparent that quartic[6] and similar operations will also provide the desired tone.

Returning to the formulation of our system model, we let the dynamic evolution of the timing jitter be given by the difference equation

$$\Delta_{n+1} = \Delta_n + w_n + \alpha_n v_n, \tag{7}$$

where $\{w_n\}$ and $\{v_n\}$ are sequences of mutually and self-independent Gaussian random variables with variances σ^2 and μ^2 respectively, and where $\mu^2 \gg \sigma^2$. The variable α_n is governed by

$$\alpha_n = \begin{cases} 0, & \text{with probability } 1 - p_0 \\ 1, & \text{with probability } p_0 \end{cases}, \tag{8}$$

where $0 \leq p_0 \ll 1$. Note that $\{\Delta_n\}$ is a Markov sequence where the mutually independent sequences $\{w_n\}$ and $\{v_n\}$ model the steady-state and the transient (delay-hit) modes, respectively. The initial value Δ_0 will be assumed to be uniformly distributed on $(0,T)$. Clearly, most of the time there are no delay hits; i.e., $\alpha_n = 0$, and the timing phase wanders about the correct value. Thus, p_0 is the probability that a timing discontinuity (which would follow a protection switch) occurs during a symbol interval. A typical sample path, or realization, of $\{\Delta_n\}$ is shown in Fig. 2, where the relative frequency of delay hits is determined by p_0. This simple two-mode model for the timing phase will be used to derive the optimum and various suboptimum data detectors, where an integral component of these detectors will be the timing-recovery loop. The steady-state jitter, w_n, is incorporated so that the timing loop will continually adjust the receiver's timing phase; note that this mechanism allows the receiver to presume nominal knowledge of the timing frequency, and any inaccuracy or drift in this quantity will be compensated for by the timing-phase tracking system.

With $\{\Delta_n\}$ specified by (7), the joint probability density function (pdf) of the $\{\Delta_n\}$ sequence is given by

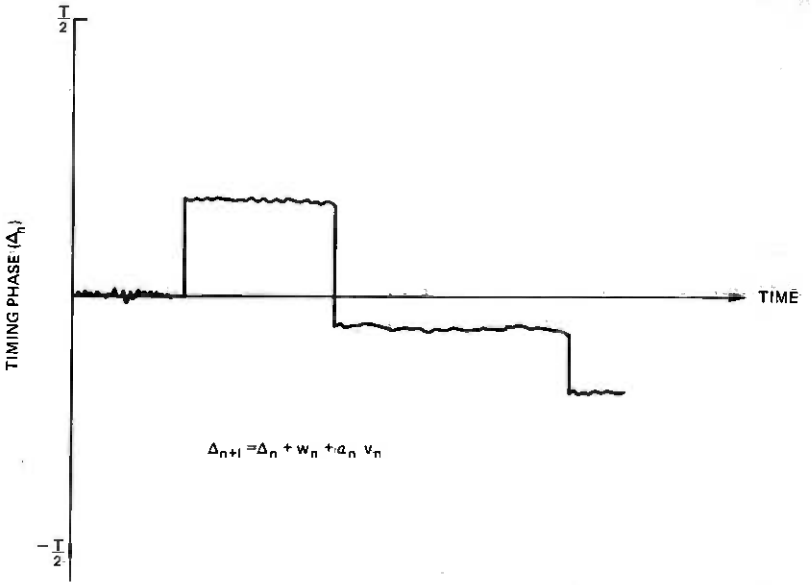


Fig. 2—A typical timing-phase trajectory ($\alpha_n = 1$ signifies that a timing hit has occurred).

$$
\begin{aligned}
p(\Delta) \equiv p(\Delta_0, \Delta_1 \cdots, \Delta_n) &= p(\Delta_0) p(\Delta_1, \Delta_2, \cdots, \Delta_n \mid \Delta_0) \\
&= p(\Delta_0) p(\Delta_1 \mid \Delta_0) p(\Delta_2, \cdots, \Delta_n \mid \Delta_1, \Delta_0) \\
&= p(\Delta_0) p(\Delta_1 \mid \Delta_0) p(\Delta_2 \mid \Delta_0, \Delta_1) p(\Delta_3, \cdots, \Delta_n \mid \Delta_0, \Delta_1, \Delta_2) \\
&= p(\Delta_0) p(\Delta_1 \mid \Delta_0) p(\Delta_2 \mid \Delta_1) p(\Delta_3 \mid \Delta_2) \cdots p(\Delta_n \mid \Delta_{n-1}) \\
&= p(\Delta_0) \prod_{i=1}^{n} p(\Delta_i \mid \Delta_{i-1}), \qquad (9)
\end{aligned}
$$

where $\Delta \equiv (\Delta_0, \Delta_1, \cdots, \Delta_n)$.

Using (7) and (8), the conditional density is given by the mixture

$$
\begin{aligned}
p(\Delta_i \mid \Delta_{i-1}) = {} & \frac{(1 - p_0)}{\sqrt{2\pi}\,\sigma} e^{-(\Delta_i - \Delta_{i-1})^2/2\sigma^2} \\
& + \frac{p_0}{\sqrt{2\pi}\,(\mu^2 + \sigma^2)^{1/2}} e^{-(\Delta_i - \Delta_{i-1})^2/2(\mu^2 + \sigma^2)}, \qquad (10)
\end{aligned}
$$

and thus the joint pdf is given by (9) and (10) where Δ_0 is distributed uniformly over $(0,T)$.

Now that the system model has been specified, we turn to our professed goal of deriving optimum and suboptimum receivers.

III. OPTIMUM RECEPTION OF TIME-JITTERED PAM DATA SIGNALS

It is well known that the optimum (minimum probability of error) data-sequence detector *maximizes* the *a-posteriori* (MAP) probability density of the received signal with respect to the data sequence.* Thus the MAP receiver will supply the end-user with a sequence of decisions which maximize the probability density function $p[\{\hat{a}_m\}|r(t), 0 \le t \le \mathcal{T}]$, where the observation interval is $(0,\mathcal{T})$. By virtue of (4) and the properties of MAP receivers, we may estimate $\hat{a}_m$ via $\hat{a}_m = \hat{c}_m - \rho(-1)^m$, i.e., the estimates of $\{c_m\}$ and $\{a_m\}$ are related as above. Since all the data sequences $\{a_m\}$ are equiprobable, the relevant probability density can be obtained by averaging over the jittered timing phases $\{\Delta_i\}$, i.e., the MAP density is proportional to

$$p[r(t)|\{\hat{a}_m\}, 0 \le t \le \mathcal{T}] = \int p[r(t)|\{\hat{a}_m\}, \{\hat{\Delta}_m\}, 0 \le t \le \mathcal{T}]p[\hat{\Delta}]d\hat{\Delta}, \quad (11)$$

where the conditional density in the integrand is given by the standard formula for the probability density functional of a known signal in white Gaussian noise,

$$p[r(t)|\{\hat{a}_m\}, \{\hat{\Delta}_m\}, 0 \le t \le \mathcal{T}] = k \exp\left\{-\frac{1}{2N_0}\int_0^{\mathcal{T}} [r(t) - \sum_m \hat{c}_m h(t - mT - \hat{\Delta}_m)]^2 dt\right\}. \quad (12)$$

In the above equation, k is a constant independent of both $\{a_m\}$ *and* $\{\Delta_m\}$. The maximization of (11) with respect to $\{\hat{c}_m\}$, or equivalently $\{\hat{a}_m\}$, can be facilitated by writing (11) as

$$p[r(t)|\{\hat{a}_m\}, 0 \le t \le \mathcal{T}] = k \int d\hat{\Delta} \exp\left\{-\frac{1}{2N_0}\left[\int_0^{\mathcal{T}} [r(t) - \sum_m \hat{c}_m h(t - mT - \hat{\Delta}_m)]^2 dt - 2N_0 \ln p(\hat{\Delta})\right]\right\}. \quad (13)$$

It can be shown that, in a high signal-to-noise-ratio environment [i.e., as $N_0 \to 0$], the above integral with respect to $\hat{\Delta}$ can be replaced by the maximum value of the integrand, i.e., as $N_0 \to 0$

$$p[r(t)|\{\hat{a}_m\}, 0 \le t \le \mathcal{T}] \sim \exp\left\{-\frac{1}{2N_0}\left[\int_0^{\mathcal{T}} [r(t) - \sum_m \hat{c}_m h(t - mT - \hat{\Delta}_m^*)]^2 dt - 2N_0 \ln p(\hat{\Delta}^*)\right]\right\}, \quad (14)$$

where $\{\hat{\Delta}_n^*\}$ is the maximizing sequence. Thus, under the asymptotic

* The bit-optimum detector has been shown to be asymptotically approximated (at high signal-to-noise ratio) in performance by the optimum sequence detector (Ref. 7).

condition described above, the optimum receiver computes the *joint* minimum* of

$$\Lambda[r(t)|\{\hat{a}_m\}, \{\hat{\Delta}_m\}, 0 \le t \le \mathcal{T}] \equiv -\ln p[r(t)|\{\hat{a}_m\}, \{\hat{\Delta}_m\}, 0 \le t \le \mathcal{T}] - 2N_0 \ln p(\hat{\Delta}) \quad (15a)$$

$$= \int_0^{\mathcal{T}} \left[r(t) - \sum_m \hat{c}_m h(t - mT - \hat{\Delta}_m) \right]^2 dt - 2N_0 \left[\sum_{m=0}^{\infty} \ln p(\hat{\Delta}_m | \hat{\Delta}_{m-1}) + \ln p(\Delta_0) \right] \quad (15b)$$

$$\equiv \Lambda_1[r(t)|\{\hat{a}_m\}, \{\hat{\Delta}_m\}] + \Lambda_2[\{\hat{\Delta}_m\}], \quad (15c)$$

where $\Lambda_1[\cdot]$ and $\Lambda_2[\cdot]$ are defined in the obvious manner from (15b). For convenience in notation, we drop the $\hat{a}_m$ and $\hat{\Delta}_m$ symbols in favor of a_m and Δ_m whenever there is no possibility of confusion. We also adopt the notational shorthand

$$\ell[\{a_m\}, \{\Delta_m\}] \equiv \Lambda[r(t)|\{a_m\}, \{\Delta_m\}, 0 \le t \le \mathcal{T}], \quad (16a)$$

$$\equiv \ell_1[\{a_m\}, \{\Delta_m\}] + \ell_2[\{\Delta_m\}], \quad (16b)$$

where the ℓ_i corresponds to the appropriate Λ_i $(i = 1, 2)$ in (15c). Thus, our task is to jointly minimize $\ell[\{a_m\}, \{\Delta_m\}]$ with respect to the discrete-valued variables $\{a_m\}$ and the continuous-range variables $\{\Delta_m\}$. Since $\int_0^{\mathcal{T}} r^2(t)dt$ is independent of $\{c_m\}$ and $\{\Delta_m\}$, the relevant portion of $\ell_1[\{a_m\}, \{\Delta_m\}]$ is given by

$$\ell_1[\{a_m\}, \{\Delta_m\}] = -2 \sum_m c_m z(mT + \Delta_m) + \sum_m \sum_k c_m c_k g((m-k)T + \Delta_k - \Delta_m), \quad (17)$$

where the matched-filter output $z(t)$ is given by

$$z(t) = \int_{-\infty}^{\infty} h(t' - t) r(t') dt' \quad (18)$$

and

$$g(t) = \int_{-\infty}^{\infty} h(t') h(t + t') dt' \quad (19)$$

is the channel correlation function. Thus the sufficient statistics are the set of matched-filter output samples $\{z(mT + \Delta_m)\}$, where the sampling phases $\{\Delta_m\}$ are still to be determined. The other component of the likelihood is given by

* It should be clear from (15b) and (10) that, in the absence of a noise or timing-phase hit, $\hat{\Delta}_n \to \Delta_n$ and $\hat{a}_n \to a_n$; i.e., the estimates tend to the true parameter values.

$$\ell_2[\{\Delta_m\}] = -2N_0 \ln p(\Delta_0) - 2N_0 \sum_{m=1} \ln \left\{ \frac{(1-p_0)}{\sqrt{2\pi}\,\sigma} e^{-(\Delta_m - \Delta_{m-1})^2/2\sigma^2} + \frac{p_0}{\sqrt{2\pi}\,(\mu^2+\sigma^2)^{1/2}} e^{-(\Delta_m-\Delta_{m-1})^2/2(\mu^2+\sigma^2)} \right\}, \quad (20)$$

and the optimum receiver minimizes $\ell = \ell_1 + \ell_2$, where ℓ_1 and ℓ_2 are given by (17) and (20), respectively, with respect to $\{a_m\}$ and $\{\Delta_m\}$. Optimization with respect to $\{\Delta_m\}$ is via differentiation and gives

$$(\Delta_1 - \Delta_0)G[\Delta_1 - \Delta_0] + \frac{\partial}{\partial \Delta_0} \ell_1[\{a_m\}, \{\Delta_m\}] = 0 \quad (21)$$

and

$$(\Delta_{k+1} - \Delta_k)G[\Delta_{k+1} - \Delta_k] - (\Delta_k - \Delta_{k-1})G[\Delta_k - \Delta_{k-1}] + \frac{\partial}{\partial \Delta_k} \ell_1[\{a_m\}, \{\Delta_m\}] = 0, \qquad k = 1, 2, \cdots \quad (22)$$

where the function $G[\cdot]$ is defined by*

$$G[x] = \left[\frac{\frac{(1-p_0)}{\sigma(\sigma^2/2N_0)} \exp\{-x^2/2\sigma^2\} + \frac{p_0}{\mu(\mu^2/2N_0)} \exp\{-x^2/2\mu^2\}}{\frac{1-p_0}{\sigma} \exp\{-x^2/2\sigma^2\} + \frac{p_0}{\mu} \exp\{-x^2/2\mu^2\}} \right]. \quad (23)$$

As we see in the next section, the nature of $G[\ ]$ will impart a dual-mode character to the various tracking loops described in the sequel. It is convenient to define the weighted differential-epoch

$$\eta_k \equiv (\Delta_k - \Delta_{k-1})G[\Delta_k - \Delta_{k-1}], \quad (24)$$

and (21) and (22) can thus be written as

$$\eta_{k+1} = \eta_k - \frac{\partial \ell_1}{\partial \Delta_k}[\{a_m\}, \{\Delta_m\}] \qquad k = 0, 1, 2, \cdots, \quad (25)$$

where $\eta_0 = 0$.

Several difficulties associated with the "iteration" prescribed by (25) preclude incorporation in a realistic detector: (*i*) as already mentioned, Δ_0 is unknown, (*ii*) as it stands, the optimization over Δ_k is for a *given* set of $\{a_m\}$, (*iii*) from (17) it is clear that $\partial \ell_1/\partial \Delta_k$ depends on all the $\{a_m\}$ and $\{\Delta_m\}$, and (*iv*) optimization of (17) with respect to the $\{a_m\}$ requires a Viterbi-related dynamic programming algorithm.[8] (The state size is

* We have assumed that $\mu \gg \sigma$ so that $\mu^2 + \sigma^2 \approx \mu^2$. The detailed nature of the function $G[\cdot]$ is discussed in the next section.

somewhat ambiguous, since the presumably finite memory of $g(t)$ is enhanced by the arbitrarily large size of $\Delta_k - \Delta_m$.)

Because of the above factors, the level of complexity associated with the optimum receiver is prohibitive, and thus even for our simplified model we must resort to suboptimum reception. In a sense, this is not surprising since the maximum likelihood receiver has at its disposal the entire observation record, and it is only in special cases that the optimum procedure can be implemented in a sequential manner.

IV. SUBOPTIMUM RECEPTION

In this section, we indicate several reasonable receivers suggested by the optimum receiver of the previous section.

4.1 Data-directed receiver

Our approach to deriving a useful suboptimum receiver is to remove, via approximation, the difficulties associated with implementing the optimum receiver—the principal simplification we will make is to take a decision-directed approach. We begin by noting that (25) would be a practical and realizable recursion if: (*i*) $\partial \ell_1 / \partial \hat{\Delta}_k$ depended only on $\hat{\Delta}_k$ and $\hat{a}_k$, and (*ii*) the optimum value of $\hat{a}_k$ depends only on $z(kT - \hat{\Delta}_k)$ and $\hat{\Delta}_k$. With these desiderata in mind, we note from (17) that

$$\frac{\partial \ell_1}{\partial \hat{\Delta}_k}[\{\hat{a}_m\}, \{\hat{\Delta}_m\}] = -2c_k \dot{z}(kT + \hat{\Delta}_k) - \hat{c}_k \times \sum_{n \neq k} \hat{c}_n [\dot{g}((k-n)T - \hat{\Delta}_k + \hat{\Delta}_n) - \dot{g}((n-k)T + \hat{\Delta}_n - \hat{\Delta}_k)] \quad (26)$$

where the "dot" indicates the time derivative. The first concession we make to realizability is to neglect the second term in (26). Note that if a tone is not transmitted and the data levels are uncorrelated, then the expected value of this term is zero. We realize, of course, that this term would make a contribution whenever a timing-phase hit occurs; however, we are relying on the $\ell_2[\cdot]$ component of the likelihood to provide the dominant indication of this event. With this approximation, (25) reduces to

$$\eta_{k+1} = \eta_k + 2\hat{c}_k \dot{z}(kT + \Delta_k), \quad (27)$$

and from (24) we have

$$\eta_{k+1} = (\Delta_{k+1} - \Delta_k) G[(\Delta_{k+1} - \Delta_k)].$$

The value of Δ_{k+1} may be generated from η_{k+1} and Δ_k via the inverse relation

$$\Delta_{k+1} = \Delta_k + F^{-1}[\eta_{k+1}], \quad (28)$$

where if $\eta = xG[x] \equiv F[x]$, then $F^{-1}[\ ]$ is defined by

$$x \equiv F^{-1}[\eta]. \quad (29)$$

Thus (27) and (28) provide a second-order system of iterative equations for generating the desired estimates of $\{\Delta_k\}$ provided: (*i*) that either the sequence $\{c_k\}$ is known or reliable decisions are available, and (*ii*) that the initial phase estimate Δ_0 is known. These equations may be viewed as a second-order, discrete-time phase-locked loop (PLL) with a nonlinearity $F^{-1}[\ ]$ necessitated by the dual-mode nature of the timing phase.

The function $F^{-1}[x]$ is plotted in Fig. 3 for $p_0 \approx 0$ and $\sigma^2 \ll \mu^2$. Note that this function is odd, exhibits hysteresis, and is multivalued over a certain range, and as shown in Fig. 3, $F^{-1}[\eta]$ can be approximated by the two straight-line segments

$$F^{-1}(\eta) = \begin{cases} \dfrac{\sigma^2}{2N_0}\eta, & |\eta| \le \eta^{(1)} \approx \sqrt{2}\sigma \sqrt{\log_e \left(\dfrac{p_0}{1-p_0}\dfrac{\mu}{\sigma}\right)} \\ \dfrac{\mu^2}{2N_0}\eta, & |\eta| \ge \eta^{(2)} \approx \sqrt{2}\sigma \sqrt{\log_e \left(\dfrac{p_0}{1-p_0}\dfrac{\mu^3}{\sigma^3}\right)} \\ \dfrac{\sigma^2}{2N_0}\eta \text{ or } \dfrac{\mu^2}{2N_0}\eta, & \eta^{(1)} \le |\eta| \le \eta^{(2)}. \end{cases} \tag{30}$$

With regard to the recursion (28), the parameters σ^2/N_0 and μ^2/N_0 can

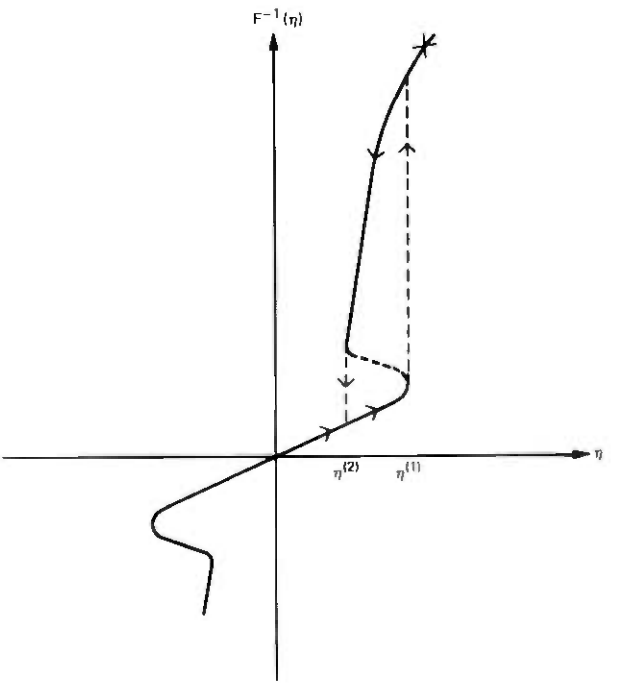


Fig. 3—The nonlinearity $F^{-1}[\eta]$ of (30).

be interpreted as a small and a large "step-size," respectively. The magnitude of the step-size depends on the "old" η_k and the "processed" observation $\hat{c}_k \dot{z}(kT + \Delta_k)$, and a typical trajectory of step-sizes is indicated by the arrows in Fig. 3.

Returning to (17), we rewrite this expression as

$$\ell_1[\{a_m\}, \{\Delta_m\}] = -2 \sum_m c_m z(mT + \Delta_m) + \sum_m c_m^2 g_0 + \sum_{m \neq k} \sum_k c_m c_k g((m-k)T + \Delta_k - \Delta_m). \quad (31)$$

If the system pulse shape is Nyquist $[g(n-k)T = g_0 \delta_{n-k}]$, then the third summation will be close to zero when the $\{\Delta_n\}$ are approximately equal over the duration of $g(t)$. With this approximation in mind, we neglect the cross term $(m \neq k)$ and complete the square to obtain

$$\ell_1[\{a_m\}, \{\Delta_m\}] \cong \sum_m \left\{ g_0 \left(c_m - \frac{z_m}{g_0} \right)^2 - \frac{z_m^2}{g_0} \right\} = \sum_m \left\{ g_0 \left(a_m - \left[\frac{z_m}{g_0} - \rho(-1)^m \right] \right)^2 - \frac{z_m^2}{g_0} \right\},$$

and thus

$$\hat{a}_m = Q\left[\frac{z_m}{g_0} - \rho(-1)^m \right], \quad (32a)$$

where $Q[\ ]$ is a function which quantizes its argument to the nearest symbol level. Observe that, with the assumptions we have made, the optimum value of $\hat{a}_m$ depends only on $z_m \equiv z(mT + \hat{\Delta}_m)$ and is in fact the symbol level closest to $(z_m/g_0) - \rho(-1)^m$. The receiver sketched in Fig. 4 implements (27), (28), and (32a).

If the system pulse shape is of the partial response type,[2] then a modified procedure is called for. We illustrate this technique when the received pulse $g'(t)$ is a Class IV partial response pulse and the shaping

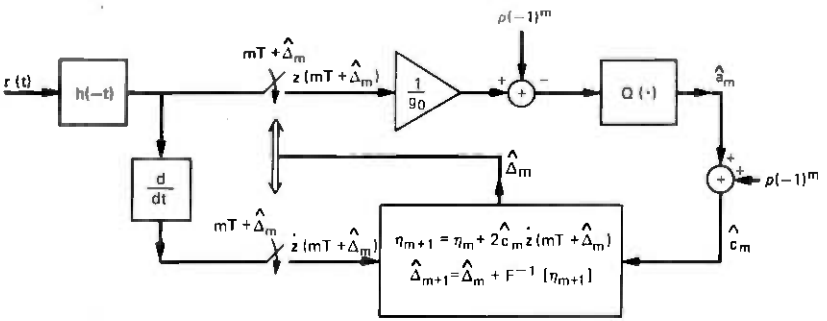


Fig. 4—Data-directed timing loop and receiver.

is split between the transmitter and receiver, the overall* characteristic being $j2T \sin \omega T$. In this case, $g_k' = g_0'(\delta_k - \delta_{k-2})$ and, again neglecting the delay differences $\Delta_k - \Delta_m$, the likelihood becomes

$$\ell_1[\{a_m\}, \{\Delta_m\}] = -2 \sum_m c_m z(mT + \Delta_m) + g_0' \left(\sum_m c_m^2 + \sum_m c_m c_{m-2} \right). \tag{32b}$$

Because of the coupling between c_m and c_{m-2}, the optimization procedure to determine the $\{\hat{c}_m\}$, from (32b), requires the use of dynamic programming (the Viterbi algorithm[8]). While the Viterbi algorithm (VA) can be implemented in a rather straightforward manner for Class IV partial response systems, the decoding delay in the VA makes tracking of the timing phase rather unwieldy, and consequently practical receivers would probably employ a suboptimum technique which directly examines the output of the receiving filter,

$$z(kT) = \sum_n c_n g'(kT - nT - \Delta_n) + \nu(kT). \tag{32c}$$

In the above equation, the samples are obtained from the output of the receiver filter and, neglecting timing jitter, we have

$$z(kT) = c_k - c_{k-2} + \nu_k = d_k + \nu_k, \tag{32d}$$

where $\{d_k\} = \{c_k - c_{k-2}\}$ is the dependent or correlated data sequence. For example, if the input data symbols c_k assume the values $\pm 1, \pm 3$, then d_k would be one of the seven output values $0, \pm 2, \pm 4, \pm 6$. Practical detectors would quantize z_k to one of the seven allowed output values, and the desired data $\{\hat{c}_k\}$ is recovered from the relation $\hat{c}_k = \hat{d}_k - \hat{d}_{k-2}$, where the data are typically precoded[2] to prevent an erroneous decision from propagating. Note that the partial response waveform can be written either as (32c) or as $\Sigma_n d_n' g(t - nT - \Delta_n)$, where d_n' are the correlated output levels and $g(t)$ is the minimum-bandwidth symmetric Nyquist pulse, $\sin(\pi t/T)/(\pi t/T)$. If we adopt this latter representation, then our maximum likelihood development can proceed as before—the only additional approximation being that, while the various sequences of $\{d_n'\}$ are not all equally likely, we have implicitly taken them to be equiprobable. Thus, an approximation to the optimum receiver shown in Fig. 5 would be to quantize z_k, using (32c), to the nearest output level and to use the corresponding $\hat{d}_n$ in (27) and (28).

Returning to Fig. 4, we recall that (27) to (29) have the appearance of a second-order, discrete-time, phase-locked loop with a bi-variable step-size. The choice of step-size is dictated by the current value of η_k which provides a measurement of the "jump" $\Delta_k - \Delta_{k-1}$. A large value of η_{k+1} is indicative of a large jump, while a small value of η_{k+1} reassures

* Note that, in this case, $g'(t)$ is antisymmetric and the receiver filter is not matched to the transmitter filter.

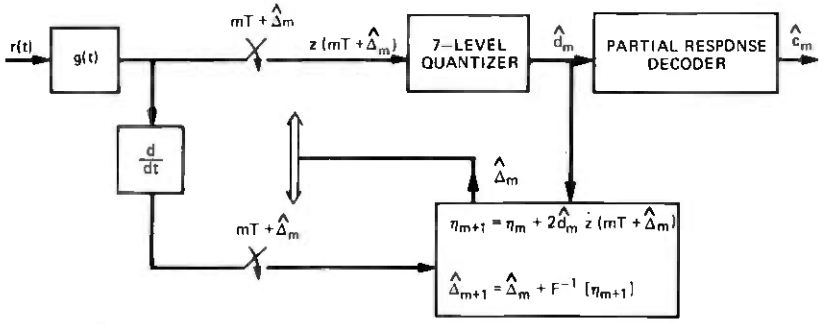


Fig. 5—Data-directed timing loop and receiver for partial-response signaling.

the tracking loop that its estimate of Δ_k is close to the correct value. The nonlinearity shown in Fig. 3 is interpreted as providing hysteresis, since via (30) we know that in the range $\eta^{(1)} \leq |\eta| \leq \eta^{(2)}$

$$\Delta_{k+1} = \Delta_k + \beta_{k+1}\eta_{k+1}, \tag{33}$$

where

$$\beta_{k+1} \approx \begin{cases} \sigma^2/N_0, & \text{if } |\eta_{k+1}| < \eta^{(2)} \text{ and } |\eta_k| < \eta^{(2)} \\ \mu^2/N_0, & \text{if } |\eta_{k+1}| > \eta^{(2)} \text{ or if } |\eta_{k+1}| > \eta^{(1)} \text{ and } |\eta_k| > \eta^{(2)}. \end{cases}$$

The bi-variable step-size appearing in the tracking loop of Fig. 6 has the effect of adaptively varying the loop's bandwidth and thus accelerating recovery from a delay jump. The omission of the quadratic term appearing in (31) may prolong this recovery by several symbol intervals, but this is a small price to pay for the resulting simplicity in implementation.* The receiver shown in Fig. 6 quantizes the filtered and sampled sequence with the aid of a decision-directed phase-locked tracking loop

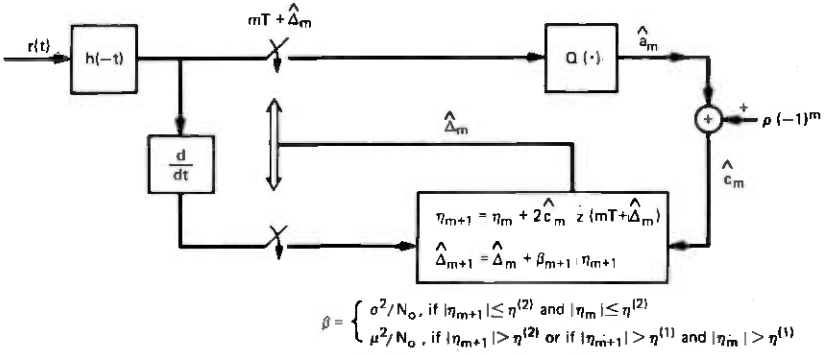


Fig. 6—Simplified data-directed receiver.

* Omission of the quadratic term is tantamount to neglecting the *amplitude* transient, caused by the delay jump, which propagates through the channel and receiver filters. In other words, if one accepts the model given by (6), then any amplitude transients are implicitly neglected.

which provides the sequence of sampling phases, given knowledge of the maximum-likelihood, initial-phase estimate, Δ_0. As described in a later section, the algorithm can be suitably modified to incorporate an estimate of Δ_0. The above decision-directed timing loop is similar to that described by Gitlin and Salz,[9] but the novel aspects here are the bivariable step-size and the hysteresis associated with the tracking loop nonlinearity.

4.2 Modifications to the decision-directed receiver

A drawback of the receiver described in Section 4.1 is the possibility of a relatively long error burst following a timing-phase jump. Suppose such a delay jump causes a large deviation from the optimum sampling time; in a bandlimited system with a narrow eye-opening,[2] this results in a large amount of intersymbol interference and consequently a high probability of error in the next symbol interval. The resulting incorrect decision, used in the decision-directed timing recovery loop, may move the estimated sampling phase in the wrong direction, further increasing the intersymbol interference. This type of effect is called *runaway* and is possible in nearly all decision-directed parameter tracking systems. Runaway is of particular concern in Class IV partial-response systems since the eye is open only for a small fraction of the symbol interval.[10] The possibility of runaway is further enhanced by our neglecting the quadratic cross-term appearing in (31).

To diminish the possibility of error proliferation due to delay jumps in a bandwidth-limited system, we propose the *coarse-quantized* timing recovery system shown in Fig. 7. The incoming signal is sampled at times $\{mT + \Delta_m + (iT/M),\ 0 \le i \le M - 1\}$, where M is some integer, i.e., the receiver samples are taken at the rate M/T instead of $1/T$ samples/s. The sampling phase is still controlled by a *single* tracking loop and, as before, after suitable filtering, each sample is quantized to the nearest data level. The number of samples M is chosen large enough that T/M is less than the width of the eye-opening corresponding to the pulse $g(t)$. Thus for a system with an open eye,[2] in the absence of noise, at least one of the sampling phases $\{\Delta_m + (iT/M)\}$ will result in a correct output* decision. Expressed mathematically, for at least one integer "i," the maximum possible interference,

$$\max_{|c_m|} \sum_{m \neq 0} \left| c_m g\left(mT + \Delta_m + \frac{iT}{M}\right) \right|,$$

is less than the minimum distance between two possible received signal levels.

The idea behind the increased sampling rate (which might be readily

* For a four-input level, Class IV partial response system, recall that the output sequence is chosen from one of seven levels.

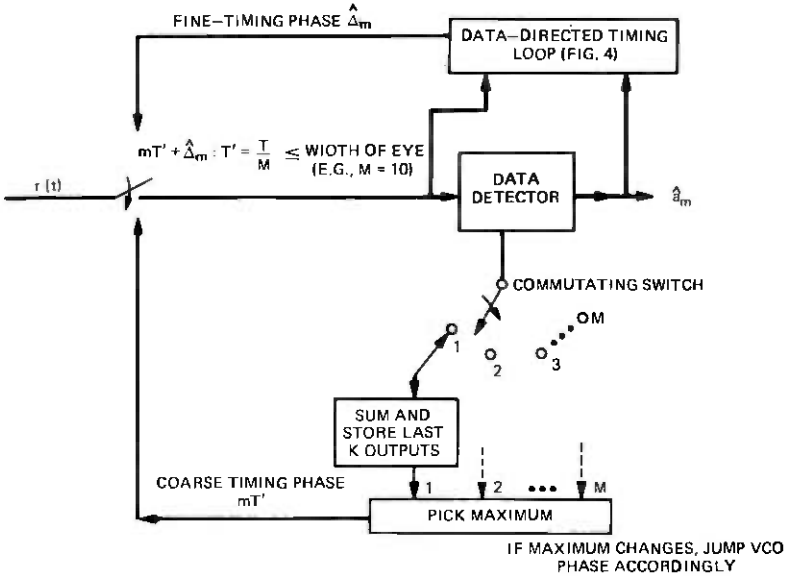


Fig. 7—Coarse-quantized timing recovery system.

available in a digital receiver) is that one of the M possible coarsely quantized sampling phases results in an open eye, and therefore in the absence of noise, supplies a correct sequence $\{c_n\}$ suitable for updating the decision-directed phase-tracking loop. The particular (coarse) timing phase chosen to supply the decision sequence used by the tracking loop is determined by reformulating the problem as picking the static timing epoch $\{iT/M\}$ which maximizes the *a posteriori* likelihood over the recent past. This strategy is mechanized from (32a) by computing a running likelihood

$$\ell_k^{(i)}\left[\{c_m^{(i)}\}, \{\Delta_m\}\right] = g_0 \sum_{j=k-K}^{k} \left[c_j^{(i)} - \frac{1}{g_0} z\left(jT + \Delta_j + \frac{iT}{M}\right)\right]^2 - \frac{1}{g_0} \sum_{j=k-K}^{k} z^2\left(jT + \Delta_j + \frac{iT}{M}\right), \quad (34)$$

where K is some suitably chosen number, $\{c_j^{(i)}\}$ are the decisions† corresponding to the sampling phase $iT/M + \Delta_j$, and the decisions are obtained by quantizing the appropriate output sample. The coarse-timing phase, i^*T/M, is chosen if $\ell_k^{(i^*)} \leq \ell_k^{(i)}$ for all $i \neq i^*$. In practice, one might use a small number of sampling epochs; e.g., three, which would bracket the correct phase. The sum in (34) is truncated to run over a finite span of duration KT seconds so that the effects of ancient delay jumps do not affect the *current sampling epoch*. Once the best coarse timing phase

† The Δ_j are supplied by the phase-locked loop driven by the decisions corresponding to the current most likely coarse-quantized timing phase.

is determined, the bi-variable step-size PLL previously described is used to determine the exact sampling phase. Whenever a new sampling phase becomes the most likely, the current estimate of the timing phase is incremented by the appropriate amount.† It should be pointed out that additive noise following a delay jump may prolong the recovery somewhat. This effect is difficult to assess analytically, and a similar statement can be made concerning the size of M (the number of timing epochs) and K (the memory of the running likelihood). The sensitivity of system performance to these parameters is best determined experimentally.

The realization shown in Fig. 7, which incorporates the coarse-quantized timing recovery scheme, has the related mechanization depicted in Fig. 8, which is specialized to a Class IV partial-response system. Here the quantized seven-level outputs are computed for each sampling phase, and each sequence is monitored for partial-response violations. The coarse-quantized sampling phase used to control the timing tracking loop is chosen as the phase which has the fewest associated partial response violations. Again, the actual logic which dictates when and how switches to a new timing phase are accomplished is probably best determined by an experimental and/or simulation study of the actual system.

4.3 A refined loop which estimates Δ_0

As it stands, the bi-modal phase-locked loop described by (27) and (28) is initialized from a random or arbitrary initial condition, Δ_0. A consequence of this initialization is that in either mode (delay hit/no delay hit) the loop exhibits a double integration (or direct second difference) structure. We now show that, when a constant step-size is used, the algorithm is potentially unstable.‡ We begin by recalling that in either mode the tracking loop is governed by equations of the form

$$\eta_{k+1} = \eta_k + c_k \dot{z}(kT + \Delta_k) \tag{27}$$

$$\Delta_{k+1} = \Delta_k + \beta_k \eta_{k+1}, \tag{33}$$

where β_k is a positive nonincreasing sequence.

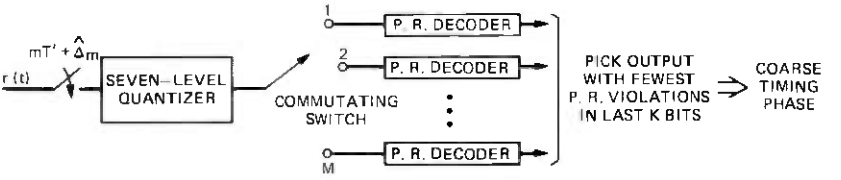


Fig. 8—Coarse-quantized timing recovery for partial-response system.

† For example, suppose $\ell_k^{(i)}$ was the maximum and the new maximum is $\ell_k^{(i^*)}$, then the timing phase should be incremented by $(i^* - i)\ T/M$ s.

‡ This instability can be regarded as a manifestation of the sensitivity of the system equations to the initial unknown phase; i.e., the effect of a wrong choice of this phase propagates endlessly. This is a consequence of viewing the system of simultaneous equations for the timing-phase estimate, (27) and (28), as a recursion with an arbitrary initial condition.

For the purpose of this discussion, we consider a fixed but unknown delay, Δ, and examine the average values of the above equations, i.e.,

$$\bar{\eta}_{k+1} = \bar{\eta}_k + \dot{g}(\Delta_k - \Delta) \tag{35a}$$

$$\bar{\Delta}_{k+1} = \bar{\Delta}_k + \beta_k \bar{\eta}_{k+1}, \tag{35b}$$

where the overbar denotes expectation. We can combine the above equations to obtain the recursion

$$\bar{\Delta}_{k+1} - \left(1 + \frac{\beta_k}{\beta_{k-1}}\right) \bar{\Delta}_k + \frac{\beta_k}{\beta_{k-1}} \bar{\Delta}_{k-1} = \beta_k \dot{g}(\bar{\Delta}_k - \Delta), \tag{36}$$

and if we denote the tracking error by

$$\epsilon_k = \overline{\Delta_k - \Delta}, \tag{37}$$

then we have

$$\epsilon_{k+1} - \left(1 + \frac{\beta_k}{\beta_{k-1}}\right) \epsilon_k + \frac{\beta_k}{\beta_{k-1}} \epsilon_{k-1} \approx \beta_k [\dot{g}(0) + \epsilon_k \ddot{g}(0)]$$

$$= \beta_k \ddot{g}(0) \epsilon_k, \tag{38}$$

where we have used a Taylor Series expansion which is valid for small ϵ_k. Note that the solution to the above time-varying difference equation will decay when the product of the "instantaneous roots," β_k/β_{k-1} is less than unity. However, if we were to use a constant step-size, i.e., $\beta_k = \beta_{k-1} = \beta$, and if $\ddot{g}(0) \approx 0$, then the solutions are of the form $\epsilon_k = \epsilon_0 \sin k\theta$; i.e., the error does not decay to zero but oscillates as soon as the error penetrates the linear region (clearly, this is an unacceptable situation). The existence of oscillations can be deduced directly from (27) and (28). Note from (28) that Δ_k is the accumulated sum of the past errors. When a phase-hit occurs, this sum will become large and the loop will enter the large step-size mode. In order that the loop ultimately converge to the correct phase, it is clear that the accumulator will have to "see" many terms opposite in sign to the original accumulants, i.e., the loop can oscillate.

In the light of the above discussion, we now derive an estimate of Δ_0 and indicate how this estimate may be incorporated into the existing timing loop to produce a stable loop. We first let

$$\Delta'_m \equiv \Delta_m - \Delta_0, \tag{39}$$

and in terms of $\{\Delta'_m\}$ we have from (20)

$$\ell_2[\{\Delta'_m\}] = -N_0 \sum_{m=1}^{\infty} \log \left\{ \frac{(1 - p_0)}{\sqrt{2\pi}\,\sigma} \exp\{-(\Delta'_m - \Delta'_{m-1})/2\sigma^2\} \right.$$

$$\left. + \frac{p_0}{\sqrt{2\pi}\,\mu} \exp\{-(\Delta'_m - \Delta'_{m-1})^2/2\mu^2\} \right\}, \tag{40}$$

where we note that $\Delta_0' = 0$; it is our intention to estimate $\{\Delta_n'\}$ and $\{\Delta_0\}$ separately and then to combine these quantities to construct $\{\hat{\Delta}_n\}$.

Proceeding as before, we define

$$\eta_k' \equiv (\Delta_k' - \Delta_{k-1}')G[\Delta_k' - \Delta_{k-1}'], \tag{41}$$

and taking the derivative of the likelihood with respect to Δ_k' gives

$$\begin{aligned}\eta_{k+1}' &= \eta_k' - \frac{\partial \ell_1[\{a_m\}, \{\Delta_m'\}, \Delta_0]}{\partial \hat{\Delta}_k'} \\ &= \eta_k' + 2c_k\dot{z}(kT + \Delta_k' + \Delta_0).\end{aligned} \tag{42}$$

Inverting (41) gives

$$\Delta_{k+1}' = \Delta_k' + F^{-1}[\eta_{k+1}'], \tag{43}$$

where $F^{-1}[\ ]$ has been previously defined and where (42) and (43) are initialized with $\Delta_0' = 0$.

An estimate of Δ_0 will be obtained by applying stochastic approximation theory,[11] and using as the increment the derivative of the current term in the likelihood, $\ell_1[\{a_m\}, \{\Delta_m'\}, \Delta_0]$, with respect to Δ_0.

The resulting stochastic approximation algorithm for the estimate of Δ_0 is

$$\hat{\Delta}_{0,k+1} = \hat{\Delta}_{0,k} + \gamma_k c_k \dot{z}(kT + \hat{\Delta}_k' + \hat{\Delta}_{0,k}), \qquad k = 0, 1, 2, \cdots, \tag{44}$$

where γ_k is a positive step-size sequence. Since the estimate of Δ_k is the sum of the component estimates, i.e.,

$$\hat{\Delta}_k \equiv \hat{\Delta}_k' + \hat{\Delta}_{0,k}, \tag{45}$$

adding (43) and (44) gives the structure shown in Fig. 9, which implements the recursions:

$$\hat{\Delta}_{k+1} = \hat{\Delta}_k + F^{-1}[\hat{\eta}_{k+1}] + \gamma_k \hat{c}_k \dot{z}(kT + \hat{\Delta}_k) \tag{46}$$

$$\hat{\eta}_{k+1}' = \hat{\eta}_k' + 2\hat{c}_k \dot{z}(kT + \hat{\Delta}_k). \tag{47}$$

In implementing (46) and (47), the step-size γ_k would probably be switched to a larger step-size whenever the $F^{-1}[\cdot]$ function indicates that a mode switch is taking place—this can be thought of as reinitializing the estimate of Δ_0. Contrasting (46) and (47) with (27) and (28), we see that, when the state of the system is such that $F^{-1}[\eta] = \beta\eta$, the latter system can be written as the second-order difference equation

$$\Delta_{k+1} - 2\Delta_k + \Delta_{k-1} = \beta c_k \dot{z}(kT + \Delta_k), \tag{48}$$

while the former system is equivalent to

$$\begin{aligned}\Delta_{k-1} - 2\Delta_k + \Delta_{k-1} = (\gamma_k + \beta)c_k\dot{z}(kT + \Delta_k) \\ - \gamma_{k-1}c_{k-1}\dot{z}(kT - T + \Delta_{k-1}).\end{aligned} \tag{49}$$

The effect of the direct feeding of the input, $\gamma_k c_{k-1} \dot{z}(kT - T + \Delta_{k-1})$, to the second summer in Fig. 9 can be seen by considering the evolution of the average phase error, (37). It is clear from (49) that with $\gamma_k = \gamma$ the modified tracking-loop structure can provide roots within the unit circle and hence eliminate the possibility of oscillations.

V. APPLICATION IN A SIMULATED CLASS IV PARTIAL-RESPONSE SYSTEM

The application of dual-mode, decision-directed timing recovery and coarse-quantized timing recovery to a data communication system subject to additive noise and occasional delay jumps was tested by means of a computer simulation of a digital version of the baseband data transmission system shown in Fig. 10. Transmission through radio channels was modeled by the addition of additive white Gaussian noise to the signal and the insertion of abrupt delay changes. Since the simulated system is not an exact replica of the idealized equations used for analysis, the actual receiver differed in some small details from the structure previously derived.

A seven-level, Class IV, partial-response waveform was generated in sampled form with sampling rate $10/T$ (10 times the symbol rate), and channel and receiver signal processing were also carried out digitally at this sampling rate.

Because the simulation was carried out in nonreal time on a digital computer, exact realization of time delays of other than multiples of $T/10$ was not possible. Moreover, the actual sampling phase of the oversampled input was not under the receiver's control. Instead, arbitrary channel and receiver sampling delays were approximated by linear interpolation. The samples at the output of the receiver filter were denoted $\{z(mT + iT'); m = 0, 1, 2, \cdots, \infty; i = 0, 1, \cdots, 9\}$.

The index i denotes a timing phase, quantized to a multiple of $T' = T/10$. The output sampled at $mT + iT' + \hat{\Delta}_m$ was taken to be, by linear interpolation,

$$z(mT + iT' + \hat{\Delta}_m) = \left(1 - \frac{\hat{\Delta}_m}{T'}\right) z(mT + iT') + \frac{\hat{\Delta}_m}{T'} z(mT + (i+1)T'). \quad (50)$$

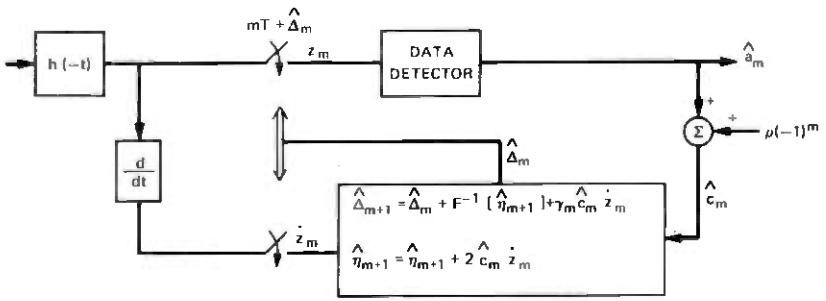


Fig. 9—Modified data-directed, fine-tuned, timing loop.

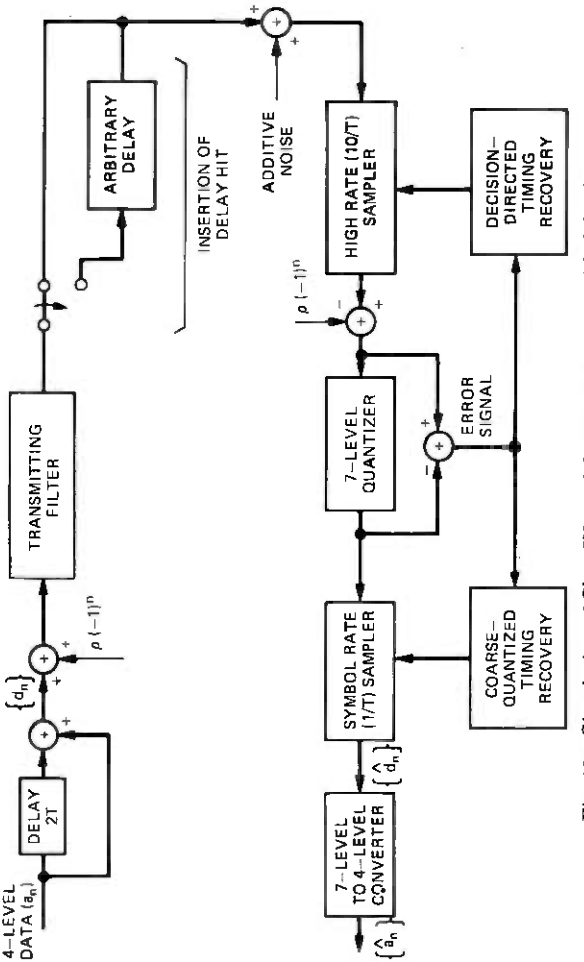


Fig. 10—Simulation of Class IV partial response system with delay hits.

The quantity $\hat{\Delta}_m$ ($0 \le \hat{\Delta}_m \le T'$) is the receiver's estimate of the sampling phase in the mth symbol interval, mod T'.

The 10-fold oversampling also permits the application of the coarse-quantized timing recovery method described in Section 4.2, with $M = 10$. Each sample[†] $z(mT + iT' + \hat{\Delta}_m)$ is quantized into $\hat{c}_m^{(i)} = \hat{d}_m^{(i)} + \rho(-1)^m$ where $\hat{d}_m^{(i)}$ is one of the seven levels $\hat{d}_m^{(i)} = 0, \pm 2, \pm 4, \pm 6$, and an error $e_m^{(i)}$ is formed as:

$$e_m^{(i)} \equiv z(mT + iT' + \hat{\Delta}_m) - \hat{c}_m^{(i)}. \tag{51}$$

For each integer i from 0 to 9, the sum of squared errors over the past K symbol intervals was formed

$$\ell_k^{(i)} = \sum_{m=k-K}^{k} e_m^{(i)2}.$$

The values of K used in the simulations were 20 and 40.

That integer $i = i^*$ which minimized $\ell_k^{(i)}$ was taken to be the current most likely coarse-quantized sampling phase. Whenever $i = i^*$ (once per symbol interval) the current decision $\hat{c}_m^{(i^*)} - \rho(-1)^m = \hat{d}_m^{(i^*)}$ is passed on as the receiver's decision on d_m. The program records the occurrence of errors (discrepancies between $\hat{d}_m^{(i^*)}$ and d_m). Note that the above definition of $\ell_k^{(i)}$ differs from that proposed in Section 4.2 in the omission of the sum of squares of z-samples. A further modification was the inhibition of a change in i^* when $\ell_k^{(i)}$ is greater than 90 percent of $\ell_k^{(i^*)}$. This "dead zone" modification reduced the occurrence of switches back and forth between two values of i for which the values of $\ell_k^{(i)}$ are nearly equal.

We remark that abrupt changes in the intervals between receiver output samples should not be passed on to the data recipient. The receiver's output samples would in practice enter an elastic buffer and be clocked out under the control of a very narrowband phase-locked loop.

The value of $\hat{\Delta}_m$ used in the interpolative sampling procedure was obtained by a digital implementation of the decision-directed, second-order timing recovery algorithm described in Section 4.3. Instead of using the correction term $\hat{c}_m^{(i^*)}\, \dot{z}(mT + \hat{\Delta}_m + i^*T')$ in the loop, we use an approximation to the negative of the derivative (gradient) of the squared error $e_m^{(i^*)2}$ with respect to $\hat{\Delta}_m$; i.e.,

$$\delta_m \equiv -e_m^{(i^*)}\, [z(mT + (i^* + 1)T') - z(mT + i^*T')]. \tag{52}$$

When the loop error is zero, the modified correction term (52) guarantees that no adjustment will be made, while the original correction term only provides this condition on the average.

[†] The overall impulse response was scaled so that the ideal sampled outputs in the absence of noise are 0, ±2, ±4, ±6.

The average value of the correction term δ_m defined by (52) can be computed from equations (51), (32c), and (32d) for an ideal Class IV partial-response system for which

$$g'(t) = \frac{\sin\left(\frac{\pi t}{T}\right)}{\left(\frac{\pi t}{T}\right)} - \frac{\sin\left(\frac{\pi(t-2T)}{T}\right)}{\left(\frac{\pi(t-2T)}{T}\right)}. \tag{53}$$

For small timing errors $(\Delta_m - \hat{\Delta}_m - i^*T')$ between the true phase Δ_m and the estimated phase $\hat{\Delta}_m + i^*T'$ (that is, neglecting quadratic and higher order terms in $(\Delta_m - \hat{\Delta}_m - i^*T')/T'$), the linearized average value of δ_m is

$$\langle \delta_m \rangle \approx 0.278 \left(\frac{\Delta_m - \hat{\Delta}_m - i^*T'}{T'}\right) \tag{54}$$

when a timing tone is not transmitted ($\rho = 0$). The corresponding linearized average value of the correction term $\langle \hat{d}_m^{(i^*)} \dot{z}(mT + \hat{\Delta}_m + i^*T') \rangle$ can similarly be shown to equal this same quantity. Note that the correction term δ_m given by (52) arises from an attempt to minimize the mean-squared error of a linear interpolation scheme applied to a digital receiver whose actual input sampling phase is not under its control. Thus, we have established a connection between this simple linear interpolation scheme with a mean-squared-error optimality criterion and the decision-directed, timing-phase recovery scheme dictated by minimum error probability considerations.

As prescribed in Section IV, a second-order decision-directed, sampling-phase, updating algorithm including a direct correction term and a cumulative correction term is used. The following equations summarize the simulated receiver's operation:

(*i*) *Interpolative sampling*:

$$z(mT + iT' + \hat{\Delta}_m) \equiv \left(1 - \frac{\hat{\Delta}_m}{T'}\right) z(mT + iT') + \frac{\hat{\Delta}_m}{T'} z(mT + (i+1)T').$$

(*ii*) *Quantization*:

$$\hat{d}_m^{(i)} = \text{quantization of } [z(mT + iT' + \hat{\Delta}_m) - \rho(-1)^m]$$

$$c_m^{(i)} = \hat{d}_m^{(i)} + \rho(-1)^m.$$

(*iii*) *Error*:

$$e_m^{(i)} \equiv z(mT + iT' + \hat{\Delta}_m) - \hat{c}_m^{(i)}.$$

(*iv*) *Coarse-quantized timing recovery*:

i^* = value of i which minimizes

$$\ell_k^{(i)} = \sum_{m=k-K}^{k} e_m^{(i)2}.$$

(v) *Decision-directed fine timing recovery:*

$$\frac{\hat{\Delta}_{m+1}}{T'} = \frac{\hat{\Delta}_m}{T'} + \gamma_m \delta_m + \beta_m \eta_{m+1},$$

where

$$\delta_m \equiv e_m^{(i^*)} [z(mT + (i^* + 1)T') - z(mT + i^*T')]$$

and

$$\eta_{m+1} \equiv (1 - \alpha)\eta_m + \delta_m,$$

with

$$\hat{\Delta}_0 = \eta_0 = 0.$$

Since in the simulated and real systems, the timing transient does not instantaneously affect the received signal, the recursion defining η_{m+1} introduces a small amount of "leakage," represented by $\alpha = 0.0005$. Suitable values of the second-order loop parameters γ_m and β_m in the narrowband and wideband modes were established by loop-bandwidth considerations and observations of the transient response of $\hat{\Delta}_m$ to simulated delay jumps. The parameter values picked for the narrowband mode were

$$\gamma_m = 0.005$$

and

$$\beta_m = 3.42 \times 10^{-6}.$$

Assuming the linearized average correction term of (54), we have a discrete-time linear model of the second-order fine-timing recovery loop shown in Fig. 11. This loop's bandwidth, for the above values of γ_m and β_m and $1/T = 772$ kHz, is readily shown to be 240 Hz.

The wideband mode is initiated first whenever the value of i^* is changed by the coarse-quantized algorithm, or second whenever the following recursively generated quantity exceeds a threshold:

$$S_{m+1} = 0.99S_m + \delta_m. \tag{55}$$

The quantity S_m is a weighted average of past correction terms, in contrast to the *cumulative sum* of all past correction terms envisaged in (47).

With the initiation of the wideband mode, γ_m is set to 1 and β_m to 3.42×10^{-4}. Thereafter,

$$\gamma_m = \max(0.005, 1/L)$$

$$\beta_m = 3.42 \times 10^{-4}/L \quad \text{up to} \quad L = 200,$$

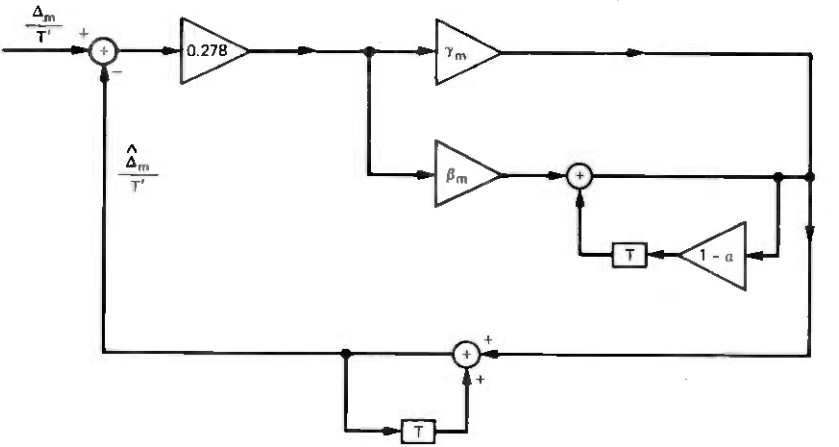


Fig. 11—Discrete time linearized, second-order loop model.

where L is the number of symbol intervals elapsed since initiation of the wideband mode. Thus, the initiation of the wideband mode restarts a stochastic approximation algorithm, with step-size decreasing toward a fixed minimum value. The duration of the wideband mode is 200 symbol intervals, after which the narrowband mode resumes.

VI. SIMULATION RESULTS

The channel model and receiver structure described in Section V were simulated with a 24-dB signal-to-noise ratio and with the insertion of occasional delay hits.

The value of K (the number of past squared errors stored by the coarse-quantized timing recovery algorithm) was set to either 20 or 40. Transmission both with and without a −18-dB ($\rho = 0.554$) $1/2T$ tone was simulated. The results are summarized in Fig. 12, which displays the observed average number of symbol errors (errors in $\hat{d}_m^{(i^*)}$) vs the delay hit expressed as a fraction of a symbol interval.

Each average plotted in Fig. 12 is only over five delay hits of the same magnitude, and thus the curves display considerable variability. Nevertheless, it is clear that a receiver employing two-mode decision-directed and coarse-quantized timing recovery can tolerate delay hits of up to almost half a symbol interval, while sustaining error bursts on the order of a dozen or less, rather than several thousand, which might be expected in a conventional tracking loop with a bandwidth on the order of 100 Hz. Greater delay hits unavoidably cause the deletion or repetition of data.

The number of errors sustained roughly doubled as K was doubled from 20 to 40. This is understandable, since the delay in detecting a phase change, by the coarse-quantized timing recovery system, is proportional to K. The risk of "false-alarm switching" decreases with K, and therefore

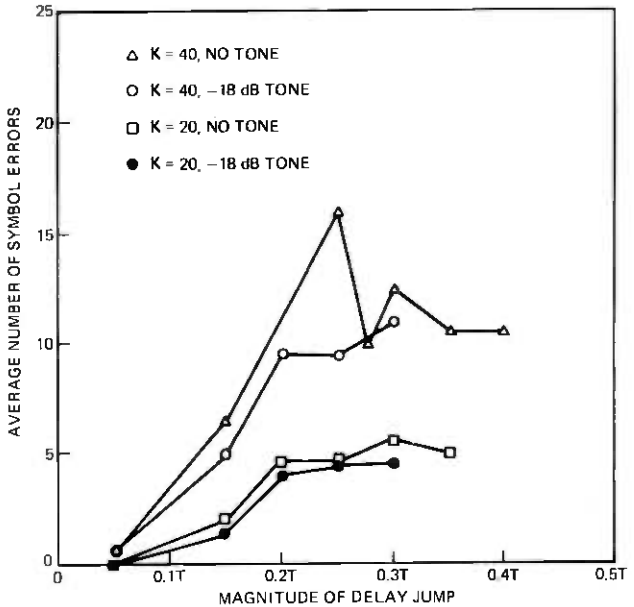


Fig. 12—Average number of errors produced by delay jumps in the simulated system employing two-mode, decision-directed and coarse-quantized timing recovery.

a relatively large value of K such as 40 may be worth the price of, say, a dozen extra errors sustained per delay jump. On the other hand, decreasing K will increase the number of errors due to "false-alarm" switching. The optimum value of K can best be determined by experience with a real system.

It is interesting to note from the curve that the presence or absence of a transmitted timing tone 18 dB below the data signal does not make a dramatic difference in the robustness of the system against delay hits. It therefore appears safe to omit the tone in a system employing decision-directed and coarse-quantized timing recovery. We note that the simulated system displayed rapid start-up characteristics in the coarse-quantized, decision-directed mode. The timing phase, correct to within $T/20$, was acquired in 20 to 25 symbol intervals in the absence of a transmitted tone. The transmission of a −18-dB tone unaccountably delayed timing-phase acquisition during start-up.

Figure 13 shows the evolution of the receiver's sampling-phase estimate following a delay jump of $-1.5T'$. The horizontal coordinate is the number of elapsed symbol intervals. The dotted curves show the phase estimate $i*T' + \hat{\Delta}_m$ (quantized by the limited resolution of the computer plotting routine). The x's at height 8.5 indicate the occurrence of symbol errors. In this example, the errors occur after the delay hit but before initiation of a coarse-quantized timing jump.

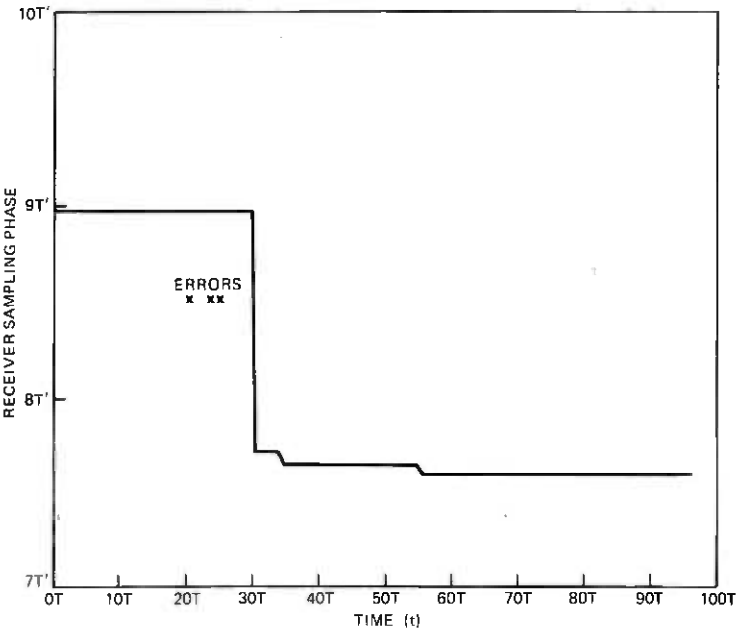


Fig. 13—Response of system to delay jump applied at $t = 0$.

VII. CONCLUSIONS

The technique of data-directed, coarse-quantized, dual-mode timing recovery has been derived and applied to the rapid acquisition of timing phase in systems subject to delay hits. In a simulated system, typical error-burst lengths, following a timing discontinuity of up to a half symbol interval, have been reduced to a dozen or so—two orders of magnitude less than that expected with a conventional phase-locked tracking system. Furthermore, the derivation and simulations have demonstrated the viability of these timing-recovery techniques in the absence of a transmitted pilot tone.

REFERENCES

1. K. L. Seastrand and L. L. Sheets, "Digital Transmission Over Analog Microwave Radio Systems," Proc. IEEE 1972 Int. Conf. Comm. (ICC'72), Philadelphia.
2. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
3. W. R. Bennett, "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.
4. R. D. Gitlin and J. F. Hayes, "Timing Recovery and Scramblers in Data Transmission," B.S.T.J., *54*, No. 3 (March 1975), pp. 569–593.
5. B. R. Saltzberg, unpublished work.
6. J. E. Mazo, "Jitter Comparison of Tones Generated by Squaring and by Fourth-Power Circuits," B.S.T.J., *57*, No. 5 (May-June 1978), pp. 1489–1498.
7. G. D. Forney, Jr., "Maximum Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-18* (May 1972), pp. 363–378.
8. G. D. Forney, Jr., "The Viterbi Algorithm," Proc. IEEE, *61*, No. 3, pp. 268–278.
9. R. D. Gitlin and J. Salz, "Timing Recovery in PAM Systems," B.S.T.J., *50*, No. 5 (May–June 1971), pp. 1645–1669.

10. J. Steel and B. M. Smith, "The Effects of Equalization, Timing and Carrier Phase on the Eye Patterns of Class-4 Partial Response Data Signals," IEEE Trans. Comm., February 1975.
11. D. J. Sakrison, "Stochastic Approximation: A Recursive Method for Solving Regression Problems," in *Advances in Communications*, Vol. 2, A. V. Balakrishran, ed., New York: Academic Press, 1966.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Caustic Patterns Associated With Melt Zones in Solidified Glass Samples—Part I: Symmetric Cases

By T. D. DUDDERAR, J. B. SEERY, and P. G. SIMPKINS

(Manuscript received March 17, 1978)

Ray-tracing algorithms have been developed to follow the propagation of a collimated beam of light traveling along and refracting out of a glass rod in a region of monotonically decreasing cross section. These algorithms have been used to study the formation and distribution of caustics as a function of the changing cross-section area. Axial profile data taken from the melt, or drawdown, zone of a solidified fiber-drawing sample provide the geometrical information needed to predict the loci of two major and two minor families of caustics. General principles for relating the observable far-field caustic patterns to the actual shapes of symmetric melt zones in glass samples are discussed.

I. INTRODUCTION

When a plane light wavefront propagates along a cylindrical glass rod in which a rapid monotonic decrease in cross section occurs, some light may be refracted from the glass and become externally visible. For a sample with homogeneous optical properties, the amount of emerging light and its intensity distribution are strongly influenced by the rate at which the cross section decreases. In a previous study of melt, or drawdown, zones of solidified samples taken from a laser-heated fiber-drawing system, the boundaries between the regions of emitted light and shadow were seen to be loci of intense illumination properly identified as "caustics."[1] These caustics were shown, by both experiment and analysis, to arise from various internal reflections and a refraction of the light from the surface. It was noted that the number of caustics increases as the rate of change of the cross section increases. No light is emitted from a very gradually tapered sample, while a great deal of light and numerous caustics are emitted from a sample with a very rapid taper. Also, as the rate of change of the cross section increases, the propagation

vectors of the light rays which form the far-field caustic patterns rotate; i.e., from the pulling or downstream direction through the radial and toward the upstream direction.

During the initial investigation, ray-tracing algorithms were developed that permitted accurate calculation of the light paths necessary to identify the two principal caustics experimentally observed. These routines utilize actual melt zone profile data and polynomial spline-fitting procedures to provide the geometrical information necessary to describe the caustics for a given value of the index of refraction n. Experimental results from four radically different samples compared very favorably with those obtained from the algorithm within the limitations of the accuracy of the profile data itself.

The present study uses the algorithms to investigate the detailed response of the caustic loci to systematic changes in the melt zone geometry. Only rotationally symmetric homogeneous examples were considered. Results were obtained for a far greater range of melt zone tapers than were originally investigated experimentally.

II. PROCEDURE

None of the four samples discussed in the earlier reports was in fact rotationally symmetric. Three were specifically selected because of their existing asymmetries. It was found that, for certain cases, these geometric asymmetries influenced the far-field caustic patterns quite strongly. In the present study, symmetric profile data were generated by averaging the least asymmetric coplanar profiles of the most gradually tapered sample (Sample 4 with $\beta = 52.3$ degrees). These data were then fitted by the same polynomial spline-fitting routine used earlier[1] to simulate a symmetric version of the original sample as shown in Fig. 1a. There, β is the angle between an axial ray and the outer normal to the profile at the inflection point I. Taking the slope of the outer normal as dx/dy, then

$$\beta = -\arctan dx/dy|_I ,$$

where x is the axial location parameter and y is the radial location parameter. The magnitude and location of the derivative at the inflection point are determined from the spline-fitting routine.

It is worth comparing the calculated values of the caustic half-angles for the symmetric melt zone assuming $n = 1.46$ and the corresponding angles originally reported in Ref. 1. For example, the symmetric 2-intercept caustic half-angle, $\theta_A^{U_1}$, of 98.2 degrees, compares well with asymmetric half-angles of 86.8 and 104.7 degrees, for an average of 95.7 degrees. Similarly, with the 3-intercept family, the symmetric data give a caustic half-angle, $\theta_A^D = 148.5$ degrees, whereas the half-angles of

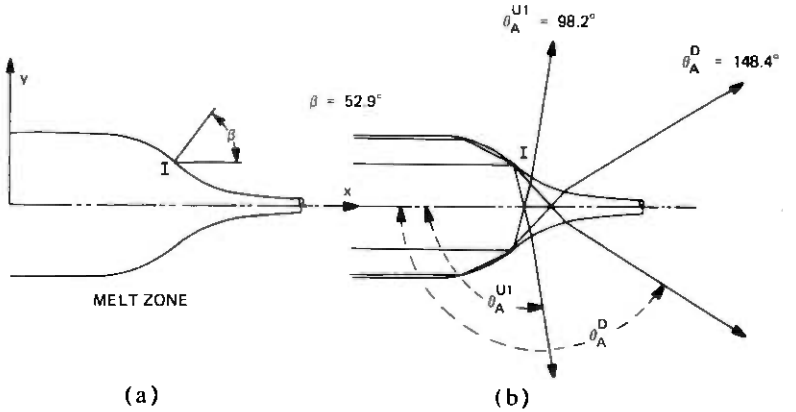


Fig. 1—(a) Symmetric profiles of Sample 4 showing coordinates and inflection point. Here β = 52.92 degrees, comparable to original data. (b) Symmetric profiles of Sample 4 (as above) showing limiting ray paths for 2-intercept caustics with θ_A^{U1} = 98.2 degrees and for 3-intercept caustics with θ_A^{D} = 148.5 degrees.

138.1 and 154.5 degrees yield an average of 146.3 degrees from the original unsymmetric data.*

The effects of changes in the taper of a symmetric melt zone were revealed by stretching one of the data coordinates by a scale factor before each calculation. With this procedure, increasing the scale factor produces an increase in β and hence a more gradually tapered melt zone. Conversely, decreasing the scale factor decreases β and increases the taper. The slope of the outer normal, dx/dy, at any point on the profile clearly varies linearly with the scale factor.

To quantitatively determine the caustic half-angle, θ_A, as a function of geometry over the maximum possible range, over 130 differently scaled melt zones were analyzed. To establish the generality of these results, the analysis was repeated using data derived from the least symmetric and most sharply tapered of the original samples (Sample 3 with $\beta \simeq 27.5$ degrees).

III. THE CAUSTICS

The two principal families of caustics are of primary interest because they appear over the greatest range of tapers. The first of these caustics is formed by light which reflects internally from a given side, crosses the axis of the melt zone, and is refracted out of the opposite side as shown in Fig. 1b. The second caustic family is due to light which makes two reflections on the initial side, then crosses the axis and is refracted out of the opposite side, as also shown in Fig. 1b. Hereafter, these first

* All angles are measured from the upstream axial direction (see Fig. 1b), whereas in Ref. 1 caustic half-angles for the 3-intercept family were measured from the opposite direction.

and second caustic families will be referred to as "2-intercept" and "3-intercept," respectively.*

Two caustics of lesser interest are also briefly discussed later in this report. The first of these is a "1-intercept" caustic which, as its name implies, is refracted from the glass on its first interception with the surface. The second arises from light which, like the 3-intercept family, reflects twice from the first side before crossing the sample. However, its interception with the opposite side results in an initial reflection and it then refracts from the glass upon its second interception with that side. This is referred to as a "4-intercept" caustic.

Figure 2 presents a plot of the rays which are refracted from the drawdown zone in a sample with $\beta = 69.3$ degrees when 2-intercept light only is emitted. Illuminating rays propagating at greater radial distances than ray 1 or lesser radial distances than ray 3 intercept the second side at angles greater than the critical angle.† Consequently, they are continuously internally reflected and propagate on down the fiber. All the rays between bounding rays 1 and 3 refract out. Ray 2 represents that ray which is incident at the point of maximum slope (labeled *I* in Fig. 2) and is therefore turned through the greatest angle. That ray consequently forms a catacaustic, i.e., a caustic by reflection, within the glass. This caustic travels across the melt zone and forms a visible external caustic when refracted out on the opposite side. From Fig. 2 we see that rays originating on either side of ray 2 are refracted out at angles greater than the ray initially incident at the inflection point.

It should be observed in Fig. 2 that the rays between 1 and 3 which are initially distributed evenly become concentrated near the caustic ray

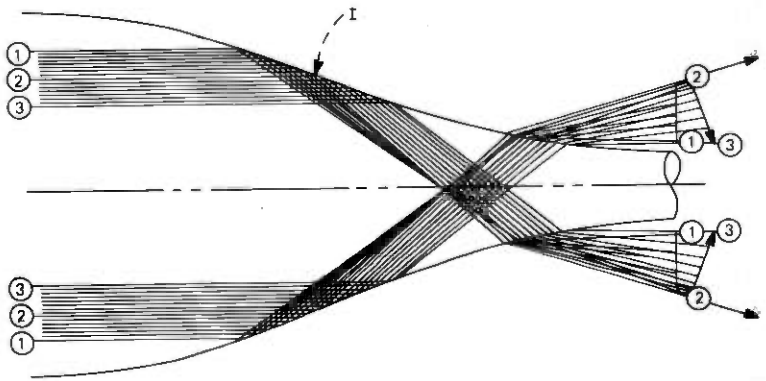


Fig. 2—Symmetric profiles of Sample 4 scaled to give only 2-intercept caustics. The figure shows all trajectories for rays refracting out of the sample between bounding rays 1 and 3 on either side of limiting caustic ray 2. Here $\beta = 69.3$ degrees.

* In Ref. 1, these were referred to as "upstream" and "downstream" caustics because this described their far-field propagation directions as observed in the first of the four original samples studied experimentally.

† Taken as 43.2 degrees, assuming an index of refraction of 1.46 (at the reference wavelength) for fused silica.

as it develops. This concentration of rays symbolizes the intensification of light found along the far-field caustic loci associated with ray 2. Conversely, the rays become widely separated as the bounding rays 1 and 3 are approached, representing a decrease in intensity.

It will be seen that, as β decreases, other 2-intercept caustics appear which are not associated with the internal catacaustic formed at the inflection point. These are due to the refraction of an internal fan of light initially produced by reflection. When the final refraction causes the light to gather into a caustic, it is called a diacaustic.

The 3-intercept caustics also involve an internal catacaustic, due to the interplay between the two initial reflections rather than from the inflection point. In this case, the internal caustic rays originate from rays propagating near the surface of the sample and finally emerge as the far-field caustic rays after refraction. To the best of our understanding, no other 3-intercept rays form externally visible caustics in symmetric melt zones of homogeneous glass.

IV. RESULTS

This section describes the development of the caustic field as a function of the melt zone geometry for the fourth sample. The history of Sample 3 is similar. The reader who is not interested in specific details should proceed to Section V.

We begin by considering melt zone examples with gradual tapers that emit little light and a single caustic. By systematically increasing the taper, we observe an increase in the amount of light emitted and corresponding increases in the number and complexity of the associated caustics. We interpret these results to give the reader a detailed understanding of their significance.

Figure 2 shows a typical distribution of rays throughout the sample including the caustic and bounding ray trajectories. Hereafter, for clarity, we show only the limiting rays (i.e., caustic and bounding rays). However, the reader is reminded that the ultimate intensity distribution is always nonuniform, being much brighter at a caustic and decaying severely as an extinction (e.g., by internal reflection) boundary is approached. Further simplification is effected by separating the graphical information as follows: The melt zone profile and the internal and external caustic rays as they appear are shown in Fig. 3. No other rays are shown. Figures 4 and 5 are "polar plots" of all the far-field limiting rays, for the 2-intercept and 3-intercept families, respectively. The scales are greatly magnified so that the collimated beam is represented as a single horizontal arrow propagating from left to right, and the sample profile is represented as a point. The far-field caustic rays are shown as solid lines and the bounding rays as broken lines. The circumferential arrows follow the continuous fan of light from the outermost bounding ray to the innermost bounding ray, including all caustic rays.

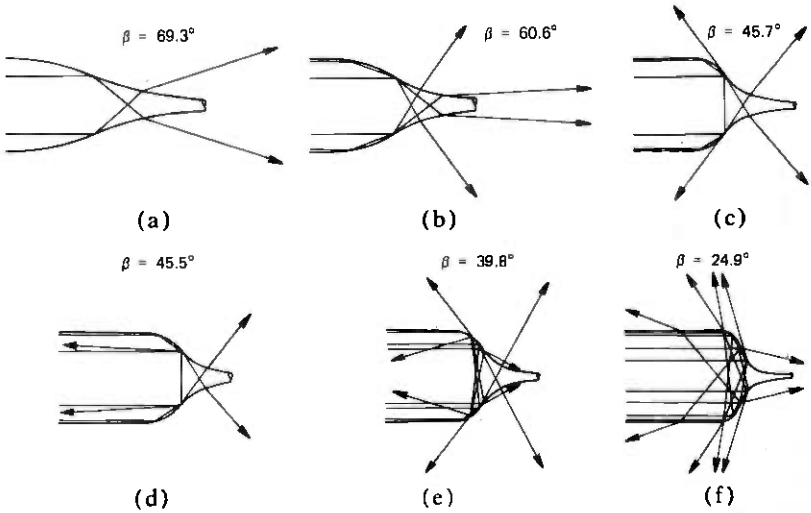


Fig. 3—Symmetric profiles of Sample 4 showing internal and external ray paths for the limiting caustic rays as they appear. (a) $\beta = 69.3$ degrees. Only one 2-intercept caustic is present. (b) $\beta = 60.6$ degrees. Both 2- and 3-intercept caustics are present. (c) $\beta = 45.7$ degrees. Both 2- and 3-intercept caustics are present. (d) $\beta = 45.5$ degrees. Only the 3-intercept caustic is present. (e) $\beta = 39.8$ degrees. 1-, 2-, and 3-intercept caustics are present. (f) $\beta = 24.9$ degrees. 1-, multiple 2-, and 3-intercept caustics are present.

In Fig. 4, the 2-intercept rays are numbered sequentially beginning with the outermost ray 1, which originates near the surface, and increasing inward. The highest numbered ray represents that which initially propagates nearest the sample core and ultimately refracts out.* The 3-intercept rays are lettered A through C or D, as shown in Fig. 5. Each of Figs. 3, 4, and 5 are repeated for a succession of scaled profiles, as shown. The complete range covered in this investigation extends from $\beta = 72$ to $\beta = 6.8$ degrees. In Ref. 1, the corresponding range extended from 53 to about 31.3 degrees.

Figure 3a is the same as Fig. 2, but with the bounding and intermediate rays omitted. The corresponding far-field caustic and bounding ray trajectories are shown in Fig. 4a for the 2-intercept family and Fig. 5a for the 3-intercept family (where the light simply propagates along the sample). Figure 3b shows the emergence of the 3-intercept caustic and a partially rotated 2-intercept caustic. (See also Figs. 4b and 5b.) The 2-intercept caustic originates at the inflection point, as it always must, while the 3-intercept ray originates from a point close to the initial change in the sample cross section, as reported in Ref. 1.

As β continues to decrease, both caustics rotate in an upstream di-

* Actually, of course, the light is not made up of discrete rays but comprises a continuum, including rays on either side of the limiting rays shown. Because of the concentration of rays at the caustic, discreteness errors made in identifying the caustic angles are very small, while those associated with locating the extinction boundaries are substantially larger. The ray spacing used in the present study was optimized so that the caustic angles could be located within +0.1 degree, while the bounding rays are accurate to within no better than ±4 degrees.

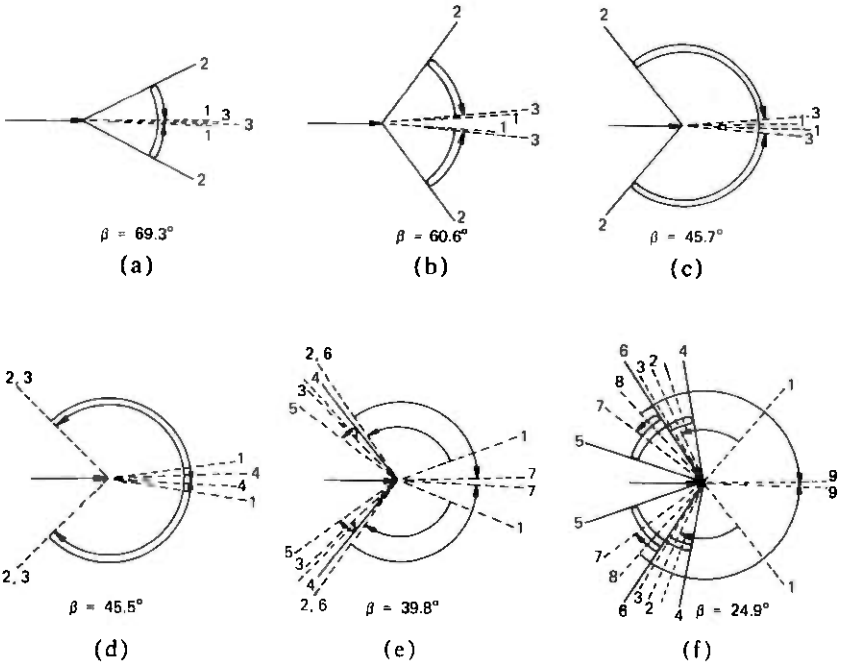


Fig. 4—Schematic diagram showing the external limiting rays for the 2-intercept light paths. The symmetric profiles are omitted, but the orientation is the same as in Fig. 3. The horizontal arrow represents the incoming beam of collimated light illuminating the sample which in turn deflects the light into the plotted ray paths in the far field. Here, the solid lines represent caustics and the broken lines represent bounding rays limited by internal reflection. The circumferential arrows follow the continuous fan of light from the outermost bounding ray to the innermost bounding ray including all caustic limiting rays. The rays are numbered in sequence beginning with 1 as the outermost ray, which originates nearest the surface, and increasing inward. The unnumbered ray in Figs. 4e and 4f represent the last ray after ray 3 which is fully reflected on its first interception with the profiles (except the last two rays, which are again fully reflected). Figures 4a through 4f relate to the same geometrical parameters as Figs. 3a through 3f.

rection, or opposite the rotation of the outer normal at I. Figures 3b and 3c show this rotation quite clearly. Simultaneously, the fans of light forming both caustics broaden (see Figs. 4b and c and 5b and c). However, there is an important distinction between these caustics which may be seen by comparing any two, e.g., Fig. 4c with Fig. 5c. The 2-intercept fan angle is larger and "evenly developed" about the caustic. That is, in Fig. 4c, the fan angle from ray 1 to ray 2 is comparable to the fan angle from ray 2 to ray 3. In contrast, Fig. 5c shows an "unevenly developed" fan of light for the 3-intercept caustic; i.e., the fan angle from ray A to ray B is much greater than the fan angle from ray B to ray C. The reasons for this can be understood by considering how the fans of light are formed internally. The rays forming the 2-intercept caustics are bounded by rays 1 and 3 defined by the internal reflection conditions on the second incident side (see Fig. 2), and therefore give rise to an evenly developed fan. Hereafter, we shall refer to these bounding rays as extinction rays

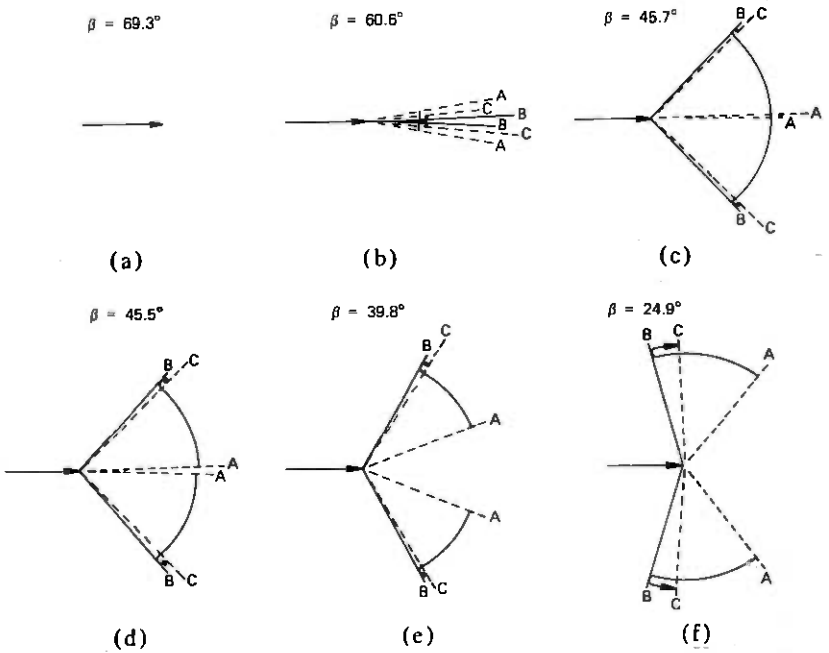


Fig. 5—Schematic diagram showing limiting rays for the 3-intercept light paths. Figure 5 is otherwise the same as Fig. 4, except that the ray sequence is lettered rather than numbered.

because the internal reflection causes extinction of the corresponding external illumination. However, for the 3-intercept family only the outermost bounding ray, A, is an extinction ray defined by the reflection condition on the second side. In contrast, the innermost bounding ray, C, is determined by the presence of the inflection point, which limits the distribution of possible double reflections. In other words, ray C is not an extinction ray as are rays 1, 3, and A, but rather C is a bounding ray determined by the geometrical conditions required for double reflections on the first incident side.

Figures 3d and 4d represent a change of only 0.2 degree from the β of Fig. 3c and show the abrupt extinction of the external 2-intercept caustic due to a second incidence angle in excess of the critical angle, 43.2 degrees. In this case, the 2-intercept catacaustic ray is internally reflected back up the sample, as shown. In Fig. 4d, it can be seen that most of the 2-intercept rays are still incident at angles less than the critical angle and therefore are refracted from the glass. However, two separate fans are formed which are bounded by extinction rays on all sides: rays 1 and 2 and rays 3 and 4. At the same time the 3-intercept caustic simply continues to rotate upstream as seen in Figs. 3d and 5d. As the taper is increased further, β reaches the critical angle at a minimum outer normal slope of 0.98, causing most of the light in the catacaustic ray to refract

from the sample on the first intercept. This results in the family of 1-intercept caustics associated with the inflection point mentioned earlier. Since the optical mechanism by which the light is concentrated is refraction, this is also a diacaustic. Initially, the only light emitted is the ray at the inflection point which exits tangent at the surface (at a half-angle of 226.8 degrees) and follows the glass surface downstream. As β increases, the amount of light refracted increases and the caustic ray refracts from the glass at decreasing angles. It is important to note that this caustic half-angle is always greater than 180 degrees; i.e., it is the only caustic that always propagates toward the axis of the sample.

At $\beta = 40.7$ degrees, a new 2-intercept caustic emerges. Here the two fans of light bounded by extinction rays are still present. In addition, there is a small bundle of rays, including a caustic ray lying between them. This caustic is termed a second 2-intercept caustic because it differs considerably from the original 2-intercept caustic, as may be understood from Figs. 3e and 4e. In Fig. 3e, the usual 3-intercept caustic can be seen originating from the outermost illuminating ray. At this point, it refracts out almost normal to the surface of the melt zone. Therefore, it depends only on the geometry of the melt zone and it is nearly independent of the magnitude of the refractive index. The innermost illuminating ray is incident at the inflection point; it gives rise to two caustics: the external diacaustic (the 1-intercept family discussed above) and the catacaustic associated with the original 2-intercept caustic which has now been reflected internally. A new caustic ray is now shown between these illuminating rays. This new ray gives rise to the new 2-intercept caustic heading upstream in Fig. 3e and is designated ray 4 in Fig. 4e. Then one obvious difference between the second 2-intercept caustic and the original 2-intercept caustic is that the second 2-intercept caustic is not generated from the inflection point. Figure 4e illustrates a second difference: namely, that extinction rays 3 and 5 propagate at smaller angles than caustic ray 4, rather than at larger angles as before. Recall, for example, Fig. 4c, where extinction rays 1 and 3 propagate at much larger angles than caustic ray 2. In other words, the new caustic is formed by a folding in the opposite direction.

Simultaneously, the fan of light formed by the 1-intercept caustic broadens as the band of illuminating rays refracting at the first interception broadens. This band is evenly developed about the inflection point ray, or I ray, and has now expanded to include the rays giving rise to the new 2-intercept caustic. Consequently, most of the light from rays forming this caustic has refracted at the first interception as part of the fan of light which forms the 1-intercept caustic; thus, the second 2-intercept caustic is much less intense than the original. When the second 2-intercept caustic first emerges, however, the band of illuminating rays refracting at the first interception was so small that it did not yet include the rays contributing to the new caustic. This is the case

at $\beta = 40.7$ degrees, where this caustic was still relatively intense. The loss of intensity is illustrated in Fig. 4e by an unnumbered ray representing the last one that is fully reflected on its first interception, shown between rays 3 and 4. Referring to fan 3, 4, and 5, the rays between 3 and the unnumbered ray are totally internally reflected on their first interception and are therefore intense. But the rays between the unnumbered ray and ray 4, then continuing to extinction ray 5, all are less intense since much light is lost on their first interception. Note that all other fans of light are totally reflected on the first interception and are also intense.

Concurrently, it was found that the first extinction rays 1 and A rotate away from their original downstream directions as β decreases from 40.7 degrees. This is not the case for final extinction ray 7, while extinction ray C continues to stay close behind the 3-intercept caustic as it also swings upstream.

At $\beta = 37.3$ degrees, the 2-intercept catacaustic ray, partially reflected at the inflection point from the second side, has re-emerged propagating in the upstream direction. At this stage, it does not form an external caustic as it did earlier. This ray, referred to as the "I ray," is not associated with any stationary point in the angular ray distribution and is located in the intermediate fan of rays. Though theoretically originating as a catacaustic limiting ray within the glass, its caustic character is momentarily lost to external observation because of the dominant effects of the severe spread in refraction angles approaching the extinction ray. Beginning with Fig. 4e, this unnumbered I ray appears between extinction ray 3 and the new caustic ray 4, such that both the caustic and the I ray can be expected to be rather less intense. While the 1- and 3-intercept caustics rotate upstream in the directions of smaller θ_As with decreasing βs, the new 2-intercept caustic rotates in the opposite direction—downstream toward greater caustic angles. In addition, its radial position in the illuminating beam has moved farther out from the center of the sample, away from the I ray and toward the illuminating ray for the 3-intercept caustic. This outward displacement of its illuminating ray and its retrograde rotation are two other properties which make the second 2-intercept caustic different from the first.

When β is decreased further to 34.5 degrees, the first 2-intercept caustic reforms from the I ray, along with yet another 2-intercept caustic. The illuminating ray for the newest caustic propagates just inside the I ray and emerges in the far field barely downstream from the re-emergent first 2-intercept caustic ray. Both these rays are of reduced intensity, since they originate within the band of rays which lose most of their light by refraction on the first interception with the boundary. As the taper is decreased still further, a continued folding of this intermediate fan takes place, and the three 2-intercept caustics separate in the far field. The illuminating rays for the second and third 2-intercept caustics move

away from the I ray. At the same time, the illuminating ray for the third caustic moves in toward the surface. The third 2-intercept caustic, like the second, is folded in a direction opposite that of the first 2-intercept caustic and shows an initial rotation toward smaller θ_As.

One additional point should be noted. In Fig. 3e, soon after the second 2-intercept caustic appeared, its point of emergence moves out along the profile with decreasing β and becomes stationary at the "start" of the melt zone, i.e., the location where the sample begins its decrease in cross section. At the same time, it continues retrograde rotation. On the other hand, the first 2-intercept caustic reappeared as soon as the emergent point of the I ray moved downstream to the same point, so that the outer normal angle becomes constant along the surface and the internal catacaustic could emerge intact with only a change in direction. Subsequently, as β is decreased, the point of emergence of the first 2-intercept caustic moves upstream along the sample surface where the diameter is constant, while the emergent points of both the second and third 2-intercept caustics merge together at the start of the melt zone. Ultimately, both the 3-intercept and the first 2-intercept caustics approach extinction by internal reflection, the latter for a second time as its propagation angle approaches 0 degree. Consequently, their rates of rotation seem to accelerate. From $\beta = 24.9$ degrees (Fig. 3f) to $\beta = 22.3$ degrees, both original caustics disappear by internal reflection, leaving only the second and third 2-intercept and the single 1-intercept caustics in the emerging light field. Figure 4f shows the complicated 2-intercept far-field light ray pattern prior to extinction, with many overlaps between the various fans of light contributing to the structure of the observed illumination. Here the first caustic, ray 5, is once again the extreme upstream ray while all other caustics and extinction boundaries are spread out in the light regions behind it. The unnumbered I ray still lies between ray 3 and the second 2-intercept caustic, ray 4, such that all of the 2-intercept caustics may appear relatively less intense than either the 1- or 3-intercept caustics. At this point, the third 2-intercept caustic, ray 6, has reversed its initial upstream rotation and, like the second 2-intercept caustic, assumed a retrograde rotation with changing slope. In Fig. 5f, $\beta = 24.9$ degrees, the fans of light behind the 3-intercept caustic can be seen to start to broaden slightly just prior to the disappearance of this caustic, while by $\beta = 22.3$ degrees the caustic ray is internally reflected and only the overlapping fans, bounded by extinction boundary rays, and the one geometrical boundary ray, remain. The two remaining 2-intercept caustics (with the original caustic extinguished) continue to rotate toward the downstream direction.

At this geometry, we have probably passed beyond the practical limit for most melt zone profiles drawn in a laser furnace. However, if it were possible to draw even blunter profiles, we would observe that the second 2-intercept caustic angle would approach 90 degrees while the 1-intercept

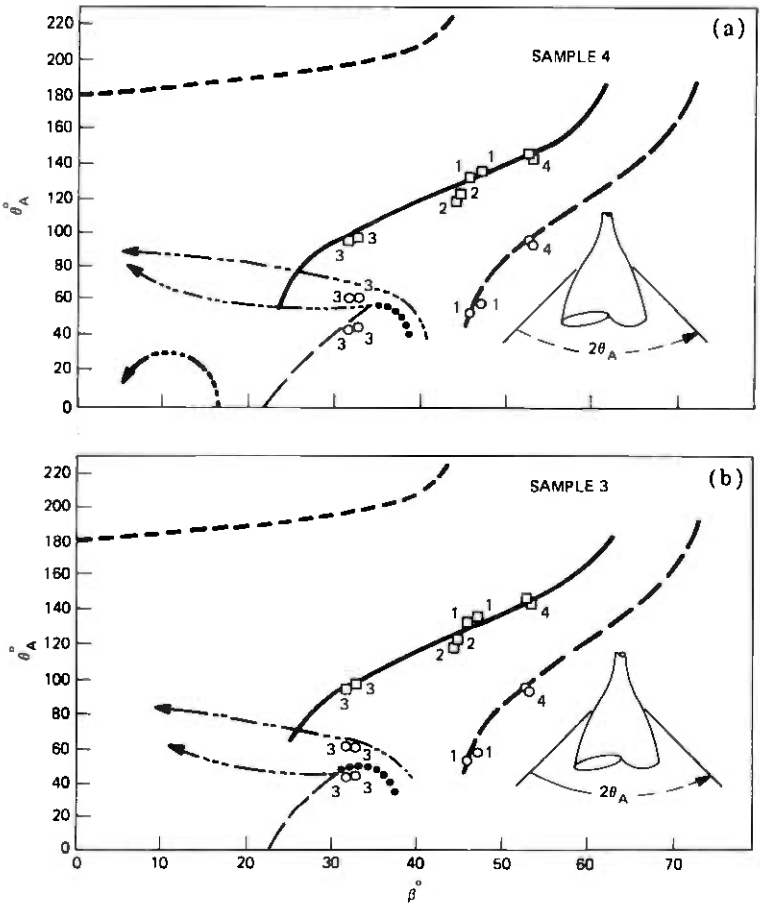
caustic angle would approach 180 degrees. At the same time, the internally reflected 3-intercept caustic would reappear, at $\beta \cong 16.4$ degrees, although somewhat altered. That is, following the two reflections on the first side of the sample and the initial reflection on the second side, the family forms an internal catacaustic which refracts out of the sample on its next interception with the second side. Since this is a fourth interception overall, it represents the emergence of a 4-intercept caustic rather than a re-emergence of the original 3-intercept caustic. Further blunting of the profile causes the caustic to rotate in the retrograde direction to a maximum caustic half-angle of about 50 degrees, somewhere between $\beta = 14.0$ and $\beta = 13.4$ degrees. It then reverses direction and disappears once more by internal reflection, at $\beta = 6.8$ degrees, leaving only the 1-intercept caustic, which would be very bright, and the second and third 2-intercept caustics, the first of which would also have become bright. Intensification of the light associated with the second 2-intercept caustic results from the outward migration of the associated illumination ray. It eventually reaches a point so near the start of the melt zone surface as to fall incident on the now rapidly curving profile at an angle greater than the critical angle. Consequently it is wholly reflected and all its light is transmitted across the sample to refract out and form the caustic.

V. SUMMARY

The variations of the angles of various 1-, 2-, and 3-intercept caustics as functions of β for both Sample 4 and Sample 3 are given in Figs. 6a and 6b. Significant values are also listed in Table I. In this study, we have identified two additional 2-intercept caustics. Comparison of Figs. 6a and 6b and the tabulated results show that, with the exception of these two caustics, the results for the morphologically different samples are numerically quite similar. Figure 6 also offers a comparison with the results of the original study.[1] For all four samples, these results compare very favorably. Since in Fig. 6 we are comparing results from symmetric samples with the average values obtained from often rather unsymmetric samples, this agreement is quite remarkable.* It is worth noting that in Ref. 1 both the experiments and the original analysis identified only two 2-intercept caustics for Sample 3. In fact, there are three such caustics present. The taper of Sample 3 was such that the first and third 2-intercept caustics were barely separated and consequently appeared to the observer as one.

Turning to the more general results, we see that, for a shallow taper, no light emerges. When β falls below 73 degrees, the 2-intercept caustic emerges, initially directed downstream toward the x-axis and along the surface at an angle, $\theta_A \simeq 190$ degrees. As the taper becomes steeper

* It indicates that, while melt zone asymmetry may have a strong influence on caustic asymmetry, it may have relatively little influence on determining the total caustic cone angle, $2\theta_A$.



CAUSTIC ANGLE, θ_A, VS. INFLECTION POINT ANGLE, β

- SINGLE INTERCEPT CAUSTIC
- DOUBLE INTERCEPT CAUSTIC
- SECOND DOUBLE INTERCEPT CAUSTIC
- THIRD DOUBLE INTERCEPT CAUSTIC
- TRIPLE INTERCEPT CAUSTIC
- QUADRUPLE INTERCEPT CAUSTIC
- I-RAY ALONE
- Ⓡ DOWNSTREAM DATA, SAMPLE R
- Ⓡ UPSTREAM DATA, SAMPLE R

Fig. 6—Plots of the limiting caustic half-angle, θ_A, vs the angle of the outer normal at the inflection point, β. (a) For symmetric Sample 4. (b) For symmetric Sample 3. The numbered data points represent averaged values of the caustic angles measured on the four original samples described in the earlier study (Ref. 1). Heavy lines denote bright caustic; light lines denote dim caustic.

(β decreases), the caustic ray swings toward the upstream direction. It reaches a half-angle, θ_A, normal to the surface at $\beta \simeq 58$ degrees and a $\theta_A = 90$ degrees at a $\beta \simeq 51$ degrees. Thereafter, its rate of rotation accelerates as it approaches extinction at $\theta_A \simeq 52$ degrees or $\beta \simeq 45.7$ degrees. Recalling that this caustic is the one formed at the inflection point,

Table I—Significant values of limiting caustic half angle

Station \ Caustic Sample Function	2-Intercept				3-Intercept			
	4		3		4		3	
	$\beta°$	$\theta_A^{U1°}$	$\beta°$	$\theta_A^{U1°}$	$\beta°$	$\theta_A^{D°}$	$\beta°$	$\theta_A^{D°}$
At Maximum β	71.7	189.6	72.7	188.5	61.3	186.7	62.6	184.0
At $\theta_A \perp$ Surface	58.5	118.9	58.0	115.0	38.5	116.0	41.6	120.0
At $\theta_A \perp$ X-Axis	51.0	90.0	51.2	90.0	28.4	90.0	29.8	90.0
At Minimum β	45.7	51.9	45.8	52.3	23.7	55.2	25.7	67.2
	2-Intercept				Third 2-Intercept			
	4		3		4		3	
	$\beta°$	$\theta_A^{U1°}$	$\beta°$	$\theta_A^{U1°}$	$\beta°$	$\theta_A^{U3°}$	$\beta°$	$\theta_A^{U3°}$
At Maximum β	34.5	58.5	31.0	46.5	34.5	58.5	31.0	46.5
At Minimum θ_A	23.4	0.0	22.8	0.0	29.6	55.5	26.9	45.3
At Minimum β	23.4	0.0	22.8	0.0	(0.0)*	(90.0)	(0.0)	(90.0)
	Second 2-Intercept				1-Intercept			
	4		3		4		3	
	$\beta°$	$\theta_A^{U2°}$	$\beta°$	$\theta_A^{U2°}$	$\beta°$	$\theta_A^{1°}$	$\beta°$	$\theta_A^{1°}$
At Maximum β	40.7	40.0	38.4	48.0	43.2	226.8	43.2	226.8
At Minimum β	(0.0)	(90.0)	(0.0)	(90.0)	0.0	180.0	0.0	180.0

* Numbers in parentheses were extrapolated and not actually calculated.

it should be noted that, at $\beta \simeq 37.5$ degrees, the internal catacaustic ray, or I ray, again emerges. However, it is not seen as an external caustic until $\beta < 34$ degrees. A different 2-intercept caustic does become visible when $\beta \simeq 40$ degrees. This caustic is not associated with the internal caustic formed by reflection at the inflection point. It also differs from the original 2-intercept caustic in that, as the taper increases, it rotates in the opposite direction, i.e., toward greater angles. It may also emerge initially as an intense caustic, but it becomes quite dim once the incidence angle of the illuminating beam falls below the critical angle. When $\beta < 16.7$ degrees, this new caustic ray is again illuminated by a ray initially at an angle greater than the critical angle and so it again becomes bright. The half-angle, θ_A, of this caustic approaches 90 degrees asymptotically as $\beta \to 0$ degree. Referring again to the I ray, we see that its rotation is also initially retrograde as β decreases. However, when it again becomes the leading 2-intercept caustic ray, its direction of rotation reverses. From this point, until its extinction at $\beta \simeq 23$ degrees, the reemergent 2-intercept caustic is much less intense. Simultaneously, a third, dim, 2-intercept caustic appears which ultimately rotates in the direction of increasing caustic angle, like the second 2-intercept caustic that preceded it.

Careful study of Figs. 6a and 6b and Table I shows that over most of its range above $\beta = 46$ degrees the 2-intercept caustic θ_A vs β relationship is very nearly the same for both samples. However, this is not true of the second and third 2-intercept caustics, which differ from the first in a number of significant ways. First, the far-field ray trajectories fold in opposite directions. Second, both new caustics rotate to larger rather than smaller caustic angles with decreasing β. Third, they are independent of the internal catacaustic originating at the inflection point. This

latter observation is significant, because while the functional forms of both samples are similar, the additional caustics are quantitatively the most highly differentiated. This differentiability is a result of their originating from rays initially incident far away from the inflection point. There they are more strongly influenced by other aspects of melt zone morphology than just β. More of this will be discussed later in this paper.

The 3-intercept caustic is also less involved with the inflection point and shows some differentiation between the two samples over its entire range (see Table I). Since its initial emergent caustic angle depends on the downstream surface geometry in much the same way as it did for the original 2-intercept caustic, it is not surprising that it too should result in an initial caustic half-angle, θ_A, significantly greater than 180 degrees, although for a 10-degree smaller value of β. As shown in Table I, the final 3-intercept caustic angles are quite different. These extinction angles appear at βs about 22 degrees smaller than their respective 2-intercept caustic's initial extinction β angles and at very nearly the same final extinction β angles as those of the re-emergent first 2-intercept caustic. Actually, the light rays associated with the 3-intercept caustic also eventually make a reappearance. As shown in Fig. 6a for Sample 4, a bright 4-intercept caustic resulting from the internally reflected 3-intercept caustic does emerge briefly headed in the upstream direction and then disappears, all at βs too small to be physically significant.

There is a diacaustic which emerges when β falls below the critical angle. This caustic is formed by the light which refracts from the surface on its first interception. The I ray which forms the internal catacaustic and, ultimately, the 2-intercept caustic, also attains an external extremum when it forms the 1-intercept caustic. Because little of the light in the illuminating rays is reflected, once the 1-intercept caustic appears, many (but not all) of the 2-intercept rays which appear at βs below the critical angle are quite dim. Consequently, over most of this range the 2-intercept caustics are also dim. This is indicated by the light line weights used to represent them in Fig. 6.

While the second and third 2-intercept caustics are unique in that they never propagate downstream or toward the fiber axis, the 1-intercept caustic is also unique because it always propagates downstream toward the fiber axis. Consequently, the 1-intercept caustic either follows the surface or reflects off the fiber at some station downstream, such that its fundamental rotation as a function of β depicted in Fig. 6 becomes reversed. Since this caustic depends only on β, its angle is a unique function of β for all melt zone profiles,

$$\theta_A^1 = 180° - \beta + \arcsin(n \sin \beta),$$

where n is the index of refraction. Hence, for $n = 1.46$ it always originates following the surface downstream at an angle of 226.8 degrees at $\beta = 43.2$ degrees and then rotates toward a limit of 180 degrees.

We conclude that the most useful caustics for studying fiber-drawing melt zones are the first 2-intercept caustic and the 3-intercept caustic. Between them, these cover a broad range of melt zone geometries. Except at very large βs, the first 2-intercept caustic depends mostly on the melt zone profile near the inflection point and does little to distinguish between samples. In contrast, the 3-intercept caustics readily distinguish between the different samples. Sample 3 is on the average larger by about 1.5 degrees in β than Sample 4, a difference originating from variations in geometry near the shoulder and heel of the profiles. Figure 7 presents plots of both sample profiles scaled to a common β of 51.3 degrees. It can be seen that Sample 4 is significantly larger than Sample 3 both upstream, at the shoulder, and downstream, at the heel of the profile. In addition, the inflection point of Sample 4 is at a greater radius, and upstream, of that for Sample 3. Since the β angles are the same, these differences produce a less than 0.7-degree change in the 2-intercept caustic angle θ_A^{U1}. However, the 3-intercept caustic involves reflections at the shoulder region of the melt zone and a refraction at the heel region downstream, and it manifests a change in the caustic angle θ_A^D more than three times that in θ_A^{U1}. As reported in Ref. 1, this difference is detectable experimentally.

Though less intense and of rather limited range, the second and third 2-intercept caustics are also potentially useful as a means of studying melt zone geometries. These are the most distinguishable caustics because they involve light which is initially incident either at the shoulder of the melt zone, as with the second 2-intercept caustic, or at the heel

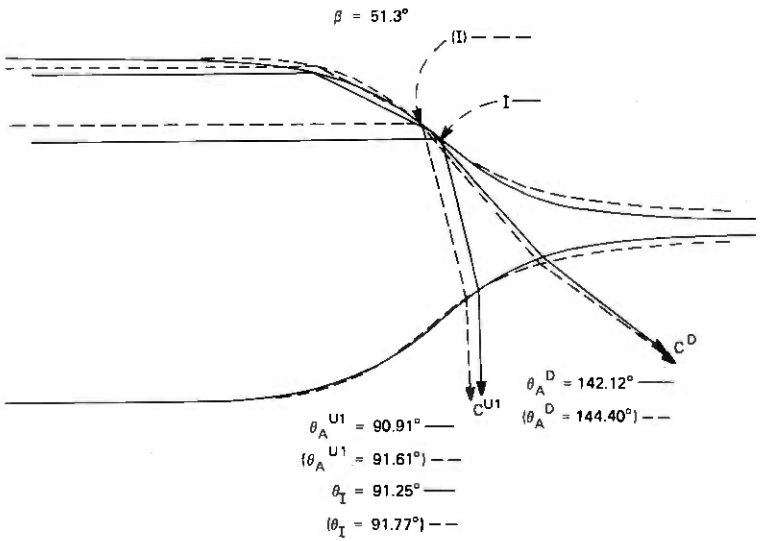


Fig. 7—Superimposed symmetric profiles from Sample 4 (dashed lines) and Sample 3 (solid lines) showing limiting ray paths for both 2- and 3-intercept caustics, one side only. Here, β = 51.3 degrees for both samples with data from Sample 3 shown in parentheses.

of the melt zone as with the third. Both caustics exhibited numerically greater angles for Sample 3 than for Sample 4 whose shoulder and heel profiles induce greater rotations in the reflected rays. Indeed, the averaged experimental results for these caustics from Sample 3 agree reasonably well with the values computed from the symmetric Sample 3 profile data and rather poorly with the values computed from the symmetric Sample 4 profile data, Figs. 6a and 6b, respectively. These caustics are the only primary ones still visible for $\beta < 22$ degrees, and together with the 1-intercept caustic provide some means of studying extremely blunt melt zones.

It should be realized that, since each of these six caustics under discussion involves a refraction from the sample, they all may depend on the index of refraction, n, as well as on the surface geometry. In Ref. 1, another caustic generated wholly by reflection and dependent only on β was discussed. That caustic is the externally illuminated equivalent of the internal catacaustic that forms at the inflection point and can easily be generated for all physically reasonable melt zone profiles. Since it does not depend on n, information from this externally illuminated catacaustic may be used to separate the β dependence of the data obtained from an appropriate internally illuminated caustic. This could provide otherwise unavailable information concerning the index of refraction. For example, consider a melt zone sample of unknown n exhibiting a first 2-intercept caustic at an angle θ_A^{U1} somewhere between 52 and 110 degrees. It may be matched with a computer analysis of the present data at a β determined from direct measurement of the externally illuminated catacaustic angle, θ_A^{EXT}. Adjustment of the n value used in the computer analysis to yield an equal theoretical θ_A^{U1} value should provide a good estimate of the unknown index of refraction. The quality of this estimate would depend on how uniform n was across the sample and how close the propagation vector of the emerging caustic was to tangency with the surface. Obviously, if β is such that the caustic emerges normal to the surface, it can provide no information about n.

VII. ACKNOWLEDGMENT

The authors wish to acknowledge the collaboration of J. McKenna, who contributed to the development of the numerical algorithm.

REFERENCE

1. P. G. Simpkins, T. D. Dudderar, J. McKenna, and J. B. Seery, "On the Occurrence of Caustics in the Drawdown Zone of Silica Fibers," B.S.T.J., *56*, No. 4 (April 1977), pp. 535–560.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Caustic Patterns Associated With Melt Zones in Solidified Glass Samples—Part II: Asymmetric Cases

By J. B. SEERY, T. D. DUDDERAR, and P. G. SIMPKINS

(Manuscript received March 17, 1978)

Two types of geometrical asymmetries have been analyzed numerically to determine their influence on the externally visible caustic patterns from a drawdown zone. The first of these asymmetries involved axial displacement of identical opposing profiles of the drawdown zone, while the second involved opposing profiles of differing slope positioned with their inflection points in axial alignment. The results of these studies are presented graphically as an aid to interpreting the asymmetries of actual caustic patterns generated by illuminating real drawdown zone samples in optical fiber drawing.

I. INTRODUCTION

In the present study, we demonstrate how deviations from rotational symmetry in the melt zone give rise to asymmetries in the resulting caustic pattern. In Ref. 1, rotationally symmetric profiles of melt zones were generated by averaging measurements of data taken from solidified melt zone samples. These data provided symmetrically idealized data sets for the study of the two principal types of caustics as functions of the rate of change in the cross section. Changes in the maximum rate of change in the cross section were simulated by scaling* the data sets, and the analysis was carried out using the algorithms developed earlier in Ref. 2. These algorithms describe families of 2- and 3-intercept caustics as shown in Fig. 1a, where the scale factor is unity. The incident angle, β, for an axial light ray at the inflection point, I, is 52.9 degrees on both sides of the cross section. However, in Ref. 2 it was shown that the samples were usually asymmetric, resulting in pronounced asymmetric caustic patterns, especially when the caustic rays emerged nearly tangent to the surface.

* That is, by scaling the axial and transverse data values differently with the scale factor, SF = (axial scale)/(transverse scale).

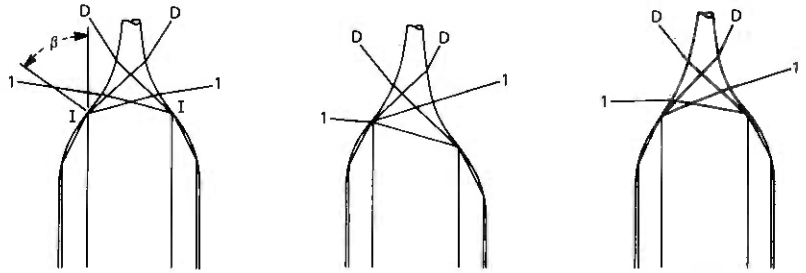


Fig. 1—Drawdown profiles for a melt zone with an average γ of 1.323 and a preform radius, R, of 0.1217 in. (3.09 mm). The data scale factor, *SF*, is 1. The rays labeled *D* represent the 3-intercept caustic trajectories and the rays labeled 1 represent the first 2-intercept caustic trajectories for the following cases: (a) Symmetric drawdown profiles with $\gamma = 1.323$ ($\beta = 52.9$ degrees) on both sides and no shift ($\Delta/\gamma R = 0.0$, $\delta\gamma = 0$ percent). (b) Asymmetric drawdown profiles with $\gamma = 1.323$ ($\beta = 52.9$ degrees) on both sides and an axial shift of $\Delta = 0.05$ in. (1.27 mm) of the left side ahead of the right side ($\Delta/\gamma R = 0.311$, $\Delta/R = \pm 0.41$). Here $\delta\gamma = 0$ percent. (c) Asymmetric drawdown profiles with $\gamma = 1.455$ ($\beta = 55.5$ degrees) on the left side and $\gamma = 1.1907$ ($\beta = 50.0$ degrees) on the right side (average $\gamma = 1.323$, $\delta\gamma = \pm 10$ percent). Here $\Delta/\gamma R = 0.0$.

Two types of asymmetric drawdown zones were studied, based upon the symmetric data set generated from the fourth sample.[1,2] In the simplest case, the asymmetry was introduced by axially shifting the profile data for one side of the cross section with respect to the other. An asymmetric cross section with the caustic rays is shown in Fig. 1b, where the relative axial shift between the two profiles is 41 percent of the preform radius* and the scale factor is 1.0. The asymmetry produces significant counterclockwise rotations in the propagation directions of the emerging caustic rays when viewed with the downstream profile as shown on the left.

A second type of asymmetry was produced by scaling the opposite sides differently. In this study, the inflection points of the opposite (differently scaled) profiles were kept in axial alignment so that the resulting shift in the caustic ray pattern would be due solely to the difference in the slopes, as shown in Fig. 1c. There the scale factors of opposite sides are 0.9 and 1.1. These produce counterclockwise rotations of the emerging caustic rays comparable to those shown in Fig. 1b and toward the elongated profile as plotted.

In addition, for both types of asymmetry, the overall scaling was changed so that the effects of these asymmetries could be assessed for greater or lesser average gradients than those of the original data. These scale factors and associated average slopes, γ, and incident angle values, β_A, are tabulated in Table I. Here γ is the average of the slopes of the two outer normals at the inflection points, dx/dy_I, and $\beta_A = -\text{arc tan } \gamma$.

The next two sections present the results from the computer simulations including diagrams of significant cases for both types of asymme-

* $R = 0.1217$ in. or 3.09 mm.

Table I—Scale factors and average slopes and angles

Scale Factor	Average Outer Normal Slope γ	Average Incidence Angle β_A (degrees)
1.300	1.720	59.8
1.200	1.588	57.8
1.000	1.323	52.9
0.775	1.025	45.7
0.700	0.926	42.8
0.520	0.688	34.5
0.450	0.595	30.8
0.340	0.450	24.2

tries. The reader who requires a less detailed synopsis of these results is referred to the discussion portions of these sections and to Section V, Summary.

II. SHIFT ASYMMETRY

The effects of axial misalignments in the profiles of the drawdown zone were studied by computing the ray paths for examples with the profiles of opposite sides shifted various amounts, Δ. The results of these computations provide a complete description of the caustic patterns for an asymmetric drawdown zone with a range of profile shifts for each γ given in Table I.

To compare the varied scalings, an asymmetry parameter, $\Delta/\gamma R$, was established, where R is the preform radius. The asymmetry was varied from $0 \leq |\Delta/\gamma R| \leq 1.20$, which represents a broader range than is ever likely to actually occur in fiber drawing. The specific incremental changes in asymmetry were chosen to illustrate what occurs as the asymmetry increases.

An extreme example of the effects of this type of asymmetry can be seen by comparing Fig. 2a with Fig. 2b, which has the same average slope but a large asymmetry. Emergent 2-intercept caustics are labeled numerically to indicate which of the family they represent; the 3-intercept caustics are labeled D. The unlabeled rays represent caustic rays of the 1-intercept family which emerge at the inflection point for γs less than 0.939. In Fig. 2b, the 3-intercept caustics have rotated counterclockwise, or in a positive direction, toward the leading side. Extra 2-intercept caustics (labeled 2′, 2″, 3′) have merged on the leading side, while all 2-intercept caustics have disappeared from the trailing side. Those 2-intercept caustics on the $+\Delta$ side have rotated clockwise, in a negative direction, away from the leading side.

2.1 Graphical results

Figures 3a to 3d are simplified versions of the format used in Figs. 1 and 2 which show both the internal ray paths and the external caustic trajectories. In these simplified diagrams, the internal ray paths have

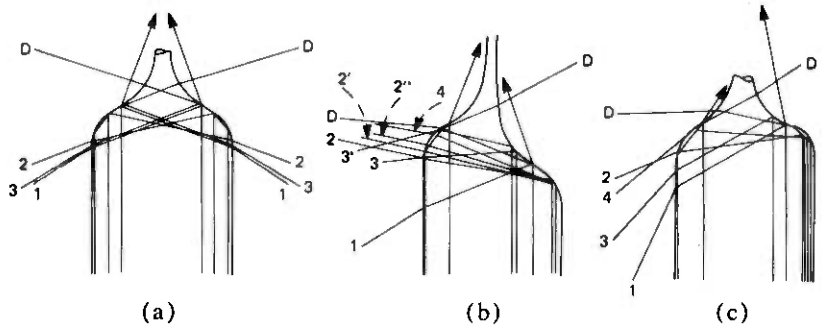


Fig. 2—Drawdown profiles for a melt zone with an average γ of 0.688 (SF = 0.52). The rays labeled D represent the 3-intercept caustic trajectories, while the rays labeled 1, 2, 3, etc., represent the first, second, third, etc., 2-intercept caustic trajectories for the following cases; (a) Symmetric drawdown profiles with γ = 0.688 (β = 34.5 degrees) on both sides and no shift, $\Delta/\gamma R$ = 0.0, $\delta\gamma$ = 0 percent. (b) Asymmetric drawdown profiles with γ = 0.688 (β = 34.5 degrees) on both sides and a shift, Δ = 0.075 in. (1.905 mm) of the left side ahead of the right side ($\Delta/\gamma R$ = ±0.896, Δ/R = ±0.616). Here, $\delta\gamma$ = 0 percent. (c) Asymmetric drawdown profiles with γ = 0.894 (β = 41.8 degrees) on the left side and γ = 0.482 (β = 25.7 degrees) on the right side (γ = 0.688, $\delta\gamma$ = ±30 percent). Here, $\Delta/\gamma R$ = 0.0.

been eliminated, and the drawdown profiles have been reduced to tangent lines at the inflection points. The profiles, shown as thick lines, meet in the symmetric case and progressively move farther apart to depict increasing ±Δs. The external caustic trajectories are shown schematically as rays radiating from the point Δ = 0. The 3-intercept caustics are again labeled D and, in the unusual case of a second 3-intercept caustic, $D2$ is used. Whenever the ray which internally reflects from the inflection point, I, is refracted from the opposite side without forming an external caustic, it is shown by an unlabeled dashed line. In Figs. 3a to 3d, the diagrams are drawn with the leading side as the left profile, consistent with Figs. 1b and 2b. We then define a positive, or forward, rotation of caustic angles to be counterclockwise, or toward the leading side.

We begin with gradually tapered profiles that emit little light. As the taper is increased, we observe an increase in the amount of light emitted and in the complexity of the caustic structure. The first case, with γ = 1.72, is shown in Fig. 3a. Both caustics present initially rotate in a forward direction as the asymmetry increases. However, the 3-intercept caustics quickly disappear from the trailing side by internal reflection. The 2-intercept caustics continue to rotate in a positive direction on the trailing side, but on the leading side they reverse direction and rotate in a negative direction at an asymmetry $\Delta/\gamma R$ of about 0.40. The situation is similar for γ = 1.588, not shown. The 3-intercept caustics also rotate forward. Since the emergent angle for the symmetric case is smaller this time (163.48 degrees as opposed to 172.66 degrees), the caustic does not internally reflect as soon. The 2-intercept caustic initially rotates forward. This time, the direction reversal on the leading side occurs earlier at a $\Delta/\gamma R$ of 0.33 and is more pronounced. Reducing γ to 1.323 does not

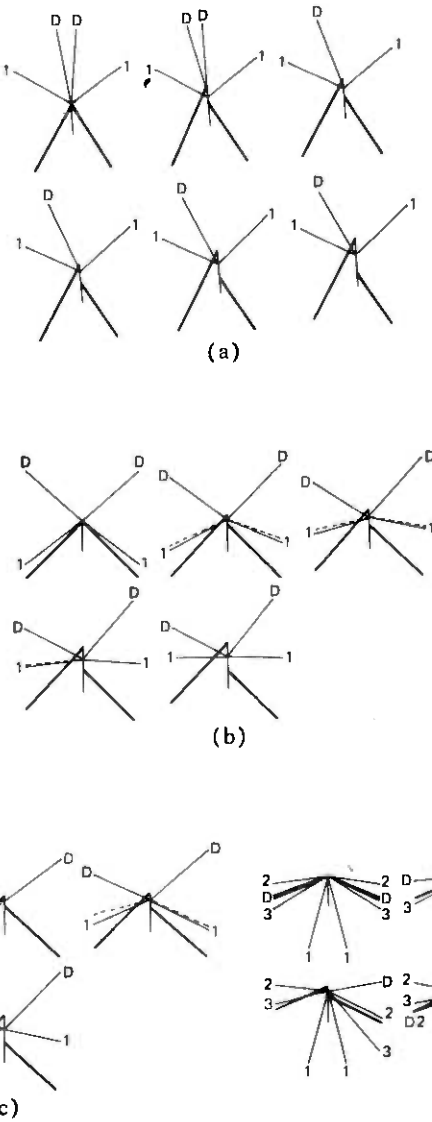


Fig. 3—Diagrams of external caustic trajectories for drawdowns of varying asymmetry due to axial shifts, $\Delta/\gamma R$, of varying amounts as follows: (a) For $\gamma = 1.720$ ($\beta = 59.8$ degrees), $\Delta/\gamma R = \pm 0.000, \pm 0.118, \pm 0.236, \pm 0.354, \pm 0.472, \pm 0.596$. (b) For $\gamma = 1.025$, ($\beta = 45.7$ degrees), $\Delta/\gamma R = \pm 0.000, \pm 0.199, \pm 0.410, \pm 0.596, \pm 0.801$. (c) For $\gamma = 0.926$ ($\beta = 42.8$ degrees); $\Delta/\gamma R = \pm 0.000, \pm 0.223, \pm 0.441, \pm 0.664, \pm 0.888$. (d) For $\gamma = 0.450$ ($\beta = 24.2$ degrees); $\Delta/\gamma R = \pm 0.000, \pm 0.183, \pm 0.365, \pm 0.548, \pm 0.730, \pm 0.913$. Here the sides of the melt zone are represented by heavy lines sloped β degrees below horizontal and shifted vertically in proportion to the axial shift, Δ. The external caustic trajectories are shown as rays emanating from a common center at their appropriate angles and labeled with D or $D2$ to indicate a first or second 3-intercept caustic and 1, 2, or 3, etc. to indicate a first, second or third, etc. 2-intercept caustic. Dashed lines represent emergent I rays which do not form caustics.

alter the situation significantly, except that both caustics are present for the entire range of asymmetry. In all cases, the rotation of all caustics on the trailing side has been consistently positive. On the leading side, the 2-intercept caustic's forward rotation reverses at a $\Delta/\gamma R$ of only 0.25. At the same time, a reversal in the forward rotation of the 3-intercept caustic also becomes apparent at an asymmetry of 0.75.

Figure 3b shows the case of $\gamma = 1.025$, where $\beta = 45.7$ degrees, i.e., very close to the 2-intercept caustic's extinction angle. This caustic emerges at a minimum angle, θ, for the symmetric case, and progressively forms larger angles as Δ increases; i.e., it rotates negatively on the leading side and positively on the trailing side. As discussed in Part I,[1] the rays which are only incident once on the first side form a catacaustic at the inflection point. If the catacaustic ray then refracts from the second side, it is *usually* seen as the first 2-intercept caustic ray. However, the dashed lines shown in the diagrams for the smaller asymmetries in Fig. 3b depict emergent rays which form internal catacaustics at the inflection points but are *not* seen as external caustics. We call these rays the *I* rays. Their angles are at least half a degree different from the 2-intercept caustic angles. Once the asymmetry has increased to a $\Delta/\gamma R$ of 0.80, however, these two rays have recombined. The 3-intercept caustic rotates in the forward direction over the asymmetry range studied.

Continuing further, the symmetric case with $\gamma = 0.926$ has no emergent 2-intercept caustic (see Fig. 3c). However, once the asymmetry has increased to $\Delta/\gamma R = 0.441$, the original 2-intercept caustic has re-emerged and subsequently rotates to larger θ angles on both sides. Again, the (noncaustic) *I* ray is initially distinct from the 2-intercept caustic ray, and they merge as the asymmetry increases. It should be noted that the *I* ray initially emerges on the forward side without forming a caustic before the 2-intercept caustic emerges. The 3-intercept caustic continues to rotate in a forward direction up to the largest calculated asymmetry, at which point it reverses direction on the forward side.

When γ is 0.688, there are three 2-intercept caustics for the symmetric case caused by multiple-folding in the fan of emerging light. This is the case shown earlier in Fig. 2a ($\Delta/\gamma R = 0.0$) and Fig. 2b ($\Delta/\gamma R = 0.896$). The folding geometry is illustrated in the left and center diagrams in Fig. 4a. Here the caustic rays are labeled 1, 2, 3, etc., consistent with the present convention, while the bounding rays are labeled L1, L2, L3, etc. The lower L numbers represent bounding rays originally propagating nearest the surface of the sample, with successive higher numbers proceeding inward. The left diagram in Fig. 4a shows that the fan of light which lies between bounding rays L3 and L4 is folded to form the three 2-intercept caustics with the first and third very nearly superimposed. As the asymmetry is increased, most of the 2-intercept caustics disappear on the trailing side while several additional ones appear on the leading

side. All these caustics rotate in a negative direction except for the original 2-intercept caustic on the leading side, which remains stationary. This unusual circumstance occurs because the I ray does not form an external caustic until its point of emergence has moved upstream of the melt zone to where the radius is constant.

In Fig. 2b, we can see that, since the profile on the initial incident side does not change, the internal trajectory of the I ray remains the same. Consequently, as long as caustic ray 1 refracts from the leading side its external angle will be constant, regardless of the magnitude of Δ. On the trailing side, the internal I ray is incident along that portion of the melt zone profile for which the radius is changing very rapidly; therefore, although it emerges for asymmetries of less than 0.3 it does not form an external caustic. However, this I ray reappears along with a first 2-intercept caustic when the asymmetry $\Delta/\gamma R$ reaches 1.2. In addition, a new 2-intercept caustic emerges on the leading side. This fourth 2-intercept caustic is associated with the closing of the break between the L1–L2 fan of light and the L3–L4 fan of light, illustrated in the center diagram in Fig. 4a. Consequently, this new caustic is folded downstream in the same way as the original first 2-intercept caustic. It originates from a coaxial ray initially propagating down the sample at a greater radius than the second 2-intercept caustic, but well inside the ray which originates the 3-intercept caustic, as shown in Fig. 2b.

Several other minor 2-intercept caustics appear and are labeled 2′, 2″, and 3′ in Fig. 4a center. These lie very close—in terms of both originating axial rays and emergent angles—to the second or third 2-intercept caustics. Each is associated with another fold in the fan of light.* The 3-intercept caustics in this case ($\gamma = 0.688$) initially rotate positively. Eventually, on the leading side this caustic reverses its rotation at $\Delta/\gamma R \simeq 0.6$ and then extinguishes at an asymmetry of about 0.8. On the trailing side, it continues to rotate in a forward direction. The additional 2-intercept caustics also appear in the case where $\gamma = 0.595$. Again the first 2-intercept caustic, when it emerges, remains stationary. The others in this family all rotate in the negative direction. The first, second, and third 2-intercept caustics emerge on the leading side for all asymmetries and are extinguished when the asymmetry is between 0.83 and 1.1 on the trailing side. Minor caustics associated with the second and third major caustics were found on the trailing side, but not on the leading side. The fourth caustic emerges at greater asymmetry than in the previous case of $\gamma = 0.688$ and again only on the leading side. This case, with $\gamma = 0.595$, is of particular interest because it exhibits for the first time a second 3-intercept caustic. The new caustic is associated with a folding

* It should be noted that, since the algorithm computes the ray trajectories for an equally spaced sequence of axial rays, finding these minor perturbations in emergent ray angles depends somewhat on the ray spacing chosen.

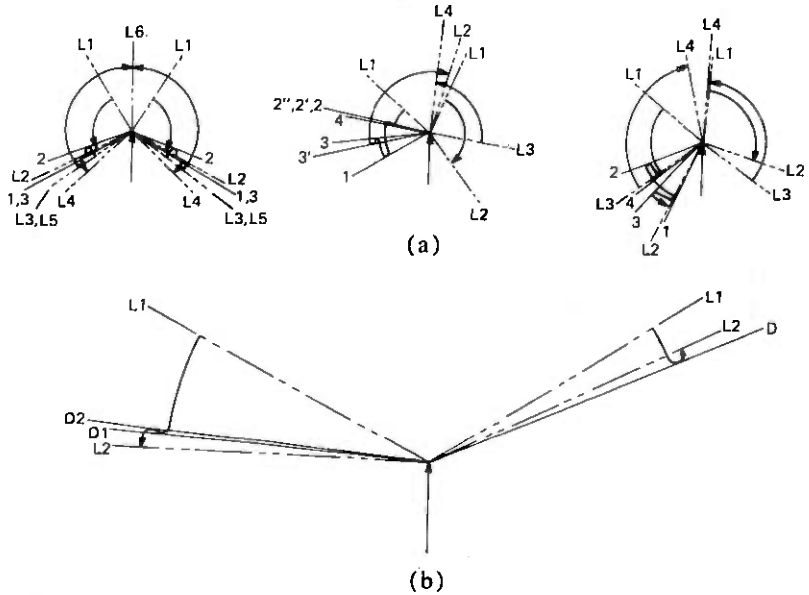


Fig. 4—(a) Schematic diagrams showing the external limiting rays for the 2-intercept light paths. The drawdown melt zone profiles are omitted, but the orientation is the same as in Figs. 1 and 2. The vertical arrow represents the incoming beam of collimated light illuminating the sample which in turn deflects the light into the plotted ray paths in the far field. Here, the solid lines represent the caustic trajectories and the broken lines represent the bounding rays limited by internal reflection (see Ref. 1). The circumferential arrows follow the continuous fan of light from the outermost bounding ray to the innermost bounding ray including all caustic rays. The bounding rays are numbered in sequence, beginning with L1 as the ray which originates nearest the surface of the sample and increasing inward L2, L3, etc. The caustic rays themselves are numbered 1, 2, 3, etc. to indicate that they represent the first, second, or third, etc. of the 2-intercept caustics. The asymmetries are as follows: *Right*—Symmetric, with $\gamma = 0.688$ ($\beta = 34.5$ degrees) on both sides and $\Delta/\gamma R = 0.0$, $\delta\gamma = 0$ percent. *Center*—Asymmetric with $\gamma = 0.688$ ($\beta = 34.5$ degrees) on both sides and a shift $\Delta = 0.075$ in. (1.905 mm) of the left side ahead of the right side ($\Delta/\gamma R = \pm 0.896$, $\Delta/R = \pm 0.616$). Here, $\delta\gamma = 0$ percent. *Left*—Asymmetric with $\gamma = 0.894$ ($\beta = 41.8$ degrees) on the left side and $\gamma = 0.482$ ($\beta = 25.7$ degrees) on the right side ($\gamma = 0.688$, $\delta\gamma = \pm 30$ percent). Here, $\Delta = 0.0$. (b) Schematic diagram showing external limiting rays for the 3-intercept light paths. The same ray identifications apply to the diagram as apply to the diagrams in Fig. 4a except that the caustics are numbered D and $D2$, representing the original and second 3-intercept caustics. The asymmetry is given by $\gamma = 0.595$ ($\beta = 30.8$ degrees) on both sides and a shift, $\Delta = 0.08$ in. (2.032 mm) of the left side ahead of the right side ($\Delta/\gamma R = \pm 1.105$, $\Delta/R = \pm 0.657$). Here, $\delta\gamma = 0$ percent.

in the fan of light in the direction opposite that of the original 3-intercept caustic. It is formed from a ray initially propagating down the fiber at a radius smaller than that ray which forms the first 3-intercept caustic. The folding pattern is shown in Fig. 4b, where the new ray is labeled $D2$. The new caustic emerges only for large asymmetry and only on the leading side. Meanwhile, the original 3-intercept caustic, as before, rotates initially in a positive direction but eventually reverses with increasing asymmetry.

The last case studied, $\gamma = 0.450$, is shown in Fig. 3d. Here the first three 2-intercept caustics emerge over the whole range of asymmetries.

The second and third caustics rotate in a negative direction; the first is again stationary. The fourth caustic emerges on the leading side and only for the largest asymmetry. The first 3-intercept caustic appears only on the trailing side and rotates downstream with increasing asymmetry. The second 3-intercept caustic appears alone on the leading side and only for larger asymmetries.

2.2 Discussion

The diagrams in Fig. 3 illustrate the far-field caustic response to increasing axial asymmetry $\Delta/\gamma R$ for fixed γ. We now present a general quantitative description of the caustic behavior due to changing geometry. Figure 5 shows the propagation angles, θ, of the major caustics as functions of $\Delta/\gamma R$ for various γ. The values, θ_-, for the trailing side are plotted to the left and the values, θ_+, for the leading side are plotted to the right of the zero asymmetry line. Figures 5a and 5b show the first 2-intercept and the 3-intercept caustics, respectively. The behavior of the remaining major 2-intercept caustics, the second, third, and fourth, are plotted in Fig. 5c. Each curve on these plots represents a fixed outer normal slope γ over the range of asymmetry tested. Those curves with arrowheads at the ends indicate caustics which emerge for higher asymmetry but were not calculated. The absence of an arrowhead indicates that the caustic ceases at or just beyond that point.

Other measures of caustic behavior are given in Fig. 6, for cases where the caustics emerge on both sides of the drawdown. In terms of actual three-dimensional drawdown samples, the caustic trajectories form cones of light whose total included angle is $\theta_- + \theta_+$. The average sum, $\theta_S = \frac{1}{2}(\theta_- + \theta_+)$, is plotted to show the half cone angle as a function of asymmetry. The average difference, $\theta_D = \frac{1}{2}(\theta_- - \theta_+)$, is plotted to show the rotation of the cone as a function of asymmetry. Figures 6a and 6c refer to various 2-intercept caustics, while Fig. 6b refers to the 3-intercept caustics.

Consider the first 2-intercept caustic shown in Figs. 5a and 6a. In Fig. 5a, starting with large γ, on the leading side θ^{U1} decreases, reaches a minimum, and slightly increases as $\Delta/\gamma R$ increases. As the profile becomes blunter, the minimum becomes more pronounced and moves toward the symmetric case, $\Delta/\gamma R \to 0$. When $\gamma = 0.926$, the drop in θ^{U1} as $\Delta/\gamma R \to 0$ becomes so large that the caustic is extinguished for small asymmetries and only the ends of the curve remain. This means that, although the first 2-intercept caustic is extinguished by internal reflection in the symmetric sample of $\gamma = 0.926$, the development of sufficient asymmetry causes an almost simultaneous reappearance on both sides of the sample. Figure 6a also shows that the average angle, θ_S^{U1}, of the first 2-intercept caustic decreases as $\Delta/\gamma R \to 0$. This decrease is greatest for profiles with the lowest γs. In other words, the caustic cone angle increases as the asymmetry increases, especially in blunt samples.

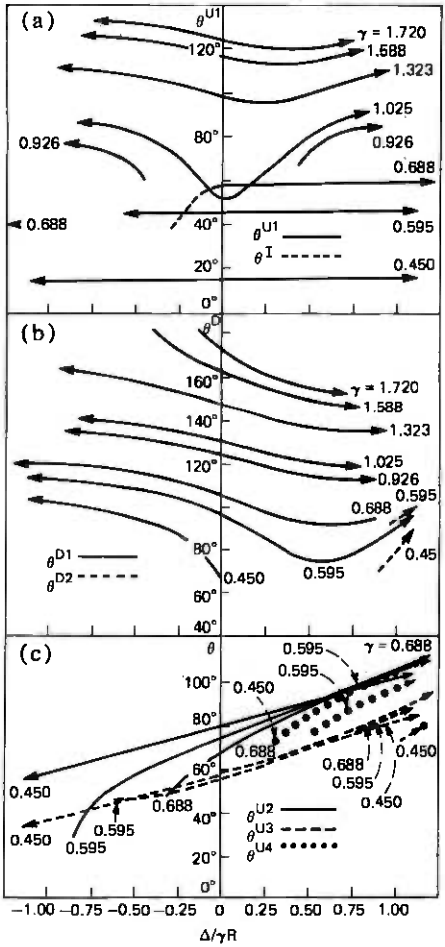


Fig. 5—Plots of the θ angles of the caustic trajectories for various γs as functions of the axial shift, Δ. Here the values of the angles, θ_+, of the caustics emergent on the leading side are plotted on the right and the values of the caustic angles, θ_-, of those emergent on the trailing side are plotted on the left. These results are arranged as follows: (a) The first 2-intercept caustic angles, designated θ^{U1}, and the I ray propagation angle, θ^{I}, vs $\Delta/\gamma R$. (b) The first and second 3-intercept caustic angles, θ^{D} and θ^{D2}, vs $\Delta/\gamma R$. (c) The second, third, and fourth 2-intercept caustic angles, θ^{U2}, θ^{U3} and θ^{U4}, vs $\Delta/\gamma R$.

At the same time, the θ_D^{U1} vs $\Delta/\gamma R$ curves in Fig. 6a show that the first 2-intercept caustic cone, which is initially symmetric about the drawing axis, first rotates toward the leading side, reverses, and then rotates back toward the axis as the asymmetry increases.

The 3-intercept caustic exhibits similar behavior, as shown in Figs. 5b and 6b, although there are significant differences. Again, there is a minimum in the caustic angle vs asymmetry curve that appears on the leading profile side of the θ^{D} vs $\Delta/\gamma R$ plot. However, the minimum does

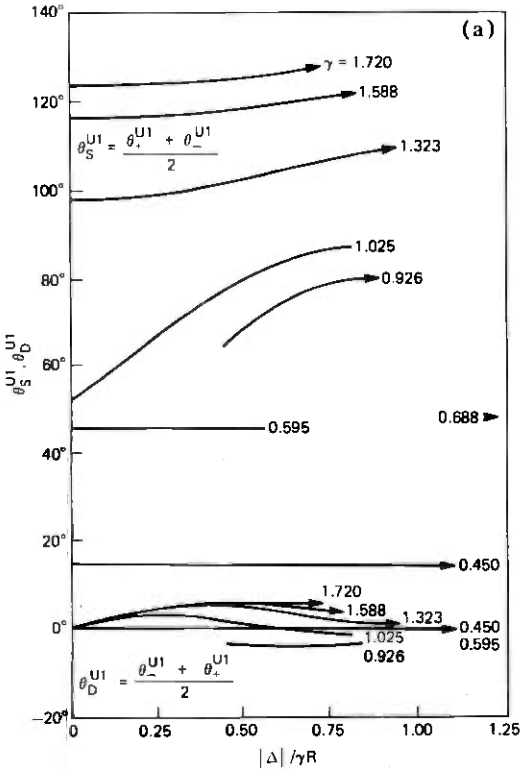


Fig. 6—Plots of the averaged sums, $\theta_S = \frac{1}{2}(\theta_- + \theta_+)$, and the averaged differences, $\theta_D = \frac{1}{2}(\theta_- - \theta_+)$, vs the magnitude of the axial shift $|\Delta|/\gamma R$ for various γs. These results are arranged as shown in 6a, 6b, and 6c. (a) The first 2-intercept caustic angles, θ^{U1}, averaged sums and differences vs $|\Delta|/\gamma R$. (*Figs. 6b and 6c on following pages.*)

not move toward the symmetry axis as γ decreases. Also, except in the region near the extinction points, the θ^D vs $\Delta/\gamma R$ plots are morphologically similar. In Fig. 5b, the magnitude of the depression in the curve still grows as γ decreases until, by $\gamma = 0.450$, the caustic is extinguished on the leading side. Figure 6b shows that the caustic cone half-angles, θ_S^D, are less influenced by the asymmetry than they were for the 2-intercept caustics (Fig. 6a). In addition, the 3-intercept caustic cone rotates more rapidly toward the leading side before it reverses direction, as shown by the θ_D^D curves in Fig. 6b. The behavior of the second, third, and fourth 2-intercept caustics is illustrated in Figs. 5c and 6c. These caustics are distinct because their directions of rotation are opposite those of the first two major caustics. This backward rotation, with increasing $\Delta/\gamma R$, of the second and third 2-intercept caustics parallels the reverse θ_A rotation with increasing γ noted in the earlier "symmetric" study.[1] At the same time, the θ_S values in the second and third 2-intercept caustics remain almost constant, indicating that the shapes of these caustic cones are

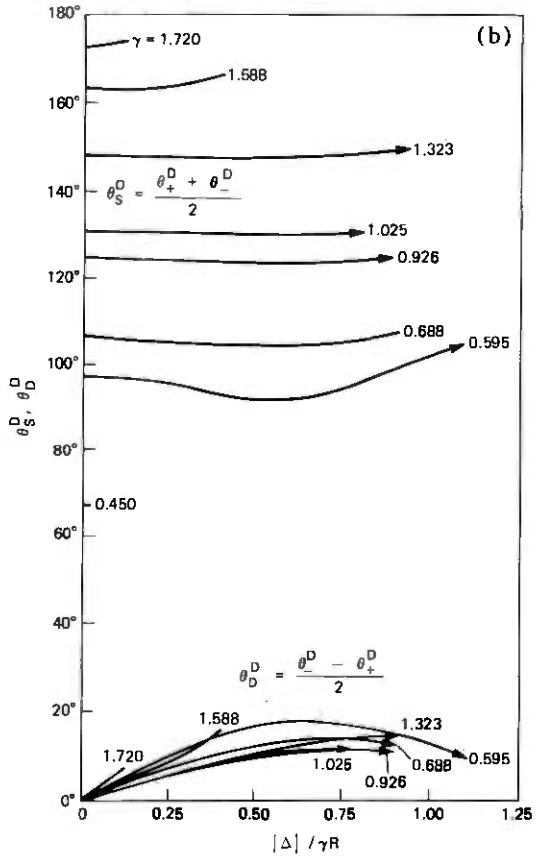


Fig. 6(b)—The 3-intercept caustic angles, θ^D, averaged sums and differences vs $|\Delta|/\gamma R$.

almost independent of the asymmetry, except near the extinction points.

III. DIFFERENTIAL SLOPE ASYMMETRY

The effects of asymmetric profiles of differing slopes were studied by changing the scaling of opposite sides of the profile by equal positive and negative percentages. These changes produce equal percentage changes in the slopes of the outer normals to the profiles, dx/dy. As shown in Part I,[1] the outer normal slope defines a unique incidence angle and is the most significant single parameter controlling the formation of the caustics. Therefore, we can define asymmetry by a ± change in γ, i.e., $\delta\gamma$.

An extreme example of the effects of this type of asymmetry is found in the comparison between Figs. 2a and 2c where in the latter $\delta\gamma = \pm 30$ percent. It can be seen that the external trajectories of the 3-intercept

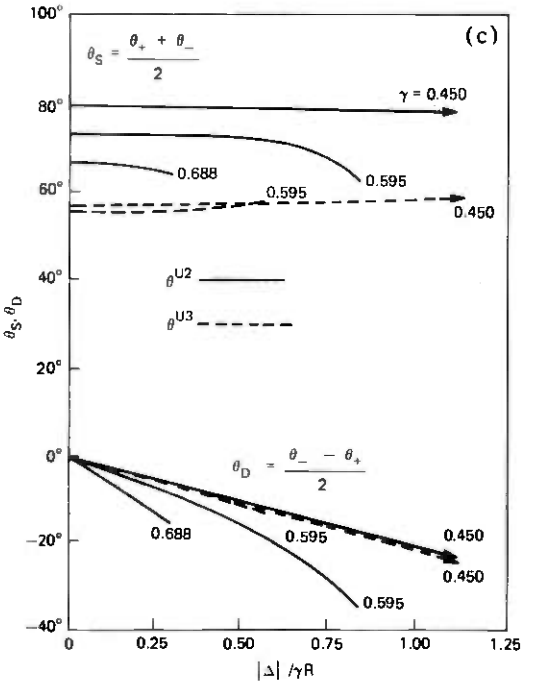


Fig. 6(c)—The second and third 2-intercept caustic angles, θ^{U2} and θ^{U3}, averaged sums and differences vs $|\Delta|/\gamma R$.

caustics have rotated in the direction of increasing γ, i.e., downstream on the shortened side and upstream on the lengthened side, which will be referred to as a forward (positive) rotation. At the same time, the external 2-intercept caustics have disappeared from the side of diminished γ, while an additional 2-intercept caustic, the fourth, appears on the opposite side. Like the 3-intercept caustic rays, the first and third 2-intercept caustic rays have rotated forward, while the direction of the second ray is unchanged. Indeed, only the insignificant 1-intercept caustics exhibit a backward or negative rotation. A better understanding of this behavior can be obtained only in the context of a range of changing asymmetries for the various average γs in Table I, which are discussed below.

3.1 Graphical results

A graphical description of the effects of this asymmetry covering a range of typical average γs is presented in Fig. 7. As in Fig. 3, these diagrams are simplified versions of the earlier formats showing only the external caustic trajectories. In most cases, the range of asymmetry, $\delta\gamma$, is from ± 0 to ± 30 percent. These represent a much greater range of asymmetry than might be expected to arise with actual samples. The

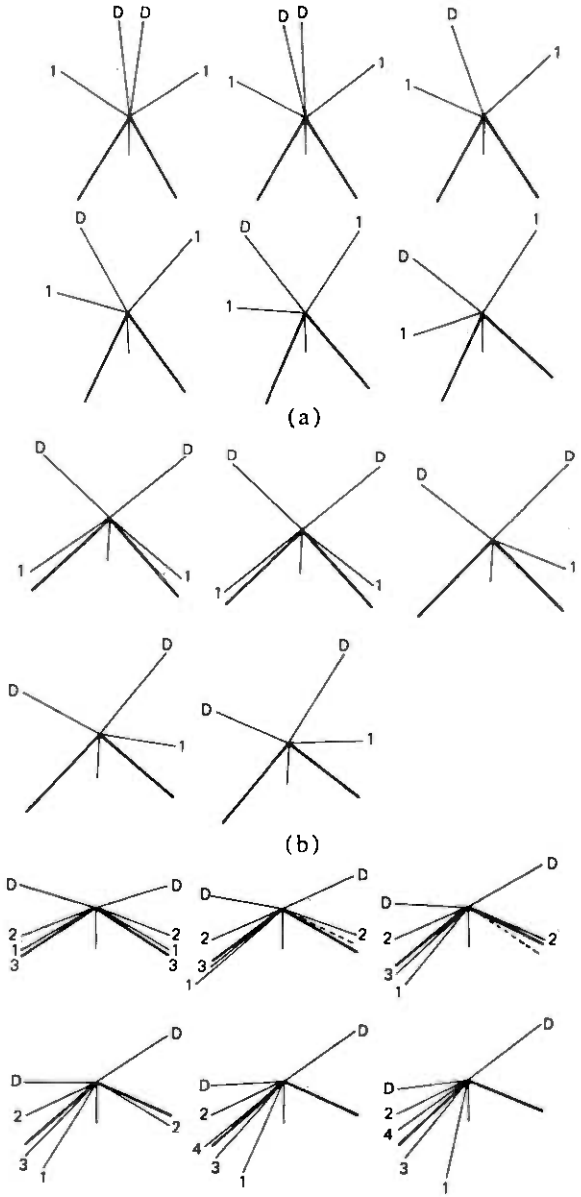


Fig. 7—Diagrams of external caustic trajectories for drawdowns of varying asymmetry due to percentage differences in γ, $\delta\gamma$, of varying amounts as follows: (a) $\gamma = 1.720 \pm 0\%$, $\pm 5\%$, $\pm 10\%$, $\pm 20\%$, $\pm 30\%$, $\pm 40\%$. (b) $\gamma = 1.025 \pm 0\%$, $\pm 1\%$, $\pm 10\%$, $\pm 20\%$, $\pm 30\%$. (c) $\gamma = 0.688 \pm 0\%$, $\pm 10\%$, $\pm 20\%$, $\pm 25\%$, $\pm 30\%$, $\pm 35\%$. Here the sides of the melt zone are represented by heavy lines sloped β degrees below horizontal. The external caustic trajectories are shown as rays emanating from a common center at their appropriate angles and labeled D to indicate a 3-intercept caustic and 1, 2, or 3, etc. to indicate a first, second, or third, etc. 2-intercept caustic. Dashed lines represent emergent I rays which do not form caustics.

specific incremental changes in asymmetry shown are not uniform, but were chosen to illustrate what occurs as the asymmetry increases. In every case, the diagram is drawn with the side of reduced γ to the left so that a forward (positive) rotation of caustic rays is consistently counterclockwise on the diagram. In general, on the *decreasing* γ side, θ *increases* and on the *increasing* γ side, θ *decreases* with increasing asymmetry.

The first two cases, for which $\gamma = 1.720$ (Fig. 7a) and 1.588, are similar; the rotation of all the caustic rays is forward and the 2-intercept caustic on the right-hand, or decreasing γ, side disappears by internal reflection. It initially propagates at a greater angle for $\gamma = 1.720$ than it does for $\gamma = 1.588$, and its rotation due to increasing asymmetry is positive. Therefore, the right-hand 2-intercept caustic extinguishes at a $\delta\gamma$ of only ± 10 percent for the case of $\gamma = 1.720$, but must reach a $\delta\gamma$ of ± 14 percent before it extinguishes for $\gamma = 1.588$. Decreasing γ to 1.323 reduces θ for all trajectories and permits forward rotations of both caustic rays on both sides for asymmetries up to $\delta\gamma = \pm 30$ percent without loss by internal reflection.

Internal reflection can also cause a caustic ray to disappear when its propagation direction swings too far upstream (θ decreasing). This is shown in Fig. 7b where the 2-intercept caustic ray on the increased γ side disappears as soon as the asymmetry begins to develop, because it emerges near the extinction region.[1] Very little asymmetry is then required to rotate the left-hand ray below the critical angle.

The case of $\gamma = 0.926$ lies within the range for which no 2-intercept caustic rays emerge from a symmetric drawdown. However, an asymmetry of ± 24 percent yields an external 2-intercept caustic ray on the side of reduced γ. Hereafter, the ray rotates in the forward direction, as do the 3-intercept rays. It is interesting to note that the I ray initially emerges on the reduced γ side at ± 22 percent asymmetry without forming an external caustic and becomes the externally visible caustic ray at $\delta\gamma = \pm 24$ percent. A similar ray, initially incident at the inflection point on the reduced γ side as a catacaustic ray, emerges even earlier on the increased γ side, at an asymmetry of $\delta\gamma = \pm 20$ percent. However, within the range of asymmetries considered, this ray never develops into an external caustic.

Reducing the γ to 0.688 creates a far more complicated structure even for the symmetric case, for which there are three 2-intercept caustics associated with multiple folds in the fan of emerging light. As shown earlier in Figs. 2a and 2c and further in Fig. 7c, the rotation is in general forward, although the second 2-intercept caustic appears to be relatively less responsive. On the decreased γ side, the fold in the fan of emerging light disappears by $\delta\gamma = \pm 10$ percent, leaving the I ray still emergent. No first or third 2-intercept caustics are formed. At $\delta\gamma = \pm 25$ percent, even the I ray has disappeared. At $\delta\gamma = \pm 30$ percent, the second

2-intercept caustic is extinguished on the reduced γ side and a fourth 2-intercept appears on the opposite side. A better understanding of this new caustic, which is the only one that rotates backward as shown by the last two diagrams of Fig. 7c, may be obtained from consideration of the left and right diagrams of Fig. 4a. As mentioned above, the left diagram in Fig. 4a shows that the fan of light which lies between bounding rays L3 and L4 is folded to form the three 2-intercept caustics with the first and third very nearly superimposed. The right side of Fig. 4a shows the same situation on the increased γ side for the asymmetric case represented by Fig. 2c with the folds of the first and third 2-intercept caustics widely separated. The break in the 2-intercept fan of light shown lying between bounding rays L2 and L3 in the left diagram in Fig. 4a (which separates the initial fan L1–L2 from the folded fan L3–L4) is closed in the right diagram by the new or fourth 2-intercept caustic. This new caustic is folded downstream in the same way as the original first 2-intercept caustic and originates from a coaxial ray initially propagating axially down the sample at an even greater radius than the second 2-intercept caustic but well inside the ray which originates the 3-intercept caustic (see Fig. 2c).

At a further reduction of scale to a γ of 0.595, all rays rotate forward due to increasing asymmetry, with the first 2-intercept the fastest and the second 2-intercept the slowest. In this case, the former disappears for the increased γ side at an asymmetry of almost ± 30 percent, while the fold in the fan of light separating the first and third 2-intercept caustic rays on the opposite side is almost completely suppressed by an asymmetry of only ± 20 percent.

A final case, at a γ of 0.450, behaves in much the same way, although since the θ angles are even smaller, the first 2-intercept caustic disappears from the increased γ side of the drawdown zone at an asymmetry of only ± 4 percent. In this case the 3-intercept caustic propagates at a lesser angle than the first 2-intercept caustic and disappears at around the same percent asymmetry and on the same side.

3.2 Discussion

Study of the collective response yields two general observations. First, it is apparent that all of the significant caustic rays rotate in a forward direction with increasing asymmetry: the first 2-intercept caustic being most responsive, the 3-intercept caustic slightly less so, and the third and second 2-intercept caustics even less so, in that order. Second, the angles between the opposing rays for each caustic appear to be fairly constant with changing asymmetry, except when one of the rays nears extinction by internal reflection.

Figures 8a, 8b, and 8c illustrate these points. Each of these figures is a plot of the propagation angle θ as a function of the percentage of γ

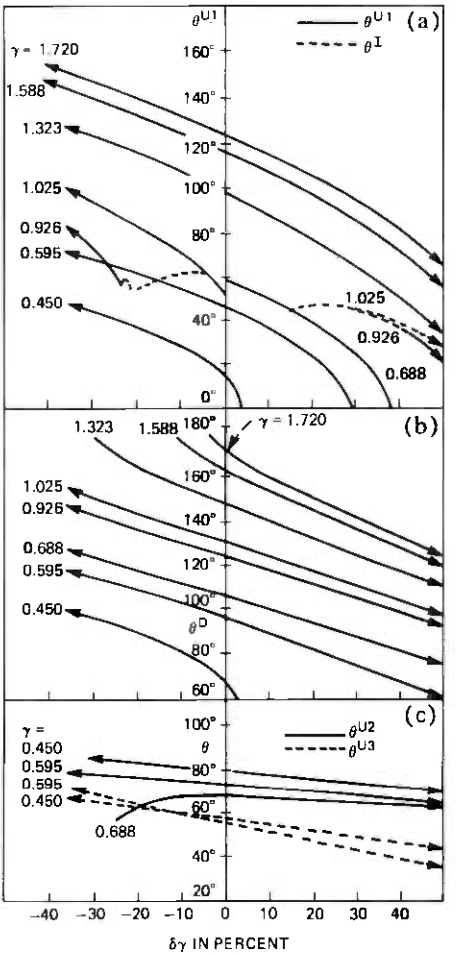


Fig. 8—Plots of the θ angles of the caustic trajectories for various γs as functions of the asymmetry, $\delta\gamma$. Here the values of the angles, θ_-, emergent on the side where γ is decreased are plotted on the left; the values of the angles, θ_+, emergent on the side where γ is increased are plotted on the right. These results are arranged as follows: (a) The first 2-intercept caustic angles, designated θ^{U1}, and the I ray propagation angle, θ^{I}, vs the asymmetry, $\delta\gamma$, in percent. (b) The 3-intercept caustic angle, θ^{D}, vs the asymmetry, $\delta\gamma$, in percent. (c) The second and third 2-intercept caustic angles, θ^{U2} and θ^{U3}, vs the asymmetry, $\delta\gamma$, in percent.

asymmetry, $\delta\gamma$. The values for the side where γ is decreased, θ_-, are plotted to the left; and the values for the side where γ is increased, θ_+, are plotted to the right of the zero asymmetry line. Figures 8a and 8b represent the response of the first 2-intercept and the 3-intercept caustics respectively, while the responses of both the second and third 2-intercept caustics are presented in Fig. 8c. The θ response of the I ray is represented by a broken curve where it appears alone on Fig. 8a, and broken curves are used in Fig. 8c to differentiate between the second and third

2-intercept caustics. Each curve shown is labeled to indicate the average γ for which the range of asymmetries is now determined. Thus, to determine the θ angles for a first 2-intercept caustic ray on a melt zone of average γ of 1.323 with an asymmetry of $\delta\gamma = \pm 30$ percent, one finds a value of 115 degrees on the left side of the plot as the value of θ_-^{U1} on the side of the melt zone where γ is reduced 30 percent, and a value of 77.4 degrees on the right of the plot as the value of θ_+^{U1} on the side of the melt zone where γ is increased 30 percent.

Except for those portions of the curves which are near points of extinction, the curves* for both the first 2-intercept caustic and the 3-intercept caustic are similar in shape for all γs. In fact, over the range of most realistic asymmetries, which means the first ± 10 or ± 15 percent, these curves are almost linear. This indicates that for these caustics the total included angle between opposing caustic rays, $\theta_+ + \theta_-$, remains constant while the angle of its bisector increases linearly with the increasing asymmetry. In terms of actual drawdown samples, these caustic rays form cones of light which remain fairly stable in shape but which tilt in the direction of these asymmetries and become distorted in those regions where they approach extinction by internal reflection. This is also true of the second 2-intercept caustic except that it seems to be relatively less sensitive to this asymmetry.

Another way of illustrating this behavior is shown in Fig. 9. Here the average sum, $\theta_S = \frac{1}{2}(\theta_- + \theta_+)$, is a direct measure of the constancy of the cone half-angle as a function of asymmetry, while the average difference, $\theta_D = \frac{1}{2}(\theta_- - \theta_+)$, gives a direct measure of the tilt or rotation of the cone as a function of asymmetry. Figure 9a shows that the cone half-angle for the first 2-intercept caustic, θ_S^{U1}, remains fairly constant but does decrease a few degrees at high asymmetries. This may be partially accounted for by the fact that the average β angle does not remain constant with increasing asymmetry but decreases as shown in Fig. 10. Consequently, if the first 2-intercept caustic is somewhat more β-dependent, it too may be expected to decrease, since in Part I[1] on symmetric drawdowns it was shown that decreasing β decreases θ. On the other hand, Fig. 9b shows fairly constant 3-intercept caustic-ray cone angles, θ_S^D, for all circumstances except those approaching extinction, as does Fig. 9c for the θ_S angles of the second and third 2-intercept caustics.

In all plots of Fig. 9, the curves for the average differences, θ_D, vs

* Many more calculations were made to produce these plots than were illustrated by the diagrams of Fig. 7, with particular attention given to accurate determination of the points where each caustic disappeared, if it did so within the range of interest. These terminations are shown exactly where they occur on the plots of Fig. 8, with arrows marking a continuation where the computation was stopped for other reasons, such as reaching the limits of the maximum range of interest or reaching the limits of the computable range of the input data.

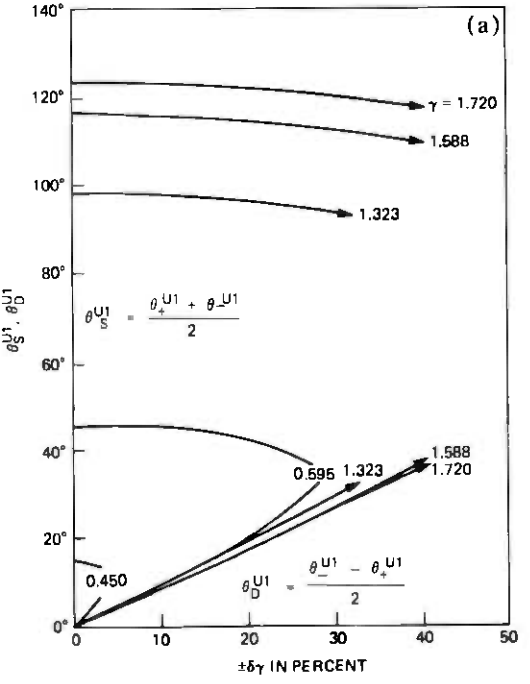


Fig. 9—Plots of the averaged sums, $\theta_S = \frac{1}{2}(\theta_- + \theta_+)$, and the averaged differences, $\theta_D = \frac{1}{2}(\theta_- - \theta_+)$, vs the asymmetry, $\delta\gamma$, for various γs. These results are arranged as shown in 9a, 9b, and 9c. (a) The first 2-intercept caustic angle's averaged sums and differences vs the asymmetry, $\pm\delta\gamma$, in percent. (*Figs. 9b and 9c on following pages.*)

asymmetry, $\delta\gamma$, are nearly linear within the same limitation and indicate a constant slope or rate of rotation, r_D. The greater slopes, r_D, of the average difference curves in Fig. 9a indicate the higher response of the first 2-intercept caustic to changes in asymmetry, which is about 0.8 degree/±1 percent. Also, their closer grouping indicates a significant degree of uniformity of this rate over a range of γs. The θ_D vs $\delta\gamma$ curves for the 3-intercept caustic ray angles are also nearly linear but shallower, with rates of rotation, r_D, of about 0.6 degree/±1 percent. They also show considerably more variation from γ to γ.

Figure 9c shows the same two functions (θ_S vs $\delta\gamma$ and θ_D vs $\delta\gamma$) for the second and third 2-intercept caustics. For both, the half-cone angles, θ_S, are reasonably constant except near extinction (i.e., the second caustic at $\gamma = 0.68$ degree), but the rates of rotation are smaller. These run from 0.4 degree/±1 percent and less for the third caustic to 0.17 degree/±1 percent and slightly negative for the second caustic, and with much greater variation from γ to γ.

The results provide a means by which the caustic angles measured on opposite sides of a section through the melt zone may be used to estimate both average γ and its asymmetry for that section. For example, since

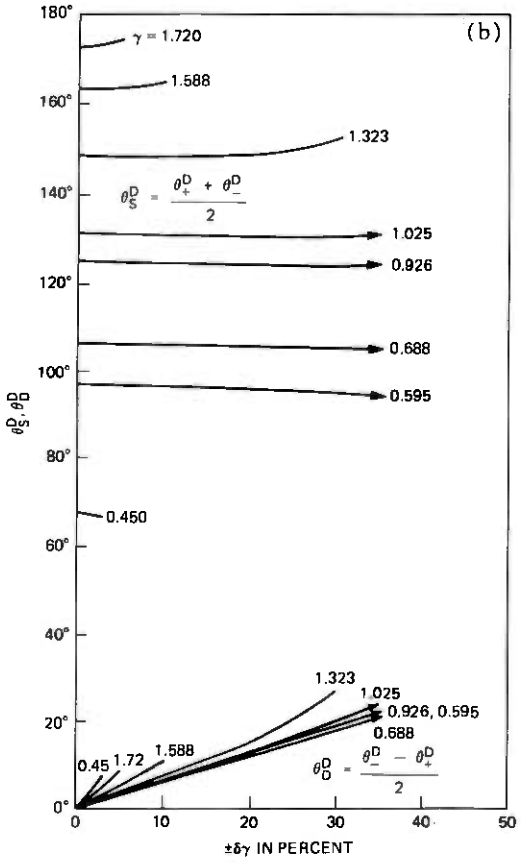


Fig. 9(b)—The 3-intercept caustic angle's averaged sums and differences vs the asymmetry, ± δγ, in percent.

the averaged sum of the 3-intercept caustic angles, θ_S^D, is relatively insensitive to the asymmetry in γ (Fig. 9b), this parameter may be used to estimate γ. In other words, θ_S^D will be very nearly the same as θ^D for the symmetric case. In turn, the dependence on γ may be derived directly from the θ^D vs β relationship given in Fig. 6 of Part I.[1] Figure 11 shows a plot of θ_S^D vs γ adapted from this original plot of θ^D vs β using $\gamma = \tan \beta$ and $\theta_S^D \approx \theta^D$. The figure also shows a plot of the inverse rate of rotation, r_D^{-1} vs γ. Therefore, for a drawdown where melt zone asymmetry is the result of an asymmetry in γ, measurement of θ_+^D and θ_-^D will provide a means of estimating γ and its asymmetry as follows:

From θ_-^D and θ_+^D, compute $\theta_D^D = \frac{1}{2}(\theta_-^D - \theta_+^D)$ and $\theta_S^D = \frac{1}{2}(\theta_-^D + \theta_+^D)$. Then use Fig. 11 to estimate a value of γ and to find the corresponding value of r_D^{-1}. The latter may then be used to compute the asymmetry, $\delta\gamma$, by multiplying by θ_D^D. Consider the example of a sample whose 3-intercept caustic angles measure 134 and 115 degrees. These yield a

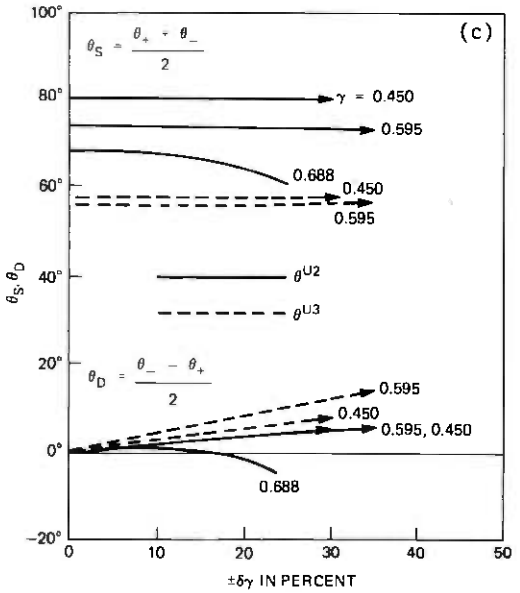


Fig. 9(c)—The second and third 2-intercept caustic angles' averaged sums and differences vs the asymmetry, $\pm\ \delta\gamma$, in percent.

θ_S^D = 124.5 degrees which, by Fig. 11, gives an estimated γ of 0.930. At this γ, Fig. 11 also gives an r_D^{-1} = 1.56. Since θ_D^D = 9.5 degrees, the asymmetry, $\delta\gamma$, is then 1.56 ×.9.5 = 14.8 percent. Computer-generated data for a γ of 0.926 with an asymmetry of ±15 percent gives θ_-^D = 134.1 degrees and θ_+^D = 115.0 degrees, which is within an absolute error of only 0.004 in estimating γ and 0.2 percent in estimating the asymmetry. This order of accuracy is consistent throughout the region for which θ_S^D is independent of γ.

The development of additional data for the various 2-intercept caustics would permit similar analyses of data taken over the appropriate γ or β ranges for which these caustics appear on both sides of the melt zone and manifest linear dependences on $\delta\gamma$ or $\delta\beta$.

IV. COMPARISON WITH AN ACTUAL DRAWDOWN SAMPLE

Empirical profile data from Sample 1 discussed in Ref. 2 provide a comparison for the present technique. Figure 12 shows data taken from the 3 and 9 o'clock profiles superimposed on smooth curves computed from the averaged data used in the present report (from Sample 4). These synthesized profiles have been asymmetrically scaled (γ = 1.006 ± 15 percent) to match the β angles of the empirical data (β_+ = 48.9 degrees (3-side) and β_- = 44.6 degrees (9-side)). It can be seen that the empirical and computed values are very similar and superimposed (in this case) with no axial shift asymmetry ($\Delta/\gamma R$ = 0). Various caustic

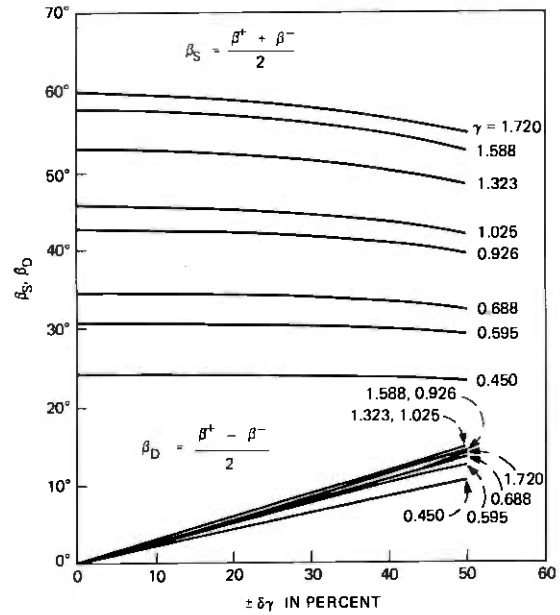


Fig. 10—Plots of the averaged sums, $\beta_S = \frac{1}{2}(\beta_- + \beta_+)$, and the averaged differences, $\beta_D = \frac{1}{2}(\beta_- - \beta_+)$, vs the asymmetry, $\delta\gamma$, in percent, for various γs. Here β_- is the incident angle on the side of reduced γ and β_+ is the incident angle on the side of increased γ.

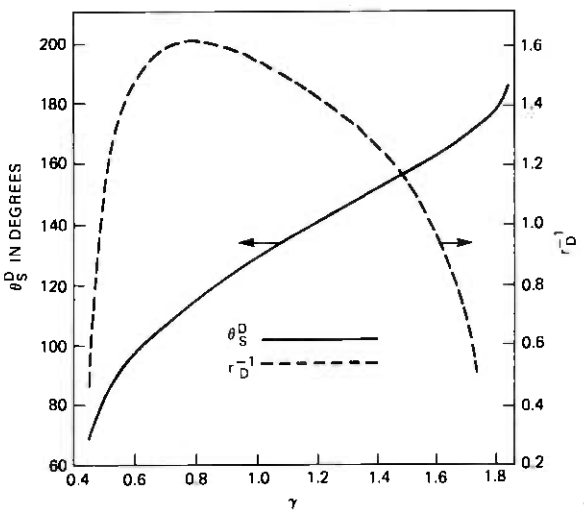


Fig. 11—Plots of the averaged sum of the caustic angles, θ_S^D, and the inverse rate of rotation (from their averaged difference curves), r_D^{-1}, vs γ for the 3-intercept caustics.

angles were computed from these profiles and are shown in Table II along with the angles measured in the original experiment and the values calculated at that time using the empirical profile data. In some cases,

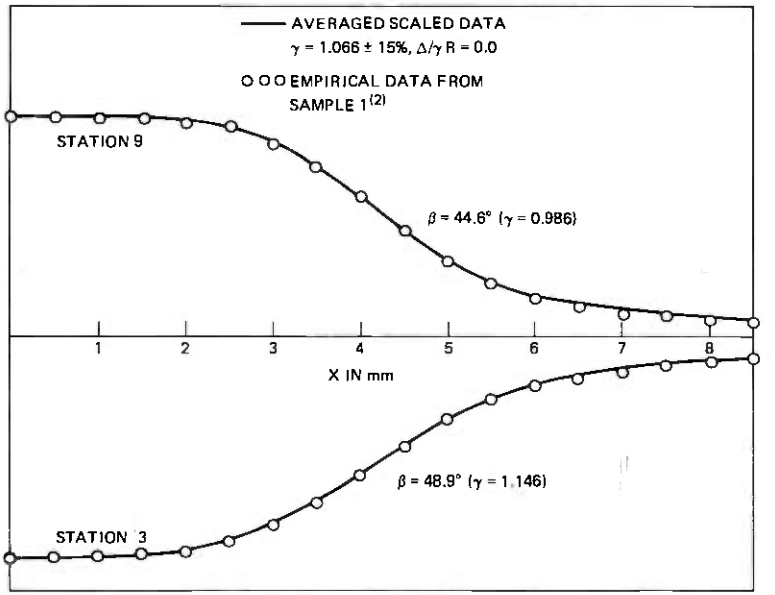


Fig. 12—Plots of the empirical profile data from Sample 1 (Ref. 2) and the computed profiles obtained using the present technique for synthesizing asymmetric profiles (γ = 1.006 ± 15 percent, $\Delta/\gamma R = 0$) with differing slopes ($\beta_+ = 48.9$ degrees, $\beta_- = 44.6$ degrees). For Sample 1, $R_0 = 3.24$ mm, and for Sample 4, $R_0 = 3.09$ mm.

Table II—Caustic angles from profile data

	Measured Values From Original Experiments (Ref. 2) (degrees)	Calculated Values Using Empirical Profile Data (Sample 1) (degrees)	Computed Values Using Smoothed Avg'd Data (Sample 4) (degrees)
θ_+^U	46.1	54.5	EXT (43.2)*
θ_-^U	72.5	75.3	75.0
θ_+^D	130.6	133.8	122.1
θ_-^D	140.0	142.7	145.4
θ_S^U	59.3	64.9	(59.1)*
θ_S^D	135.3	138.3	133.8
θ_D^U	13.2	10.4	(15.9)*
θ_D^D	4.7	4.5	11.7

* In this case the caustic is extinguished by internal reflection, so the extinction angle itself is used in computing effective θ_S^U and θ_D^U values.

the newer values are in better agreement with the observed angles than are those computed using the actual profiles, and certainly they appear to be about as good overall.

V. SUMMARY

Two types of asymmetry have been studied using computer simulation for predicting the trajectories of caustics emitted by a coaxially illuminated, fused-silica melt zone. The first type of asymmetry involves an

axial shift of one side of the melt zone with respect to the other. Increases in this type of asymmetry produce changes in the trajectories of the various types of caustics which appear. The appearance of these caustics also depends very strongly upon the average taper of the melt zone.[2] The following effects have been observed:

(*i*) The first 2- and 3-intercept caustics rotate downstream when they appear on the trailing side of the melt zone.

(*ii*) The first 2-intercept caustic initially starts to rotate upstream and then reverses direction when it appears on the leading side. This effect is more pronounced in blunter melt zones.

(*iii*) The first 3-intercept caustic generally rotates upstream on the leading side but may reverse its rotation for large asymmetries on very blunt melt zones.

(*iv*) The second 3-intercept caustic appears on the leading side only for blunt melt zones ($\gamma = 0.450$) and only at large asymmetries for which it rotates downstream.

(*v*) The second and third 2-intercept caustics rotate downstream on the leading side and upstream on the trailing side with increasing asymmetry. This is the reverse rotation of the first two major caustics.

(*vi*) The fourth 2-intercept caustic appears only on the leading side of melt zones which are moderately blunt ($\gamma \leq 0.688$) and rotates downstream with increasing asymmetry.

(*vii*) When a particular caustic appears on opposite sides of a melt zone, its intersecting trajectories form an angle which may be thought of as the included angle of a cone of light emitted from a three-dimensional sample.* Cone angles of the first 3-intercept caustic and the second and third 2-intercept caustics are moderately insensitive to the asymmetry.

(*viii*) The cone angle of the first 2-intercept caustic increases significantly with increasing asymmetry.

(*ix*) The cones of light for most of these caustics rotate forward with increasing asymmetry. The rotation of the cones of light for the 3-intercept caustic and the first 2-intercept caustic is initially forward or toward the leading side and then reverses.

(*x*) The light cones for the second and third 2-intercept caustics generally rotate backward at a constant rate with increasing asymmetry for a given γ.

(*xi*) The rate of rotation of the first 3-intercept caustic is generally greater than that of the first 2-intercept caustic, while the rates of rotation for the second and third 2-intercept caustics are always negative.

* It should be kept in mind that the cone angle, as we have defined it here, is the sum of the particular caustic angles on opposite sides of the drawdown measured from the upstream axial direction.

The second type of asymmetry investigated involves scaling opposite sides of the melt zone differently to produce asymmetries in slope. Increases in slope asymmetry produce changes in the external caustic trajectories, which are in many ways simpler than those observed with the axial shift asymmetry. These changes are as follows.

(*i*) All caustics generally rotate downstream on the trailing (decreased) γ side and upstream on the leading (increased) γ side. This is comparable to the behavior of the first 2- and 3-intercept caustic response to the shift asymmetry.

(*ii*) The cone angles of the 3-intercept caustic and the second and third 2-intercept caustics are almost independent of the asymmetry over significant ranges.

(*iii*) The cone angle of the first 2-intercept caustic is only weakly dependent on the asymmetry in γ and may be varying in response to changes in the average value of β, which itself changes with asymmetry.

(*iv*) There is no second 3-intercept caustic, as was the case previously.

(*v*) All the caustics rotate forward at constant rates for any particular γ.

(*vi*) The rates of rotation are ordered as follows:

$$r_{U1} > r_D > r_{U3} > r_{U2}.$$

For both types of asymmetry, sufficient rotation in any axial plane usually causes one of the caustic trajectories to be extinguished by internal reflection. The approach of a caustic trajectory to its extinction usually results in an acceleration in its rate of rotation, thereby distorting the general pattern of movement of the cone of light for that caustic. Nevertheless, over much of the range of reasonable geometries there will be caustics whose rotations are sufficiently regular as to provide a direct means of estimating the geometrical asymmetry from the asymmetry of the caustic cone itself, if the caustic and type of asymmetry are properly identified.

VI. ACKNOWLEDGMENT

The authors wish to acknowledge the collaboration of J. McKenna, who contributed to the development of the numerical algorithm.

REFERENCES

1. T. D. Dudderar, J. B. Seery, and P. G. Simpkins, "Caustic Patterns Associated With Melt Zones in Solidified Glass Samples—Part I: Symmetric Cases," B.S.T.J., this issue, pp. 3209–3225.
2. P. G. Simpkins, T. D. Dudderar, J. McKenna, and J. B. Seery; "On the Occurrence of Caustics in the Drawdown Zone of Silica Fibers," B.S.T.J., *56*, No. 4 (April 1977), pp. 535–560.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Measurement of Echoes Due to Spurious TE_{0n} Modes in a Long-Distance 60-mm Waveguide Communication System

By J. L. DOANE

(Manuscript received May 19, 1978)

To estimate the seriousness of spurious TE_{0n} mode generation in a long-distance millimeter waveguide communication system requires a measurement of small echoes in the impulse response with delays up to about 500 ns. We report the results of such a measurement made as part of a recent 14-km field test of the WT4 communication system with a novel test set operating near 41 GHz. The results show that the WT4 system can perform satisfactorily without any filters of spurious TE_{0n} modes.

I. INTRODUCTION

The performance of a long-distance millimeter waveguide communication system can be seriously degraded by slowly decaying echo trains in the impulse response. Slowly decaying echo trains are caused by coupling over long distances of the desired signal mode (TE_{01}) with forward traveling spurious TE_{0n} modes. When many echoes in such an echo train act in phase, the resulting peak distortion on a received signal can be quite large, even with relatively small total echo power.

A previous paper[1] described an echo test set (Fig. 1) designed to detect such slowly decaying echo trains. This test set transmits a 20-ns pulse into a 60-mm waveguide line with a shorting plate at the end, samples the signal in 4-ns increments for 1 μs near the returning pulse, transforms the data to loss and delay versus frequency, and overlaps data taken at different center frequencies, f_c, to extend the frequency range of the data. Measurements were made with this test set as part of a recent 14-km field test of the WT4 system[2] between Netcong and Long Valley, New Jersey. These measurements indicated indirectly that the power in slowly decaying echo trains generated in the 60-mm waveguide itself was too small to degrade system performance.[1]

To demonstrate overall system performance, however, we must de-

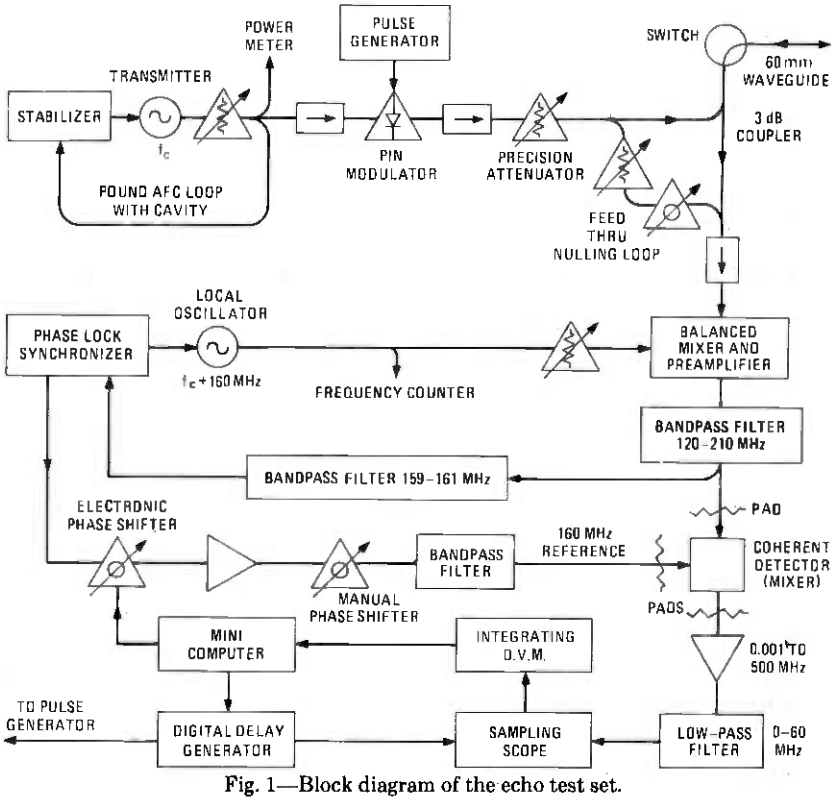


Fig. 1—Block diagram of the echo test set.

termine the impulse response for the entire WT4 system, including the band diplexer trees at each end of the 60-mm waveguide. A band diplexer tree[3] contains several miter elbows and hybrids, each of which converts a relatively large amount of energy to and from spurious TE_{0n} modes.[4] Energy that is converted to TE_{0n} modes in the long-distance 60-mm waveguide medium and then reconverted at the band diplexer tree (or vice versa) can cause echoes much larger than those generated in the waveguide medium alone. Miter elbows used for sharp waveguide bends in an experimental Japanese system caused serious ripples in the measured loss versus frequency[5] and presumably serious echoes in the impulse response also.

In this paper, we evaluate the seriousness of spurious TE_{0n} mode generation in the WT4 system from the levels of echoes in the impulse response. We explain first, in Section II, how to obtain the impulse response from measurements of loss and delay versus frequency. To establish credibility for the results, we consider round-trip loss and delay measurements of the field test without the diplexer tree[1] and show that the level of the large echoes in the impluse response agrees with independent data. Since the echo test set could not conveniently measure

through the diplexer tree, we made a discrete TE_{0n} generator to simulate the tree. In Section III, we describe theoretical and experimental characterizations of the mode conversion levels in this TE_{0n} generator and compare the level of TE_{02} with that expected from the diplexer tree. In Section IV, we compare the impulse response of Section II with that obtained from round-trip measurements of the field test with the TE_{0n} generator in place. From the data in Sections III and IV, we also estimate in Section V the level of TE_{02} generated within the waveguide itself and show that this level is consistent with an estimate based on mechanical measurements. We conclude, in Section VI, with a discussion of the implications of all these measurements on the WT4 system design.

II. DETERMINATION OF THE IMPULSE RESPONSE

2.1 Theory

The levels of echoes in the impulse response can be determined from the waveguide forward transfer function. First we write the total waveguide transfer function I_o for the TE_{01} mode (the desired signal mode) as:

$$I_o(f) = P(f)H_o(f), \tag{1}$$

where $P(f)$ is the transfer function for ideal waveguide and contains the effects of dispersion and ohmic losses. For a length z of waveguide, this transfer function is

$$P(f) = e^{-\Gamma_o z} = e^{-\alpha z}e^{-j\beta z}. \tag{2}$$

Here $\alpha(f)$ and $\beta(f)$ are the real and imaginary parts, respectively, of the propagation constant Γ_o for the TE_{01} mode. Reflections and mode conversion cause the normalized transfer function $H_o(f)$ to differ from unity. In the absence of reflections and when there is mode conversion from TE_{01} to only one other mode, $H_o(f)$ is identical to the function $G_o[\Delta\Gamma(f)]$ used in studies of mode conversion.[6,7,8] There, $\Delta\Gamma$ is the differential propagation constant

$$\Delta\alpha + j\Delta\beta = \Delta\Gamma = \Gamma_o - \Gamma_n, \tag{3}$$

where Γ_n is the propagation constant of the spurious mode.

Since we know the ideal transfer function $P(f)$, we consider now only the normalized impulse response $h(t)$, which is the Fourier transform of $H_o(f)$:

$$h(t) = \int_{-\infty}^{+\infty} H_o(f)\exp(j2\pi ft)df. \tag{4}$$

To obtain the normalized transfer function $H_o(f)$, we first remove the characteristics of ideal waveguide from the loss and delay $\tau(f)$ data. We integrate the residual delay data to obtain the residual phase versus frequency $\phi(f)$ and then find $H_o(f)$ by exponentiation of the residual loss $A(f)$ and phase:

$$H_o(f) = \exp(-A(f) + j\phi(f)). \quad (5)$$

To reduce echo sidelobe levels, we multiply the real and imaginary parts of $H_o(f) - 1$ by a data window and transform the result using an FFT (fast Fourier transform). This procedure produces $h(t)$ minus the delta function at zero delay. Finally, we square and sum the real and imaginary parts of $h(t)$ to obtain the power in each echo as a function of delay. The echo power is corrected for the attenuation introduced by the data window.

2.2 Example

For illustration, we consider round-trip measurements of the 14-km waveguide medium without the TE_{0n} generator. The baseline for these measurements is the round trip to a closed shutter[9] at the manhole (see Fig. 2). Data over a 350-MHz band near 41 GHz are shown in Fig. 3a. These data are the same as reported previously, except that the delay dispersion characteristic of ideal waveguide has been subtracted out.

Most of the features of the data in Fig. 3a exhibit the expected behavior. The decreasing loss versus frequency is characteristic of the ohmic loss in the waveguide wall. The level of the very rapid ripples is about at the level of the test set noise (0.01 dB rms).[1]

The magnitude of the normalized impulse response shown in Fig. 3b reveals a pair of equal magnitude echoes separated from the main impulse by $t_0 = \pm 10$ ns. These echoes correspond to the ripple seen in the delay data with period $1/t_0 = 100$ MHz. Since these echoes produce a ripple in the delay but not in the loss, they must be 180 degrees out of phase.[10]

This pair of echoes is caused by reflections from a pressure window located 1.5 m in front of the shorting end cap in the test station (Fig. 2). This pressure window is a short length of foam with low relative dielectric constant ϵ embedded in the 60-mm waveguide. Some energy is reflected from the pressure window without ever reaching the shorting end cap and returns to the test set 10 ns before the main pulse. Other energy, after being reflected from the end cap, is reflected from the pressure window back to the end cap a second time and finally arrives at the test set 10 ns after the main pulse.

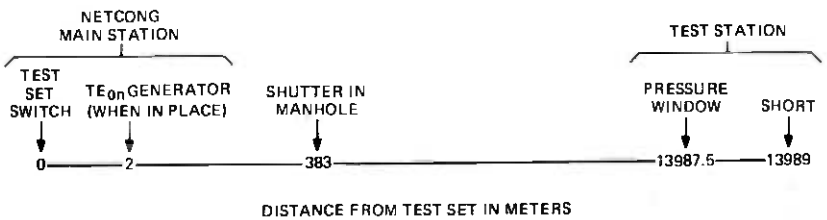


Fig. 2—Distances in the WT4 field test.

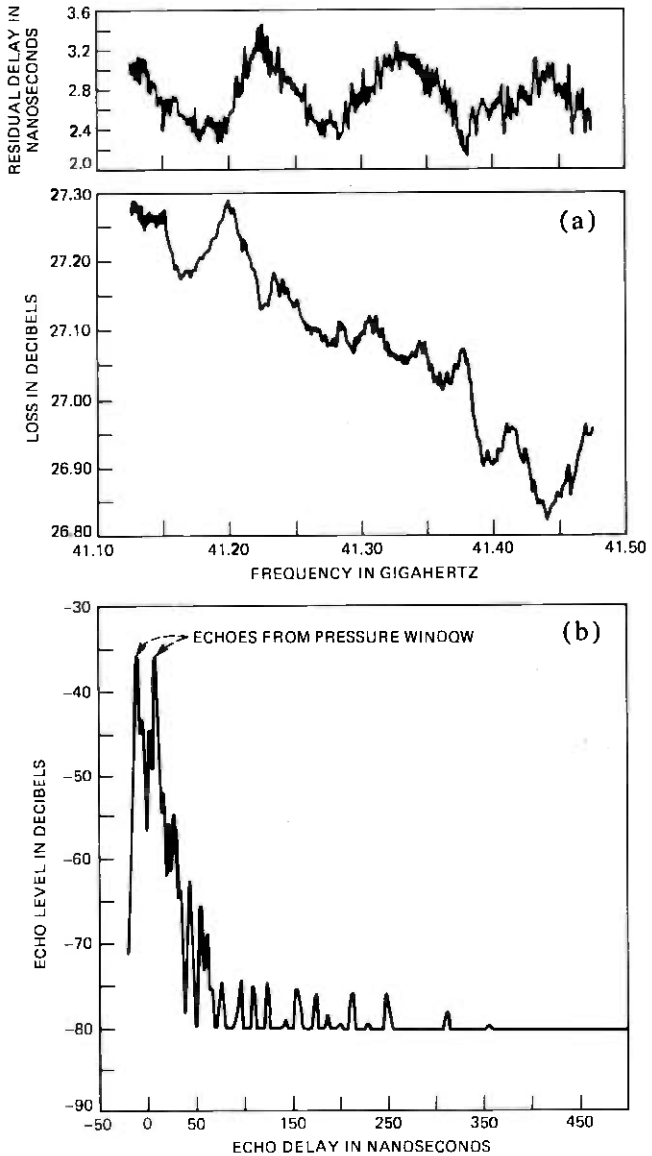


Fig. 3—(a) Measured loss and residual delay for the 27.21-km round trip between the manhole and the test station. (b) Normalized impulse response corresponding to Fig. 3a.

The magnitude of this echo pair seen in Fig. 3b agrees with other measurements. Fault location measurements made at 41.24 GHz with the shorting end cap removed also gave $20 \log_{10}|R| = -37$ dB for the total reflection coefficient R at the location of the pressure window.[11] Furthermore, since the two echoes are 180 degrees out of phase, it is easy to

show that the reflection coefficient R_o for each end of the window must satisfy the relation $R_o = |R|/2 = 0.007$. Then we infer that $\sqrt{\epsilon} - 1 \simeq 0.014$, or $\epsilon = 1.03$, which is exactly the dielectric constant specified for the foam.[12]

The remaining echo power in the impulse response shown in Fig 3b is quite small. We conclude that there is negligible power in long echo trains caused by mode conversion and reconversion within the 60-mm waveguide medium itself. Considerably stronger echoes appear, however, when the TE_{0n} generator is inserted to simulate the diplexer tree.

III. TE_{0n} GENERATOR CHARACTERIZATION

The TE_{0n} generator was designed for insertion on the 60-mm side of the circular waveguide taper at the Netcong test set platform (Figs. 2 and 4). The interior of the TE_{0n} generator is a 76-mm (3-in.) length of reduced diameter waveguide. The inside diameter is 54 mm (2.125 in.) over the central 50.8-mm (2 in.) region. To reduce reflections, the diameter is tapered linearly up to 60 mm (2.362 in.) in 12.7-mm (0.5 in.) tapers on each end of the generator.

Perturbation theory[6] applied to the above diameter variation predicts excitation of spurious modes with the levels shown in Table I. We may expect these estimates to be somewhat in error, because the diameter change from 60 to 54 mm is a fairly large fraction of the original diameter and therefore may violate the assumption inherent in the perturbation theory.

We determined experimentally the levels of spurious TE_{0n} modes excited by the generator from the levels of ripples in the round-trip loss from the test set to the shutter in the manhole. We obtained a baseline

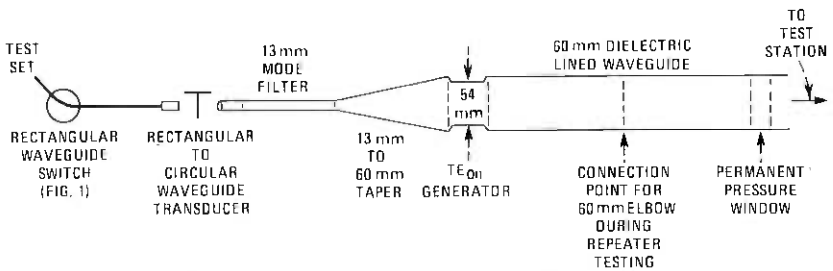


Fig. 4—Layout of test components at Netcong main station including the TE_{0n} generator.

Table I — Mode levels relative to TE_{01} at the TE_{0n} generator output (perturbation theory)

	40 GHz	45 GHz
TE_{02}	−13 dB	−14 dB
TE_{03}	−16 dB	−15 dB
TE_{04}	−24 dB	−34 dB

measurement in this case by closing the rectangular waveguide switch in the test set (see Figs. 1 and 4). The net round-trip measurement sees the effects of two equal discrete mode converters of strength x separated by a distance 2ℓ, where ℓ is 381 m (see Fig. 2). The measured mode conversion loss A due to one spurious TE_{0n} mode is then:[6]

$$A = x^2(1 + e^{-|\Delta\alpha|2\ell} \cos \Delta\beta 2\ell) \text{ nepers.} \tag{6}$$

Here $\Delta\alpha$ and $\Delta\beta$ are the differential attenuation and phase, respectively, of the TE_{0n} mode relative to TE_{01} (see eq. (3)). The first term in eq. (6) is the round-trip insertion loss of the generator. Because $\Delta\beta$ varies with frequency, the second term in eq. (6) represents a loss ripple.

The period of the ripple in eq. (6) corresponds to a change in $\Delta\beta$ of $2\pi/2\ell$. The period δf in frequency is:

$$\delta f = \left(\frac{df}{d\Delta\beta}\right)\left(\frac{2\pi}{2\ell}\right). \tag{7}$$

Equivalently,

$$t_R = \frac{1}{\delta f} = \left(\frac{d\Delta\beta}{d\omega}\right)2\ell, \tag{8}$$

where t_R is clearly the round-trip differential group delay of TE_{0n} relative to TE_{01}.

The theoretical parameters for the lowest order TE_{0n} modes are shown in Table II for 41.3 GHz, which is close to the center frequency of our measurements. We also list for reference the beat wavelengths $2\pi/\Delta\beta$ and wall coupling coefficients Ξ. For TL_{0n} modes, these parameters are not very sensitive to the thickness or properties of the dielectric liner on the wall of the waveguide.

The data in Fig. 5 show clearly that the TE_{0n} generator excites TE_{02}, TE_{03}, and TE_{04}. The loss versus frequency data in Fig. 5a show the 14-MHz ripple period expected from TE_{02} in a 762-m round trip. The presence of TE_{03} and TE_{04} is seen more easily after Fourier transformation of the loss.

With the help of eq. (6), we determined the mode conversion levels plotted in Fig. 5b. Defining $A(k)$ as the N-point discrete Fourier transform of $A(f)$ in eq. (6), we find that the peak magnitudes in $A(k)$ are:

Table II — TE_{0n} parameters at 41.3 GHz in 60-mm waveguide*

	TE_{02}	TE_{03}	TE_{04}	
$\lvert\Delta\alpha\rvert$	0.26	0.70	1.38	nepers/km
$\Delta\beta$	22.73	59.76	113.23	radians/m
$2\pi/\Delta\beta$	0.276	0.105	0.056	m
Ξ	1216	−1804	2447	m^{-2}
$d(\Delta\beta)/d\omega$	−0.092	−0.253	−0.514	ns/m
t_R (in 762 m)	70	193	392	ns
δf (for 762 m)	14.3	5.2	2.6	MHz
$8.686\lvert\Delta\alpha\rvert\ell$ (in 381 m)	0.86	2.32	4.57	dB

* Assuming a 179-μm dielectric liner with $\epsilon_r = 2.28$ and $\tan\delta = 1.25 \times 10^{-4}$.

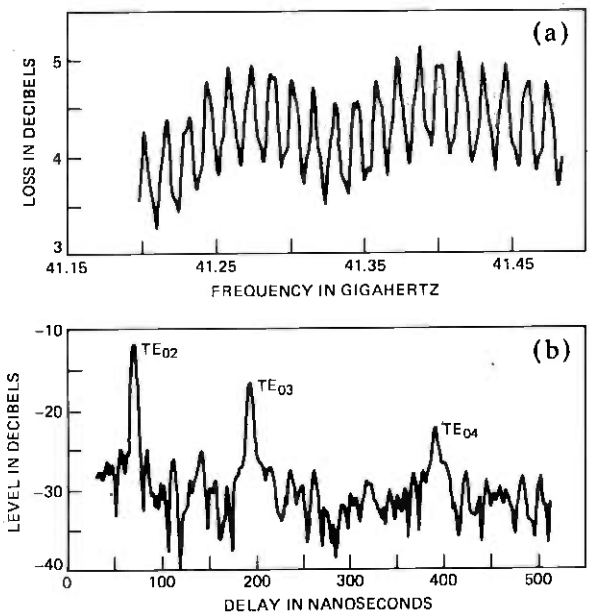


Fig. 5—(a) Measured loss for the 766-m round trip between the test set and the manhole with the TE_{0n} generator in place. (b) Modified Fourier transform of Fig. 5a, showing levels of TE_{0n} modes generated at the TE_{0n} generator.

$$|A(k)| = \frac{N}{2} x^2 e^{-|\Delta\alpha|2\ell}, \qquad k = Nt_R\Delta f, \tag{9}$$

where Δf is the sampling interval in frequency, and t_R is defined in eq. (8). For these data, $\Delta f = 1$ MHz, corresponding to the 1-μs total sampling interval of the original time domain measurements.[1] Inverting eq. (9) to find the mode conversion levels x, we obtain:

$$a_k \equiv 20 \log_{10} x = 10 \log_{10} \left(\frac{2|A(k)|}{N}\right) + 8.686|\Delta\alpha|\ell, \; k = Nt_R\Delta f. \tag{10}$$

The ratio of $8.686|\Delta\alpha|\ell$ to round-trip delay t_R is very nearly constant for TE_{0n} modes and is within 3 percent of 0.012 dB/ns for the three modes and conditions of Table II. Using this fact to evaluate the last term in eq. (10), we then plotted eq. (10) in Fig. 5b from the data in Fig. 5a.

The data in Fig. 5b agree quite well with the theoretical expectations for the TE_{0n} generator listed in Table I. The peaks corresponding to TE_{02}, TE_{03}, and TE_{04} occur at delays within 1 ns of those predicted in Table II. The mode levels inferred from Fig. 5b are −12 dB, −17 dB, and −23 dB relative to TE_{01} for TE_{02}, TE_{03}, and TE_{04}, respectively. In particular, the level of TE_{02} generated is roughly equal to the worst possible TE_{02} mode level that could be generated in the band diplexer tree between 40 and 110 GHz.[13]

IV. LONG-DISTANCE ECHO TRAIN MEASUREMENTS WITH THE TE_{0n} GENERATOR

The presence of the TE_{0n} generator causes the round-trip measurements of the 14-km field test to differ significantly from the measurements presented in Section II. In this section, we examine these differences, first in the frequency domain and then in the time domain. The time domain data provide an estimate of the impulse response of the entire waveguide system including the diplexer tree.

Data from measurements with and without the TE_{0n} generator are displayed together in Fig. 6a. For these measurements, the shorting switch in the echo test set was used to obtain the baseline. The difference in average loss with and without the TE_{0n} generator is due mainly to the mode conversion loss for a round trip through the generator. For the mode conversion levels observed in Fig. 5 for TE_{02}, TE_{03}, and TE_{04}, this loss is theoretically 0.77 dB. The additional discrepancy is consistent with the test set's ±1 percent accuracy in absolute loss.[1] A minor difference in average delay is caused by a slight difference in the pressure of nitrogen filling the waveguide. [We used only the nominal pressure to estimate the effect of the refractive index of nitrogen on the ideal transfer function $P(f)$ of eq. (1).]

The level of the rapidly varying ripples increased substantially after insertion of the TE_{0n} generator. An additional slowly varying loss ripple has also appeared in Fig. 6a that is not present in the measurement made with the baseline in the manhole (Fig. 3a). The origin of these ripples is clarified by the time domain data in Fig. 6b, which was obtained from Fig. 6a by the method described in Section II.

An additional echo at a 10-ns delay is evident in Fig. 6b and is due to a reflection from the precision attenuator in the test set. The echoes leading and lagging the main signal are no longer equal in level, as they were when the baseline was taken at the manhole (see Fig. 3b and Section II). Correspondingly, a loss ripple with a period of about 100 MHz appears in Fig. 6a.

The large echoes in Fig. 6b with delays of 35 and 96 ns may be due to mode conversion at the TE_{0n} generator followed by reconversion near the manhole (and vice versa). The differential delays for the 381 meters between the generator and the manhole are indeed 35 and 96 ns for TE_{02} and TE_{03}, respectively (see Table II).

The power in echoes with long delays generally is substantially greater with the TE_{0n} generator in place than without it. For example, the integrated total power in the echoes shown in Fig. 6b with delays greater than 50 ns, excluding the echo at 96 ns, is 44 dB below the signal. In contrast, the power in such echoes (and test set noise) shown in Fig. 3b is 57 dB below the signal. These levels of echo power are consistent with the expected levels of spurious TE_{0n} modes generated in the waveguide itself, as discussed next.

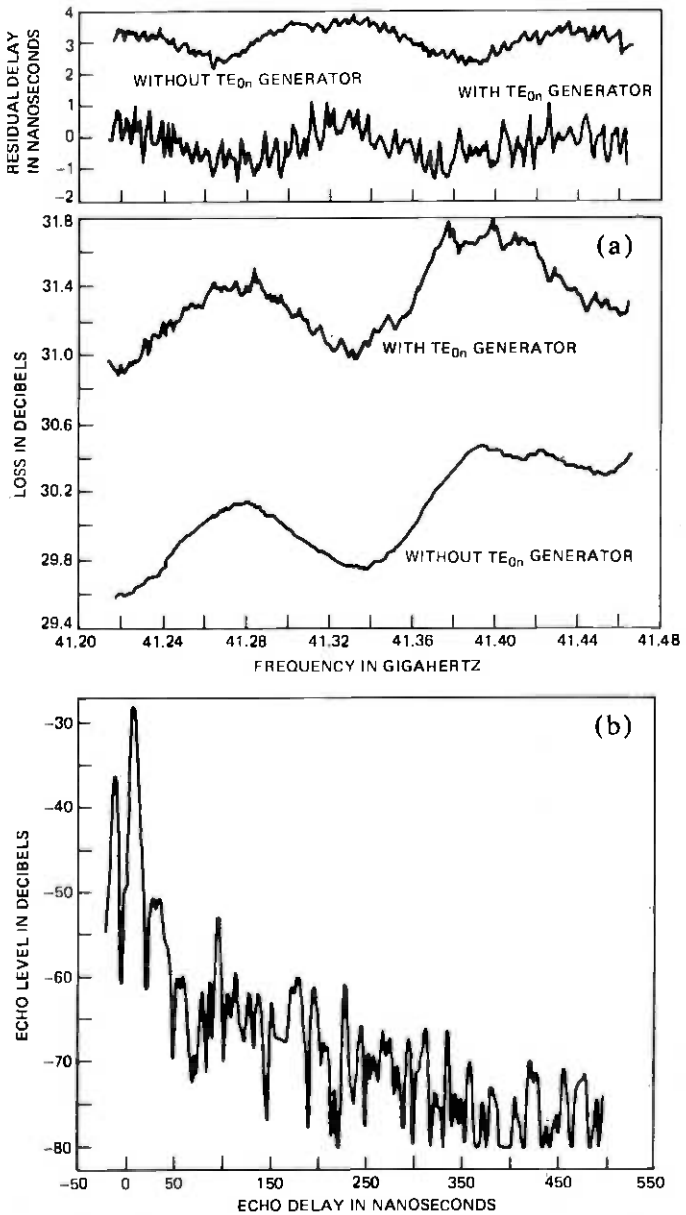


Fig. 6—(a) Measured loss and residual delay for the 27.98-km round trip between the test set and the test station. Measurements with and without the TE_{0n} generator in place are shown. (b) Normalized impulse response corresponding to the data in Fig. 6a for the TE_{0n} generator in place.

V. COUPLING TO TE_{0n} MODES IN THE BURIED 60-mm WAVEGUIDE MEDIUM

5.1 Levels estimated from electrical measurements

From the measurement data presented in Sections III and IV, we can now estimate the amount of spurious TE_{0n} mode power generated in the buried 60-mm waveguide. First, from Section III (Fig. 5 and Table II), there is no evidence that the TE_{0n} generator excites any modes but those in the TE_{0n} family. Therefore, the increase in power in echoes with long delay seen in Fig. 6b is due solely to mode conversion to and from TE_{0n} modes.

For simplicity, let us assume that spurious TE_{0n} energy generated in the buried waveguide is predominantly in the TE_{02} mode. Let us also assume that the level of TE_{02} generated per unit length does not vary greatly with position along the 14-km route. Because TE_{02} has a low differential attenuation per unit length $|\Delta\alpha|$, almost all the TE_{02} energy reconverted at the TE_{0n} generator is then in echoes with delays longer than, say, 50 ns.

To calculate the level of TE_{02} incident on the TE_{0n} generator from the waveguide, we argue in reverse. If the TE_{02} level returning from a round trip in the 60-mm waveguide and incident on the TE_{0n} generator were -38 dB, then the reconverted power seen in TE_{01} would be 12 dB lower, or -50 dB. However, the total reconverted power also contains a contribution from energy converted to TE_{02} at the generator at the beginning of the round trip and then reconverted along the waveguide. Since the coupling is completely reciprocal, the echoes caused by coupling to and from TE_{02} on the forward and reverse legs of the round trip will add in phase. Hence the total reconverted power in echoes would be 6 dB greater, or -44 dB, as was measured (excluding the echo at 96 ns and those with delays less than 50 ns).

Coupling to higher order TE_{0n} modes probably causes a smaller fraction of the total echo power than does coupling to TE_{02}, because those modes have a higher differential attenuation and also less coupling at the TE_{0n} generator (see Tables I and II). The level of TE_{02} generated through continuous coupling in a one-way trip through the buried waveguide is therefore roughly 40 dB below the level of TE_{01}.

5.2 Levels estimated from mechanical measurements

From independent measurements, we can estimate the level of TE_{02} that is generated by various coupling mechanisms. These mechanisms include continuous coupling at diameter distortions in helix and dielectric-lined waveguide and coupling at localized diameter distortions at the waveguide flanges.

The average TE_{02} spurious mode level generated by continuous cou-

pling in dielectric-lined waveguide can be calculated from the spectrum of waveguide diameter variation with distance.[6] At the mechanical frequency $\Delta\beta/2\pi$ appropriate for coupling to TE_{02} near 41 GHz, the average spectrum measured on individual waveguide sections corresponds to an average TE_{01} mode conversion loss in 14 km of about 5×10^{-5} neper (0.0004 dB). According to the coupled power equations,[14] the average random spurious mode power P builds up to a steady state independent of the length L of waveguide, when the total differential attenuation $|\Delta\alpha|L$ is large. This steady-state level is

$$P = \frac{\langle A \rangle}{|\Delta\alpha|}, \tag{11}$$

when $\langle A \rangle$ is the average TE_{01} loss per unit distance. Taking $|\Delta\alpha|$ from Table II, we find $10 \log_{10} P$ to be about -48 dB for continuous conversion in dielectric-lined waveguide.

Analysis of diameter data for some helix mode filters in the field test showed that TE_{02} levels of about -51 dB could be generated in some of these filters near 50 GHz.[15] If we assume every mode filter generates that much TE_{02} and take into account the effect of differential attenuation, we find that the steady-state level P after 14 km is about 3 dB higher, or $P = -48$ dB.

The TE_{02} level generated at the flanges can be very serious, but the randomized waveguide length scheme used in the field test reduces this level considerably.[16] Without randomization of lengths, the steady-state TE_{02} power after a long distance would be only 16 dB below the power in TE_{01}. (This level was calculated from measured diameter distortions near flanges.) Randomization of the waveguide lengths destroys the coherent build-up of spurious mode power and thus keeps the average random TE_{02} mode power P at a steady state 45 dB below the power in TE_{01}.

The sum of the calculated TE_{02} mode power generated by the above three sources of coupling is -42 dB, which is close to the value of -40 dB inferred above from the electrical measurements. While the levels calculated from the mechanical and electrical measurements are only rough estimates, the agreement between them is reasonable.

Reconversion of TE_{02} mode power at strong couplers such as the TE_{0n} generator or the diplexer tree is evidently always more serious than continual conversion and reconversion along the waveguide alone. The calculated reconverted power caused by the above three sources of mode coupling without the TE_{0n} generator in place is less than -80 dB and thus is completely invisible in Fig. 3. For the case without randomized lengths, the total power reconverted from TE_{02} with and without the TE_{0n} generator would be -22 dB and -50 dB, respectively.

VI. CONCLUSIONS

The power in long echo trains caused by coupling between TE_{01} and spurious TE_{0n} modes is too small to affect the performance of the WT4 system. In the two- and four-phase WT4 systems, the designed thermal noise levels are 22 and 28 dB below the carrier, respectively.[17] By comparison, the estimated maximum power in long echo trains in the WT4 field test is about 44 dB below the signal power after a 28-km round trip including the diplexer tree. The system thermal noise will therefore certainly swamp out the power in long echo trains.

Assuming the waveguide lengths are randomized as in the field test, the power in long echo trains should be no greater in a 50-km or 60-km repeater hop than in the field test. After a certain distance in 60-mm waveguide, the average random power in any spurious TE_{0n} mode reaches a steady state independent of the length of the waveguide. Echo power generated by continuous mode conversion and reconversion in the 60-mm waveguide alone does increase indefinitely with the length of the waveguide. The level of this echo power, however, is generally much smaller than the echo power caused by reconversion at the band diplexer tree of spurious TE_{0n} mode power generated in the 60-mm waveguide.

Filters of spurious TE_{0n} modes are therefore not necessary in the WT4 system. Such filters are difficult to fabricate without introducing significant TE_{01} insertion loss, and they also require obstructions within the waveguide[18] that may interfere with waveguide maintenance.

VII. ACKNOWLEDGMENTS

Helpful suggestions and comments from D. A. Alsberg, H. E. Rowe, and D. T. Young are acknowledged.

REFERENCES

1. J. L. Doane, "Measurement of the Transfer Function of Long Lengths of 60-mm Waveguide with 1-MHz Resolution and High Dynamic Range," Digest of the Conference on Precision Electromagnetic Measurements, Boulder, Colorado, June 28 to July 1, 1976, IEEE Publication No. 76, CH1099-1 IM, pp. 153–155.
2. D. A. Alsberg, J. C. Bankert, and P. T. Hutchison, "The WT4/WT4A Millimeter-Wave Transmission System," B.S.T.J., *56*, No. 10 (December 1977), pp. 1829–1848.
3. E. T. Harkless, A. J. Nardi, and H. C. Wang, "Channelization," B.S.T.J., *56*, No. 10 (December 1977), pp. 2089–2101.
4. E. A. Marcatili, "Miter Elbow for Circular Electric Mode," Proc. Symp. on Quasi-Optics, Polytechnic Institute of Brooklyn, June 1964, p. 535.
5. K. Yamaguchi, K. Kondoh, F. Nihei, and M. Kikushima, "Transmission Characteristics of an Experimental Millimeter-Waveguide Line including Miter Elbows," Rev. Elec. Commun. Lab., *20*, Nos. 11-12 (November–December 1972), pp. 1114–1118.
6. H. E. Rowe and W. D. Warters, "Transmission in Multimode Waveguide with Random Imperfections," B.S.T.J., *41*, No. 5 (May 1962), pp. 1031–1170.
7. H.E. Rowe and D. T. Young, "Transmission Distortion in Multimode Random Waveguides," IEEE Trans. Microwave Theory and Techniques, *MTT-20*, No. 6 (June 1972), pp. 349–365.
8. H. E. Rowe and D. T. Young, "Minimum Phase Behavior of Random Media," IEEE Trans. Microwave Theory and Techniques, *MTT-23*, No. 5 (May 1975), pp. 411–416.

9. M. A. Gerdine, L. W. Hinderks, S. D. Williams, and D. T. Young," Electrical Transmission Measurement System," B.S.T.J., *56*, No. 10 (December 1977), pp. 2025–2034.
10. H. A. Wheeler, "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," Proc. IRE, *27*, No. 6 (June 1939), pp. 359–384.
11. J. L. Doane, unpublished work.
12. S. Shapiro, private communication.
13. J. L. Doane and D. N. Zuckerman, unpublished work.
14. S. E. Miller, "The Nature of and System Inferences of Delay Distortion Due to Mode Conversion in Multimode Transmission Systems," B.S.T.J., *42*, No. 9 (November 1963), pp. 2741–2760.
15. S. C. Moorthy, unpublished work.
16. J. L. Doane, unpublished work.
17. S. Cheng, unpublished work.
18. K. Hashimoto, "Circular TE_{0n} Mode Filters for Guided Millimeter-Wave Transmission," IEEE Trans. Microwave Theory and Techniques, *MTT-24*, No. 1 (January 1976), pp. 25–31, and references cited therein.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Dependence of Depolarization on Incident Polarization for 19-GHz Satellite Signals

BY H. W. ARNOLD and D. C. COX

(Manuscript received July 20, 1978)

Rain and ice crystals depolarize radio waves along earth-satellite propagation paths. The magnitude of this depolarization is a function of incident polarization angle and is minimized when polarization and depolarizer symmetry axes coincide. A technique is presented for a direct determination of the medium's attenuation and depolarization for any incident polarization, based on measurements taken at two orthogonal polarizations. Some sample results from this technique are presented, using data collected at Crawford Hill, New Jersey using the 19-GHz COMSTAR satellite beacon.

I. INTRODUCTION

Depolarization caused by rain and ice crystals is an important factor in the design of future satellite communication systems operating at frequencies above 10 GHz.[1,2] These systems will likely use dual, orthogonal polarizations to increase transmission capacity. Quantitative knowledge of depolarization is necessary for determining the pair of polarizations that experience the least depolarization, for determining whether the isolation between any two polarizations is adequate during rain and ice depolarizing conditions, or for guiding the design of circuits for canceling crosstalk resulting from depolarization if the isolation is not adequate.

Only a few measurements have been made of depolarization along earth-space propagation paths.[3–5] These measurements were not necessarily made at incident polarizations that produce minimum depolarization. Aerodynamic forces acting on vertically falling raindrops are expected to orient raindrop symmetry axes vertical and horizontal on the average. Thus, rain depolarization is expected to minimize at linear horizontal and vertical polarizations. Maximum depolarization is expected for linear polarizations oriented 45 degrees to horizontal and for circular polarizations.[5,6] These expectations are supported by mea-

surements for terrestrial propagation paths, but there is no experimental evidence available to either confirm or refute them for earth-space paths. Depolarization caused by ice crystals with unknown orientations raises additional uncertainty in the expected behavior of depolarization along earth-space paths.[7]

It is difficult and expensive to instrument satellites to measure depolarization directly at several incident polarizations. A technique involving direct measurement of elements in the polarization transmission matrix and direct calculation of depolarization for other incident polarizations was proposed for the COMSTAR beacon propagation experiments.[8,9] This calculation method has the potential for answering many of the questions regarding rain and ice depolarization along earth-space propagation paths. This paper presents what are believed to be the first experimental results obtained using the technique.[10] A minimum in depolarization occurs for vertical and horizontal incident polarizations for the attenuation event analyzed.

II. THE MEASUREMENTS

2.1 Equipment and signal parameters

The transmitting source for the propagation measurements is the 19-GHz beacon[11] on the COMSTAR satellite located at 95°W longitude. This beacon output is switched at a 1-kHz rate between two linear orthogonal polarizations. These 19-GHz signals, among others, are received at Crawford Hill, New Jersey with a 7-m diameter antenna and precision receiving electronics described in Refs. 12 and 13.

The polarizations of the beacon signals received at Crawford Hill are rotated 21 degrees from horizontal and vertical (for simplicity, these signals are referred to as H and V). Amplitudes of copolarized, V and H, and cross-polarized, XV and XH, signal components are measured for the two transmitted polarizations. Phase differences, ϕ, referenced to the V signal are measured for the H, XV, and XH signals. The receiver 3-dB predetection bandwidths of 10 Hz in co- and crosspolarized channels yield a dynamic range of 60 dB between the clear air copolarized signal level and the receiver noise level; crosspolarized signal amplitudes are also measured in 1-Hz bandwidths yielding a 70-dB dynamic range. Post-detection bandwidths are 1 Hz for all amplitude and phase measurements.

Residual signals in the crosspolarized signal channels are produced by transmitting and receiving antenna imperfections. These residuals are <-35 dB below the copolarized signal levels everywhere within the -3 dB beamwidth of the receiving antenna. Residuals are <-40 dB everywhere within the beam for the XV channel (XV is the channel for transmit vertical and receive horizontal); on-axis residuals are typically <-45 dB for XV and -36 to -40 dB for XH.

2.2 Calibration

The crosspolarized signal channels are calibrated with the antenna feed in the normal receiving position. The receiver polarization switches (see Figs. 1 and 14 of Ref. 13) are temporarily configured to connect the H and V copolarized signals through the XV and XH channels respectively. These known signal levels are then used to establish the gains of the crosspolarized signal channels. These channel gains remain within ±0.1 dB over many months. The H and V signals are within 0.05 dB of each other as measured by physically rotating the receiving antenna feed assembly 90 degrees. The phase scale for the differential phase, ϕV-H, is calibrated by reversing the polarity of one signal input to produce a 180-degree phase change.

Determination of the zero on the phase scales for the crosspolarized signals, ϕV-XV and ϕV-XH, requires simultaneous insertion of signals with known phase relationship into the receiving feeds. Rotating the feed assembly inserts projected spatial components of the copolarized signals into both feeds. These components have the same phase and will produce signals with either 0- or 180-degree phase difference in the crosspolarized signal channels; the sense, 0 or 180, depends on the direction of feed rotation. Contamination by the residual crosspolarized signals produces an uncertainty in this zero of about ±1 degree. This zero calibration remains within ±2 degrees over many months. The phase scale is calibrated by reversing the polarity of one signal input while rotating the feed assembly for the zero determination. Note that the phases of the clear-air residual crosspolarized signals are not suitable for determining phase scale and zero because they are low level and their phases are extremely sensitive to the accuracy of the nulling of the residuals with feed assembly rotation.[7]

2.3 Baseline removal

The amplitudes and phases of co- and crosspolarized beacon signals exhibit small diurnal variations. These diurnal baseline variations repeat very closely from day to day, but change somewhat from season to season. The variations result from thermal changes in the satellite beacon circuits, waveguides, and antennas as the solar illumination changes. Day-to-day receiver contributions to baseline variations are insignificant. Because of the repeatability of these baseline variations, they can be removed from the experimental data.

V and H attenuations and differential phase, ϕV-H, are determined by subtracting, from values measured during propagation events, the values measured at the same clock times on clear days within a few days of the events. The differential values of attenuation and phase that are critical in the calculation of depolarization for other incident polarizations are determined within ±0.2 dB and ±0.5 degree.

Baseline removal of residual crosspolarized components is done by vectorially subtracting, from the values measured during events, the weighted residual components at the same clock times. These vector subtractions can be made since both amplitudes and phases are measured. Prelaunch measurements of the beacon antennas, pattern range measurements on the receiving antenna, and measurements with the receiving feed at normal polarization and rotated 90 degrees all suggest that the XH residual is dominated by the beacon antenna. Thus, the XH residual is weighted by the event attenuation and differential phase as though it were all contributed by the beacon. Weighting of the XV residual assumes equal contribution by beacon and receiving antennas. The vector subtraction of clear-air crosspolarized channel residuals suppresses the residuals to −45 to −50 dB below copolarized signal levels before and after events and on adjacent clear days. This accuracy degrades with the intensity of the propagation event because of the uncertainty in assigning residual contributions to beacon and receiving antennas. However, during weak events when depolarization is low and residual suppression is needed, the residual is suppressed to −45 to −50 dB; when the event is stronger and the suppression degrades, the depolarization is strong and the relative errors contributed by the residual are less significant anyway.

III. POLARIZATION ROTATION TECHNIQUE

This section introduces the concept of the medium transmission matrix for a given set of incident and received polarizations. Receiver outputs directly generate the transmission matrix for the incident polarizations; results for other incident polarizations are generated through rotations of the measured matrix.

Let us assume two orthogonal signals, E_{Ta} and E_{Tb}, incident on the medium. The medium outputs in the same reference frame are E_{Ra} and E_{Rb}. The medium is then completely described by four coupling coefficients, as shown in Fig. 1. Coefficients A and D represent the copolarized signal attenuations of the medium, while B and C indicate the medium's depolarization. Note that these coefficients are, in general, complex. The above relation may be written in matrix form:

$$\begin{bmatrix} E_{Ra} \\ E_{Rb} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E_{Ta} \\ E_{Tb} \end{bmatrix}. \tag{1}$$

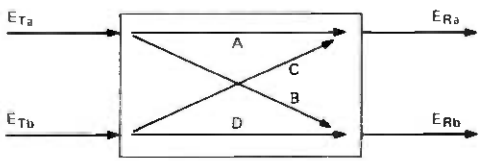


Fig. 1—Description of propagation medium by coupling coefficients A, B, C, D (E_{Ta} and E_{Tb} transmitted, E_{Ra} and E_{Rb} received).

Since the 19-GHz COMSTAR beacon transmits alternately on two orthogonal polarizations, E_{Ta} and E_{Tb} are alternately 0. Receiver outputs[8,13] are thus directly proportional to A and C (for $E_{Tb} = 0$) or B and D ($E_{Ta} = 0$). The transmission matrix T_M for the actual incident polarizations is given by these measured values.

$$[T_M] = \begin{bmatrix} A_M & B_M \\ C_M & D_M \end{bmatrix}. \tag{2}$$

This measured transmission matrix may now be used to generate the transmission matrix for any other orthogonal set of incident polarizations. This operation is done without recourse to path models. Linear, circular, or elliptical polarizations are admissible; the following discussion considers only linear polarization. As shown in Fig. 2, the transmission matrix for the 1–2 frame is desired, based on measurements obtained in the $M_a - M_b$ frame. Components in the 1–2 frame are obtained from those in the $M_a - M_b$ frame through the following relation:

$$\begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = [R_\theta] \begin{bmatrix} E_{Ma} \\ E_{Mb} \end{bmatrix}, \tag{3}$$

where

$$[R_\theta] = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{4}$$

To generate the transmission matrix for incident polarizations along the 1–2 axes, the 1–2 inputs are rotated into the measurement frame by $[R_\theta]$, passed through the transmission matrix $[T_M]$, and rotated back to the 1–2 frame by $[R_\theta]^{-1}$, i.e.,

$$\begin{bmatrix} E_{R1} \\ E_{R2} \end{bmatrix} = [R_\theta]^{-1} [T_M][R_\theta] \begin{bmatrix} E_{T1} \\ E_{T2} \end{bmatrix}. \tag{5}$$

The transmission matrix $[T_{M,\theta}]$ for two orthogonal polarizations rotated an angle θ from the measurement frame is thus

$$[T_{M,\theta}] = [R_\theta]^{-1} [T_M][R_\theta]. \tag{6}$$

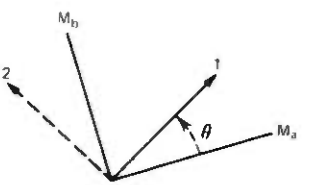


Fig. 2—Relation between reference frames of measured ($M_a - M_b$) and desired (1 – 2) transmission matrices.

IV. OBSERVATIONS

The accuracy of the polarization rotation technique described here is critically dependent on amplitude and phase calibration of all four receiver channels.[9] To verify the end-to-end operation of the receiver and rotation software, a depolarizer with known symmetry axes was inserted near the receiving antenna focus in clear weather. Data were taken at several orientations of the depolarizer with respect to the actual received polarizations. For each case, depolarization was computed as a function of incident polarization angle relative to the depolarizer symmetry axes. All results were identical, with minimum depolarization along the depolarizer symmetry axes. This agreement validates both the receiver calibration and rotation software.

Data from a recent rainstorm were analyzed to determine the dependence of depolarization and attenuation on incident polarization angle. Since little depolarization was observed before or after the high-attenuation portion of the event, it appeared that rain, rather than ice, was the predominant depolarizing agent.[7,10]

Some effects of incident polarization angle are shown in Fig. 3 for three points in this storm. These points had maximum vertical copolarized attenuations of 5.4, 10.4, and 20.3 dB. The lower three curves show the effect of vertical depolarization (the ratio of vertical crosspolarized to copolarized levels) as a function of polarization rotation from local ver-

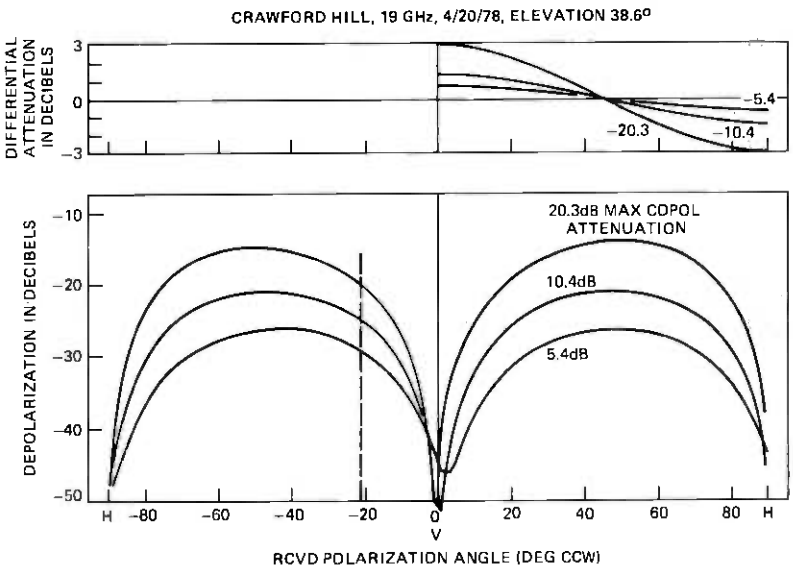


Fig. 3—Variations in depolarization and differential attenuation with incident polarization angle for three points during April 20, 1978 rainstorm. Solid curves are calculated from measurements made at an incident polarization of −21 degrees indicated by the dashed line.

tical to horizontal. The experimental data, taken at a 21-degree rotation angle, appear along the dashed line. All three curves exhibit a sharp null within 2 degrees of local vertical, indicating a mean raindrop canting angle close to 0 degree.[5,6,10]

The upper three curves indicate the variation in differential attenuation (the ratio of horizontal to vertical copolarized attenuations) with polarization angle. Maximum differential attenuation coincides with minimum depolarization, since the raindrops then exhibit their maximum oblateness along the two incident polarizations. Differential attenuation changes sign at 45 degrees from this point, as "vertical" and "horizontal" interchange roles.

The temporal history of a portion of this rainstorm is shown in Fig. 4. The upper two curves indicate maximum depolarization (for 45-degree polarization) and the concurrent copolarized attenuation. A maximum depolarization of −11 dB was observed at 30-dB copolarized attenuation. Maximum depolarization exceeded −20 dB for all copolarized attenuations exceeding 11 dB.

The lower curves indicate time variations for polarization angles in the vicinity of the nulls shown in Fig. 3. The central curve indicates the polarization angle exhibiting minimum depolarization, while the outer curves show the −20 and −30 dB depolarization contours around this

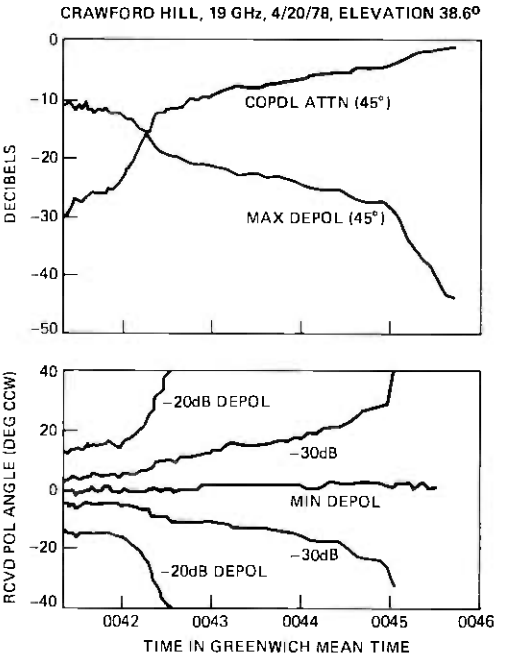


Fig. 4—Temporal history of portion of April 20, 1978 rainstorm. Upper curves show copolarized attenuation and maximum depolarization. Lower curves show angular location and extent of depolarization minima.

minimum. For this portion of this event, the polarization angle exhibiting minimum depolarization remained within 2 degrees of local vertical. Depolarization remained below − 30 dB for all polarizations within 3.5 degrees of local vertical, and below −20 dB within 12 degrees of local vertical.

The relation between attenuation and depolarization appeared to be well-behaved for the section shown of this rainstorm. Other regions, however, appeared less homogeneous. Figure 5 shows the polarization angle dependence of depolarization and differential attenuation for two points in the storm separated by 19 minutes. Both points had 10.5-dB copolarized attenuation. The earlier point, however, had 2.5-dB lower maximum depolarization, almost 50 percent lower differential attenuation, and more than 50 percent greater maximum differential phase shift between copolarized channels. In addition, the depolarization null was filled in below −41 dB. All these changes are consistent with lower average raindrop ellipticity (or less preferential drop orientation) and the presence of a small amount of depolarization due to ice crystals.[7,10] Very preliminary observations of other events have suggested that depolarization nulls occasionally deviate far from local vertical during periods of predominantly ice depolarization.

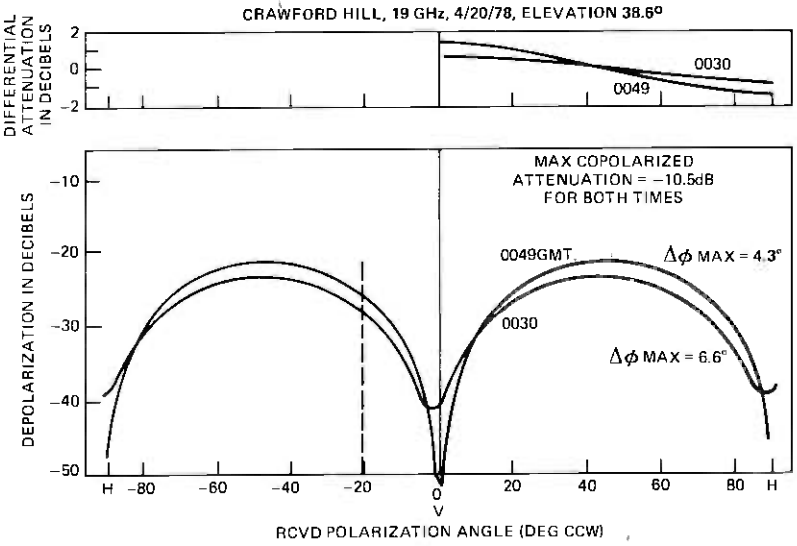


Fig. 5—Variations in depolarization and differential attenuation with incident polarization angle for two points in April 20, 1978 rainstorm exhibiting identical copolarized attenuation but different depolarization characteristics. Solid curves are calculated from measurements made at an incident polarization of −21 degrees indicated by the dashed line.

V. CONCLUSIONS

A method has been presented, based on transmission measurements at one set of incident polarizations, for direct calculation of attenuation and depolarization of the transmission medium for any incident polarization. This technique has been applied to 19-GHz data from the Crawford Hill COMSTAR beacon propagation receiver. Test results indicate that the receiver calibration accuracy is sufficient to provide useful information on polarization-dependent effects along earth-space propagation paths.

The rainstorm analyzed exhibited a sharp depolarization minimum for incident polarizations within 2 degrees of vertical and horizontal. This implies that the mean raindrop symmetry axis was nearly vertical. The minimum depolarization was sufficient for communication systems employing polarization reuse.

Similar observations of ice-produced depolarization events suggest that the depolarization minima for these events occur occasionally at polarizations other than vertical and horizontal. This implies ice particle alignment by other than gravity. The simultaneous presence of ice and rain with differing symmetry axes could destroy the depolarization nulls observed with either separately. Such events, if frequent, could have a serious impact on dual-polarization communication systems. The technique described here will allow further investigation of this and other polarization-dependent propagation phenomena on earth-space paths.

REFERENCES

1. D. O. Reudink, "A Digital 11/14 GHz Multibeam Switched Satellite System," AIAA/CASI 6th Communication Satellite Systems Conference, Montreal, April 5–8, 1976.
2. L. C. Tillotson, "A Model of a Domestic Satellite Communication System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.
3. D. A. Gray, Fig. 35 in ref. 5.
4. A. J. Rustako, Jr., "An Earth-Space Propagation Measurement at Crawford Hill Using the 12-GHz CTS Satellite Beacon," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1431–1448.
5. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE, *63* (September 1975).
6. T. S. Chu, "Rain-Induced Cross-Polarization at Centimeter and Millimeter Wavelengths," B.S.T.J., *53*, No. 8 (Octoher 1974), pp. 1557–1579.
7. D. C. Cox, H. W. Arnold, and H. H. Hoffman, "Depolarization of 19- and 28-GHz Earth-Space Signals by Ice Particles," Radio Science, *13* (May–June 1978).
8. D. C. Cox, "Design of the Bell Laboratories 19- and 28-GHz Satellite Beacon Propagation Experiment," IEEE International Conference on Communications (ICC '74) Record, June 17–19, 1974, Minneapolis, Minnesota, pp. 27E1–27E5.
9. D. C. Cox, "Some Effects of Measurement Errors on Rain Depolarization Experiments," B.S.T.J., *54*, No. 2 (February 1975), pp. 435–450.
10. D. C. Cox and H. W. Arnold, "COMSTAR Beacon Measurements at Crawford Hill: Attenuation Statistics and Depolarization," USNC/URSI Spring Meeting, May 15–19, 1978, University of Maryland, College Park, Maryland.
11. D. C. Cox, "An Overview of the Bell Laboratories 19- and 28-GHz COMSTAR Beacon Propagation Experiments," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1231–1255.

12. T. S. Chu, R. W. Wilson, R. W. England, D. A. Gray, and W. E. Legg, "The Crawford Hill 7-Meter Millimeter-Wave Antenna," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1257–1288.
13. H. W. Arnold, D. C. Cox, H. H. Hoffman, R. H. Brandt, R. P. Leck, and M. F. Wazowicz, "The 19- and 28-GHz Receiving Electronics for the Crawford Hill COMSTAR Beacon Propagation Experiment," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1289–1329.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, November 1978
Printed in U.S.A.

Signal Design for PAM Data Transmission to Minimize Excess Bandwidth

By A. D. WYNER

(Manuscript received December 9, 1977)

In a conventional PAM data transmission system, the transmitted signal is $x(t) = \Sigma\, \alpha_n g(t - nT)$, *where* $\{\alpha_n\}$ *is a* 2^L*-level data sequence, and* $g(t)$ *is a Nyquist pulse* ($g(0) \neq 0$, $g(mT) = 0$, $m \neq 0$). *Ideally, the bandwidth of the pulse* $g(t)$ *and, therefore, the bandwidth of* $x(t)$ *can be made equal to* $1/2T = \rho/2L$, *where* ρ *is the data rate. In practice, however, an "excess bandwidth" of at least 10 to 20 percent is required.*

Using a class of real sequences called "discrete prolate spheroidal sequences," we show how to construct a modulated signal with bandwidth just slightly in excess of the optimal $\rho/2L$ *(say, by 2 to 4 percent). The new signal is similar in many ways to a conventional PAM signal, and in particular an ad-hoc receiver structure is suggested for which the resulting error performance is about the same as for a conventional PAM system operating in the same environment.*

I. INTRODUCTION

To fix ideas, consider the following conventional (baseband) PAM data-transmission scheme (see, for example, Ref. 1). The data to be transmitted is a sequence $\{\alpha_k\}_{-\infty}^{\infty}$. The α_k are independent identically distributed copies of the random variable α, which is uniformly distributed on the set $\{\pm 1, \pm 3, \cdots, \pm 2^L - 1\}$. Thus, α takes 2^L equally likely values, where $L = 1, 2, \cdots$, is a fixed parameter. The modulated signal is

$$x_0(t) = \sum_{k=-\infty}^{\infty} \alpha_k g_0(t - kT_0), \tag{1}$$

where $g_0(\cdot)$ is a real-valued "Nyquist pulse"—i.e.,

$$g_0(0) \neq 0,$$
$$g_0(kT_0) = 0, \qquad k \neq 0. \tag{2}$$

Since (2) implies that $x_0(kT_0) = \alpha_k g(0)$, $k = 0, \pm 1, \pm 2, \cdots$, the data

sequence $\{\alpha_k\}$ can be obtained from $x_0(\cdot)$ simply by sampling. We assume that the Fourier transform

$$G_0(f) = \int_{-\infty}^{\infty} g_0(t)e^{-i2\pi ft}dt, \qquad -\infty < f < \infty, \tag{3}$$

of $g_0(\cdot)$ has support on the interval $[-F_0, \pm F_0]$, where $F_0 \leq 1/T_0$. Under this assumption, the Nyquist condition (2) is known[1] to be equivalent to

$$G_0(f) + G_0\left(f - \frac{1}{T_0}\right) = Tg_0(0), \qquad 0 \leq f \leq \frac{1}{T_0} \tag{4}$$

(except perhaps in a set of measure zero). Figure 1 is an example of a real $G_0(f)$ which satisfies (4). An often-used Nyquist pulse $G_0(f)$ is the so-called raised-cosine pulse. (See Ref. 1, pp. 50–51.) The bandwidth of $x_0(\cdot)$, which is the same as the bandwidth of $g_0(\cdot)$, is taken as F_0. The difference $F_0 - 1/2T_0$ is called the "excess bandwidth."

To conserve bandwidth, it is desirable to make F_0 as close to $1/2T_0$ as possible, but typically $(F_0 - 1/2T_0)/F_0 \geq 10$ to 20 percent in real systems. Further reduction in the excess bandwidth is difficult, since the very sharp cutoff filter used to generate $g_0(t)$ with F_0 close to $1/2T_0$ will introduce either phase distortion or ripples in the amplitude characteristic.

In a practical data transmission system for the voice-grade telephone channel, a reduction in bandwidth is also desirable, since the channel characteristics at the band edges are poor.

In this paper, we suggest another approach to the signal design problem which will allow a further reduction in the excess bandwidth, perhaps to as little as 2 to 4 percent. The technique involves a family of sequences called "discrete prolate spheroidal sequences" (DPSS) and is also intimately tied up with notions concerning the space of square summable sequences (l_2). Therefore, before presenting our scheme, we must digress to review some notions about the space l_2 and to introduce the DPSSs. We do this in Sections II and III, respectively. In Section IV we discuss our new scheme.

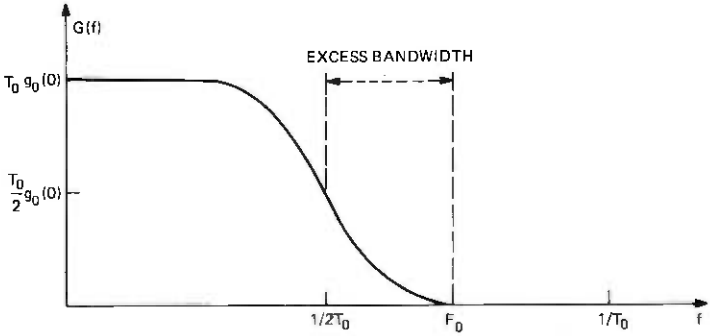


Fig. 1—Example of real $G(f)$ satisfying eq. (4).

At this point, I would like to acknowledge with thanks three of my colleagues without whose help this project could never have gotten off the ground. D. Slepian introduced me to DPSSs, and with much kindness and no small amount of work helped me to get numerical values for these sequences and their associated eigenvalues. J. Mazo taught me most of what I know about data communication, and J. Salz's interest and enthusiasm stimulated me to obtain a full understanding of the properties of the modulation scheme.

II. REVIEW OF THE SPACE l_2

The space l_2 of square-summable, real-valued sequences is the set of sequences $\{a(n)\}_{n=-\infty}^{\infty}$ (or $a(\cdot)$) such that

$$\sum_{n=-\infty}^{\infty} a^2(n) < \infty. \tag{5}$$

Let $a(\cdot), b(\cdot) \in l_2$; then the *inner product* of $a(\cdot)$ and $b(\cdot)$ is

$$\langle a, b \rangle = \sum_{n=-\infty}^{\infty} a(n)b(n). \tag{6}$$

Also, the norm of $a(\cdot)$ is

$$\|a\| = \langle a, a \rangle^{1/2}. \tag{7}$$

We will need the following facts. For $a(\cdot), b(\cdot), c(\cdot) \in l_2$, and any real number γ,

$$\langle \gamma a, b + c \rangle = \gamma \langle a, b \rangle + \gamma \langle a, c \rangle, \tag{8}$$

which implies that, for $a_j \in l_2, j = 1, 2, \cdots$,

$$\| \sum_j a_j \|^2 = \sum_j \|a_j\|^2 + 2 \sum_{j<k} \langle a_j, a_k \rangle. \tag{9}$$

Further, the Schwarz inequality is, for $a, b \in l_2$,

$$|\langle a, b \rangle| \leq \|a\| \, \|b\|. \tag{10}$$

For $a(\cdot) \in l_2$, the (sequence) *Fourier transform* is defined by

$$A_T(f) = \sum_{n=-\infty}^{\infty} a(n)e^{-i2\pi fTn}, \qquad -\infty < f < \infty, \tag{11}$$

where $T > 0$ is a fixed parameter. Of course, $A_T(f)$ is periodic with period $1/T$, and usually we will be concerned only with its values on the interval $[-(1/2T), (1/2T)]$. The sequence $\{a(n)\}$ can be recovered from $A_T(\cdot)$ by the formula

$$a(n) = T \int_{-1/2T}^{1/2T} A_T(f)e^{i2\pi fTn} df, \qquad -\infty < n < \infty. \tag{12}$$

The *convolution theorem* states that if

$$c(n) = \sum_{m=-\infty}^{\infty} a(m)b(n - m)$$

(which we denote $c = a*b$), then

$$C_T(f) = A_T(f)B_T(f), \qquad -\infty < f < \infty, \tag{13}$$

where A_T, B_T, and C_T are the transforms of a, b, and c, respectively.

The Parseval relation is, for $a, b \in l_2$,

$$\langle a, b \rangle = T \int_{-1/2T}^{+1/2T} A_T(f)B_T^*(f)df, \tag{14}$$

where "$*$" denotes complex conjugate. Thus, in particular,

$$\|a\|^2 = \langle a, a \rangle = T \int_{-1/2T}^{+1/2T} |A_T(f)|^2 df. \tag{15}$$

We say that a sequence $a(\cdot) \in l_2$ is *bandlimited* to $[0, F]$, $0 \le F \le 1/2T$, if its transform $A_T(f) = 0$, for $F \le |f| \le 1/2T$. Thus, a bandlimited $a(\cdot)$ can be written

$$a(n) = \int_{-F}^{F} A(f)e^{i2\pi fTn}df. \tag{16}$$

A sequence $a(\cdot)$ has *support* on the interval $[N_1, N_2]$, $-\infty \le N_1 \le N_2 \le \infty$, if $a(n) = 0$, for $n \notin [N_1, N_2]$. A sequence with support on $[N_1, N_2]$, where $|N_1|, |N_2| < \infty$, cannot be bandlimited to $[0, F]$ with $F < 1/2T$.

It is convenient to define the *bandlimiting* operator on l_2, $\mathcal{B} = \mathcal{B}_F$, $0 \le F \le 1/2T$, by (for $a \in l_2$)

$$\mathcal{B}a = b, \tag{17a}$$

where

$$b(n) = \int_{-F}^{F} A_T(f)e^{i2\pi fTn}df. \tag{17b}$$

In other words, the transform of $b(\cdot)$ is

$$B_T(f) = \begin{cases} A_T(f), & |f| \le F, \\ 0, & F \le |f| \le \dfrac{1}{2T}. \end{cases} \tag{17c}$$

A sequence $a \in l_2$ is bandlimited to $[0, F]$ iff $\mathcal{B}_F a = a$. Corresponding to the operator $\mathcal{B}_F$, we also define the complementary operator $\mathcal{B}' = \mathcal{B}_F' = I - \mathcal{B}_F$, where I is the identity operator.

We also define the *index-limiting* (or *time-limiting*) operator $\mathcal{D} = \mathcal{D}_N$ $(1 \le N < \infty)$, by (for $a \in l_2$)

$$\mathcal{D}a = b, \tag{18a}$$

where

$$b(n) = \begin{cases} a(n), & 1 \le n \le N, \\ 0, & \text{otherwise.} \end{cases} \tag{18b}$$

Thus $a \in l_2$ has support on $[1, N]$ iff $\mathcal{D}_N a = a$. We will need the following easily established propositions.

Proposition 1: Let $x(t)$, $-\infty < t < \infty$, be a real-valued function with ordinary Fourier transform (as defined by (3)) $X(f)$, $-\infty < f < \infty$. Let the sequence $a(\cdot)$ be defined by $a(n) = x(nT)$. Then, the sequence Fourier transform of $a(\cdot)$ is

$$A_T(f) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X\left(f - \frac{k}{T}\right), \qquad -\infty < f < \infty.$$

In particular, if $x(t)$ is bandlimited to F Hz, and $1/T > 2F$, then

$$A_T(f) = \frac{1}{T} X(f), \qquad |f| \leq \frac{1}{2T}.$$

Thus the sequence $a(\cdot)$ is bandlimited to $[0,F]$.

Proposition 2: Let $a(\cdot) \in l_2$, and let $g(t)$ be a real-valued function of the continuous variable t. Let

$$x(t) = \sum_{n=-\infty}^{\infty} a(n) g(t - nT), \qquad -\infty < t < \infty.$$

Then the ordinary Fourier transform of $x(t)$ is

$$X(f) = A_T(f) G(f), \qquad -\infty < f < \infty.$$

where $A_T(f)$ is the sequence Fourier transform of $a(\cdot)$ and $G(f)$ is the ordinary Fourier transform of $g(t)$.

III. DISCRETE PROLATE SPHEROIDAL SEQUENCES

Let $T, F > 0$ (where $W = FT \leq 1/2$) and $N > 1$ be given; let the operators $\mathcal{B} = \mathcal{B}_F$, $\mathcal{D} = \mathcal{D}_N$ be as defined in Section II. The following theorem is proved in Appendix A.

Theorem 3: There exists a set of real sequences $\{\phi_j(\cdot)\}_{j=1}^N$, called "discrete prolate spheroidal sequences" (DPSS), with support on $[1, N]$ and a corresponding set of real numbers $\{\lambda_j\}_1^N$, called "eigenvalues," with the following properties.

(A) $1 \geq \lambda_1 \geq \lambda_2 \cdots \geq \lambda_N > 0$, *and* $\sum_{j=1}^{N} \lambda_j = 2FTN$.

(B) $\mathcal{D}\mathcal{B}\phi_j = \lambda_j \phi_j$, $1 \leq j \leq N$.

(C) $\langle \phi_j, \phi_k \rangle = \delta_{jk}$

(D) $\langle \mathcal{B}\phi_j, \mathcal{B}\phi_k \rangle = \lambda_j \delta_{jk}$

(E) *With $\delta > 0$, and F, T held fixed, as $N \to \infty$,*

$$\frac{1}{N}\left\{\begin{array}{c}\text{number of } j \text{ such that} \\ \delta < \lambda_j < 1 - \delta\end{array}\right\} \to 0.$$

(F) *With $\epsilon > 0$, and F, T held fixed, as $N \to \infty$,*

$$\lambda_{2FTN(1-\epsilon)} \to 1,$$

$$\lambda_{2FTN(1+\epsilon)} \to 0.$$

(G) *(Slepian [Ref. 4, eq. (63)]), with $\epsilon > 0$ and F, T held fixed, as $N \to \infty$,*

$$1 - \lambda_{2FTN(1-\epsilon)} = \exp\{-C(\epsilon)N + o(N)\},$$

where $C(\epsilon) > 0$.

(H) *The $\phi_j(\cdot)$ and λ_j, $1 \le j \le N$, depend on F, T only through their product $W = FT$.*

Remarks:

(*i*) In the course of giving the proof of Theorem 3, we will show explicitly how to find the DPSSs $\{\phi_j\}$ and the corresponding $\{\lambda_j\}$.

(*ii*) Theorem 3A, F implies that, with N large, about $2FTN$ of the λ_js are about 1, and that the remainder (about $(1 - 2FT)N$) of the λ_js are about 0. Theorem 3G indicates that the convergence as $N \to \infty$ is quite rapid. Since this fact is crucial to our modulation scheme, we list some of the λ_js for $FT = 1/4$, and various values of N in Table I. Here $2FTN = N/2$, so that about half of the λ_js are 1 and the remainder are about 0.

Table I — $\{\lambda_j\}$ for $W = 0.25$, for $N = 5, 10, 20, 50, 100$

	j	λ_j		j	λ_j
$N = 5$	1	0.9976686		15	0.0000212
	2	0.9244132		16	0.0000008
	3	0.5000000		17	0.0000000
	4	0.0755868		18	0.0000000
	5	0.0023143		19	0.0000000
				20	0.0000000
$N = 10$	1	0.9999994			
	2	0.9999490	$N = 50$	1–21	>0.9997
	3	0.9980787		22	0.998
	4	0.9650286		23	0.985
	5	0.7326630		24	0.914
	6	0.2673371		25	0.680
	7	0.0349714		26	0.320
	8	0.0019213		27	0.086
	9	0.0000510		28	0.015
	10	0.0000005		29	0.002
				30–50	<0.00023
$N = 20$	1	0.9999999			
	2	0.9999999	$N = 100$	1–45	>0.9998
	3	0.9999999		46	0.9993
	4	0.9999999		47	0.996
	5	0.9999992		48	0.976
	6	0.9999788		49	0.892
	7	0.9995798		50	0.664
	8	0.9940340		51	0.336
	9	0.9435514		52	0.108
	10	0.7070557		53	0.024
	11	0.2929445		54	0.004
	12	0.0564486		55	0.0007
	13	0.0059659		56–100	<0.0001
	14	0.0004201			

(*iii*) Theorem 3C implies that $\|\phi_j\|^2 = 1$, and Theorem 3D implies that $\|\mathcal{B}\phi_j\|^2 = \lambda_j$. Thus, the fraction of the energy of ϕ_j within the band $[0, F]$ is λ_j. Therefore, when λ_j is close to unity, ϕ_j is a sequence with support on $[1, N]$ with most of its energy in the band $[0, F]$. Theorem 3 implies that, with N large, there are about $2FTN$ orthogonal sequences, i.e., ϕ_j $(j = 1, 2, \cdots, 2FTN(1 - \epsilon))$, with support on $[1, N]$ which are approximately bandlimited to $[0, F]$.

(*iv*) Slepian has made an exceptionally detailed study of DPSSs and their properties. Reference 4 contains most of his results, and Ref. 5 describes a Fortran program for computing the DPSSs and their eigenvalues.

IV. HEURISTIC DESCRIPTION OF THE MODULATION SCHEME

Let the data to be transmitted be as in Section I, the 2^L-level sequence $\{\alpha_j\}_{-\infty}^{\infty}$. We break this sequence into blocks of length ν, where the kth block is $\alpha_{k\nu+1}, \cdots, \alpha_{(k+1)\nu}$, $-\infty < k < \infty$, and where ν is an integer to be chosen later. Consider the 0th block $\alpha_1, \cdots, \alpha_\nu$. Let $N > \nu$ be another integer parameter, and let F, $T > 0$, with $FT < \frac{1}{2}$, be given. Let ϕ_j, λ_j, $1 \le j \le N$, be the DPSSs and eigenvalues guaranteed by Theorem 3, with parameters N, F, T. Then define the sequence

$$a(n) = \sum_{j=1}^{\nu} \alpha_j \phi_j(n), \qquad -\infty < n < \infty. \tag{19}$$

Observe that $a(\cdot)$, like the $\phi_j(\cdot)$, has support on the interval $[1, N]$. Further, if we take $\nu = 2FTN(1 - \epsilon)$, with N sufficiently large so that $\lambda_\nu \approx 1$, then from Theorem 3 (see remark iii), $a(\cdot)$ is approximately bandlimited to $[0, F]$.

Now the modulated waveform corresponding to the 0th data block is

$$x_0(t) = \sum_{n=1}^{N} a(n) g(t - nT), \qquad -\infty < t < \infty, \tag{20}$$

where the pulse $g(t)$ has Fourier transform $G(f)$ which satisfies

$$G(f) = \begin{cases} T, & |f| \le F, \\ 0, & |f| > \dfrac{1}{2T}, \end{cases} \tag{21a}$$

$$|G(f)| \le T, \qquad F \le |f| \le \frac{1}{2T}. \tag{21b}$$

Thus, we do not specify $G(f)$ in the interval $[F, 1/2T]$, except by (21b). Since $G(f)$ need not have sharp transitions, it is not difficult to implement in practice. For the kth data block $(-\infty < k < \infty)$, $\alpha_{k\nu+1}, \cdots, \alpha_{(k+1)\nu}$, we set

$$a(n) = \sum_{j=1}^{\nu} \alpha_{k\nu+j} \phi_j(n - Nk), \qquad Nk + 1 \le n \le N(k + 1), \tag{22}$$

and let the modulated waveform be

$$x_k(t) = \sum_{n=Nk+1}^{N(k+1)} a(n)g(t - nT). \tag{23}$$

The entire modulated signal is

$$x(t) = \sum_{k=-\infty}^{\infty} x_k(t) = \sum_{n=-\infty}^{\infty} a(n)g(t - nT). \tag{24}$$

Since the number of bits in each data block is $L\nu$ and each data block "occupies" NT seconds, the transmission rate is

$$\rho = \left(\frac{\nu}{N}\right)\frac{L}{T} \text{ bits/s.} \tag{25}$$

We now give an intuitive, though imprecise, explanation of the properties of the modulation scheme. Consider $x_0(t)$ given by (20). Its Fourier transform is, from Proposition 2,

$$X_0(f) = A_T(f)G(f), \tag{26a}$$

where

$$A_T(f) = \sum_{n=1}^{N} a(n)e^{-i2\pi fTn}$$

$$= \sum_{j=1}^{\nu} \alpha_j \Phi_{jT}(f), \tag{26b}$$

where Φ_{jT} is the sequence Fourier transform of $\phi_j(\cdot)$. In the light of remark iii following Theorem 3, the $\{\Phi_{jT}(f)\}_{j=1}^{\nu}$ and therefore $A_T(f)$ are approximately zero for $|f| \in [F, 1/2T]$ provided $\nu \leq 2FN(1 - \epsilon)$. Since $G(f)$ is bounded in this interval and 0 for $|f| > 1/2T$, we see that $X_0(f)$ is approximately bandlimited to $|f| \leq F$. Further, if we take $\nu = 2FTN(1 - \epsilon)$, we have from (25) that the transmission rate ρ is $2FL(1 - \epsilon)$. Thus in our scheme we can transmit $2F(1 - \epsilon)$, 2^L-level data symbols per second with bandwidth F. If $\epsilon = 0$, then we would have effectively constructed a PAM system with no excess bandwidth. Since, in practice, ϵ can be made very small, we can in fact come quite close to the ideal.

So far so good. But we still must show that the data symbols $\{\alpha_j\}_1^{\nu}$ can be recovered conveniently from $x_0(t)$. In fact, we claim that the samples $x_0(nT) \approx a(n)$, $1 \leq n \leq N$. The key observation here is that, since $A_T(f) \approx 0$, $|f| \in [F, 1/2T]$, then $X_0(f)$ is not appreciably changed when $G(f)$ is replaced by $G_I(f)$ ("I" for "ideal") where

$$G_I(f) = \begin{cases} T, & |f| \leq 1/2T, \\ 0, & |f| > 1/2T. \end{cases} \tag{27}$$

The inverse transform of G_I is $g_I(t) = (\sin \pi t/T)/(\pi t/T)$. Let us therefore

define $x_I(t)$ by replacing $g(t)$ by $g_I(t)$ in the definition of $x_0(t)$. We obtain

$$x_I(t) = \sum_{n=1}^{N} a(n) g_I(t - nT),$$

so that

$$x_I(nT) = a(n), \qquad 1 \le n \le N.$$

It follows that

$$x_0(nT) - a(n) = x_0(nT) - x_I(nT), \qquad 1 \le n \le N.$$

Now define the sequence $c(\cdot)$ by

$$c(n) = x_0(nT) - x_I(nT), \quad 1 \le n \le N.$$

Since $x_0(t) - x_I(t)$ is bandlimited to $1/2T$ Hz, we have from Proposition 1 that

$$C_T(f) = \frac{1}{T}[X_0(f) - X_I(f)], \qquad |f| \le \frac{1}{2T},$$

and from Proposition 2 that

$$C_T(f) = \frac{1}{T}[X_0(f) - X_I(f)]$$

$$= \frac{1}{T} A_T(f)[G(f) - G_I(f)].$$

From the Parseval relation (15),

$$\sum_{n=1}^{N} [x_0(nT) - a(n)]^2 \le \|c\|^2$$

$$= T \int_{-1/2T}^{1/2T} |C_T(f)|^2 df$$

$$= \frac{1}{T} \int_{F < |f| \le 1/2T} |A_T(f)|^2 |G(f) - G_I(f)|^2 df$$

$$\le 4T \int_{F < f \le 1/2T} |A_T(f)|^2 df = 4 \| \mathcal{B}_F' a \|^2 \approx 0.$$

The inequality follows from (21b), which implies that

$$|G(f) - G_I(f)| \le 2T.$$

Thus, $a(n)$, $1 \le n \le N$ can be recovered from $x_0(t)$ simply by sampling. From (19) and the orthonormality of the $\{\phi_j\}_1^r$,

$$\alpha_j = \sum_{n=1}^{N} a(n) \phi_j(n), \tag{28}$$

so that the $\{\alpha_j\}_1^\nu$ can be recovered from samples $x(nT)$, $1 \le n \le N$.

Aside from imprecision, the above arguments completely ignored the effects of the other data blocks ($k \neq 0$) and the effects of channel distortion and noise. In the next section, we give a precise definition of the modulation scheme and of a proposed receiver, and then state theorems that give bounds on the error introduced by linear channel distortion and a channel noise. We will also bound the instantaneous power $x^2(t)$.

V. PRECISE STATEMENT OF RESULTS AND DISCUSSION

Let $\{\alpha_j\}_{-\infty}^{\infty}$ be, as in Section IV, a sequence of independent, identically distributed copies of the 2^L-valued random variable α, where

$$\Pr\{\alpha = m\} = 2^{-L}, \qquad m = \pm 1, \pm 3, \cdots, \pm (2^L - 1). \tag{29}$$

The sequence $\{\alpha_j\}$ is the data sequence to be transmitted. Let $\nu, N, F, T > 0$ be parameters such that ν, N are integers, and

$$\nu < N, \tag{30a}$$

$$W \triangleq FT < 1/2. \tag{30b}$$

Partition the data sequence into blocks of length ν, such that the kth block is

$$\alpha_{\nu k+1}, \cdots, \alpha_{\nu(k+1)},$$

$k = 0, \pm 1, \pm 2, \cdots$. Let $\{\phi_j(\cdot), \lambda_j\}_{j=1}^N$ be the quantities (DPSSs and eigenvalues) whose existence is guaranteed by Theorem 3 with parameters N, W. Corresponding to the kth data block, define the sequence $a_k(\cdot)$ by

$$a_k(n) = \sum_{j=1}^{\nu} \alpha_{\nu k+j} \phi_j(n - Nk), \tag{31a}$$

and let

$$a(\cdot) = \sum_{k=-\infty}^{\infty} a_k(\cdot). \tag{31b}$$

Since $\phi_j(\cdot)$ has support on $[1, N]$, $a_k(\cdot)$ has support on $[Nk + 1, N(k + 1)]$. Finally, the modulated signal is

$$x(t) = \sum_{k=-\infty}^{\infty} x_k(t), \qquad -\infty < t < \infty, \tag{32a}$$

where

$$\begin{aligned} x_k(t) &= \sum_{n=-\infty}^{\infty} a_k(n) g(t - nT), \\ &= \sum_{n=Nk+1}^{N(k+1)} a_k(n) g(t - nT), \end{aligned} \tag{32b}$$

and the pulse $g(t)$ is the inverse Fourier transform of $G(f)$, which we leave unspecified for now.

A block diagram for the modulator described above is given in Fig. 2. The data symbols appear at a rate of ν/NT per second. Box A takes the data symbols ν at a time and calculates the numbers $\{a(n)\}$—producing N outputs for every ν inputs. Thus, the $a(n)$ appear at a rate of $1/T$ per second. Box B produces $x(t)$ by modulating the amplitude of a pulse train with the $\{a(n)\}$. Although we will not specify the ratio ν/N and the pulse $g(t)$ now, it will be useful to informally think of

$$\nu/N = 2FTN(1 - \epsilon) = 2WN(1 - \epsilon),$$

and $g(t) \leftrightarrow G(f)$ as in (21). We will allow the possibility of non-physically realizable pulses $g(t)$, with the usual understanding that a close approximation to $g(t)$ can be obtained with a finite delay (which we shall ignore).

The received signal is taken as

$$y(t) = w(t) + z(t), \tag{33a}$$

where

$$w(t) = \int_{-\infty}^{\infty} x(\tau)\, h_c(t - \tau) d\tau, \tag{33b}$$

and where $h_c(t)$ is the impulse response of the channel ($H_c(f)$, the channel transfer function, is the transform of $h_c(t)$), and $z(t)$ is noise with zero mean and power spectral density $N_Z(f)$.

We now turn to the receiver. We will postulate a simple receiver structure which, though not optimum, has the virtues of simplicity and amenability to analysis. Furthermore it is probably not very far from being optimal itself. Refer to Fig. 3. The received waveform $y(t)$ is first sampled at $t = nT$, to produce the sequence $\{y(nT)\}_{n=-\infty}^{\infty}$. These samples are the input to box C, a tapped delay line with $2M + 1$ taps. The output of box C is the sequence $\{\hat{a}(n)\}$ given by

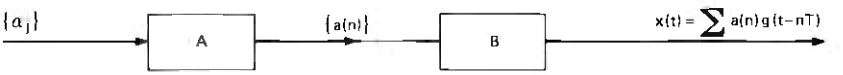


Fig. 2—The modulator.

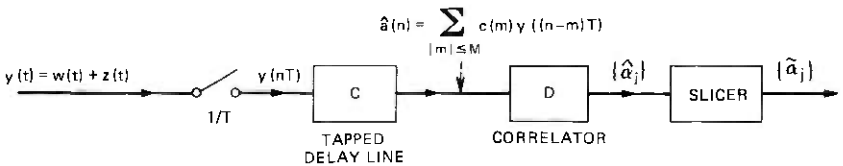


Fig. 3—The receiver.

$$\hat{a}(n) = \sum_{m=-M}^{+M} c(m)y((n-m)T), \qquad -\infty < n < \infty, \tag{34}$$

where $\{c(m)\}_{-M}^{+M}$ are the tap weight coefficients. As we shall see, the tap weights should be chosen so that the sequence $\hat{a}(\cdot)$ is the receiver's best guess of the sequence $a(\cdot)$.

Consider the 0th block of $\{\hat{a}(n)\}$, i.e., $\hat{a}(1), \cdots, \hat{a}(N)$. If $\hat{a}(n) \equiv a(n)$, then the 0th block of data symbols $\alpha_1, \cdots, \alpha_\nu$ could be recovered from $\{\hat{a}(n)\}$ using (28). Our approach will be to use these same formulas to obtain an estimate $\{\hat{\alpha}_j\}$ of the $\{\alpha_j\}$, i.e.,

$$\hat{\alpha}_j = \sum_{n=1}^{N} \hat{a}(n)\phi_j(n), \qquad 1 \le j \le \nu. \tag{35a}$$

For the remaining blocks ($k \neq 0$), we proceed analogously, viz.,

$$\hat{\alpha}_{k\nu+j} = \sum_{n=Nk+1}^{N(k+1)} \hat{a}(n)\phi_j(n - Nk), \qquad 1 \le j \le \nu. \tag{35b}$$

This is box D. The final step in the demodulation process is a "slicer," which examines $\hat{\alpha}_j (-\infty < j < \infty)$ and emits $\tilde{\alpha}_j$ where $\tilde{\alpha}_j$ equals a value of $m \in \{\pm 1, \pm 3, \cdots, \pm 2^L - 1\}$ which minimizes $|\hat{\alpha}_j - m|$.

As in (25), the transmission rate is $\rho = (\nu/N) \cdot (L/T)$ bits/s.

We are now ready to state the properties of the modulation scheme in the form of theorems. The proofs of these theorems are given in Section VI. Theorem 4 gives an upper bound on the average power of the transmitted signal $x(t)$. Theorem 5 gives an upper bound on the expected instantaneous power $E\,x^2(t)$, as a function of t. Finally, Theorem 6 gives an upper bound on the mean-squared error. We state these results with no restrictions on $G(f)$, $H_c(f)$, and $N_Z(f)$. In the remarks which follow the statement of the theorems, we will look at some interesting special cases.

We begin with a bit of notation. Denote the variance of the data random variable α, defined in (29), by

$$\sigma_\alpha^2 = E\,\alpha^2 = 2^{-L} \sum_{\substack{|m| \le 2^L - 1 \\ m \text{ odd}}} m^2 = \frac{(2^L - 1)(2^L + 1)}{3}. \tag{36}$$

Also, for $\nu, N, W = FT$ satisfying (30), let

$$Q = Q(\nu, N, W) = \frac{1}{\nu} \sum_{j=1}^{\nu} (1 - \lambda_j). \tag{37}$$

Of course, if we set $\nu = 2WN(1 - \epsilon)$, and let $N \to \infty$, with $W, \epsilon > 0$ held fixed, then $Q \to 0$. It will be helpful to think of Q as a small quantity. Here are the theorems. Although they may seem formidable at first glance, please stick with it! In the extensive discussion following the theorem statements, you will see that they can be easily applied and yield useful information.

Theorem 4: (*average power*)

$$P_{AV} \triangleq \frac{1}{NT} E \int_0^{NT} x^2(t)dt$$
$$\leq \frac{\sigma_\alpha^2}{T}\left[\int_{-F}^{F} \sum_{k=-\infty}^{\infty} \left|G\left(f-\frac{k}{T}\right)\right|^2 df + \frac{\nu}{N} A_1 Q\right], \tag{38a}$$

where

$$A_1 = \sup_{F \leq |f| \leq 1/2T} \frac{1}{T^2} \sum_{k=-\infty}^{\infty} \left|G\left(f-\frac{k}{T}\right)\right|^2. \tag{38b}$$

Theorem 5: For $-\infty < t_1 < \infty$,

$$E\, x^2(t_1) \leq \frac{\sigma_\alpha^2}{T} \int_{-1/2T}^{1/2T} \left| \sum_{k=-\infty}^{\infty} G\left(f-\frac{k}{T}\right) e^{i2\pi f t_1}\right|^2 df. \tag{39}$$

Theorem 6: (mean-squared error)

$$\epsilon^2 \triangleq \frac{1}{\nu} E \sum_{j=1}^{\nu} (\hat{\alpha}_j - \alpha_j)^2 = \epsilon_N^2 + \epsilon_I^2, \tag{40}$$

(N stands for noise, and I for intersymbol interference). The noise error ϵ_N^2 *is bounded by*

$$\epsilon_N^2 \leq \frac{N}{\nu} \int_{-F}^{F} |C_T(f)|^2 \left(\sum_{k=-\infty}^{\infty} N_Z\left(f-\frac{k}{T}\right)\right) df + A_2 Q, \tag{41a}$$

where

$$C_T(f) \triangleq \sum_{n=-M}^{M} c(n) e^{-i2\pi f T n}, \tag{41b}$$

$N_Z(f)$ *is the power spectral density of the noise, and*

$$A_2 \triangleq \sup_{F \leq |f| \leq 1/2T} \frac{|C_T(f)|^2}{T} \sum_{k=-\infty}^{\infty} N\left(f-\frac{k}{T}\right). \tag{41c}$$

The intersymbol interference error ϵ_I^2 *is bounded by*

$$\epsilon_I^2 \leq \sigma_\alpha^2 \left(\frac{N}{\nu}\right)\left[T \int_{-F}^{F} |C_T(f) B_T(f) - 1|^2 df + A_3 Q\right], \tag{42a}$$

where

$$B_T(f) \triangleq \frac{1}{T} \sum_{k=-\infty}^{\infty} G\left(f-\frac{k}{T}\right) H_c\left(f-\frac{k}{T}\right), \tag{42b}$$

$C_T(f)$ *is given by (41b), and*

$$A_3 \triangleq \sup_{F \leq |f| \leq 1/2T} |C_T(f) B_T(f) - 1|^2. \tag{42c}$$

Remarks:

(*i*) The reason that P_{AV} as defined by (38a) is the "average power" is that the random function $x(t)$ is cyclostationary with period NT. In other words, the shifted sequence $\tilde{x}(t) \triangleq x(t - kNT)$ has the same statistical properties as $x(t)$ itself (for $k = 0, \pm 1, \pm 2, \cdots$). Thus, it follows that

$$\lim_{\tau \to \infty} E \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} x^2(t)dt = \frac{1}{NT} E \int_0^{NT} x^2(t)dt = P_{AV}. \tag{43}$$

(*ii*) When Q is small, the upper bound of (38a) on P_{AV} depends essentially on the folded power spectrum

$$\sum_{k=-\infty}^{\infty} G\left(f - \frac{k}{T}\right), \qquad |f| \leq F.$$

Furthermore, if $G(f) = 0$ for $|f| \leq 1/2T$, then

$$\sum_{k} G\left(f - \frac{k}{T}\right) = G(f), \qquad |f| \leq F,$$

so that Theorem 4 becomes

$$P_{AV} \leq \frac{\sigma_\alpha}{T}\left[\int_{-F}^{F} |G(f)|^2 df + \frac{\nu}{N} A_1 Q\right], \tag{44a}$$

where

$$A_1 = \sup_{F \leq |f| \leq 1/2T} |G(f)|^2. \tag{44b}$$

(*iii*) If we assume, as in Remark *ii*, that $G(f) = 0$, $|f| > 1/2T$, then Theorem 5 becomes

$$E\,x^2(t) \leq \frac{\sigma_\alpha^2}{T}\int_{-1/2T}^{1/2T} |G(f)|^2 df. \tag{45}$$

Thus the upper bound on $E\,x^2(t)$ depends on $G(f)$ for $|f| \leq 1/2T$.

(*iv*) Saltzberg's[3] bound can be applied (see Appendix B) to our problem to show that the distribution for the instantaneous power satisfies

$$\Pr\{|x(t)|^2 > r^2\} \leq 2\exp\left\{-\frac{r^2}{2E\,x^2(t)}\right\}.$$

The bound of Theorem 5 can be applied here to further overbound this probability.

(*v*) We now explain the rationale for using the mean-squared error $\epsilon^2 = (1/\nu)\sum_{j=1}^{\nu} E(\alpha_j - \hat{\alpha}_j)^2$. First note that with $\epsilon_j^2 \triangleq E(\hat{\alpha}_j - \alpha_j)^2$, Saltzberg's bound (see Appendix B) can again be used to show that, if the noise is Gaussian, then

$$P_{ej} = \Pr\{\tilde{\alpha}_j \neq \alpha_j\} \leq 2\exp\left\{-\frac{1}{2\epsilon_j^2}\right\}.$$

Since the sequence of random pairs $\{\alpha_j, \tilde{\alpha}_j\}_{j=-\infty}^{\infty}$ is cyclostationary with period ν, the overall error probability is

$$P_e = \frac{1}{\nu}\sum_{j=1}^{\nu} P_{ej} \leq \frac{2}{\nu}\sum_{j=1}^{\nu} \exp\left\{-\frac{1}{2\epsilon_j^2}\right\}. \tag{46}$$

Now let $\epsilon_{\max}^2 = \max_{1 \leq j \leq \nu} \epsilon_j^2$. Ineq. (46) yields

$$P_e \leq 2 \exp\left\{-\frac{1}{2\epsilon_{\max}^2}\right\}.$$

Thus it would appear that $\epsilon_{\max}^2$ rather than ϵ^2 is the appropriate criterion. Nevertheless the use of ϵ^2 (which was, of course, chosen for its mathematical tractability) can be justified by the following argument.

Let R be the $\nu \times \nu$ covariance matrix with (j, k)th entry $E(\hat{\alpha}_j - \alpha_j)\cdot(\hat{\alpha}_k - \alpha_k)$, $1 \leq j, k \leq \nu$. Let M_0 be a $\nu \times \nu$ orthogonal matrix such that the diagonal elements of $M_0^t R M_0$ are all identical.* One choice of M_0 is $\tilde{M}_0 \triangleq M_1 M_2$, where M_1 is a $\nu \times \nu$ orthogonal matrix which diagonalizes R (i.e., $M_1^t R M_1$ is a diagonal matrix), and M_2 is a normalized Hadamard matrix (i.e., a $\nu \times \nu$ orthogonal matrix with entries $\pm 1/\sqrt{\nu}$). A normalized Hadamard matrix is known to exist for all ν which are multiples of 4 up to 200.

Now modify the communication system as follows. Preceding box A in Fig. 2, insert a device which multiplies the data symbols $\{\alpha_j\}$, taken in blocks of ν, by the matrix M_0. Then, following box D in Fig. 3, insert a device which multiplies the input $\{\hat{\alpha}_j\}$, taken in corresponding blocks of length ν, by $M_0^{-1} = M_0^t$. Let the output of this device be $\{\alpha_j'\}$. It is easy to show that (i) the analysis which yields Theorems 4 to 6 is unchanged in the modified system and (ii) $E(\alpha_j' - \alpha_j)^2 \equiv \epsilon^2$, $1 \leq j \leq \nu$.

We must emphasize that the choice of the matrix M_0 depends on R which in turn depends on the channel which is usually unknown or variable. Although it is undoubtedly possible to find an adaptive procedure for finding a good matrix M_0, our conjecture is that, in most real situations, M_0 can be chosen to be any normalized Hadamard matrix with fairly good results.

(vi) To obtain more insight into our scheme, let us assume that $Q \approx 0$, and $G(f) = H_c(f) = N(f) = 0$, $|f| > 1/2T$. Then (38), (41), (42) become

$$P_{AV} \leq \frac{\sigma_\alpha^2}{T} \int_{-F}^{F} |G(f)|^2 df, \tag{47a}$$

$$\epsilon_N^2 \leq \frac{N}{\nu} \int_{-F}^{F} |C_T(f)|^2 N(f) df, \tag{47b}$$

$$\epsilon_I^2 \leq \sigma_\alpha^2 T \left(\frac{N}{\nu}\right) \int_{-F}^{F} \left|\frac{1}{T} C_T(f) G(f) H_c(f) - 1\right|^2 df. \tag{47c}$$

We see immediately that our bounds on the important quantities P_{AV}, ϵ_N^2, ϵ_I^2 depend on $G(f)$, $H_c(f)$, $N(f)$ only for $|f| \leq F$ and not for $F \leq |f| \leq 1/2T$. In particular, we need a channel whose bandwidth is F. Since the transmission rate

$$\rho = \frac{\nu}{N} \cdot \frac{L}{T} = \frac{\nu}{N} \cdot \frac{L}{2FT} \cdot 2F = \frac{\nu}{2WN} \cdot L \cdot 2F$$

* Witsenhausen (Ref. 6) has shown that there always exists an M_0 with the desired property for all $\nu = 1, 2, \cdots$.

we have that

$$F = \left(\frac{\rho}{2L}\right)\left(\frac{2WN}{\nu}\right). \tag{48}$$

Now the ideal bandwidth in a conventional 2^L-level PAM system with rate ρ is $(\rho/2L)$. We pointed out in Section I that the required channel bandwidth in real systems is usually no less than 10 to 20 percent in excess of this. For our system with $\nu = 2WN(1-\epsilon)$, the ratio of the required bandwidth given by (48) to $\rho/2L$ is $(1-\epsilon)^{-1}$, which can be made very small. See the numerical example in remark *vii*.

(*vii*) Roughly speaking, Theorem 6 tells us that there are three sources of error. The first is given by the integral in (41a) which is a bound on the error introduced by the noise in the band $[-F, F]$. The second is given by the integral in (42a), which is a bound on the error introduced by the imperfections of the channel, as compensated by $C_T(f)$, in the band $[-F, F]$. The third source of error is the fact that $Q > 0$. The first two of these sources appear in conventional PAM systems, but the last is unique to our system. The following numerical example shows that Q can in fact be made small.

Let $W = 2FT = 0.415$, $N = 80$, $\nu = 64$. Then $Q(\nu, N, W) = (1/\nu)\,\Sigma_{j=1}^{\nu}(1-\lambda_j) = 1.01 \times 10^{-4}$, which corresponds to -40.0 dB. The ratio $2WN/\nu = 1.0375$, so that the required bandwidth is 3.7 percent in excess of the ideal $\rho/2L$.

Continuing with this example, let us say that $1/T = 6 \times 10^3$/s, $F = W/T = 2490$ Hz, and $L = 2$. Then the transmission rate $\rho = (\nu/N) \cdot (L/T) = 9.6$ kb/s. Note that $1/2T = 3$ kHz, so that the system performance is essentially independent of the channel characteristics or noise in the band [2.49 kHz, 3 kHz]. Of course, we are assuming that $H_c(f) = N(f) = 0$, $|f| > 3$ kHz.

Finally, observe that the receiver-correlator (box D in Fig. 3) must perform $N \cdot \nu$ multiplications every $N \cdot T$ second, or ν/T multiplications per second. For $\nu = 64$, and $1/T = 6 \times 10^3$, this works out to one multiplication every 2.6 μs. To store the $N \times \nu$ $\phi_j(n)$, $1 \le n \le N$, $1 \le j \le \nu$, to say 10-bit accuracy, we need a ROM with capacity $10 \cdot N \cdot \nu = 51.2$ kb.

(*viii*) Continuing with the assumptions made in remark *vi*, let us further assume that $G(f) \equiv T$, $|f| \le F$, and $H_c(f) \equiv 1$, $|f| \le F$, i.e., a perfect channel response (in band). Also, let $N(f) = N_0/2$, $|f| \le 1/2T$. Then, from (47),

$$P_{AV} \le 2FT\,\sigma_\alpha^2.$$

Further, the upper bound on the total mean-squared error is minimized for $C_T(f) \equiv 1$ (i.e., $c(0) = 1$, $c(m) = 0$, $m \ne 0$). Then

$$\epsilon^2 \le \frac{N}{\nu}\int_{-F}^{F}\frac{N_0}{2}\,df = \frac{N}{\nu}(N_0F).$$

If $\nu/N \approx 2FT$, the total mean-squared error

$$\epsilon^2 \leq \frac{(N_0F)}{P_{AV}}\,\sigma_\alpha^2 \tag{49}$$

Note that (N_0F) is the noise power in the band $[-F, F]$. Observe that in a conventional PAM system with a perfect Nyquist equivalent channel and additive white noise, the mean-squared error is given the right member of (49) with F = the Nyquist bandwidth (see Ref. 1, Chap. 5).

(ix) Suppose that it is desired to transmit our data using a modulated signal $x(t)$ which is bandpass in the band $[F_1, F_2]$. Then, using quadrature amplitude modulation (QAM) in a straightforward manner, we can modify the present scheme to achieve a bandpass signal. We will now outline the procedure.

Let $0 < F_1 < F_2$ be given. Set $F = (F_2 - F_1)/2$, and choose $T < 1/2\,F$. With F, T so chosen, form two modulated signals (with independent data) according to our (baseband) prescription. Denote these baseband signals by $x^{(1)}(t)$, $x^{(2)}(t)$. Their rates are each $\nu L/NT$. Then form a bandpass signal

$$x(t) = x^{(1)}(t)\cos 2\pi F_c t + x^{(2)}(t)\sin 2\pi F_c t,$$

where $F_c = (F_2 + F_1)/2$. The signal $x(t)$ is essentially bandpass with lower frequency

$$F_c - F = \left(\frac{F_2 + F_1}{2}\right) - \left(\frac{F_2 - F_1}{2}\right) = F_1\,,$$

and upper frequency $F_c + F = F_2$. The transmission rate for $x(t)$ is

$$\rho = \frac{2\nu L}{NT} = 2\,\frac{\nu}{N}\,\frac{L}{2FT}\,(2F) = \left(\frac{\nu}{N}\right)\frac{L}{W}\,(F_2 - F_1).$$

Thus the required channel bandwidth to pass $x(t)$ is

$$(F_2 - F_1) = \frac{\rho}{2L}\left(\frac{2WN}{\nu}\right),$$

exactly as in the baseband case (48). It is a fairly simple matter to analyze the QAM system and obtain results analogous to Theorems 4 to 6.

Another way of accomplishing the synthesis of a bandpass signal is to use "bandpass" DPSSs instead of the conventional DPSS characterized in Section III.

(x) Combining (31) and (32), we can rewrite the modulated signal as

$$x(t) = \sum_{j=-\infty}^{\infty} \alpha_j \tilde{g}(t, j),$$

where for $k\nu < j \leq (k+1)\nu$,

$$\tilde{g}(t, j) = \sum_{n=Nk+1}^{N(k+1)} \phi_{j-k\nu}(n - Nk)\,g(t - nT).$$

Note that, for $s = 0, \pm 1, \pm 2, \cdots$,

$$\tilde{g}(t, j + s\nu) = \tilde{g}(t + sNT, j),$$

so that there are only ν possible shapes for $\tilde{g}(t, j)$.

We conclude from this that the present system is a kind of PAM, with the data $\{\alpha_j\}$ modulating the amplitude pulses $\{\tilde{g}(t, j)\}$.

(*xi*) *Computation of the Tap Weights:* Let us again make the assumptions of remark *vi*. Then, to minimize ϵ^2, it is a reasonable strategy to choose the coefficients $\{c(m)\}_{-M}^{M}$ so that $C_T(f)G(f)H_c(f)$ is as close as possible to unity for $|f| \leq F$. Of course, we must take care not to enhance the noise by making $C_T(f)$ too large. In fact, it is a simple matter to solve for the optimal set $\{c(m)\}_{-M}^{M}$ which minimizes our bound on the total mean-squared error (with $Q \approx 0$)

$$= \left(\frac{N}{\nu} T\right) \left[\int_{-F}^{F} |C_T(f)|^2 \frac{N(f)}{T} df + \sigma_\alpha^2 \int_{-F}^{F} \left| C_T(f) \frac{G(f)\, H_c(f)}{T} - 1 \right|^2 df \right]. \quad (50)$$

Let the sequences $\xi_0(\cdot)$, $\xi_1(\cdot)$, $\xi_2(\cdot)$ be the inverse transforms of

$$\frac{N^{1/2}(f)}{T^{1/2}} \Gamma(f), \qquad \frac{G(f)H_c(f)\Gamma(f)}{T} \Gamma(f), \qquad \Gamma(f),$$

respectively, where

$$\Gamma(f) = \begin{cases} 1, & |f| \leq F, \\ 0, & F < |f| \leq \dfrac{1}{2T}. \end{cases}$$

Then the bound of (50) is, from the Parseval relation (15) ("$*$" indicates convolution),

$$= \frac{N}{\nu} [\|c * \xi_0\|^2 + \sigma_\alpha^2 \|c * \xi_1 - \xi_2\|^2]$$

$$= \frac{N}{\nu} \sum_{n=-\infty}^{\infty} \left\{ \left[\sum_{m'=-M}^{+M} c(m')\xi_0(n - m') \right]^2 + \sigma_\alpha^2 \left[\sum_{m'=-M}^{M} c(m')\xi_1(n - m') - \xi_2(n) \right]^2 \right\}.$$

Differentiating with respect to $c(m)$, $-M \leq m \leq M$, and setting the result equal to zero, yields

$$\sum_{m'=-M}^{M} c(m') \left\{ \sum_{n=-\infty}^{\infty} \xi_0(n - m)\xi_0(n - m') + \sigma_\alpha^2 \sum_{n=-\infty}^{\infty} \xi_1(n - m)\xi_1(n - m') \right\} = \sum_{n=-\infty}^{\infty} \xi_1(n - m)\xi_2(n),$$

or

$$\boxed{\sum_{m=-M}^{M} c(m)\mu_0(m-m') = \mu_1(m)}, \qquad m = 0, \pm 1, \cdots, \pm M, \tag{51a}$$

where

$$\mu_0(m) = \sum_{n=-\infty}^{\infty} [\xi_0(n)\xi_0(n-m) + \sigma_a^2 \xi_1(n)\xi_1(n-m)] \tag{51b}$$

$$\mu_1(m) = \sum_{n=-\infty}^{\infty} \xi_1(n-m)\xi_2(n). \tag{51c}$$

Clearly, $\mu_0(\cdot)$ is the inverse transform of

$$\left[\frac{N(f)}{T} + \sigma_a^2 \frac{|G(f)|^2 |H_c(f)|^2}{T^2}\right] \Gamma(f),$$

and $\mu_1(\cdot)$ is the inverse transform of $(\sigma_a^2/T)\, G^*(f) H_c^*(f)\Gamma(f)$. The tap weights $\{c(m)\}_{-M}^{M}$ are found by solving the linear equations (51a).

Of course, the above computation of the tap weight coefficients is possible only when the channel transfer function $H_c(f)$ and the noise spectrum $N(f)$ are known. In most real applications, these quantities are unknown or changing, so that an adaptive learning technique is required.

VI. PROOF OF THEOREMS

Proof of Theorem 4: Using (32), we have

$$\begin{aligned} P_{AV} &\triangleq \frac{1}{NT} E \int_0^{NT} x^2(t)dt \\ &= \frac{1}{NT} E \int_0^{NT} \left(\sum_{k=-\infty}^{\infty} x_k(t)\right)^2 dt \\ &\overset{(1)}{=} \frac{1}{NT} \sum_{k=-\infty}^{\infty} E \int_0^{NT} x_k^2(t)dt \\ &\overset{(2)}{=} \frac{1}{NT} \sum_{k=-\infty}^{\infty} E \int_{-kNT}^{-kNT+NT} x_0^2(t)dt \\ &= \frac{1}{NT} E \int_{-\infty}^{\infty} x_0^2(t)dt = \frac{1}{NT} E \int_{-\infty}^{\infty} |X_0(f)|^2 df. \end{aligned} \tag{52}$$

Step (1) follows from the independence of the $\{\alpha_j\}_{-\infty}^{\infty}$, which implies [see (31) and (32b)] that $E\, x_k(t)x_{k'}(t) = 0$, $k \neq k'$. Step (2) follows from (32b), and the fact that $\{a_k(n)\}_{n=Nk+1}^{N(k+1)}$ has the same statistics as $\{a_0(n)\}_{n=1}^{N}$. Thus, $x_k(t)$ has the same statistics as $x_0(t - kNT)$.

We next apply Proposition 2, which implies that

$$X_0(f) = A_T(f)G(f), \tag{53a}$$

where

$$A_T(f) = \sum_{n=1}^{N} a(n)e^{-i2\pi fTn}. \tag{53b}$$

Substituting (53a) into (52), we obtain

$$P_{AV} = \frac{1}{NT} E \int_{-\infty}^{\infty} |A_T(f)|^2 |G(f)|^2 df$$

$$= \frac{1}{NT} \sum_{k=-\infty}^{\infty} E \int_{k/T}^{(k+1)/T} |G(f)|^2 |A_T(f)|^2 df$$

$$= \frac{1}{NT} \sum_{k=-\infty}^{\infty} E \int_{0}^{1/T} \left| G\left(f - \frac{k}{T}\right) \right|^2 |A_T(f)|^2 df,$$

where we have used the fact that $A_T(f)$ is periodic with period $1/T$. Thus

$$P_{AV} = \frac{1}{NT} \int_0^{1/T} \left(\sum_{k=-\infty}^{\infty} \left| G\left(f - \frac{k}{T}\right) \right|^2 \right) (E|A_T(f)|^2) df$$

$$= \frac{1}{NT} \int_{-1/2T}^{+1/2T} \left(\sum_{k=-\infty}^{\infty} \left| G\left(f - \frac{k}{T}\right) \right|^2 \right) (E|A_T(f)|^2 df,$$

where the last step follows from the fact that the integrand is periodic with period $1/T$, so that we can change the interval of integration from $[0, 1/T]$ to $[-1/2T, 1/2T]$. Continuing, we have

$$P_{AV} = \frac{1}{NT} \int_{-F}^{F} \left(\sum_{k=-\infty}^{\infty} \left| G\left(f - \frac{k}{T}\right) \right|^2 \right) (E|A_T(f)|^2) df$$

$$+ \frac{1}{NT} E \int_{F \le |f| \le 1/2T} \left(\sum_{k} \left| G\left(f - \frac{k}{T}\right) \right|^2 \right) |A_T(f)|^2 df$$

$$= I_1 + I_2. \tag{54}$$

Now the second integral I_2 can be overbounded by

$$I_2 \le \frac{1}{NT} (A_1 T^2) E \int_{F \le |f| \le 1/2T} |A_T(f)|^2 df = \frac{1}{N} A_1 T E \frac{1}{T} \| \mathcal{B}' a_0 \|^2,$$

where A_1 is defined by (38b). From (31), with $k = 0$, and $\langle \mathcal{B}'\phi_j, \mathcal{B}'\phi_{j'} \rangle = (1 - \lambda_j)\delta_{jj'}$, we have

$$I_2 \le \frac{1}{N} A_1 E \left\| \sum_{j=1}^{\nu} \alpha_j \mathcal{B}' \phi_j \right\|^2$$

$$= \frac{1}{N} A_1 E \sum_{j=1}^{\nu} \alpha_j^2 (1 - \lambda_j)$$

$$= \frac{\nu}{N} A_1 \sigma_a^2 Q, \tag{55}$$

where Q is defined by (37).

To overbound I_1, the first term in (54), we again use (31) to obtain

$$E|A_T(f)|^2 = E \left| \sum_{j=1}^{\nu} \alpha_j \Phi_{jT}(f) \right|^2$$

where $\Phi_{jT}(f)$ is the transform of $\phi_j(\cdot)$. Since $E\,\alpha_j\alpha_{j'} = \sigma_\alpha^2\delta_{jj'}$, we have

$$E|A_T(f)|^2 = \sum_{j=1}^{\nu} \sigma_\alpha^2 |\Phi_{jT}(f)|^2$$
$$\leq \sigma_\alpha^2 N,$$

by Theorem 7 (proved in Appendix C). Thus

$$I_1 \leq \frac{\sigma_\alpha^2}{T} \int_{-F}^{F} \left(\sum_{k=-\infty}^{\infty} \left| G\left(f - \frac{k}{T}\right) \right|^2 \right) df. \tag{56}$$

Substituting (55) and (56) into (54) yields Theorem 4.

Proof of Theorem 5: Let $t_1, -\infty < t_1 < \infty$, be given. Then from (32) and (31),

$$x(t_1) = \sum_{k=-\infty}^{\infty} x_k(t_1) = \sum_{k=-\infty}^{\infty} \sum_{n=Nk+1}^{N(k+1)} \sum_{j=1}^{\nu} \alpha_{k\nu+j}\phi_j(n - Nk)g(t_1 - nT)$$
$$= \sum_k \sum_j \alpha_{k\nu+j} \left[\sum_{n=Nk+1}^{N(k+1)} \phi_j(n - Nk)g(t_1 - nT) \right].$$

Using $E\,\alpha_j\alpha_{j'} = \sigma_\alpha^2\delta_{jj'}$, we obtain

$$E\,x^2(t_1) = \sum_{k=-\infty}^{\infty} \sum_{j=1}^{\nu} \sigma_\alpha^2 \left[\sum_{n=Nk+1}^{N(k+1)} \phi_j(n - Nk)g(t_1 - nT) \right]^2. \tag{57}$$

Now with t_1 held fixed, define the sequence $c(\cdot)$ by

$$c(n) = g(t_1 - nT), \qquad -\infty < n < \infty.$$

Also for $-\infty < k < \infty$, $1 \leq j \leq \nu$, define the sequences $\phi_{kj}(\cdot)$ by

$$\phi_{kj}(n) = \phi_j(n - kN), \qquad -\infty < n < \infty.$$

Thus $\phi_{kj}(\cdot)$ has support in the interval $[Nk + 1, N(k + 1)]$. Of course, for $-\infty < k, k' < \infty$, $1 \leq j, j' \leq \nu$,

$$\langle \phi_{kj}, \phi_{k'j'} \rangle = \begin{cases} 1, & k = k', j = j', \\ 0, & \text{otherwise}, \end{cases}$$

so that $\{\phi_{kj}\}_{k,j}$ is a family of orthonormal sequences. Furthermore, the term in brackets in (57) is $\langle \phi_{kj}, c \rangle$, so that (58) can be written

$$E\,x^2(t_1) = \sigma_\alpha^2 \sum_{k=-\infty}^{\infty} \sum_{j=1}^{\nu} \langle \phi_{kj}, c \rangle^2. \tag{58}$$

Letting $\mathcal{S}$ be the subspace of l_2 spanned by the $\{\phi_{kj}\}$, and $P_{\mathcal{S}}c$ the projection of the sequence c into $\mathcal{S}$, (58) is

$$E\,x(t_1) = \sigma_\alpha^2 \|P_{\mathcal{S}}c\|^2 \leq \sigma_\alpha^2 \|c\|^2. \tag{59}$$

We will now bound $\|c\|^2$.

Define the function $w_1(t)$, $-\infty < t < \infty$, by

$$w_1(t) = g(t_1 - t),$$

with t_1 still held fixed. Then

$$c(n) = w_1(nT),$$

and Proposition 1 yields

$$C_T(f) = \frac{1}{T} \sum_{k=-\infty}^{\infty} W_1\left(f - \frac{k}{T}\right),$$

where $W_1(f)$ is the ordinary Fourier transform of $w(t)$. Since $W_1(f) = G^*(f)e^{-i2\pi f t_1}$, we have, from (15),

$$\|c\|^2 = T \int_{-1/2T}^{1/2T} |C_T(f)|^2 df$$

$$= \frac{1}{T} \int_{-1/2T}^{1/2T} \left| \sum_{k=-\infty}^{\infty} G^*\left(f - \frac{k}{T}\right) e^{-i2\pi f t_1} \right|^2 df.$$

Combining this with (59) yields Theorem 5.

Proof of Theorem 6: We begin by observing that the entire system described in Section V (up to the slicer in Fig. 3) is linear and the noise is additive and independent of the data. Thus, the error sequence $\{\hat{\alpha}_j - \alpha_j\}_{-\infty}^{\infty}$ can be written as the sum of two sequences $\{\beta_j\}_{-\infty}^{\infty}$ and $\{\gamma_j\}_{-\infty}^{\infty}$. The sequence $\{\beta_j\}$ is data dependent and is of the form

$$\beta_j = \sum_{j'} a_{jj'}\alpha_{j'}.$$

In fact, the sequence $\{\beta_j\}$ is the output of box D in Fig. 3 when the noise $z(t) \equiv 0$. The sequence $\{\gamma_j\}$ is due to the noise and is, in fact, equal to the output of box D in Fig. 3 when we set $w(t) \equiv 0$. Since the data and noise are uncorrelated, so are $\{\beta_j\}$ and $\{\gamma_j\}$. Thus the mean-squared error

$$\epsilon^2 = \frac{1}{\nu} E \sum_{j=1}^{\nu} (\hat{\alpha}_j - \alpha_j)^2$$

$$= \frac{1}{\nu} E \sum_{j=1}^{\nu} (\beta_j + \gamma_j)^2$$

$$= \frac{1}{\nu} E \sum_{j=1}^{\nu} \beta_j^2 + \frac{1}{\nu} E \sum_{j=1}^{\nu} \gamma_j^2$$

$$\triangleq \epsilon_I^2 + \epsilon_N^2. \tag{60}$$

We will overbound ϵ_I^2 and ϵ_N^2 separately.

We begin with ϵ_N^2, the error due to the noise. Thus we must overbound

$$\epsilon_N^2 = E \frac{1}{\nu} \sum_{j=1}^{\nu} \gamma_j^2,$$

where $\{\gamma_j\}$ is the output sequence of box D in Fig. 3 when $w(t) \equiv 0$. Let us define the sequence $b_0(\cdot)$ to be the output of box C when $w(t) \equiv 0$. Then

$$\gamma_j = \langle b_0, \phi_j \rangle = \sum_{n=1}^{N} b_0(n)\phi_j(n), \qquad 1 \le j \le N$$

and

$$\epsilon_N^2 = \frac{1}{\nu} E \sum_{j=1}^{\nu} \gamma_j^2 = \frac{1}{\nu} \sum_{j=1}^{\nu} \sum_{n=1}^{N} \sum_{m=1}^{N} \phi_j(n)\phi_j(m) E\, b_0(n) b_0(m). \tag{61}$$

Next observe that $b_0(\cdot)$ is a stationary random sequence with $E\, b_0(n) = 0$, $E\, b_0(n)b_0(m) = R_{b_0}(n-m)$. The sequence $R_{b_0}(n)$, $-\infty < n < \infty$, is the inverse Fourier transform of its spectral density $S_{b_0}(f)$, which is, from Proposition 1,

$$S_{b_0}(f) = \frac{1}{T} |C_T(f)|^2 \sum_{k=-\infty}^{\infty} N_Z\left(f - \frac{k}{T}\right). \tag{62}$$

Returning to (61), we write

$$\epsilon_N^2 = \frac{1}{\nu} \sum_{j=1}^{\nu} \sum_{n=1}^{N} \phi_j(n) \left[\sum_{m=1}^{N} R_{b_0}(n-m)\phi_j(m) \right]. \tag{63}$$

The quantity in brackets in (63) is $d_j(n)$, where the sequence $d_j(\cdot)$ is the convolution of the sequences $R_{b_0}(\cdot)$ and $\phi_j(\cdot)$. Further, (63) can be written as

$$\epsilon_N^2 = \frac{1}{\nu} \sum_{j=1}^{N} \sum_{n=1}^{N} \phi_j(n) d_j(n) = \frac{1}{\nu} \sum_{j=1}^{N} \langle \phi_j, d_j \rangle.$$

From the Parseval relation (14), we have

$$\epsilon_N^2 = \frac{T}{\nu} \sum_{j=1}^{N} \int_{-1/2T}^{1/2T} D_{jT}(f) \Phi_{jT}^*(f) df, \tag{64}$$

where $\Phi_{jT}(f)$, $D_{jT}(f)$ are the transforms of $\phi_j(\cdot)$, $d_j(\cdot)$, respectively. Furthermore, the convolution theorem (13) yields

$$D_{jT}(f) = \Phi_{jT}(f) S_{b_0}(f),$$

so that (64) becomes

$$\begin{aligned} \epsilon_N^2 &= \frac{T}{\nu} \sum_{j=1}^{\nu} \int_{-1/2T}^{1/2T} S_{b_0}(f) |\Phi_{jT}(f)|^2 df \\ &= \frac{T}{\nu} \int_{-F}^{F} S_{b_0}(f) \left(\sum_{j=1}^{\nu} |\Phi_j(f)|^2 \right) df \\ &+ \frac{T}{\nu} \sum_{j=1}^{\nu} \int_{F<|f|\le 1/2T} S_{b_0}(f) |\Phi_j(f)|^2 df. \end{aligned}$$

Using Theorem 7 (Appendix C), we obtain

$$\begin{aligned} &\le \frac{N}{\nu} \int_{-F}^{F} (T S_{b_0}(f)) df + \left[\sup_{F<|f|\le 1/2T} S_{b_0}(f) \right] \frac{1}{\nu} \sum_{j=1}^{\nu} \| \mathcal{B}' \phi_j \|^2 \\ &= \frac{N}{\nu} \int_{-F}^{F} |C_T(f)|^2 \sum_k N_Z\left(f - \frac{k}{T}\right) df + A_2 \frac{1}{\nu} \sum_{j=1}^{\nu} (1 - \lambda_j), \end{aligned}$$

which is (41).

It remains to verify (42), which is an upper bound on the data-dependent error or intersymbol interference error ϵ_I^2. Thus, set the noise $z(t) \equiv 0$, and

$$\epsilon_I^2 = E \frac{1}{\nu} \sum_{j=1}^{\nu} (\hat{\alpha}_j - \alpha_j)^2. \tag{65}$$

We begin by observing that, since the sequences $\phi_j(\cdot)$, $j = 1, \cdots, \nu$, are orthonormal, any sequence $c_0(\cdot)$ can be written as

$$c_0(\cdot) = \sum_{j=1}^{\nu} \gamma_j \phi_j(\cdot) + r(\cdot), \tag{66}$$

where $\langle r, \phi_j \rangle = 0$ and $\gamma_j = \langle c_0, \phi_j \rangle$, $1 \le j \le \nu$. Applying (66) to the sequence $\mathcal{D}(\hat{a} - a)$, where $\hat{a}(\cdot)$ and $a(\cdot)$ are as defined in Section V and $\mathcal{D} = \mathcal{D}_N$ is the index-limiting operator defined in Section II, we obtain

$$\begin{aligned} \mathcal{D}(\hat{a} - a) &= \sum_{j=1}^{\nu} \langle \mathcal{D}(\hat{a} - a), \phi_j \rangle \phi_j + r(\cdot) \\ &= \sum_{j=1}^{\nu} \langle \hat{a} - a, \phi_j \rangle \phi_j + r, \end{aligned}$$

where $\langle r, \phi_j \rangle = 0$, $1 \le j \le \nu$. Thus

$$\begin{aligned} \| \mathcal{D}(\hat{a} - a) \|^2 &= \sum_{j=1}^{\nu} \langle \hat{a} - a, \phi_j \rangle^2 + \| r \|^2 \\ &\ge \sum_{j=1}^{\nu} \langle \hat{a} - a, \phi_j \rangle^2. \end{aligned} \tag{67}$$

Now from (31),

$$\langle a, \phi_j \rangle = \alpha_j, \qquad 1 \le j \le \nu,$$

and from (35a),

$$\langle \hat{a}, \phi_j \rangle = \hat{\alpha}_j.$$

Thus (67) is

$$\| \mathcal{D}(\hat{a} - a)^2 \|^2 \ge \sum_{j=1}^{\nu} (\hat{\alpha}_j - \alpha_j)^2,$$

so that the mean-squared error,

$$\epsilon_I^2 \triangleq \frac{1}{\nu} E \sum_{j=1}^{\nu} (\hat{\alpha}_j - \alpha_j) \le \frac{1}{\nu} E \| \mathcal{D}(\hat{a} - a) \|^2. \tag{68}$$

Ineq. (68) relates the error ϵ_I^2 to the error which the system makes in transmitting the sequence $a(\cdot)$.

We proceed to overbound $(1/\nu) E \| \mathcal{D}(\hat{a} - a) \|^2$. We now define

$$H_T(f) = C_T(f)B_T(f) = \frac{1}{T} C_T(f) \sum_{k=-\infty}^{\infty} G\left(f - \frac{k}{T}\right) H_C\left(f - \frac{k}{T}\right). \tag{69}$$

$B_T(f)$ is defined by (42b). It is easy to verify that $H_T(f)$ is the (sequence) transfer function of the overall system from the input to box B at the transmitter (Fig. 2), through the channel [defined by (33)], and through the sampler and box C at the receiver (Fig. 3). Thus with $h(\cdot)$, the inverse transform of $H_T(f)$,

$$\hat{a}(n) = \sum_{m=-\infty}^{\infty} h(n-m)a(m), \qquad -\infty < n < \infty. \tag{70}$$

Further, the error sequence

$$\begin{aligned} \hat{a}(n) - a(n) &= \sum_{m=-\infty}^{\infty} a(m)[h(n-m) - \delta_{n,m}] \\ &= \sum_{m=-\infty}^{\infty} a(m)u(n-m), \qquad -\infty < n < \infty, \end{aligned} \tag{71a}$$

where the sequence $u(\cdot)$ is defined by

$$u(n) = h(n) - \delta_{n,0}, \qquad -\infty < n < \infty. \tag{71b}$$

The transform of the sequence $u(\cdot)$ is

$$U_T(f) = H_T(f) - 1. \tag{72}$$

Now, with $a_k(\cdot)$ as defined by (31a), we define the convolution of $a_k(\cdot)$ and $u(\cdot)$ to be

$$\begin{aligned} v_k(n) = (a_k * u)(n) &= \sum_{m=-\infty}^{\infty} a_k(m)u(n-m) \\ &= \sum_{m=Nk+1}^{N(k+1)} a_k(m)u(n-m). \end{aligned}$$

Since $a(\cdot) = \sum_{k=-\infty}^{\infty} a_k(\cdot)$, (71) is

$$\hat{a} - a = a * u = \sum_{k=-\infty}^{\infty} v_k. \tag{73}$$

We next introduce the time-truncation operators $\mathcal{D}^{(k)}$, $-\infty < k < \infty$, defined by

$$(\mathcal{D}^{(k)}b_0)(n) = \begin{cases} b_0(n), & Nk+1 \le n \le N(k+1), \\ 0, & \text{otherwise.} \end{cases} \tag{74}$$

Of course, $\mathcal{D} = \mathcal{D}^{(0)}$. Then, from (73),

$$\begin{aligned} \frac{1}{\nu} \|\mathcal{D}(\hat{a} - a)\|^2 &= \frac{1}{\nu} \|\mathcal{D}^{(0)}(\hat{a} - a)\|^2 \\ &= \frac{1}{\nu} \left\| \sum_{k=-\infty}^{\infty} \mathcal{D}^{(0)} v_k \right\|^2 = \frac{1}{\nu} \sum_{k,k'=-\infty}^{\infty} \langle \mathcal{D}^{(0)} v_k, \mathcal{D}^{(0)} v_{k'} \rangle. \end{aligned} \tag{75}$$

Now let us observe that v_k depends (through a_k) on the data symbols in and only in the kth data block

$$\alpha_{\nu k+1}, \cdots, \alpha_{\nu(k+1)}.$$

Since all the data symbols are assumed to be statistically independent, we conclude that v_k and $v_{k'}$ $(k \neq k')$ are also statistically independent and uncorrelated. Thus

$$E\langle \mathcal{D}^{(0)}v_k, \mathcal{D}^{(0)}v_{k'}\rangle = 0, \qquad k \neq k',$$

and (68) and (75) are

$$\epsilon_I^2 \leq \frac{1}{\nu} E\|\mathcal{D}(\hat{a} - a)\|^2 = \frac{1}{\nu}\sum_{k=-\infty}^{\infty} E\|\mathcal{D}^{(0)}v_k\|^2. \tag{76}$$

We now make another observation about the sequence $v_k(\cdot)$. As we observed in the proof of Theorem 4, the N random variables $\{a_k(n)\}_{n=Nk+1}^{N(k+1)}$ have the same statistics as the N random variables $\{a_0(n)\}_{n=1}^{N}$. It follows that, for $-\infty < n < \infty$, $v_k(n)$ has the same statistics as $v_0(n - Nk)$, so that

$$\begin{aligned} E\|\mathcal{D}^{(0)}v_k\|^2 &= E\sum_{n=1}^{N} v_k^2(n) \\ &= E\sum_{n=1}^{N} v_0^2(n - Nk) = E\sum_{n=-Nk+1}^{N(-k+1)} v_0^2(n) \\ &= E\|\mathcal{D}^{(-k)}v_0\|^2. \end{aligned} \tag{77}$$

Substituting (77) into (76), we have

$$\epsilon_I^2 \leq \frac{1}{\nu}\sum_{k=-\infty}^{\infty} E\|\mathcal{D}^{(-k)}v_0\|^2 = \frac{1}{\nu}E\|v_0\|^2. \tag{77a}$$

Now from the convolution formula (13) and the Parseval relation (15),

$$\begin{aligned} E\|v_0\|^2 &= E\|a_0 * u\|^2 \\ &= T\,E\int_{-1/2T}^{1/2T} |A_T(f)|^2 |U_T(f)|^2 df. \end{aligned} \tag{78}$$

Since

$$A_T(f) = \sum_{j=1}^{\nu} \alpha_j \Phi_{jT}(f)$$

and the $\{\alpha_j\}_1^\nu$ are uncorrelated,

$$E|A_T(f)|^2 = \sum_{j=1}^{\nu} \sigma_\alpha^2 |\Phi_{jT}(f)|^2. \tag{79}$$

Substituting (79) and (78) into (77), we have

$$\epsilon_I^2 \leq \frac{1}{\nu} T \int_{-1/2T}^{1/2T} |U_T(f)|^2 \sum_{j=1}^{\nu} \sigma_\alpha^2 |\Phi_{jT}(f)|^2 df$$

$$= \frac{\sigma_\alpha^2}{\nu} T \int_{-F}^{F} |U_T(f)|^2 \sum_{j=1}^{\nu} |\Phi_{jT}(f)|^2 df$$

$$+ \frac{\sigma_\alpha^2}{\nu} T \sum_{j=1}^{\nu} \int_{F \le |f| \le 1/2T} |U_T(f)|^2 |\Phi_{jT}(f)|^2 df$$

$$\le \frac{\sigma_\alpha^2}{\nu} T N \int_{-F}^{F} |U_T(f)|^2 df$$

$$+ \frac{\sigma_\alpha^2}{\nu} T \sup_{F \le |f| \le 1/2T} |U_T(f)|^2 \sum_{j=1}^{\nu} \int_{F \le |f| \le 1/2T} |\Phi_j(f)|^2 df.$$

Using (72) and (69), which define $U(f)$, and the definition of A_3 (42c), we obtain

$$\epsilon_I^2 \le \sigma_\alpha^2 \left(\frac{N}{\nu}\right) T \int_{-F}^{F} |C_T(f) B_T(f) - 1|^2 df$$

$$+ A_3 \frac{1}{\nu} \sum_{j=1}^{\nu} \|\mathcal{B}' \phi_j\|^2.$$

Since $\|\mathcal{B}'\phi_j\|^2 = 1 - \lambda_j$, we have established (42), completing the proof of Theorem 6.

APPENDIX A

Proof of Theorem 3

Let $T, F > 0$, with $W \triangleq FT < 1/2$ be given. Let $N = 1, 2, \cdots$, also be given. Define the sequence $\gamma(\cdot)$ by

$$\gamma(n) = \frac{\sin 2\pi W n}{\pi n}, \qquad -\infty < n < \infty. \tag{80}$$

The transform of $\gamma(\cdot)$ is easily seen to be

$$\Gamma_T(f) = \sum_{n=-\infty}^{\infty} \gamma(n) e^{-i2\pi f T n} = \begin{cases} 1, & |f| \le F \\ 0, & F < |f| \le \dfrac{1}{2T}. \end{cases} \tag{81}$$

The bandlimiting operator $\mathcal{B} = \mathcal{B}_F$ is therefore defined by $b = \mathcal{B}a$, where $B_T(f) = \Gamma_T(f) A_T(f)$.

Let K be the $N \times N$ matrix with (m, n)th entry $\gamma(n - m)$, $1 \le n, m \le N$. Consider the matrix eigenvalue equation

$$K\bar{\alpha} = \lambda\bar{\alpha}, \tag{82}$$

where $\bar{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)^t$. Equation (82) is equivalent to

$$\sum_{m=1}^{N} \gamma(n - m)\alpha_m = \lambda\alpha_n, \qquad 1 \le n \le N. \tag{82'}$$

Since K is a symmetric matrix, eq. (82) or (82′) has N (not necessarily

distinct) real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ and a corresponding set of N real orthogonal eigenvectors $\overline{\alpha}_j = (\alpha_{j1}, \cdots, \alpha_{jN})^t$, $1 \leq j \leq N$. We assume that the eigenvectors are normalized so that $\overline{\alpha}_j^t \overline{\alpha}_{j'} = \delta_{jj'}$. Also we can write

$$\sum_{m=1}^{N} \gamma(n-m)\alpha_{jm} = \lambda_j \alpha_{jn}, \qquad 1 \leq n \leq N. \tag{83}$$

We can now define our sequences $\{\phi_j(\cdot)\}$. Let

$$\phi_j(n) = \begin{cases} \alpha_{jn}, & 1 \leq n \leq N, \\ 0, & \text{otherwise.} \end{cases} \tag{84}$$

Equation (83) is therefore

$$\sum_{n=1}^{N} \gamma(n-m)\phi_j(m) = \lambda_j \phi_j(n), \qquad 1 \leq n \leq N, \tag{85}$$

which is equivalent to

$$\mathcal{D}\mathcal{B}\phi_j = \lambda_j \phi_j, \qquad 1 \leq j \leq N, \tag{86}$$

where $\mathcal{D} = \mathcal{D}_N$, and $\mathcal{B} = \mathcal{B}_F$. This is Theorem 3B.

Now, for $1 \leq j \leq N$, let $c_j(\cdot) = \mathcal{B}\phi_j$. Then (86) implies that

$$c_j(n) = \lambda_j \phi_j(n), \qquad 1 \leq n \leq N. \tag{87}$$

Thus

$$\begin{aligned} \langle c_j, \phi_j \rangle &= \sum_{n=1}^{N} c_j(n)\phi_j(n) = \sum_{n=1}^{N} \lambda_j \phi_j^2(n) \\ &= \lambda_j \|\phi_j\|^2 = \lambda_j. \end{aligned}$$

Further, since the transform of $c_j(\cdot)$ is $C_{jT}(f) = \Gamma_T(f)\Phi_{jT}(f)$, the Parseval relation (14) yields

$$\begin{aligned} \lambda_j = \langle c_j, \phi_j \rangle &= T \int_{-1/2T}^{1/2T} \Gamma_T(f)\Phi_{jT}(f)\Phi_{jT}^*(f)df \\ &= T \int_{-F}^{F} |\Phi_j(f)|^2 df. \end{aligned} \tag{88}$$

From (88), $\lambda_j \leq \|\phi_j\|^2 = 1$ and $\lambda_j \geq 0$. In fact, if $\lambda_j = 0$, then $\Phi_j(f) = 0$ for $|f| \leq F$. But, since $\Phi_j(f)$ is a polynomial in $e^{-i2\pi fT}$, it vanishes on an interval only if it vanishes identically, which contradicts $\|\phi_j\| = 1$. Thus $\lambda_j > 0$ for $1 \leq j \leq N$. Since $\Sigma_1^N \lambda_j = \text{trace } K = 2WN$, the $\{\lambda_j\}_1^N$ can be labeled so that they satisfy Theorem 3A.

Now Theorem 3C follows from

$$\langle \phi_j, \phi_k \rangle = \overline{\alpha}_j^t \overline{\alpha}_k = \delta_{jk}.$$

Theorem 3D is established as follows:

$$\langle \mathcal{B}\phi_j, \mathcal{B}\phi_k \rangle = \langle c_j, c_k \rangle = T \int_{-1/2T}^{1/2T} C_{jT}(f) C_{kT}^*(f) df$$

$$= T \int_{-1/2T}^{1/2T} \Gamma_T(f)\Phi_{jT}(f)\Phi_{kT}^*(f)df$$

$$= \langle \mathcal{B}\phi_j, \phi_k \rangle \overset{(1)}{=} \langle \mathcal{D}\mathcal{B}\phi_j, \phi_k \rangle$$

$$\overset{(2)}{=} \lambda_j \langle \phi_j, \phi_k \rangle = \lambda_j \delta_{jk}.$$

Step (1) follows from the fact that ϕ_k has support on $[1, N]$ so that for any $a(\cdot)$, $\langle a, \phi_k \rangle = \langle \mathcal{D}a, \phi_k \rangle$. Step (2) follows from Theorem 3B.

To prove Theorem 3F, observe* that

$$\sum_{j=1}^{N} \lambda_j = \text{trace } K = 2WN \tag{89}$$

and

$$\sum_{j=1}^{N} \lambda_j^2 = \text{trace } (K^t K) = \sum_{n,m=1}^{N} \gamma^2(n-m).$$

Substitution of the formula for $\gamma(n)$ (80) and a simple computation yields

$$\sum_{1}^{N} \lambda_j^2 \geq 2WN - 0 (\log N), \tag{90}$$

as $N \to \infty$. Combining (89) and (90), we have

$$\frac{1}{N}\sum_{1}^{N} \lambda_j(1-\lambda_j) = \frac{1}{N}\sum_{1}^{N} \lambda_j - \lambda_j^2 \leq \frac{0 (\log N)}{N} \xrightarrow{N} 0. \tag{91}$$

Let $S = \{j: \delta < \lambda_j < 1 - \delta\}$. Then (91) yields

$$\frac{1}{N}\delta^2 (\text{card } S) \leq \frac{1}{N}\sum_{1}^{N} \lambda_j(1-\lambda_j) \xrightarrow{N} 0,$$

which is, on dividing by δ^2, Theorem 3E.

Theorem 3F follows directly from Theorems 3A and 3F. Theorem 3G is established in Ref. 4. Finally, Theorem 3H follows immediately from the definition of the $\phi_j(\cdot)$, λ_j, $1 \leq j \leq N$.

APPENDIX B

Saltzberg's Bound

Saltzberg's bound[2,3] states that if ξ is a random variable defined by

$$\xi_1 = \sum_{j=-\infty}^{\infty} \eta_j \alpha_j,$$

where $\{\alpha_j\}_{-\infty}^{\infty}$ are i.i.d. copies of the r.v. α, defined by (29), and $\{\eta_j\}$ are fixed coefficients, then the moment generating function of ξ_1

$$M_{\xi_1}(s) = E\, e^{s\xi_1} \geq \exp\left\{\frac{s^2\sigma_1^2}{2}\right\}, s \geq 0, \tag{92a}$$

* This trick is used in Ref. 7.

where

$$\sigma_1^2 = \text{Var}\, \xi_1 = \sigma_\alpha^2 \sum_{j=-\infty}^{\infty} \eta_j^2. \tag{92b}$$

The right member of (92a) is the moment-generating function of a Gaussian r.v. with zero mean and variance σ_1^2.

Now let $\xi = \xi_1 + \xi_2$, where ξ_1 is as above and ξ_2 is a Gaussian r.v. with zero mean and variance σ_2^2. Let ξ_1 and ξ_2 be statistically independent. Then, from Saltzberg's bound (92), the moment-generating function of ξ,

$$M_\xi(s) = M_{\xi_1}(s) \cdot M_{\xi_2}(s) \leq \exp\left\{\frac{s^2}{2}(\sigma_1^2 + \sigma_2^2)\right\}, \qquad s \geq 0.$$

It follows from the Chernoff bounding technique that, for $r > 0$

$$\Pr\{\xi > r\} \leq \exp\left\{-\frac{r^2}{2(\sigma_1^2 + \sigma_2^2)}\right\}. \tag{93}$$

Let us now apply (93) to establish the claims made in remarks *iv* and *v* in Section V. In remark *iv*, observe that $x(t)$ is a random variable of the form of ξ_1, i.e., a linear combination of the data symbols $\{\alpha_j\}$. If we apply (93) with $\xi_1 = x(t)$, $\xi_2 \equiv 0$, we obtain the inequality of remark *iv*.

Next consider Remark *v*. Observe that, due to the linearity of the system, the error $\hat{\alpha}_j - \alpha_j$ is of the form of ξ, with $\sigma_1^2 + \sigma_2^2 = \epsilon_j^2$. Thus (93) yields

$$\Pr\{(\hat{\alpha}_j - \alpha_j) > 1\} \leq \exp\left\{\frac{-1}{2\epsilon_j^2}\right\}.$$

Since $\tilde{\alpha}_j \neq \alpha_j$, only when $|\hat{\alpha}_j - \alpha_j| > 1$, we have

$$P_{ej} = \Pr\{\tilde{\alpha}_j \neq \alpha_j\} \leq 2 \exp\left\{\frac{-1}{2\epsilon_j^2}\right\}.$$

APPENDIX C

Theorem 7: Let $\{\phi_j(\cdot)\}_{j=1}^N$ be an orthonormal set of sequences (in l_2) with support on $[1, N]$. That is, $D_N\phi_j = \phi_j$ and $\langle \phi_j, \phi_k \rangle = \delta_{jk}$, $1 \leq j,k \leq N$. Let $\Phi_{jT}(f)$, $-\infty < f < \infty$, be the Fourier transform of $\phi_j(\cdot)$, $1 \leq j \leq N$. Then

$$\sum_{j=1}^{N} |\Phi_{jT}(f)|^2 = N, \qquad -\infty < f < \infty.$$

Proof: Let $v(n) = e^{-i2\pi fTn}$ or 0 according as $n \in [1, N]$ or $n \notin [1, N]$. From the orthonormality of the $\{\phi_j\}_1^N$, we conclude that they span the N-dimensional space of complex-valued sequences with support on $[1, N]$. Thus we can write

$$v(n) = \sum_{j=1}^{N} v_j \phi_j(n), \qquad -\infty < n < \infty,$$

where

$$v_j = \langle v, \phi_j \rangle = \sum_{n=1}^{N} v(n)\phi_j(n)$$

$$= \sum_{n=-\infty}^{\infty} e^{-i2\pi fTn}\phi_j(n) = \Phi_{jT}(f), \qquad 1 \le j \le N.$$

Furthermore, the orthonormality of the $\{\phi_j\}$ also implies

$$\sum_{n=1}^{N} |v(n)|^2 = \sum_{j=1}^{N} |v_j|^2 = \sum_{j=1}^{N} |\Phi_{jT}(f)|^2.$$

Since $|v(n)| \equiv 1$, we have established the theorem.

REFERENCES

1. R. W. Lucky, J. Salz, and E. J. Weldon, *Principles of Data Communication*, New York: McGraw-Hill, 1968.
2. J. E. Mazo, "Quantizing Noise and Data Transmission," B.S.T.J., *47*, No. 8 (October 1968), pp. 1737–1753 (Section IV).
3. B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signalling," IEEE Trans. Inform. Th., *14* (July 1968), pp. 563–568.
4. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—V: The Discrete Case," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1371–1430.
5. D. Slepian, "Program to Compute Discrete Prolate Spheroidal Wave Functions, Sequences and Eigenvalues," unpublished work.
6. H. S. Witsenhausen, unpublished work.
7. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—III," B.S.T.J., *41*, No. 4 (July 1962), pp. 1295–1336.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 11, November 1978
Printed in U.S.A.

Sound Alerter Powered Over an Optical Fiber

By B. C. DeLOACH, Jr., R. C. MILLER, and S. KAUFMAN

(Manuscript received July 18, 1978)

An optically powered sound alerter has been constructed which demonstrates the feasibility of converting optical power into sound power with good efficiency and at power levels comparable to those of present telephone ringers. The alerter has an overall optical-to-acoustic efficiency of about 35 percent at 2 mW of acoustic output power. Optical power is converted to electrical power by a 52-percent efficient photovoltaic detector and then into acoustical power by a 72-percent efficient electroacoustic tone generator which uses a piezoelectric transducer. This demonstration establishes that it is technically feasible to deliver optically, via a fiber lightguide, sufficient power to operate a telephone, since all other telephone signaling functions can be accomplished, in principle, with less power and within the context of dielectric lightguide technology. For conventional usage, the design of a telephone alerter must take many factors into consideration, including background noise masking, frequencies not irritating to the customer, satisfactory performance for customers with impaired hearing, etc. These factors have not been addressed here.

I. INTRODUCTION

The potential introduction of lightguide connecting residential and commercial premises to central switching offices offers exciting possibilities for communications users. The availability to each customer of hundreds of megahertz of inexpensive switchable bandwidth could revolutionize telecommunications. One cost barrier to the employment of lightguide in the local loop would be eased if the guide could also be used to provide the essential functions of ordinary telephone service without requiring metallic wires to carry electrical power to the telephone. The possibility would then exist for introducing a lightguide telephone system; the essential functions of this system would be powered from central offices, and broadband services could subsequently be added to it in a cost-effective manner. The broadband services and non-essential telephone services could be locally powered.

The largest technical uncertainty limiting the consideration of dielectric lightguide for ordinary telephony is power: Can telephone operating power requirements realistically be met by photovoltaic conversion of optical power emergent from the lightguide? Since the largest power demands in a conventional telephone occur when the bell is rung, we have given first priority to investigating the power efficiency of an optically driven sound alerter. The other essential functions—speech signaling and recognition of the telephone hook status—will be discussed in a subsequent report.

II. AN OPTICALLY POWERED ALERTER

Electromechanical ringers of the *TRIMLINE*® and 500D-type telephones produce multitone outputs in the range from 0.4 to 0.6 mW of acoustic power; the simultaneous sounding of five ringers (extension telephones) produces 2.5 mW of acoustic power; and the S1A "hard-of-hearing" alerter produces 4 mW of acoustic power. Cost, physical size, and customer acceptability of the alerter noise are important features in the design of these ringers and in the past have dominated the economic importance of high electroacoustic efficiency. However, if the acoustic power levels just listed are to be realized (or even approached) with optically powered sound alerters, then the attainment of high optical-to-acoustic power conversion efficiency is a matter of paramount necessity—at least within the boundaries set by present laser and lightguide technology.

We have fabricated an optical sound alerter demonstration unit, shown in the photograph of Fig. 1 and block diagram of Fig. 2, in which optical power from a GaAlAs laser is transmitted through a large numerical aperture, low-loss, optical fiber and is air-coupled into a GaAlAs photovoltaic detector[1] where it generates dc electrical power at 1.0 volt. This is converted by attendant circuitry into an audio frequency waveform at the terminals of an electroacoustic tone generator which uses a piezoelectric transducer. The components are mounted on a 16-in. × 10-in. × ¼-in. Lucite board with compact X-Y-Z positioners used for optical alignments. The 0.823-in. high × 1.156-in. diameter Helmholtz acoustic cavity was constructed as an integral part of the board, with the tone generator acoustic output coupled into the surroundings via a Helmholtz air piston consisting of 97 holes of 0.078-in. diameter drilled through the ¼-in. Lucite thickness. A surface-tension microlens, formed by wetting the cut end of the optical fiber with a small drop of optical cement and curing in ultraviolet light, increased the laser-to-fiber coupling efficiency from 45 to over 60 percent. The photovoltaic detector quantum efficiency was improved by applying a thin-film, ZrO_2, antireflection coating which reduced the net reflection loss to 2.5 percent.

To simulate conventional telephone ringing, the laser transmitter is operated in a 2-second-ON, 3-second-OFF cycle (but is otherwise un-

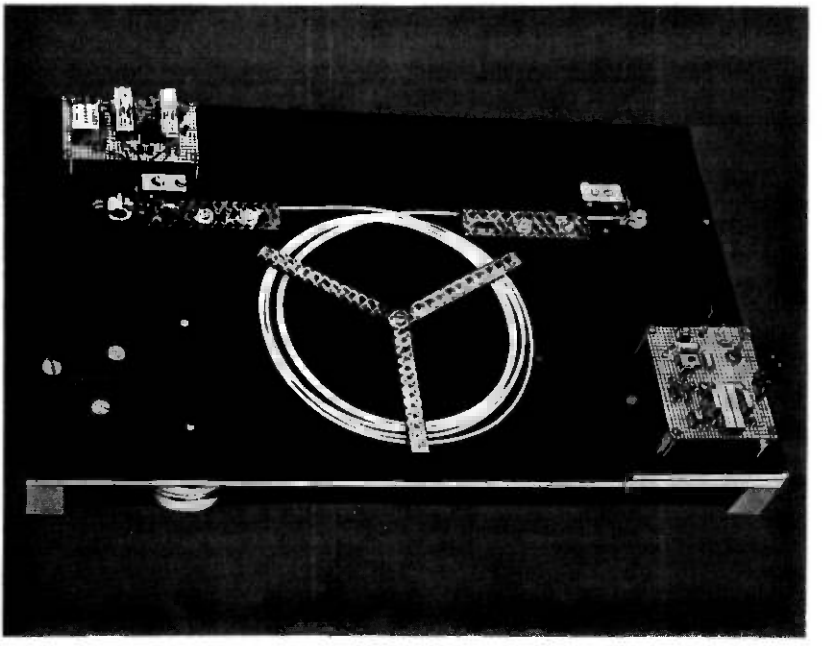


Fig. 1—Optically power sound alerter demonstration unit. Laser, on the right, is coupled by large N.A. fiber lightguide to the photovoltaic detector. Tone generator and high-Q inductors for the ringing choke circuit are at lower left.

modulated), causing the alerter to respond similarly. These periodic bursts of optical power are converted into electrical power by the photodetector, and this power is converted into an audio signal by a free-running, 2.0-kHz nominal frequency, astable multivibrator and an acoustically damped, ringing-choke circuit whose capacitive element is the tone generator. A portion of the acoustic response curve of the tone generator at room temperature is shown in the inset of Fig. 2. A raucous, buzzer-like sound is produced by a second multivibrator which sweeps the audio frequency over a ±100 Hz interval about 1.98 kHz at a 30-Hz rate. The ringing-choke and frequency-control circuits are operated directly from the photovoltaic detector output terminals.

Selection of the 2-kHz tone generator was based on measured comparisons of its acoustic output power with that of similarly designed 1.0-kHz and 1.4-kHz tone generators, the results being qualitatively consistent with acoustic efficiency scaling laws. The ±100-Hz frequency modulation ameliorates some of the more irritating effects associated with the use of a single, unmodulated tone—e.g., strong standing waves (and hence dead zones) in a room, an intensely piercing sound, and masking of the alerter sound if the noise environment contains pure tones of similar frequency. The expedient adopted for this demonstration is

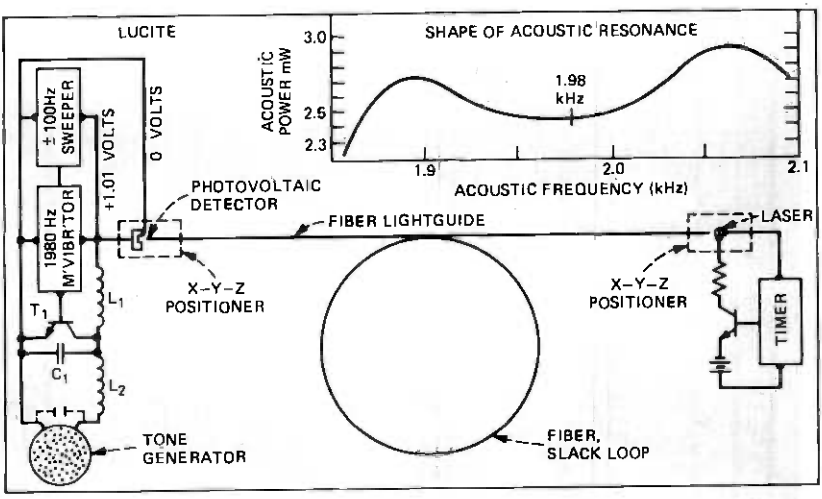


Fig. 2—Schematic of sound alert demonstrator. Laser is battery-operated, with a timing circuit controlling the 2-second-ON, 3-second-OFF ringing cycle. Optical acoustic efficiency was measured with the ±100-Hz sweeper disconnected and the optical fiber replaced by lenses. A 45F transistor is used for T_1. The values of L_1, L_2, and C_1 are 0.24h, 0.16h, and 0.0033 μf for 1.98-kHz operation. The tone generator room temperature acoustic output power at 5.0 V rms is shown in the inset for a limited frequency range.

a departure from usage in existing multitone telephone ringers, which incorporate widely spaced frequencies in the range 750 to 1600 Hz to satisfy human factors criteria. Our optical techniques can be extended to multitone systems with some loss in electroacoustic efficiency. It should also be noted that the volume occupied by the present tone generator and the high-Q inductors of the ringing choke circuit is larger than that consumed by conventional electromechanical telephone ringers.

III. EFFICIENCY

As seen from the Fig. 2 inset, the tone-generator output power varies by about ±5 percent over the ±100-Hz swept band. To obtain a definite, single-frequency result, the sound alerter optical-to-acoustical power conversion efficiency was measured at 1.98 kHz with the ±100-Hz sweeper disconnected. This sweeper consumes approximately 3 percent of the photovoltaic output power when used in the sound alerter demonstration unit. Accuracy in the efficiency measurements required that the optical fiber be replaced by a lens system which focused the $\lambda = 0.801$ μm laser output, with correction for astigmatism,[2,3] onto the photovoltaic detector. (Auxiliary measurements show that the fiber-to-detector coupling efficiency exceeds 95 percent.) Optical powers were measured with a thermopile whose calibration against a pyroelectric radiometer of ±0.5 percent absolute accuracy agreed to within ±1.1 percent with its standard lamp calibration. Acoustic powers were measured as func-

tions of frequency and rms voltage in an absolutely calibrated anechoic test chamber. Electroacoustic efficiencies were obtained from this calibration and the voltage and current waveforms and phase angle at the tone generator.

A summary of the optical, electrical, and acoustic powers and power conversion efficiencies obtained in this measurement is given in Fig. 3. The overall efficiency, defined as the total sound power divided by the optical power incident onto the photovoltaic detector, achieves a maximum value in the range 33 to 36 percent at acoustic powers in the neighborhood of 2.1 mW. With our present unsophisticated circuits, the alerter efficiency decreases slowly and uniformly as the acoustic (or optical) power is lowered, and it decreases abruptly if the power is raised too high. This behavior follows from the shape of the photovoltaic V-I curve and from the fact that the ringing choke circuit presents a load at the detector output terminals which is nearly independent of optical power and which permits optimum photovoltaic efficiency to be achieved only over a narrow optical power range. Thus, the total acoustic power of 2.1 mW might be distributed among three or four alerters ringing simultaneously, but only if the present circuit were modified to maintain an optimum efficiency impedance match to the photodetector. Realization, in the present arrangement, of a previously attained "best" value of photovoltaic efficiency[1] could raise the overall efficiency into the 37 to 40 percent range.

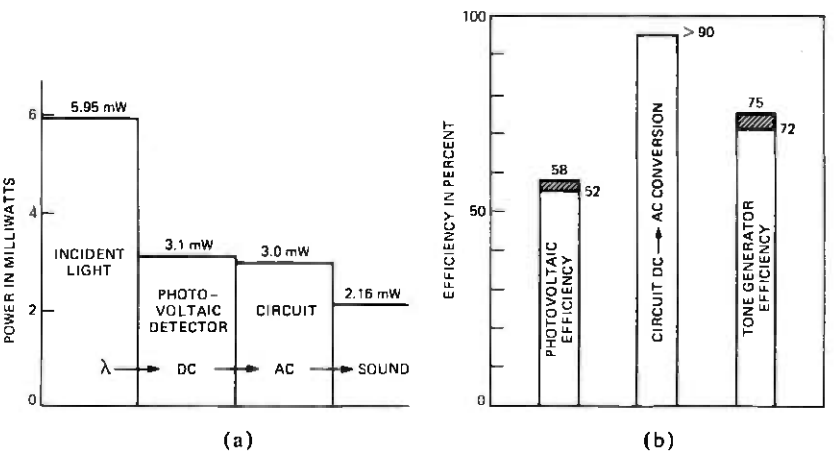


Fig. 3—Power and efficiency. (a) The indicated values represent optical power incident onto detector, detector dc output power, circuit audio frequency power into tone generator, and tone generator total radiated acoustic power at 1980 Hz. (b) The photovoltaic efficiency was 52 percent, and the tone generator efficiency was 72 percent for the results cited in this paper. A value of 58 percent has been measured on another detector of this type, and the present tone generator can provide 75 percent electroacoustic efficiency (over a narrower frequency band) by increasing its cavity height to 0.861 inch.

IV. EXTRAPOLATION TO FUTURE TECHNOLOGY A POWER ESTIMATE

The losses expected for the large N.A. fiber at the 0.801-μm wavelength used in the present demonstration alerter exceed 6 dB/km. Thus, even with perfect laser-to-fiber and fiber-to-detector coupling, at least 5 watts of optical power would be needed at the input end of a 5-km fiber to produce 5 mW of light at the photodetector, a large power indeed. However, fiber losses as low as 1.2 dB/km have been reported[4] at wavelengths near 0.9 μm, which are attainable with GaAs lasers and photodetectors, and as low as 0.5 dB/km at wavelengths near 1.3 μm, which are attainable with $In_xG_{1-x}As_yP_{1-y}$ devices.

For purposes of estimating the power that might eventually be needed at the central office to operate *one* optically driven sound alerter, we will consider the more optimistic case corresponding to 1.3-μm wavelength. In expectation that multi-junction[5] InGaAsP photovoltaic detectors can be constructed whose power conversion efficiency will be comparable to that of the present GaAs detector, we arrive at the values listed in Table I.

A 10-percent efficient InGaAsP laser driving a 5-km loop needs to draw only 0.10 watt during the alerter sounding (and considerably less during speech communication). For comparison, we note that present Bell-System-wired telephones draw at least 20 mA from a 50V battery, i.e., 1 watt, when operating, and they require more to ring.

The alerters on conventional extension telephones are rung simultaneously, but the correct engineering approach to ringing several optically powered telephones in one location is not obvious. A 1-second-ON, 2-second-OFF ringing sequence with appropriate circuits to switch between extensions would permit the ringing of three telephones with the same power as noted in Table I, provided the 2-second delay between

Table 1 — Central office power for $\lambda = 1.3\mu$m laser to operate one sound alerter

Assumed acoustic power per alerter	= 0.6 mW
Tone generator electroacoustic efficiency	= 75 percent*
DC-to-audio circuit efficiency	= 90 percent†
Photovoltaic power conversion efficiency	= 58 percent‡
Fiber-to-photodetector coupling efficiency	= 93 percent
Optical power required at fiber output end	= 1.65 mW
Fiber loss (N.A. = 0.2)	= 0.5 dB/km
Cabling losses (including splices)	= 0.8 dB/km
Total medium loss	= 1.3 dB/km
Laser-to-fiber coupling efficiency	= 80 percent
Optical power required at central office for fiber length of:	1 km 2.8 mW
	2 km 3.8 mW
	5 km 9.2 mW
	10 km 41.2 mW

* Present best value, near 2 kHz.

† Assumes that circuit load line is matched to photovoltaic optimum power point for the optical power actually used.

‡ Assumes that multi-junction photovoltaic efficiency can be made equal to single junction efficiency.

ringing the first and third telephones is tolerable. The ringing of more extensions and of a hard-of-hearing alerter would probably require the assistance of local power, as would additional services such as wideband video.

One reason for examining the possibilities of optical systems in the loop plant, in today's time frame, stems from the expanded services to telephone subscribers which are implicit in the hundreds of megahertz of switchable fiber bandwidth. Beyond this, there are many technical advantages, including freedom from lightning strikes (low voltage electronics), power line pick-up, and problems with potentially dangerous electrical pick-up from customer-provided electrical equipment. Optical loop systems would be more difficult to tap and thus somewhat more secure than wired systems. They may also exhibit compelling cost advantages and power savings, but today's technology is much too rudimentary to permit firm predictions.

V. SUMMARY

We have measured an optical-to-acoustic power conversion efficiency of 33 to 36 percent on an optically powered sound alerter driven by a 0.801-μm wavelength laser. Assuming that comparable efficiency can be attained in the 1.3-μm wavelength range of low fiber-lightguide losses, this result suggests that the optical powering of telephones over glass fibers may in the future be technically feasible for loop lengths up to at least 5 km—in the sense that the other essential station set signaling functions can be performed compatibly with an all-dielectric technology at less power than that consumed by the sound alerter.

Since the high efficiency is made possible by restricting the alerter acoustic output to a narrow range of high frequencies, the tone quality of optically driven alerters would not be expected to equal that of conventional electromechanical ringers. Also, it is unclear, with limited optical powers, how best to handle the problems of ringing extension telephones and hard-of-hearing alerters.

VI. ACKNOWLEDGMENTS

The authors are indebted to R. B. Lawry for assistance with the circuit characterization and efficiency measurements. We thank W. D. Johnston, Jr., for use of his pyroelectric radiometer, and acknowledge the design help given by W. E. Hess, Jr. in initial phases of this work. Many ideas on photovoltaic detector design grew out of conversations with D. L. Rode.

REFERENCES

1. R. C. Miller, B. Schwartz, L. Koszi, and W. R. Wagner, unpublished work.
2. D. D. Cook and F. R. Nash, J. Appl. Phys., *46* (1975), p 1660.
3. R. G. Chemelli and R. C. Miller, U. S. Patent No. 3,974,507 (August 10, 1976).
4. H. Osanai, T. Shioda, T. Moriyama, S. Araki, M. Horiguchi, T. Izawa, and H. Takata, Electron. Lett. *12* (1976), p. 549.
5. M. Ilegems, B. Schwartz, L. A. Koszi, and R. C. Miller, unpublished work.

Contributors to This Issue

H. W. Arnold, B.A., 1965, Occidental College; M.A., 1967, Sc.D., 1971, Columbia University; Bell Laboratories, 1971—. Mr. Arnold has conducted millimeter wave mobile radio propagation experiments and has investigated advanced communications satellite systems. He was involved in the design of the Crawford Hill COMSTAR beacon propagation receivers and is presently performing data analysis from that experiment. Member, IEEE.

John R. Cavanaugh, B.S. (Electrical Engineering), 1961, Ohio University; M.S. (Electrical Engineering), 1963, New York University; Bell Laboratories, 1961—. Mr. Cavanaugh has worked on determining acceptable picture quality standards for broadcast television transmission. He is currently a member of the Network Objectives Department working on subjective evaluations of digital telephone systems. Member, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu.

Donald C. Cox, B.S. (E.E.), 1959, and M.S. (E.E.), 1960, University of Nebraska; Ph.D. (E.E.), 1968, Stanford University; U.S. Air Force Research and Development Officer, Wright-Patterson AFB, Ohio, 1960–1963; Bell Laboratories, 1968—. After coming to Bell Laboratories from Stanford where he was engaged in microwave trans-horizon propagation research, Mr. Cox was engaged in microwave propagation research in mobile radio environments and in high-capacity mobile radio systems studies until 1973. He is now doing millimeter wave satellite propagation and systems research. Senior Member, IEEE and member, Commissions B, C and F of USNC/URSI, Sigma Xi, Sigma Tau, Eta Kappa Nu, and Pi Mu Epsilon; Registered Professional Engineer.

William R. Daumer, B.S.E.E., 1972, M.S.E.E., 1973, Drexel University; Bell Laboratories, 1973—. Mr. Daumer is currently a member of the Digital Network Planning Department, where he is involved in studying the effects of various digital technologies on the transmission performance of the evolving telephone network. Earlier Bell Laboratories experience included work on modeling the economic impact on digital transmission systems in the toll network.

Bernard C. De Loach, Jr., B.S., 1951, M.S. (Physics), 1952, Auburn University; Ph.D. (Physics), 1956, Ohio State University; Bell Laboratories, 1956—. Mr. De Loach has worked in the areas of microwave and millimeter wave solid-state devices and more recently in light-emitting diodes and semiconductor lasers. He is currently head of the Lightwave Sources Department. Fellow, IEEE.

John L. Doane, B.E., 1964, Yale University; S.M., 1965, Ph.D., 1970, Massachusetts Institute of Technology; Bell Laboratories 1970–1977. From 1970 to 1975, Mr. Doane worked on theoretical and experimental determinations of delay distortion in WT4 millimeter waveguide. From 1975 to 1977, he did circuit design and systems engineering for the TSS-R computerized maintenance system for AR6A single sideband radio. Since 1977, he has been designing millimeter-wave instrumentation for monitoring plasma density and temperature at the Plasma Physics Laboratory of Princeton University. Member, IEEE.

Thomas D. Dudderar, B.S.M.E., 1957, Lehigh University; Sc.M., 1961, New York University; Ph.D., 1966, Brown University; Bell Laboratories, 1966—. Mr. Dudderar works in the Materials Research Laboratory. He has published research papers on mechanical properties of materials, photoelasticity, holographic interferometry, and laser speckle photography and has been awarded patents on mechanical testing techniques and stress analysis using holographic interferometry. Member, SESA, SES.

David D. Falconer, B.A.Sc., 1962, University of Toronto; S.M., 1963, and Ph.D., 1967, Massachusetts Institute of Technology; post-doctoral research, Royal Institute of Technology, Stockholm, 1966–1967; Bell Laboratories, 1967—. Mr. Falconer has worked on problems in communication theory and high-speed data communication. During 1976–77 he was a visiting professor of electrical engineering at Linköping University, Linköping, Sweden. He presently supervises a group working on digital signal processing for speech bit rate reduction. Member, Tau Beta Pi, Sigma Xi, IEEE.

Richard D. Gitlin, B.E.E., 1964, City College of New York; M.S., 1965, and D. Eng. Sc., 1969, Columbia University; Bell Laboratories 1969—. Mr. Gitlin is presently concerned with problems in data transmission. He is a member of the Communication Theory Committee of the IEEE Communications Society and is editor for Communication

Theory of the IEEE Transactions on Communications. Senior Member, IEEE; Member, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

Stanley Kaufman, B.E., 1948, M.S., 1957, Johns Hopkins University; Bellcomm, 1968–1971; Bell Laboratories, 1972—. Mr. Kaufman has worked principally in structures and structural dynamics. At Bellcomm, he developed a model for the stability and performance of the Lunar Roving Vehicle. Mr. Kaufman is the author of numerous papers on the finite element method and has participated in the development of various piezoelectric devices. Associate Fellow, AIAA.

William A. Massey, B.A., 1977, Princeton University (Mathematics); Stanford University, 1977—. Mr. Massey spent the summers of 1977 and 1978 working at Bell Laboratories under the Cooperative Research Fellowship Program. He is currently attending Stanford University and is pursuing a Ph.D. in mathematics in the area of Stochastic Processes. Member, Sigma Xi, Phi Beta Kappa.

Stephen C. Mettler, B.S., 1962, U.S.A.F. Academy; M.S. (Physics), 1972, Ph.D. (M.E.), 1976, Purdue University; Bell Laboratories, 1976—. Mr. Mettler is presently engaged in optical fiber splicing studies. Member, Optical Society of America.

Calvin M. Miller, B.S.E.E., 1963, North Carolina State University at Raleigh; M.S.E., 1966, Akron University; Goodyear Aerospace Corporation, 1963–1966; Martin Marietta Company, 1966–1967; Bell Laboratories, 1967—. Before joining Bell Laboratories, Mr. Miller designed electronic and optical components of side-looking radar processor equipment and control systems for reentry vehicles and aircraft flying simulators. At Bell Laboratories, Mr. Miller developed equipment and methods for transmission line characterization. His present interests are in the area of fiber optics as a practical transmission medium. He is supervisor of an exploratory optical fiber splicing group. Member, Optical Society of America.

Richard C. Miller, B.A. (Mathematics), 1949, University of Chicago; Ph.D. (Physics), 1957, University of Illinois; Bell Laboratories, 1959—. Mr. Miller has been concerned with the study of the nonlinear optical properties of materials, high-resolution laser microrecording systems, and the optical properties of semiconductor lasers. He is currently a member of the Lightwave Sources Department.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. He has recently been interested in queuing problems associated with data communications networks. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester 1968. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

Judith B. Seery, B.A., 1968, College of St. Elizabeth; M.S., 1972, New York University; Bell Laboratories, 1968—. Ms. Seery does computing and analysis in the Mathematics and Statistics Research Center. She has recently participated in problems in fiber optics, minimal spanning networks, and multidimensional scaling. Member, Mathematical Association of America, Association for Women in Mathematics.

Peter G. Simpkins, Diploma in Technology, 1957, University of London; M.S., 1960, California Institute of Technology; Ph.D., 1965, Imperial College, London; AVCO Corporation, 1965–1968; Bell Laboratories, 1968—. Mr. Simpkins is currently working in the Materials Research Laboratory. He has published articles on gas dynamics, fluid mechanics, underwater acoustics, and fracture mechanics. During a leave of absence in 1974, he was Senior Research Fellow at the University of Southampton, England.

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 co-chairperson of two international symposia. In 1976, he was president of the IEEE Information Theory Group. Fellow, IEEE, member, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

Papers by Bell Laboratories Authors

CHEMISTRY

Analysis of Phosphorus-Diffused Layers in Silicon. R. B. Fay, J. Electrochem. Soc., *125* (February 1978), pp. 323–327.

Flameless Atomic Absorption Spectrometric Determination of Ultratrace Elements in Silicon Tetrachloride. T. Y. Kometani, Anal. Chem., *49*, No. 14 (1977), pp. 2289–2291.

Functional Group Analysis of Large Chemical Kinetic Systems. T. E. Graedel, J. Phys. Chem., *81* (1977), pp. 2372–2374.

Intermediate Valence in $YbAl_3$ and $EuCu_2Si_2$ by X-Ray Photoemission (XPS). K. H. J. Buschow, M. Campagna, and G. K. Wertheim. Solid State Commun., *24* (1977), pp. 253–256.

Investigation of the Ti-Pt Diffusion Barrier for Gold Beam Leads on Aluminum. S. P. Murarka, H. J. Levinstein, I. Blech, T. T. Sheng and M. H. Read, J. Electrochem. Soc., *125* (January 1978), pp. 156–162.

Noble Gas Chemistry and the Fluoride Literature—What Influences Research Directions? D. T. Hawkins and W. E. Falconer, J. Chem. Inform. Comput. Sci., *17* (1977), pp. 219–220.

Optical Absorption as a Control Test for Plasma Silicon Nitride Deposition. M. J. Rand and D. R. Wonsidler. J. Electrochem. Soc., *125* (January 1978), pp. 99–101.

Picosecond Dynamics of Conformational Changes in 1,1′ Binaphthyl Solutions. C. V. Shank, E. p. Ippen, O. Teschke, and K. B. Eisenthal. J. Chem. Phys., *67*, No. 12 (December 15, 1977), pp. 5547–5551.

Picosecond Dynamics of the Singlet Excited State of Trans and Cis-Stilbene. O. Teschke, E. P. Ippen, and G. R. Holtom. Chem. Phys. Lett., *52*, No. 2 (December 1, 1977), pp. 233–235.

Pulsed EPR Studies of Type I and Type II Copper of *Rhus Vernicifera* Laccase and Porcine Ceruloplasmin. B. Mondovi, M. T. Graziani, W. B. Mims, R. Oltzik, and J. Peisach. Biochem., *16* (1977), pp. 4198–4202.

Si Molecular Beam Epitaxy (n on n^+) with Wide Range Doping Control. Y. Ota, J. Electrochem. Soc., *124*, No. 11 (November 1977), pp. 1795–1802.

Synthesis and Structure of $BaPtO_3$. P. K. Gallagher, D. W. Johnson, E. M. Vogel, G. K. Wertheim, and F. H. Schnettler. J. Solid State Chem., *21* (1977) pp. 277–282.

COMPUTING

Book Review, "Queueing Systems, Vol. 2: Computer Applications" by L. Kleinrock. D. P. Heyman, Networks, *7* (Fall 1977), pp. 285–286.

Single Server Queues With Correlated Inputs. B. Gopinath and J. A. Morrison, Computer Performance, K. M. Chandy and M. Reiser, Eds. Proceedings of the International Symposium on Computer Performance Modeling, Measurement and Evaluation, IBM Thomas J. Watson Research Center, New York, August 16–18, 1977, pp. 263–277.

ELECTRICAL AND ELECTRONIC ENGINEERING

Chevron Detector Design Study. T. J. Nelsón, IEEE Trans. Magn., *MAG-13* (November 1977), pp. 1773–1776.

Contact Design Methodologies, Insulation Displacement Contacts for Multigauge Splicing Connectors. W. E. Pugh, III, Electrical Contacts—1977, Proceedings of the 23rd Annual Holm Conference on Electrical Contacts, *I* (November 1977), pp. 135–140.

Digital Coding of Color Video Signals—A Review. J. O. Limb, C. B. Rubinstein, and J. E. Thompson, IEEE Trans. Commun., *COM-25* (November 1977), pp. 1349–1385.

Digital Signal Processor Architecture. S. K. Tewksbury, R. B. Kieburtz, J. S. Thompson, and S. P. Verma, IEEE Commun. Soc. Mag., *16*, No. 1 (January 1978) pp. 23–27.

High Frequency Approximations For Surface Waves Propagating Along Cylinders of General Cross-Section. J. A. Morrison, Recent Advances in Engineering Science, G. C. Sih, Ed., Proceedings of the 14th Annual Meeting of the Society of Engineering Science, Inc., Nov. 14–16, Bethlehem, Pa. (1977) pp. 673–687.

High-Field Electronic Conduction in Insulators. K. K. Thornber, Solid State Electron., *21* (1978), pp. 259–266.

High Voltage Single-Ended DC/DC Converter. R. P. Massey and E. C. Snyder, IEEE Power Electronics Specialists Conference (December 1977), pp. 156–159.
MM-Wave PIN Switching Diode Fabrication Using Silicon Molecular Beam Epitaxy. Y. Ota, W. L. Buchanan, and O. G. Peterson, Technical Digest of IEEE International Conference on Electron Devices, Washington D. C. December 1977, pp. 375–378.
Multimicrophone Signal-Processing Technique to Remove Room Reverberation From Speech Signals. J. B. Allen, D. A. Berkley, and J. Blauert, J. Acoust. Soc. Amer., *62* (October 1977), pp. 912–915.
Multipath Delay Spread and Path Loss Correlation for 910 MHz Urban Mobile Radio Propagation. D. C. Cox, IEEE Trans. Veh. Technol., *26* (November 1977), pp. 340–344.
Multiple Reflection Corrections in Light-Scattering Experiments. D. F. Nelson and P. D. Lazay, J. Opt. Soc. Am., *67* (November 1977), pp. 1599–1601.
Nonionizing Radiations. G. M. Wilkening, Encyclopedia of Environmental Science and Engineering, *2* (1977), pp. 612–623.
Photons in Fibers for Telecommunications. S. E. Miller, Science, *195* (March 18, 1977), pp. 1211–1216.
Self-Contained Charge-Coupled Split-Electrode Filters Using Novel Sensing Technique. C. H. Sequin, M. F. Tompsett, P. I. Suciu, D. A. Sealer, P. M. Ryan, and E. J. Zimany, IEEE J. Solid State Circuits *SC-12* (December 1977), pp. 626–632.
Spectra of PSK Signals With Overlapping Baseband Pulses. L. J. Greenstein, IEEE Trans. Commun., *COM-25*, No. 5 (May 1977), pp. 523–530.

GENERAL MATHEMATICS AND STATISTICS

Charge Singularity at the Corner of a Flat Plate. J. A. Morrison and J. A. Lewis, SIAM J. Appl. Math., *31* (September 1976), pp. 233–250.
Charge Singularity at the Vertex of a Slender Cone of General Cross-Section. J. A. Morrison, SIAM J. Appl. Math., *33* (July 1977), pp. 127–132.
Effective Versions of the Chebotarv Density Theorem. J. C. Lagarias and A. M. Odlyzko, Algebraic Number Fields (Proceedings of the 1975 Durham Symposium), A. Fröhlich, ed., London and New York: Academic Press, 1977, pp. 409–464.
Scheduling Equal-Length Tasks under Tree-Like Precedence Constraints to Minimize Maximum Lateness. P. Brocher, M. R. Garey, and D. S. Johnson, Mathematics of Operations Research, *2* (August 1977), pp. 275–284.

MATERIALS SCIENCE

Oxide Thickness Measurements Up to 120Å on Silicon and Aluminum Using the Chemically Shifted Auger Spectra. C. C. Chang and D. M. Boulin, Surface Sci., *69* (1977), pp. 385–402.
Reactive Ion Beam Sputtering of Thin Films of Lead, Zirconium and Titanium. R. N. Castellano, Thin Solid Films, *46* (1977), pp. 213–221.
Temperature Dependence of the Orthorhombicity of Gallium Metal. W. H. Haemmerle and J. B. Lastovka, J. Appl. Crystallogr., *10* No. 3 (June 1977), pp. 180–183.
A True Swap Gate for Magnetic Bubble Memory Chips. P. I. Bonyhard, IEEE Trans. Magn., *MAG-13* (November 1977), pp. 1785.
X-Ray Fluorescence Analysis of Some Ferrite Compositions. F. Schrey and P. K. Gallagher, Ceram. Soc. Bull., *56* (November 1977), pp. 981–983.

MECHANICAL AND CIVIL ENGINEERING

Guidelines for High-Reliability Applications of Integrated Circuits. E. R. Schmid, Machine Design, *49* No. 19 (August 25, 1977) pp. 80–82.

PHYSICS

Core-Electron Line Shapes in X-Ray Photoemission Spectra from Semimetals and Semiconductors. G. K. Wertheim and D. N. E. Buchanan, Phys. Rev. B, *16* (September 1977), pp. 2613–2617.
Elastic Constants of hcp ^{4}He. D. S. Greywall, Phys. Rev. B Comments & Addenda, *16* (December 1977), pp. 5127–5128.
Getter-Sputtering at Low Temperature ($\simeq$20°K). J. J. Hauser, Appl. Phys. Lett., *32*, No. 3 (February 1, 1978), pp. 125–127.

Group Velocity in Crystal Optics. D. F. Nelson, Amer. J. Phys., *45* (December 1977), pp. 1187–1190.
The Oxidation of Ammonia, Hydrogen Sulfide, and Methane in Nonurban Tropospheres. T. E. Graedel, J. Geophys. Res., *82* (1977), pp. 5917–5922.
The Prospects for Photovoltaic Conversion. W. D. Johnston, Jr., Amer. Sci., *65* (1977) pp. 729–736.
Void Growth in the Early Stages of Aging and Electromigration. J. R. Lloyd and S. Nakahara, J. Appl. Phys., *48* (1977), pp. 5092–5095.

Bell Laboratories Scientists Named 1978 Nobel Prize Laureates

Arno A. Penzias, director of the Radio Research Laboratory, and Robert W. Wilson, head of the Radio Physics Research Department, at Bell Laboratories, Holmdel, N.J., have been named co-winners of the 1978 Nobel Prize in Physics, jointly with Professor Pyotr Kapitsa of the Academy of Sciences, Moscow. The Bell Labs scientists are receiving the award for their discovery of cosmic microwave background radiation, and Kapitsa is being cited for his basic work in low-temperature physics.

In 1964, Penzias and Wilson began using the most sensitive radio astronomy antenna available, assembled at Crawford Hill initially for satellite communications studies on Echo and Telstar and subsequently for a project they hoped would improve our understanding of the Milky Way. They concluded that what first appeared to be a faint noise signal was the background radiation (3°K) remaining from a cosmic explosion that gave birth to the universe some 20 billion years ago.

"One of the consequences of [the big bang] theory is that the heat from the explosion . . . should be left over and should be barely detectable," Penzias said.

The significance of the theory confirmed by their discovery, Wilson said, "is that the universe had a definite origin. During the first few minutes, when hydrogen and helium were being formed in the universe, there was a critical time when the ratio of certain elements was set. The present temperature and density of matter in the universe are an indication of what must have gone on billions of years ago."

Following in the tradition of Karl Jansky, the founder of radio astronomy, Wilson, Penzias, and their colleagues have recently introduced millimeter-wave technology to the study of radio astronomy spectra, thus increasing the useful spectrum. An important early consequence of this achievement was the discovery of dozens of chemical compounds in interstellar space.

Continuing Bell Labs pioneering efforts in microwave radio communications and radio astronomy, Penzias and Wilson and other members of the Radio Research Laboratory have been applying astronomical techniques to the measurement of earth-space signal propagation. Using a new millimeter-wave antenna at Crawford Hill, they are gathering more comprehensive data than ever before on the effects of weather on high-frequency signals.

Wilson and Penzias are the sixth and seventh Bell Labs scientists to receive the Nobel Prize in Physics. In 1937, Clinton C. Davisson won the award for discovery of the wave nature of matter. The Nobel Prize was awarded in 1956 to John Bardeen, Walter Brattain, and William Shockley for the transistor. Last year, Philip W. Anderson, consulting director in Physics Research, was awarded the Nobel Prize for his theoretical studies of magnetism and disordered systems.

Robert W. Wilson, left, and Arno A. Penzias, Nobel Prize Laureates