THE BELL SYSTEM

Technical Journal

Volume 52 November 1973 Number 9

THE BELL SYSTEM TECHNICAL JOURNAL

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

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

Volume 52 November 1973 Number 9

Operational Limitations of Charge Transfer Devices

By K. K. THORNBER

(Manuscript received November 15, 1972)

The incomplete transfer of charge and the existence of random noise lead to the primary operational limitations of charge transfer devices. Owing to the signal dependence of the residual charge, which accumulates as a result of the incomplete transfer, signal detection with static detection levels becomes seriously impaired before the onset of significant signal attenuation or noise degradation. A scheme using dynamic detection levels is found to greatly extend the operational range of CTD's and achieves the minimum possible error rate for detecting uncorrelated charge packet sizes. By contrast, simple coding procedures are found to be ineffective in overcoming signal degradation due to incomplete transfer. Shannon's expression for maximum information transmission capacity is transformed into an expression for maximum information storage capacity. It is found that significantly larger storage capacities are possible with CTD's than have been achieved.

PRELIMINARY REMARKS

Proposition: Devices Function at Their Limits of Operation

One has only to design or fabricate a device which "exceeds specs," and then, because of his success, receive a set of revised (and more demanding) specifications from the systems people to appreciate this

very basic principle. Or, if one fixed his goals too low, some clever fellow will come along with a new twist which makes full use of the device and revolutionizes the industry. On the other hand, devices function at their limits of operation *and no better*. This lies behind the widely recognized importance of ascertaining fundamental operational limitations in the early stages of device development.[1]

It is, therefore, the primary objective of this article to discuss the operational limitations of charge transfer devices (CTD's). At a time when the basic device structure to be developed is still uncertain, and with relatively little analysis of the charge transfer process, noise, error rates, storage capacities, etc. (and with even less experimental verification of these analyses), the results of such an article may seem very preliminary at best. In some respects this will be the case. However, sufficient progress has been made in understanding the operational features of CTD's, especially with respect to incomplete transfer and noise, that definite limitations can be placed on how large information storage capacity and how small (digital) error rates can be made.

It is my intention to outline CTD operational limitations in rather general (even perhaps philosophical) terms and to refer to appendixes, existing articles, work in preparation, etc., for mathematical details. The results are optimistic to the extent that they indicate how much better one can hope to do with CTD's than is commonly envisioned. However, at the same time, the results are pessimistic in that it is not obvious how one is to achieve this optimum use. Implementing our dynamic detection scheme, which results in the minimum possible error rate for digital signals, represents, nonetheless, a major step towards operating CTD's near full capacity.

I. INTRODUCTION

According to a most significant theorem due to Shannon,[2] the maximum information transmission capacity C_T of any channel is determined by the bandwidth and the signal/noise ratio of the channel. With a slight modification this theorem can be transformed into a theorem on the maximum information storage capacity C_s of any channel, in particular of a CTD. This C_s then places an upper limit on the number of bits of information which can be stored in an unregenerated section of a CTD. If we restrict consideration to codes in which the size of each charge packet is independent of preceding or subsequent packets, we can calculate a minimum error rate and specify an optimum detection scheme for digital signals. To obtain this result, some knowledge of the signal/noise ratio and the accumu-

lation of residual charge owing to the incomplete transfer of portions of every charge packet is necessary.

In this paper, in order to establish the minimum possible error rate and maximum possible storage capacities attainable with CTD's, it will be necessary to first review the way incomplete charge transfer determines the attenuation-versus-frequency characteristic and the total bandwidth of a CTD. The role of incomplete charge transfer in signal degradation will then be discussed. Once the dependence of the accumulated residual charge on the preceding signal is understood, it is possible to devise a dynamic detection scheme for simple digital signals and to compare this with the basic method of absolute-amplitude or fixed-threshold detection (static levels). With any form of detection, noise will introduce errors. Following a review of CTD noise, error rates are discussed. It is found that in the presence of noise this dynamic detection scheme attains the theoretical minimum error rate. Finally, maximum information storage capacities are calculated using the bandwidth and signal/noise ratio characteristic of a CTD. The methods used to obtain specific results are sufficiently general that they can be used, for example, to calculate error rates for nonoptimal detection schemes.

II. REVIEW OF INCOMPLETE CHARGE TRANSFER, SIGNAL ATTENUATION, AND BANDWIDTH

As with other electronic devices, signal attenuation and device bandwidth are extremely useful concepts with which to discuss the maximum storage capacity and minimum error rates characterizing the operational limitations of charge transfer devices. In a CTD, signal attenuation arises primarily from the incomplete transfer of charge[3–9] and only secondarily from charge losses, for example, through thermal (or other) leakage currents. Let us first consider incomplete transfer, then signal attenuation, and finally bandwidth.

2.1 *Incomplete Transfer*

It is clear that the incomplete transfer of charge from one elemental cell to the next will lead to signal degradation.[7,9] The character of this degradation can be ascertained as follows. The charge $Q_{i,t}$ in the ith elemental cell at time t will be the charge in the previous cell during the previous transfer cycle (of period τ_o) diminished by the charge $Q^b_{i-1,t-\tau_o}$ left behind in the $(i-1)$th cell but increased by the charge left behind in the ith cell during the immediately preceding transfer $Q^b_{i,t-\tau_o}$, and also less any charge lost (or gained) during the previous

storage, $Q^l_{i-1,t-\tau_o}$. Thus

$$Q_{i,t} = Q_{i-1,t-\tau_o} - Q^b_{i-1,t-\tau_o} + Q^b_{i,t-\tau_o} - Q^l_{i-1,t-\tau_o}. \tag{1}$$

In general, this equation is very difficult to solve since the Q^b_i and Q^l_i are nonlinear functions of Q_i. Nonlinearities, however, are common in electronic devices. Taking the usual approach, one makes a small-signal analysis in order to linearize the equations. Thus we write $Q_{i,t} = q_{i,o} + q_{i,t}$, where $q_{i,o}$ is the time-independent (dc) component (bias) and $q_{i,t}$ is the small (ac)-signal component. Similarly, Q^b and Q^l can be decomposed into dc and ac components. Substituting these into eq. (1) we can obtain two equations, one for the dc terms and one for the ac terms. The dc equation can give us the time-independent (dc) charge bias level at the output. While knowledge of this may be important in some applications, it does not lead to any significant operational limitations. Of greater importance is the solution of the equation for the ac terms.

The equation for the ac terms is from eq. (1)

$$q_{i+1,t+\tau_o} = q_{i,t} - q^b_{i,t} + q^b_{i+1,t} - q^l_{i,t}. \tag{2}$$

As part of the linearization, one takes $q^b_i = p\alpha q_i$ and $q^l_i = p\beta q_i$. (α and β can be calculated from the coefficients of a Taylor series expansion of Q^b and Q^l in terms of q: $p\alpha = dQ^b/dQ_o$ and $p\beta = dQ^l/dQ_o$, where p is the number of individual charge transfers within an elemental cell and Q_o is the charge to be transferred.) Substituting into (2) one obtains the basic equation for $q_{i,t}$:

$$q_{i+1,t+\tau_o} = q_{i,t} - p\alpha q_{i,t} + p\alpha q_{i+1,t} - p\beta q_{i,t} \tag{3}$$

which becomes upon taking the Fourier transform

$$q_{i+1}(\omega)e^{i\omega\tau_o} = q_i(\omega)(1 - p\alpha - p\beta) + p\alpha q_{i+1}(\omega). \tag{4}$$

Up to this point we have linearized eq. (1) and passed to the frequency domain, typical procedures in electrical engineering. We proceed to solve eq. (4) by first calculating

$$q_{i+1}(\omega)/q_i(\omega) = \frac{1 - p\alpha - p\beta}{1 - p\alpha e^{-i\omega\tau_o}} e^{-i\omega\tau_o}.$$

Then we note that since $q_{i+1}(\omega)/q_i(\omega)$ is independent of i, it follows that $q_N(\omega)/q_o(\omega) = [q_{i+1}(\omega)/q_i(\omega)]^N$, where N is the number of elementary cells in the shift register. Recognizing that $q_N(\omega)/q_o(\omega)$ is the transfer function of the shift register, $H(\omega)$, one finds that[7,9]

$$H(\omega) = e^{-i\omega N\tau_o}\left(\frac{1 - pa - p\beta}{1 - p\alpha e^{-i\omega\tau_o}}\right)^N. \tag{5}$$

As discussed in Section 2.2, $H(\omega)$ can in principle be used to determine $q_{N,t}$ for any given sequence of input charge packets $q_{t-N\tau_o}$, $q_{t-(N+1)\tau_o}$.

The first factor in $H(\omega)$ is just the phase delay in the signal present even in the limit of perfect transfer ($\alpha = \beta = 0$). The second factor contains a frequency-dependent attenuation and a further phase shift. To a good approximation we may write

$$H(\omega) = A(\omega)e^{-i\phi(\omega)}, \tag{6}$$

where the attenuation factor $A(\omega)$ is given by

$$A(\omega) = e^{-n\beta}e^{-n\alpha(1-\cos\omega\tau_o)} \tag{7}$$

and where the phase factor $\phi(\omega)$ is given by

$$\varphi(\omega) = N\omega\tau_o + n\alpha \sin\omega\tau_o. \tag{8}$$

($n = Np$, the total number of charge transfers in the shift register from input to output.) With knowledge of the device transfer function $H(\omega)$, we can discuss the attenuation $A(\omega)$ and then compute the device bandwidth.

2.2 *Attenuation*

The attenuation factor in eq. (7) can be interpreted as follows. The first factor results from charge loss. If a fraction β of charge is lost with each charge transfer, after n transfers the fraction remaining is just $(1 - \beta)^n \approx \exp(-n\beta)$ (if $n\beta^2 \ll 1$). Charge loss is clearly frequency independent. The second factor results from the incomplete transfer of charge. For $\omega\tau_o \approx 0$, very-low-frequency components, the size of adjacent charge packets is approximately the same. Thus the charge incompletely transferred at site i is nearly compensated by the charge incompletely transferred at $i + 1$. [$-p\alpha q_i + p\alpha q_{i+1} \approx 0$ in eq. (3).] Thus, apart from charge losses, $q_{i+1} \approx q_i$ and, hence, low-frequency components are expected to be attenuated very little. Equation (7) bears this out. By contrast, if $\omega\tau_o = 2\pi f/f_o \approx \pi$ ($f \approx f_o/2$ where $f_o = 1/\tau_o$ is the clock frequency), the attenuation is relatively large, $\exp(-2n\alpha)$. Again referring to eq. (3), $f \approx f_o/2$ implies that $q_{i,t}$ and $q_{i+1,t}$ are ~180 degrees out of phase and $q_{i,t} \approx -q_{i+1,t}$. Thus contributions to incomplete transfer add (rather than compensate as for low frequencies) and, again ignoring charge loss, eq. (3) predicts an attenuation of $(1 - 2\alpha)^n \approx \exp(-2n\alpha)$. Again, eq. (7) bears this out. For $\omega\tau_o = \pi/2$ ($f = f_o/4$) the attenuation is $\exp(-n\alpha)$, an intermediate case in which the phases of each successive packet differ by 90 degrees.

One further point concerning incomplete transfer should be emphasized. A charge packet which "loses" a fraction α of its charge in each of n transfers might be expected to be attenuated by a factor of $(1 - \alpha)^n \approx \exp(-n\alpha)$. However, eq. (7) for $A(\omega)$ shows how sensitive the actual degradation of a packet is to the presence and nature of the other charge packets composing the signal. Thus considering one "isolated" charge packet can be very misleading. In Appendixes A and B we discuss examples of attenuation in the time domain, and in

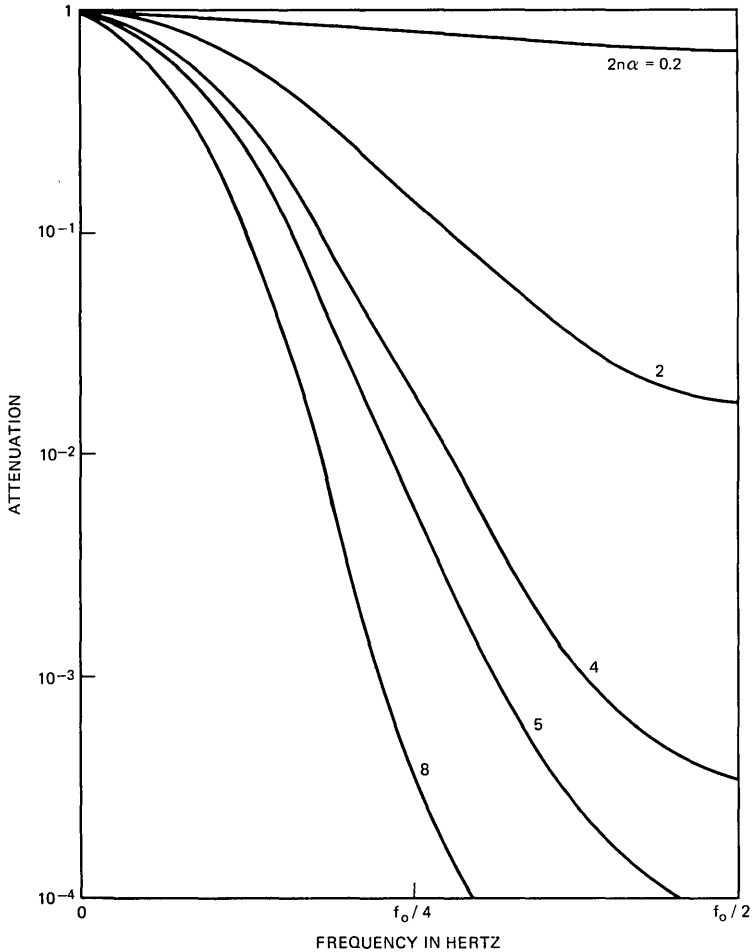


Fig. 1—Attenuation versus frequency for CTD's with $2n\alpha$ of 0.2, 2, 4, 5, 8.

Appendix C we consider $A(\omega)$ in more detail. $A(\omega)$ is plotted in Fig. 1 as a function of ω for several values of $n\alpha$. ($\beta = 0$.)

2.3 *Bandwidth*

In Fig. 1 we have plotted $A(\omega)$ for $0 \leqq f \leqq f_o/2$, although eq. (7) would seem to apply for $0 \leqq f$. The reason for this is basic. According to Shannon's sampling theorem,[2] $f_o/2$, one-half the clock frequency (one-half the sampling frequency), is the maximum frequency of the signal which can be transmitted. Thus given the clock frequency f_o, the maximum bandwidth a CTD can have is $f_o/2$.

Incomplete charge transfer clearly reduces the effective bandwidth of a CTD. This is evident from the attenuation plotted in Fig. 1. Normally one defines bandwidth by the size of the range of frequencies for which the attenuation A exceeds some fraction $\delta < 1$. A more convenient definition from the point of view of information transmission and storage capacity is that the bandwidth B be given by the following expression:

$$B = \int_0^{f_o/2} \frac{|A(f)|^2}{|A(0)|^2}\, df = \frac{f_o}{2}\, e^{-2n\alpha} I_o(2n\alpha)\,, \tag{9}$$

where I_o is a modified Bessel function.[10] In Fig. 2, B is plotted as a function of $n\alpha$. A slowly varying function, B decreases as $(f_o/2)\cdot(4\pi n\alpha)^{-\frac{1}{2}}$ for $n\alpha \gg 1$. Thus despite the rapid attenuation associated with $n\alpha \approx 10$ ($e^{-10} \approx 0.5 \times 10^{-4}$ for $f = f_o/4$ and $e^{-20} \approx 0.2 \times 10^{-8}$ for $f = f_o/2$), the bandwidth is still approximately $0.09 \times (f_o/2)$, 9 percent of its maximum value. The relative insensitivity of β to $n\alpha$

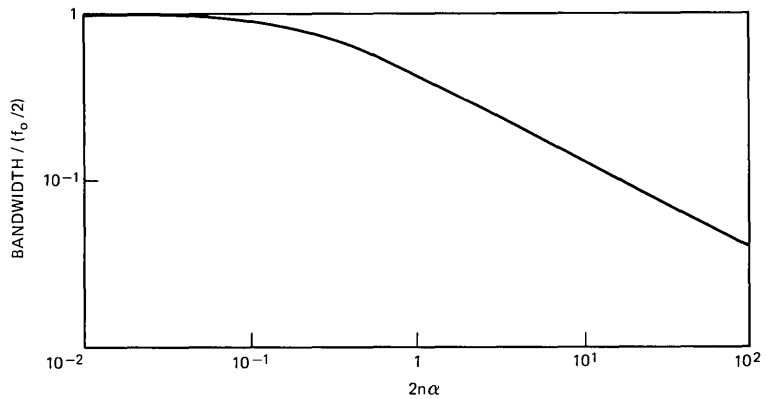


Fig. 2—Bandwidth versus $n\alpha$ for a CTD.

has an important effect on the maximum achievable CTD transmission and storage capacities as we shall see.

III. THE ROLE OF INCOMPLETE CHARGE TRANSFER IN SIGNAL DEGRADATION

Even though attenuation and bandwidth are frequency-domain (and hence analog) concepts, we shall see in Section VII that it is these quantities, along with the signal-power-to-noise-power ratio, which are needed in order to calculate the maximum information transmission and storage capacity for digital as well as analog signals in CTD's. Nonetheless, it is still helpful to discuss certain features of digital signals in the time domain in order to better appreciate certain operational limitations of CTD's. For the next few sections we shall ignore charge losses and consider only the much more important effects of incomplete charge transfer on the signal.

Suppose a charge packet of initial size Q_1 (representing a digital "one") follows some sequence of charge packets either of initial size Q_1 or of initial size Q_0 (representing a digital "zero"). If α is the coefficient of incomplete transfer, then after a single transfer the original charge packet has been reduced by a factor of $(1 - \alpha)$, and, after n similar transfers, by a factor of $(1 - \alpha)^n$. In addition, the original packet picks up some charge left behind by the preceding packets. This residual charge we shall refer to as Q_R, which is in general a function of the preceding signal. Thus the size of the charge packet representing the "one" at the output is given by

$$Q = (1 - \alpha)^n Q_1 + Q_R. \tag{10}$$

A brief analysis of eq. (10) reveals that the primary operational limitation imposed by incomplete transfer is the dependence of the residual charge Q_R on the preceding signal—the preceding sequence of zeroes and ones. Neither the attenuation of the size of the charge packet *per se* nor the accumulation of the incompletely transferred charge *per se* play a primary role. For example, suppose that we are using a very simple form of absolute-amplitude detection in which any charge packet of size $Q > \bar{Q} = (Q_1 + Q_0)/2$ is detected as a "one" and any packet of size $Q < \bar{Q}$ is detected as a "zero." Then a one preceded by a long string of zeroes will be detected as a zero if $n\alpha > 0.7$. Thus, under noiseless conditions we shall have a *non*zero error rate in cases where the signal attenuation is only one-half (see Appendix A). On the other hand, the accumulation of incompletely transferred charge can be reduced using zero-net-charge coding. As each bit of signal is

coded into the same amount of total charge (distributed between two packets), the total amount of incompletely transferred charge from each bit is the same. Nonetheless, no significant improvement is obtained. Indeed, the maximum $n\alpha$ for zero errors under noiseless conditions remains a strong function of the preceding sequence of charge packets (see Appendix B). This result is true for analog as well as digital signals, as is clear from the discussion in Appendix C. The dependence of the residual charge on the previous signal suggests that more attention should be given the detection of the signal rather than its coding. This we discuss in the next section.

IV. THE OPTIMUM DETECTION OF SIMPLE DIGITAL SIGNALS

Let us suppose that some arbitrary sequence of charge packets of size Q_1 for a "one" and Q_0 for a "zero" have preceded the charge packet which we now wish to detect. The residue or residual charge added to the charge packet of interest can be designated Q_R as before, where Q_R is a function (given in Appendix A) of the preceding signal. If the charge packet which we are detecting is in fact a one, then the size $Q(1)$ of the packet will be

$$Q(1) = (1 - \alpha)^n Q_1 + Q_R. \tag{11}$$

If, however, the charge packet is zero, then the packet's size $Q(0)$ will be given by

$$Q(0) = (1 - \alpha)^n Q_0 + Q_R. \tag{12}$$

One will clearly optimize the detection (even in the presence of noise) if one choses for the detection level Q_d of the mean of $Q(1)$ and $Q(0)$:

$$Q_d \equiv (1 - \alpha)^n \tfrac{1}{2}(Q_1 + Q_0) + Q_R. \tag{13}$$

If $Q > Q_d$, we say that we have detected a Q_1 packet, and if $Q < Q_d$ we say that we have detected a Q_0 packet. Because Q_d is a function of Q_R which in turn depends on the entire preceding signal, we shall refer to this as a dynamic detection procedure. In contrast to Q_d given in (13), the static detection procedure mentioned in Section III has $Q_s = \bar{Q}/(1 - \alpha)$, and $Q > Q_s$ implies Q_1 and $Q < Q_s$ implies Q_0.

It is shown in Appendix A that under noiseless conditions this scheme of dynamic coding is errorfree regardless of the size of $n\alpha$ or of the nature of the preceding sequence of zeroes and ones. This again illustrates the role of the dependence of Q_R on the preceding signal, which we noted at the end of Section III. It is shown in Appendix A that

$$Q(1) - Q_d = (1 - \alpha)^n \frac{Q_1 - Q_0}{2} = Q_d - Q(0) \tag{14}$$

as is clear from eqs. (11), (12), and (13) as well. The quantity $(Q_1 - Q_0)$ may be referred to as the dynamic range of the device. Thus relative to the dynamic detection level, Q_d, the signal, $Q(1)$ or $Q(0)$, is attenuated as $(1 - \alpha)^n$. [Note that $[Q(1) - Q_d]$ and $[Q_d - Q(0)]$ are independent of the residual charge Q_R.] In the presence of noise, errors will clearly be introduced if $(1 - \alpha)^n(Q_1 - Q_0)/2$ approaches the noise level. This also shows that, having eliminated the signal-dependent residual charge, attenuation now plays an important role in limiting device operation. As n increases, this signal attenuation coupled with the compounding of noise both reduce the signal-to-noise ratio and lead to a reduction in the information transmission and storage capacities of the device. However, now it will be for $n\alpha \approx 4$ rather than for $n\alpha \approx 0.7$ that attenuation becomes limiting.

To set the dynamic detection level Q_d, Q_R must be realizable. In the absence of noise, this is always possible in principle since Q_R is an explicit function of the known, preceding signal. In the presence of noise, Q_R determined by eq. (31) also yields, for most cases of interest, nearly optimal detection in spite of the possibility that some preceding packets may have been incorrectly detected.[11]

In Section V, we briefly review noise in CTD's and then in Section VI we shall see how this dynamic detection scheme minimizes the error rate in the presence of noise. This further stresses the importance of detection in optimizing the operation of "simply coded" CTD's.

V. REVIEW OF NOISE

Noise in charge transfer devices is a fascinating subject which, unfortunately, can be only highlighted in this section.[12–17] Owing to the dramatic time dependence of the current during a single charge transfer, the noise generated during a single transfer is quite nonstationary. Since nearly all theories of noise in solid-state devices assume that the noise is stationary,[18] it is necessary to redo much of the theory taking into account the nonstationary aspect. A time-domain analysis has been found to be most convenient, whereas standard treatments are carried out in the frequency domain.

In Fig. 3 the most common sources of noise in CTD's are categorized. At the input, one has full shot noise only if the electrons enter the source independently (e.g., if generated by the random arrival of phonons in an imaging device or if injected by an emission-limited diode). At the output the nonrandom coupling to the clock line is the worst source of distortion in some cases. A distinction[14] is made between noise generated from transfer processes, typically thermal and trap-

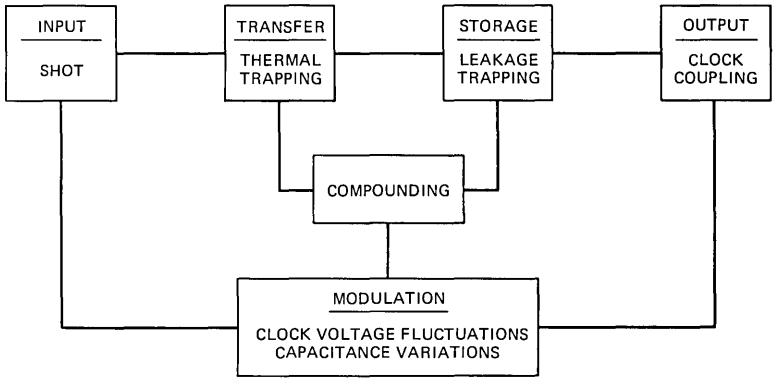


Fig. 3—Sources of noise for a CTD.

ping noise, and that from storage processes, typically leakage and again trapping noise. We shall return to this distinction shortly. Since a charge packet acquires some noise with each transfer-storage period, the noise component increases as the packet is transferred from input to output. This we refer to as compounding. Finally, the occurrence of clock voltage fluctuations in the presence of fabrication variations in the individual capacitances leads to a form of modulation noise. This type of noise is also compounded. For simplicity we have left out of Fig. 3 a number of less important noise sources.

Let us now return to the important distinction between storage process and transfer process noise.[14] It should be recalled that a CTD shift-register performs two functions simultaneously, the transfer of charge and the storage of charge. In Fig. 4 we indicate the basic distinction between the noise generated from these two processes. In the case of storage process (SP) noise, the charge fluctuation generated during each transfer cycle in each cell is essentially independent of that in any other cell. For transfer process (TP) noise this is not the case. Conservation of charge implies that if an excess of ΔQ is transferred from one storage region to the next, $-\Delta Q$ is left behind for the subsequent charge packet. This introduces a correlation in the noise in adjacent charge packets which leads to a suppression at low frequencies of the spectral density of TP noise. SP noise, by contrast, is uncorrelated and, therefore, the spectral density is flat (white). This difference between TP and SP noise is important for analog applications of CTD's and is discussed in more detail elsewhere.[14]

For digital applications we shall need the ratio S/N of the square of the signal charge to the mean-square noise charge at the detector.

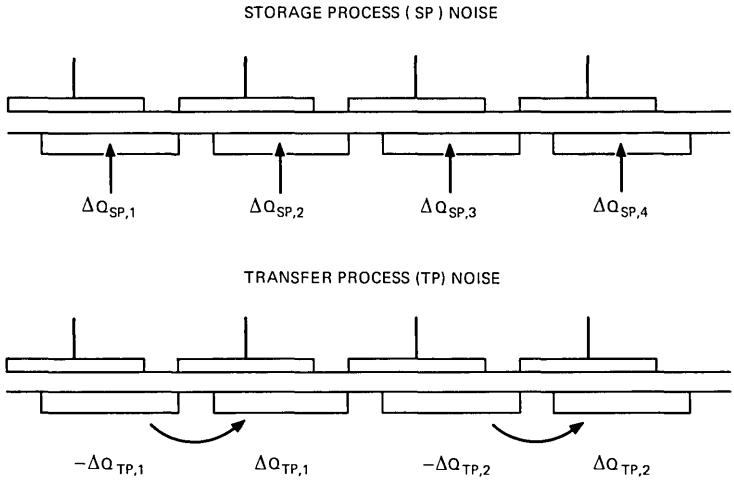


Fig. 4—The distinction between storage process and transfer process noise.

From the discussion of Section IV [see eq. (13)] the square of the effective signal charge at the detector is $(1 - \alpha)^{2n}[(Q_1 - Q_0)/2]^2$. The mean-square noise charge ΔQ^2 can be written

$$\Delta Q^2 = \Delta Q^2_{\text{Input}}(1 - \alpha)^{2n} + \Delta Q^2_{\text{SP}} H_{\text{SP}}(n) + 2\Delta Q^2_{\text{TP}} H_{\text{TP}}(n), \quad (15)$$

where $\Delta Q^2_{\text{Input}}$ is the input noise contribution, ΔQ^2_{SP} the storage process noise acquired by a single packet during a single clock period, ΔQ^2_{TP} the transfer process noise acquired by a single packet during a single charge transfer, $(1 - \alpha)^{2n}$ the attenuation from input to output, $H_{\text{SP}}(n)$ the compounding factor[12] for storage process noise, and $H_{\text{TP}}(n)$ the compounding factor[12] for transfer process noise. These compounding factors are approximately equal to n for $n\alpha \ll 1$; however, for $n\alpha \gtrsim 1$ they are suppressed[12] by incomplete transfer effects. For $n\alpha \gg 1$, $H_{\text{SP}}(\text{n}) \approx (n/\pi\alpha)^{\frac{1}{2}}$ and $H_{\text{TP}} \approx (2\alpha)^{-1}$, both of which are much less than n. Essentially the effect of incomplete transfer is to attenuate the accumulated noise as well as the signal. Owing to the correlation between the transfer-process noise components,[14] H_{TP} saturates. For shot noise, $\Delta Q^2_{\text{Input}} = qQ$, where Q is the (mean) total signal charge $(Q_1 - Q_0)$. For thermal noise, $\Delta Q^2_{\text{TP}} = \frac{2}{3}\text{k}TC$, where T is the temperature of the charge carriers and C is the storage capacitance. For our purposes here we shall ignore other noise contributions. Thus we find

$$S/N = (1 - \alpha)^{2n}[(Q_1 - Q_0)/2]^2/\Delta Q^2, \quad (16)$$

where ΔQ^2 is given by eq. (15). Equation (16) is plotted in Figs. 5 and 6 for $\alpha = 10^{-3}$ and $\alpha = 10^{-4}$ for $C = 1$, 0.1, 0.01, and 0.001 pF, for thermal noise only and for thermal and shot noise. We shall use eq. (16) in Section VI on error rates and in Section VII on storage capacity.

VI. MINIMUM DIGITAL ERROR RATES

No one would operate a CTD under conditions where errors in detection could occur even under noiseless conditions. However, in the presence of noise it is possible for a "one" to acquire sufficient net "negative" noise charge to be detected as a "zero" even under optimum conditions. It is the purpose of this section to calculate the probability of making a detection error, and to see to what degree the error rate (error probability times clock frequency) is minimized for the simple, two-level digital coding scheme by using dynamic detection.

Suppose that an arbitrary charge packet following an arbitrary sequence of charge packets would, under noiseless conditions, be of size Q_s at the output of the shift register. In the presence of noise the probability $P(Q)$ that the observed size of the packet is Q within dQ is given by

$$P(Q)dQ = \exp[-(Q - Q_s)^2/2\Delta Q^2]/(2\pi\Delta Q^2)^{\frac{1}{2}}dQ. \tag{17}$$

If we are using only zeroes and ones, then the probability P of detecting a certain "one" as a zero is given by

$$P_1 = \int_{-\infty}^{Q_d} P(Q)dQ\,, \tag{18}$$

where Q_d is the detection level (see Fig. 7). In eq. (18), P_1 depends upon $Q_s = Q(1)$ which in turn is a function of the sequence of signal charge packets preceding the one [see eqs. (11) and (12)]. To determine the average error probability, P_1 must now be averaged over all possible sequences of signals, in general a very difficult task.

Let us write eq. (18) in a slightly different form by changing variables.

$$P_1 = \int_{-\infty}^{Q_d - Q(1)} \frac{\exp(-Q^2/2\Delta Q^2)}{(2\pi\Delta Q^2)^{\frac{1}{2}}}\,dQ$$

or

$$P_1 = \int_{-\infty}^{[Q_d - Q(1)]/(\Delta Q^2)^{\frac{1}{2}}} e^{-x^2/2}dx/(2\pi)^{\frac{1}{2}}$$

or

$$P_1 = f_1\{[Q_d - Q(1)]/(\Delta Q^2)^{\frac{1}{2}}\}, \tag{19}$$

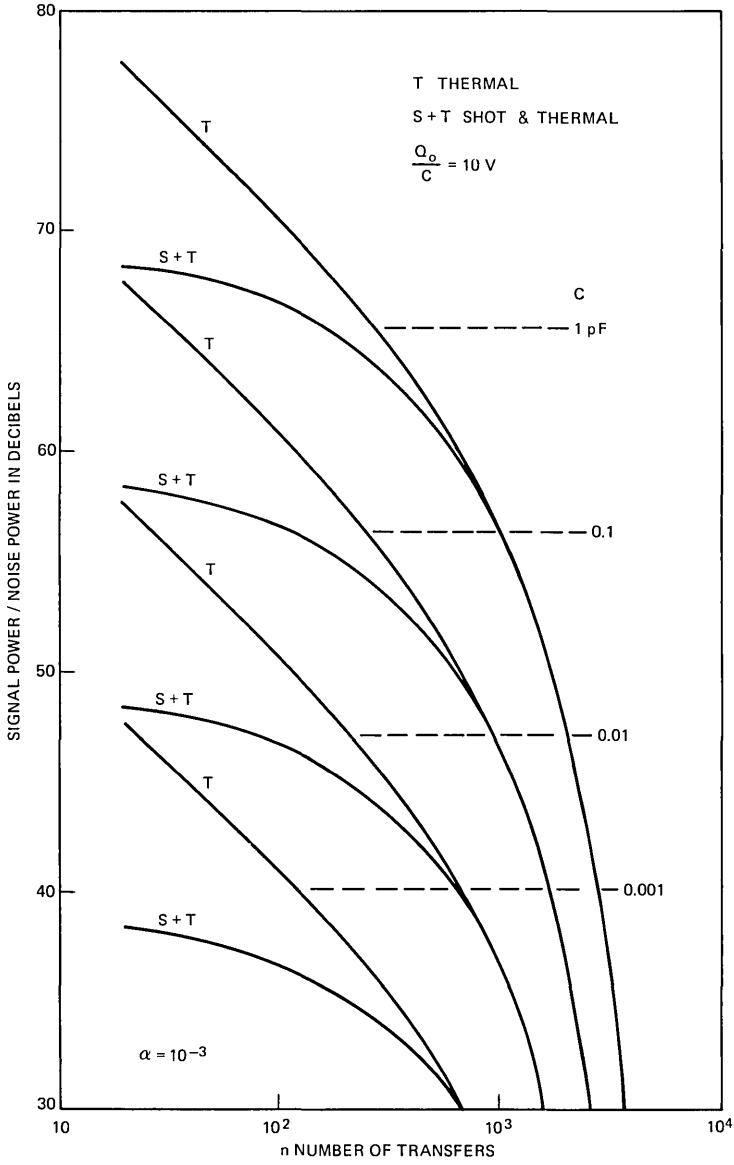


Fig. 5—Signal-to-noise ratio for CTD ($\alpha = 10^{-3}$) with storage capacitance C of 1 pF, 0.1 pF, 0.01 pF, 0.001 pF.

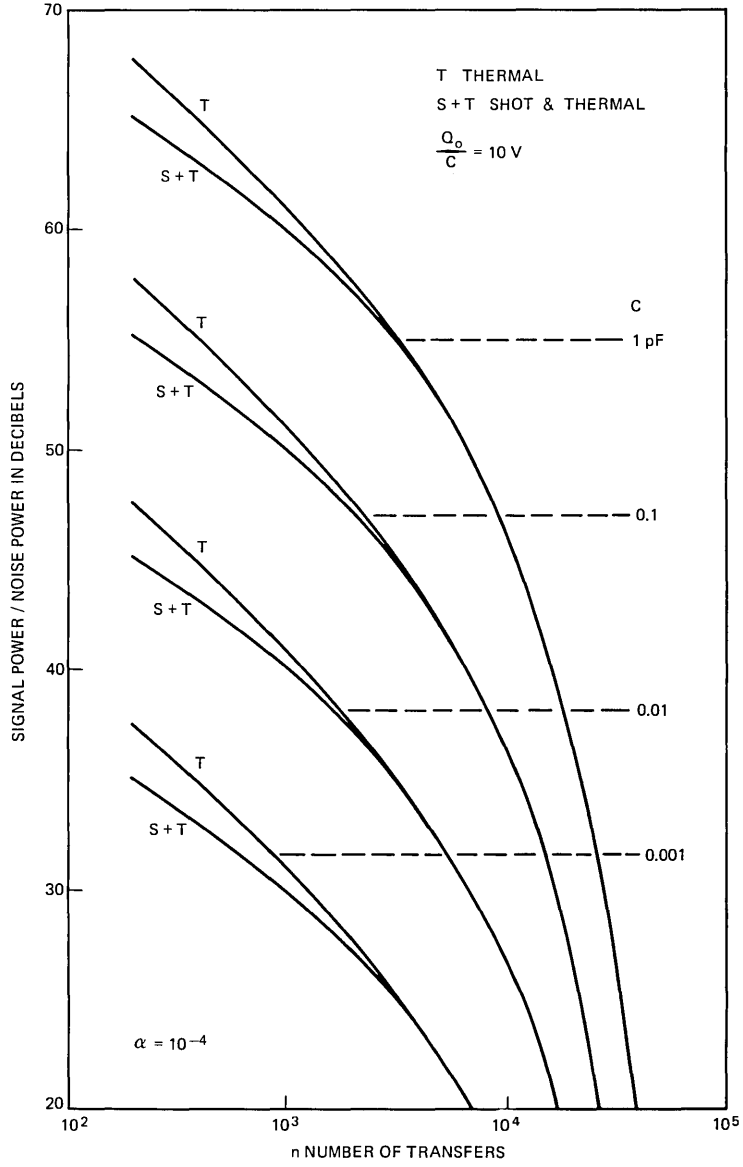


Fig. 6—Signal-to-noise ratio for CTD ($\alpha = 10^{-4}$) with storage capacitance C of 1 pF, 0.1 pF, 0.01 pF, 0.001 pF.

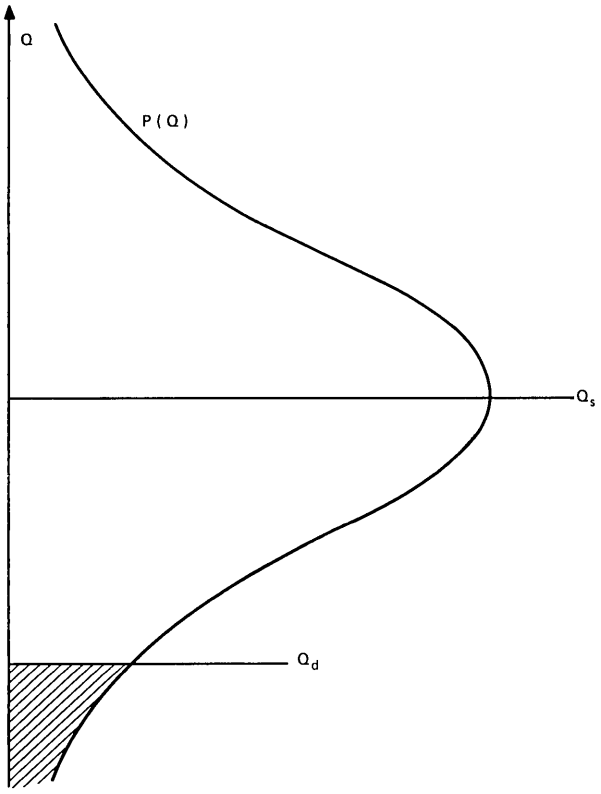


Fig 7.—A charge packet of size Q_s under noiseless conditions has a probability distribution of $P(Q)$ in the presence of noise. The probability that $Q < Q_d$ (that an error is made in detection) is the area "under" $P(Q)$ for $Q < Q_d$.

where, of course,

$$f_1(y) \equiv \int_{-\infty}^{y} e^{-x^2/2} dx/(2\pi)^{\frac{1}{2}} . \tag{20}$$

We desire $\langle P_1 \rangle = \langle f_1 \rangle$, where the brackets "$\langle\ \rangle$" denote averaging $Q(1)$ (and possibly Q_d) over all possible sequences of ones and zeroes $[Q(1) = Q(1-\alpha)^n Q_1 + Q_R$ from eq. (11)]. If $Q_s > Q_d$ for all possible Q_R (as it must if errors are to be avoided under noiseless conditions), then we show in Appendix D that $\langle P \rangle \geqq f\{\langle [Q_d - Q(1)]/(\Delta Q^2)^{\frac{1}{2}} \rangle\}$. In other words, the function f evaluated at the average of its argument is a lower bound to the average of the function, the average error probability. This permits putting a *lower* bound on the error rate.

If $[Q_d - Q(1)]$ is independent of Q_R, as is the case for the dynamic detection scheme discussed in Section IV, then

$$\langle P_1 \rangle = f_1\{[Q_d - Q(1)]/(\Delta Q^2)^{\frac{1}{2}}\} \tag{21}$$

and the average error probability is at the lower bound. For the dynamic detection scheme of Section IV, $[Q_d - Q(1)] = -(1-\alpha)^n \times (Q_1 - Q_0)/2$, which implies that

$$\langle P_1 \rangle = \int_{-\infty}^{(S/N)^{\frac{1}{2}}} e^{-x^2/2} dx/(2\pi)^{\frac{1}{2}}, \tag{22}$$

where (S/N) is given in eq. (18). If static detection had been used $[Q_d = \bar{Q}/(1-\alpha)]$, then the average error probability would always exceed the $\langle P_1 \rangle$ given in (22).

It remains to prove that the optimum (dynamic) detection scheme given in Section IV gives the lowest possible error probability for simple digital coding. The proof is as follows. We found above that for detecting a one the average error probability was at least

$$\langle P_1 \rangle_{\text{lb}} = f_1\{[\langle Q_d \rangle - \langle Q(1) \rangle]/(\Delta Q^2)^{\frac{1}{2}}\}. \tag{23}$$

(The "lb" stands for *l*ower *b*ound.) Noting the definition of f_1 [eq. (20)], we note that we can make $\langle P_1 \rangle_{\text{lb}}$ smaller by reducing $\langle Q_d \rangle$, $\langle Q(1) \rangle$ being already determined. However, we must also consider the error probability in detecting a "zero." Proceeding as for a "one," we obtain for P_0 the error probability for detecting a certain "zero,"

$$P_0 = f_0\{[Q_d - Q(0)]/(\Delta Q^2)^{\frac{1}{2}}\}, \tag{24}$$

where now

$$f_0(y) \equiv \int_y^{\infty} e^{-x^2/2} dx/(2\pi)^{\frac{1}{2}}. \tag{25}$$

Also

$$\langle P_0 \rangle_{\text{lb}} = f_0\{[\langle Q_d \rangle - \langle Q(0) \rangle]/(\Delta Q^2)^{\frac{1}{2}}\}. \tag{26}$$

From the definition of f_0, we note that we can made $\langle P \rangle_{\text{lb}}$ smaller by increasing $\langle Q_d \rangle$, $\langle Q(0) \rangle$ being already determined. Assuming that an equal number of "zeroes" and "ones" are used in the simple digital coding, then (by symmetry) choosing $\langle Q_d \rangle$ so that $\langle P_1 \rangle_{\text{lb}} = \langle P_0 \rangle_{\text{lb}}$, we shall achieve the minimum lower bounds. $\langle P_1 \rangle = \langle P_0 \rangle_{\text{lb}}$ for $\langle Q_d \rangle = [\langle Q(1) \rangle + \langle Q(0) \rangle]/2$. But for our dynamic detection scheme

$$\langle Q_d \rangle = \bar{Q}(1-\alpha)^{-1} = [\langle Q(1) \rangle + \langle Q(0) \rangle]/2 \tag{27}$$

and for our dynamic detection scheme $\langle P_1 \rangle = \langle P_1 \rangle_{\text{lb}}$ and $\langle P_0 \rangle = \langle P_0 \rangle_{\text{lb}}$. Therefore, since $\langle Q_d \rangle$ for the dynamic detection scheme produces the

lowest possible lower bounds for error probabilities, and since the error probabilities are in fact equal to these lower bounds, no other detection scheme can detect with lower error probability. (It is possible that another scheme can do just as well, however, since it is only $\langle Q_d \rangle$ and not Q_d itself which is the determining factor.)

It is clear that this theorem places an operational limitation (a minimum error rate in detection) and CTD's using simple digital two-level coding. The theorem can be extended[11] to the dynamic detection of multilevel digital codes.

VII. MAXIMUM STORAGE CAPACITY

One use of the CTD is as a memory or storage element. In other applications the CTD can be used to shift or to transfer information from one location to another. To properly access the operations of CTD's in these applications one must calculate the maximum information transmission capacity and the maximum information storage capacity of the CTD. As a result of the work of Shannon, our labors are greatly diminished.

Shannon[2] proved a most profound theorem. Let B be the bandwidth of a transmission channel, and let S/N be the signal-power-to-noise-power ratio. Then the maximum transmission capacity of the channel C_T in bits per second is given by

$$C_T = B \log_2(1 + S/N). \tag{28}$$

This result can be understood for the CTD in the $S/N \gg 1$ range as follows. The number of levels into which a digital signal can be divided and still be detected with reasonably small error is $(S/N)^{\frac{1}{2}}$. $\text{Log}_2(S/N)^{\frac{1}{2}}$ is the maximum amount of information in bits detected with each charge packet. f_o is the rate at which charge packets are detected. Thus $f_o \log_2(S/N)^{\frac{1}{2}} = \frac{1}{2} f_o \log_2(S/N) \approx B \log_2(1 + S/N)$ is the number of bits of information transmitted per second. (In Section 2.3 we noted that for $n\alpha \ll 1$, $B \approx f_o/2$.) Shannon was, of course, much more interested in the $S/N \ll 1$ range. For this case his theorem implies that no matter how noisy the transmission channel may be, it is always possible to pass information along it. We shall not make use of Shannon's result in this latter range.

A more interesting quantity from the standpoint of the CTD is the maximum information storage capacity. This can be calculated from Shannon's Theorem[2] as follows. If C_T is the number of bits per second transmitted, then if one waits a time T_o equal to the time it takes the information to be transferred from the input to the output of the

linear medium, the maximum storage capacity in bits C_s must be given by

$$C_s = T_o C_T = T_o B \log_2(1 + S/N). \tag{29}$$

For a transmission line, T_o is given by the length of the line divided by the propagation velocity. For a CTD, $T_o = N_o/f_o$ where $N_o = n/p$ and p is the number of charge transfers per clock period $T_o = 1/f_o$. Thus for a CTD we find for the maximum information storage capacity C_s in bits:

$$C_s = N_o(B/f_o) \log_2(1 + S/N). \tag{30}$$

[Strictly speaking, the maximum information storage capacity will actually be less than or equal to the C_s given in eq. (30). This is because as S/N decreases, the length of the code word increases.[2] However, for a CTD with N_o storage units, the maximum length of a code word is restricted to N_o. Thus for small S/N, the prediction of eq. (30)

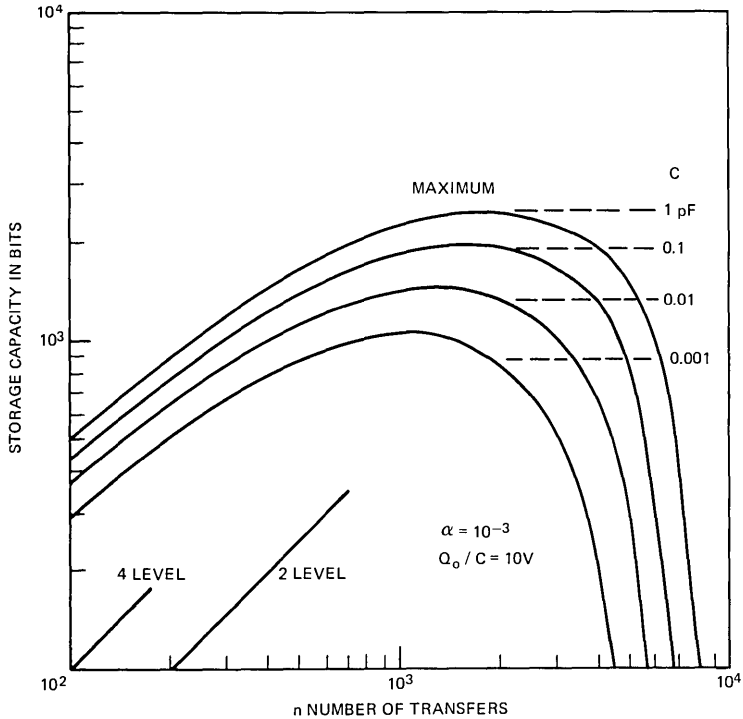


Fig. 8—Storage capacity C_s versus n, the number of charge transfers, for $\alpha = 10^{-3}$, $Q/C = 10$ volts, $C = 1$, 0.1, 0.01, 0.001 pF. Shown for comparison is C_s for 2- and 4-level digital codes.

may be in fact unrealizable. As our primary concern is for $S/N \gg 1$, the upper limit of C_s should be sufficiently accurate.]

Knowing N_o, (B/f_o), and (S/N) as functions of n, the number of charge transfers, we can calculate C_s versus n to determine the maximum C_s possible under various circumstances and for what $n\,C_s$ is maximum. This has been done in Fig. 8 for $\alpha = 10^{-3}$ and in Fig. 9 for $\alpha = 10^{-4}$. In both figures $Q/C = 10$ volts, $Q = Q_1 = 2Q_0$, and storage capacitance $C = 1$, 0.1, 0.01, 0.001 pF. Also shown is C_s for two-level and four-level codes. Here n is limited by an n(maximum) for each code at the number of transfers beyond which signal degradation due to incomplete transfer would lead to errors in absolute-amplitude detection in the absence of noise. We note (*i*) that the maximum C_s occurs for n about a factor of three larger than for n(maximum) from the examples of simple coding and detection, and (*ii*) that the maximum value of C_s is about a factor of four to five

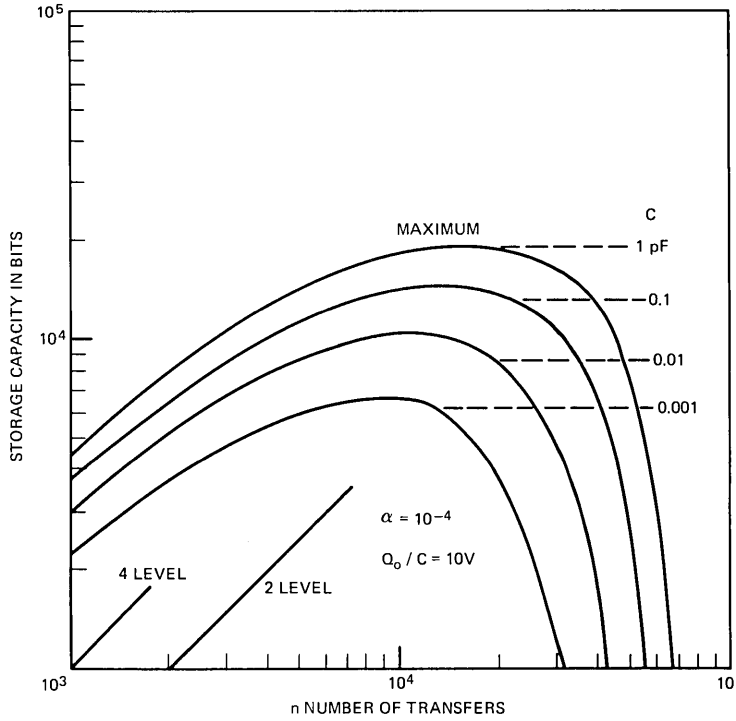


Fig. 9—Storage capacity C_s versus n, the number of charge transfers, for $\alpha = 10^{-4}$, $Q/C = 10$ volts, $C = 1$, 0.1, 0.01, 0.001 pF. Shown for comparison is C_s for 2- and 4-level digital codes.

larger than C_s at n(maximum) for these common digital codes. [It should also be noted that C_s at n(maximum) for two-level digital coding exceeds C_s at n(maximum) for multilevel coding. This is discussed in more detail elsewhere.[19]] It is encouraging to note that there exists such a margin between what is theoretically possible and what can be simply accomplished. From Section VI, however, it appears that it will be rather difficult to achieve optimum performance if this be desired (or essential).

VIII. CONCLUSIONS AND RECOMMENDATIONS

It is not surprising to find that incomplete charge transfer and random noise (especially shot and thermal) limit the bandwidth, storage capacity, and error rate of CTD's. What is surprising, however, is that the residual charge level Q_R resulting from the portions of charge (preceding the packet of interest) incompletely transferred is so strongly signal dependent that signal detection with static detection levels becomes seriously impaired prior to the onset of significant signal attenuation or noise degradation. Coding to offset the signal dependence of Q_R is found to be ineffective for the simple examples considered. On the other hand, by employing our dynamic detection scheme, which adjusts the detection levels to null out the signal dependence of the incompletely transferred charge, the operational range is significantly extended, limited only by the physically unavoidable effects of attenuation and noise. It is also shown that no detection scheme can be devised with a lower error rate than this dynamic detection scheme.

It might be concluded on the basis of the above result that more attention should be focused on detection rather than coding as a means of offsetting the worst effects of incomplete charge transfer. Noting the results shown in Figs. 8 and 9, however, it is apparent that substantial increases in storage capacity are possible with more sophisticated coding-decoding schemes.

IX. ACKNOWLEDGMENTS

It is a pleasure to thank J. R. Brews, R. J. Strain, and G. E. Smith for assistance in preparing the manuscript.

APPENDIX A

In this appendix the degradation of digital signals is discussed in general in some detail. In Appendix B these results are applied to

certain specific cases using several simple coding procedures. In particular, the signal-dependent residual charge Q_R is discussed. These mathematical details should be of some assistance in understanding several equations presented in Sections III, IV, and VI of the text. The analysis will be in the time domain. In Appendix C a frequency domain analysis is given.

We shall denote by Q_N the size of the Nth charge packet at the input which precedes the packet of interest by N clock cycles. As in the text we shall denote by Q_R the size of the accumulated residual charge originating from the incompletely transferred portions of the preceding packets, Q_N. Mathematically Q_R is given by[7,19]

$$Q_R = (1 - \alpha)^n \sum_{N=1}^{\infty} \binom{n + N}{N} \alpha^N Q_N , \tag{31}$$

where n is the total number of transfers from input to output and α is the coefficient of incomplete transfer for each transfer. [Equation (31), as well as eq. (3), are somewhat approximate. To obtain a "pure" binomial factor in (31), or equivalently to be able to write a single-transfer equation like eq. (3), one must assume that the actual transfer of charge can be approximated by simplified single transfers either on a per-cell basis as in (31) or on a storage-region basis as in eq. (3). The error involved in this approximation will be of the order of α or $n\alpha^2$, whichever is larger.]

The physical significance of (31) is the following. The portion of Q_N which will show up in Q_R are electrons incompletely transferred N times, each time introducing a factor α. The binomial factor gives the number of distinct alternative sequences of "transfer" or "no transfer" which can lead to a portion of Q_N contributing to Q_R.

Suppose now, as in the first example in Section III, we have a packet of size Q_1 preceded by an infinite string of packets of size Q_0. Then for Q_R one has

$$\begin{aligned} Q_R(000\cdots) &= (1 - \alpha)^n \sum_{N=1}^{\infty} \binom{n + N}{N} \alpha^N Q_0 \\ &= -(1 - \alpha)^n[1 - (1 - \alpha)^{-(n+1)}]Q_0 . \end{aligned} \tag{32}$$

Similarly,

$$Q_R(111\cdots) = -(1 - \alpha)^n[1 - (1 - \alpha)^{-(n+1)}]Q_1. \tag{33}$$

For a Q_1 following the string of Q_0's, the size of the charge packet Q[eq. (10)] at the output will equal $\bar{Q} = (Q_1 + Q_0)/2$ if n is such that

$$(1 - \alpha)^n Q_1 - (1 - \alpha)^n[1 - (1 - \alpha)^{-(n+1)}]Q_0 = (Q_1 + Q_0)/2$$

or if $(1-\alpha)^n = \frac{1}{2}$ [to within a factor of $(1-\alpha) \approx 1$]. Should $(1-\alpha)^n > \frac{1}{2}$, $Q < \bar{Q}$ and the one would be detected as a zero.

In Appendix B we shall discuss more complicated coding schemes to see whether Q_B (31) can be reduced or at least made less sensitive to the signal preceding the charge packet of interest. For the present let us continue to derive some of the other results stated in the text.

To cast eq. (13) into a simpler form we proceed as follows:

$$\begin{aligned} Q_d &= (1-\alpha)^n \frac{Q_1 + Q_0}{2} + Q_R \\ &= (1-\alpha)^n \bar{Q} + (1-\alpha)^n \sum_{N=1}^{\infty} \binom{n+N}{N} \alpha^N (Q_N - \bar{Q} + \bar{Q}) \\ &= \bar{Q}\left[(1-\alpha)^n - (1-\alpha)^n \left(1 - \frac{1}{(1-\alpha)^{n+1}}\right)\right] + Q'_R \\ &= \bar{Q}(1-\alpha)^{-1} + Q'_R, \end{aligned} \tag{13'}$$

where

$$Q'_R \equiv (1-\alpha)^n \sum_{N=1}^{\infty} \binom{n+N}{N} \alpha^N (Q_N - \bar{Q}) . \tag{34}$$

The static detection level, $Q_s = \bar{Q}/(1-\alpha)$, actually differs by a factor of $(1-\alpha)^{-1}$ from the $\bar{Q}$ used in Section III and in the discussion following eq. (33). The difference, while insignificant, arises from whether one takes Q_d to be the average of Q_1 and Q_0, the sizes of the charge packets at the input, or whether one takes Q_d to be the size at the output of an average charge packet [of size $\bar{Q} = (Q_1 + Q_0)/2$] following a string of similar packets. Thus for such a case

$$\begin{aligned} Q_s = Q &= \bar{Q}(1-\alpha)^n + (1-\alpha)^n \sum_{N=1} \binom{n+N}{N} \alpha^N \bar{Q} \\ &= \bar{Q}(1-\alpha)^n - (1-\alpha)^n [1 - (1-\alpha)^{-(n+1)}] \bar{Q} \end{aligned}$$

or

$$Q_s = \bar{Q}/(1-\alpha). \tag{35}$$

To show that by using the dynamic level Q_d given by eq. (13) one can have zero detection errors in the absence of noise we proceed as follows. Using eq. (11) one has at once that

$$Q(1) - Q_d = (1-a)^n (Q_1 - \bar{Q}) \tag{36}$$

independent of Q_R. As $Q_1 > Q_0$, $Q_1 > \bar{Q}$ and, therefore, $Q(1) - Q_d > 0$. Similarly using eq. (12) one finds $Q_d - Q(0) > 0$. Thus, in the absence of noise, $Q(1)$ and $Q(0)$ are always separated by Q_d, and hence no error need be made in distinguishing them.

APPENDIX B

In this appendix we use the results of Appendix A to investigate what improvement if any is possible in CTD operation by using several simple coding procedures. We shall assume noiseless absolute-amplitude detection using a static detection level at the average output charge level. Equation (31) for Q_R can be used to calculate the result of other coding procedures.

In Table I, I have enumerated four simple means of representing or coding a digital zero (0) and a digital one (1) using charge packets. The first is just to represent a 0 by a Q_0 packet and a 1 by a Q_1 packet. As calculated in Appendix A [see following (33)], a Q_1 following a long string of Q_0's [example "(a)"] will be detected as a Q_0 if $n\alpha > 0.7$. This is the "$n\alpha$ Limit" entry in the table. Finally, the size of the Q_1 packet is attenuated as $\exp(-n\alpha)$ as stated. For this coding a second example, "(b)," is given—a $\cdots Q_1Q_0Q_1Q_0 \cdots$ sequence. In this case Q_R is always sufficiently large for a Q_1 and sufficiently small for a Q_0 that under noiseless conditions $Q(1) > \bar{Q}$ and $\bar{Q} > Q(0)$ for any $n\alpha$. However, as noted in Section 2.2, such a signal is attenuated as $\exp(-n\alpha)$ attenuation.

One might hope that by preventing Q_R from becoming much differ-

TABLE I—FOUR SIMPLE MEANS OF REPRESENTING DIGITAL ZEROES AND ONES USING CHARGE PACKETS

Example	Representation	$n\alpha$ Limit*	Attenuation
	0 1 Q_0 Q_1		
(a) 1000 $\cdots$		0.7	$e^{-n\alpha}$
(b) 1010 $\cdots$		∞	$e^{-2n\alpha}$
	Q_0Q_0 Q_1Q_1		
(a) 1000 $\cdots$		(1) 0.7; (2) 1.67	$e^{-n\alpha}$
(b) 1010 $\cdots$		(1) 0.79; (2) 2.4	$e^{-n\alpha}$
	Q_1Q_0 Q_0Q_1		
(a) 1000 $\cdots$		(1) ∞; (2) 0.8	$e^{-2n\alpha}$, $e^{-n\alpha}$
(b) 1010 $\cdots$		(1) 2.36; (2) 0.785	$e^{-n\alpha}$
	$\bar{Q}\,Q_0$ $\bar{Q}\,Q_1$		
(a) 1000 $\cdots$		(1) 0.4; (2) 0.5	$e^{-n\alpha}$
(b) 1010 $\cdots$		(1) 1.6; (2) 3.1	$e^{-n\alpha}$

* The notation (1) refers to the first of the two packets forming a bit, and (2) refers to the second.

ent than $\bar{Q}$, one could increase the $n\alpha$ limit. In Table I three possibilities are given. The first consists merely of coding 0 into two adjacent Q_0 packets and 1 into two adjacent Q_1's. If one detects the second of the two Q_1 packets in sequence (a), the $n\alpha$ limit is increased to 1.67. For sequence (b) detecting the second Q_1 now has an $n\alpha$ limit of 2.4 while the signal is attenuated as $\exp(-n\alpha)$. How much of an improvement this offers, however, is questionable. To store the same amount of information n must be doubled reducing the 1.67 to an effective 0.835. To maintain the same information rate, f_o, the clock frequency must be doubled. This will increase α: if α is doubled, then the 1.67 limit, already reduced to 0.835, will be reduced further to about 0.42. Compared with the 0.7 limit of the simplest code, this is rather unfavorable. One compensation is that having two packets to detect rather than just one can be used to reduce the error rate induced by noise. However, one can do better, as the following example illustrates.

The third example in Table I is the zero-net-charge code. Here a 0 is coded as a Q_0 packet followed (in time) by a Q_1 packet. (In the register shifting from charge left to right this is represented as Q_1Q_0.) A 1 is coded as Q_0Q_1. The advantage of this procedure is that each pair, whether coding a 0 or a 1, contains the same amount of charge, $2\bar{Q}$. This prevents a buildup of charge in Q_R. The most demanding test is sequence (b) in which the $n\alpha$ limit is 2.36. This is a significant improvement over the 1.67 limit in the previous example. However, if one takes into account that to contain the same amount of information n must be doubled (as now each bit requires two charge packets) and that the clock frequency must be doubled to maintain the same data rate (which will increase α), one realizes that really very little has been achieved by increasing the upper limit on $n\alpha$ from $n\alpha < 0.7$ to $n\alpha < 2.4$ $(2.4/4 = 0.6)$. Other straightforward modifications of the basic 0, 1 code, of course, suffer from the same fault. Thus to achieve any improvement it is necessary that one still must be able to take advantage of the possibility of detecting *both* charge packets to do better than the simplest code. The reason for the failure of the zero-net-charge code in terms of frequency-domain concepts is given in Appendix C.

One final example is to follow the Q_0 or the Q_1 with an intermediate packet of size $\bar{Q}$. As seen in Table I, sequence (a) puts an $n\alpha$-limit of 0.5, which is inferior to the other codes. This attempt to reduce $|Q_B - \bar{Q}|$ by following Q_0 or Q_1 with a $\bar{Q}$ packet to "average" out the incompletely transferred charge is thus seen to be ineffective.

APPENDIX C

It is quite informative to briefly discuss in the frequency domain the effects of various digital coding schemes on the character of the signal.[7,9]

In Section 2.1 we noted that incomplete charge transfer leads to a frequency-dependent attenuation $A(\omega)$ given by

$$A(\omega) = \exp[-n\alpha(1 - \cos \omega\tau_o)] \tag{37}$$

for $\beta = 0$ in eq. (7). $A(\omega)$ is plotted in Fig. 1 for various values of $n\alpha$. As discussed in Section 2.2, low-frequency components ($f \ll f_o/2$) suffer very little attenuation, whereas components with frequency near half the clock frequency ($f \approx f_o/2$) are attenuated by $\exp(-2n\alpha)$, a large attenuation for $n\alpha \gtrsim 3$.

One can offset this high-frequency attenuation by the following scheme. If one takes every other charge packet and replaces it by a Q_1 if it originally was a Q_0, and by a Q_0 if it originally was a Q_1, then relative to $\tilde{Q}$ one essentially multiplies each packet in turn by $+1$, -1, $+1$, -1, $+1$, -1, $\cdots$. This has the effect of converting the spectrum of the signal from $F(f)$ to $F(f_o/2 - f)$: the $f = 0$ component is attenuated as $A(f_o/2)$ and the $f = f_o/2$ component as $A(0)$. To better preserve the entire signal, one can sum the outputs of a register with attenuation $A(f)$ and a register with attenuation $A(f_o/2 - f)$. The ratio of maximum attenuation to minimum attenuation is thus improved from $\exp(-2n\alpha)$ to $2\exp(-n\alpha)/[1 + \exp(-2n\alpha)]$. However, distortion near $f = f_o/4$ is still significant for $n\alpha > 2$.

To see the effect of the zero-net-charge coding scheme on the signal, consider this example. If the clock frequency is f_o, the maximum frequency the CTD can carry is $f_o/2$. However, if two charge packets are devoted to each 0 or 1 as in the second through fourth examples in Table I, then the bandwidth is reduced to $f_o/4$. If the second example is chosen, then the band extends from $f = 0$ to $f = f_o/4$; if the third example (zero-net-charge coding) is chosen, then the band extends from $f = f_o/4$ to $f_o/2$, the lower-frequency components of the signal being carried at the higher frequencies and vice versa. If amplified by $\exp(+n\alpha)$, the ultimate effect of incomplete transfer on a signal coded using zero-net-charge coding is seen to be essentially the same as that on a signal coded using the second example. What is most striking, however, is that by reducing the clock frequency by a factor of two and using simple coding, one reduces $n\alpha$ by a factor of four, greatly reducing the attenuation.

By examining the effect of other coding schemes on the spectrum of the signal, and by taking into account the frequency-dependent attenuation accompanying charge transfer in CTD's it is possible to ascertain whether an improvement in (noiseless) detection will be in fact real or only apparent.

APPENDIX D

In noise, detection, and communication theory one often encounters an integral of the form

$$I(A) = \int_{-\infty}^{-A} e^{-x^2/2} dx/(2\pi)^{\frac{1}{2}}, \tag{38}$$

where $A > 0$. This expression, while extensively tabulated numerically, is difficult to work with analytically. In this appendix we shall (*i*) bound $I(A)$ between two simple analytic functions of A which differ by only a factor of 2, and (*ii*) prove that $\langle I(A)\rangle \geqq I(\langle A\rangle)$ for $A \geqq 0$.

(*i*) Bounds on $I(A)$:

In Fig. 10 we illustrate the motivation for our approximations. $I(A)$ is the area under the Gaussian for $x = -\infty$ to $x = -A$. If we draw a line tangent to the Gaussian at $x = -A$ and extend the line from $x = -A$ to the x-axis as shown, the area of the triangle formed by this tangent, the x-axis, and the vertical line $x = -A$ is clearly less than $I(A)$. Similarly, if an exponential curve $[B \exp(+Cx)]$ also tangent to the Gaussian at $x = -A$ and decaying to the left is drawn, then the area between this curve and the x-axis for $x \leqq -A$ is clearly greater than $I(A)$. Thus, if we calculate these two areas, we will have an upper and lower bound on $I(A)$. (These curves will clearly not cross the Gaussian if $A \geqq 1$, the inflection point of the Gaussian.)

To calculate the areas we proceed as follows. The slope of $\exp(-x^2/2)$ at $x = -A$ is $A \exp(-A^2/2)$, and of course its value at $x = -A$ is $\exp(-A^2/2)$. Thus the equation of the tangent is

$$y(x) = \exp(-A^2/2) + A \exp(-A^2/2)(x + A) \tag{39}$$

(which is zero for $x = -A - 1/A$, $f(-A - 1/A) = 0$) and of the exponential is

$$y(x) = \exp[-A^2/2 + A(x + A)]. \tag{40}$$

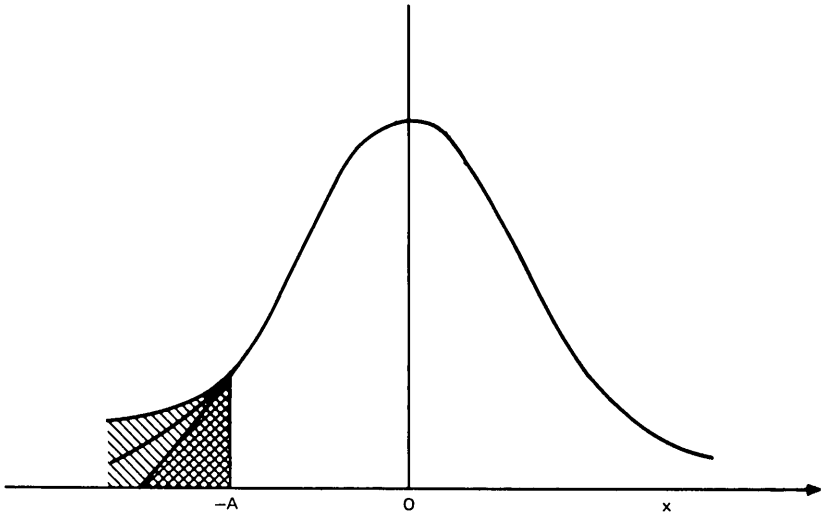


Fig. 10—Approximating the area under a portion of Gaussian curve by bounding the area between that of a right triangle (whose hypotenuse is tangent to the Gaussian at $x = -A$) and by the area under an exponential (also tangent to the Gaussian at $x = -A$).

Thus computing the appropriate areas we find that if $A \geqq 1$ then

$$D/2 < (2\pi)^{\frac{1}{2}} I(A) < D, \tag{41}$$

where

$$D = \exp(-A^2/2)/A. \tag{42}$$

Such bounds are very useful in calculating error rates, where one is seldom interested in accuracy better than a factor of two, and where upper and lower bounds are often very useful.

(*ii*) $\langle I(A) \rangle \geqq I(\langle A \rangle)$, $A \geqq 0$:

To evaluate $\langle I(A) \rangle$ is clearly very difficult even under the simplest of probability distributions of A, whereas $I(\langle A \rangle)$ is generally very easy to compute if $\langle A \rangle$ is known. $I(\langle A \rangle)$ can then be used as a lower bound for the more interesting $\langle I(A) \rangle$. We shall now prove the above inequality. (This result is reasonably well known.[20] The proof is given here for completeness.)

$I(A)$ is a function of A. According to the mean value theorem we may write

$$I(A) = I(\langle A \rangle) + \left.\frac{dI}{dA}\right|_{\langle A \rangle} (A - \langle A \rangle) + \frac{1}{2} \left.\frac{d^2I}{dA^2}\right|_{A'} (A - \langle A \rangle)^2, \tag{43}$$

where $A'(A)$ lies between A and $\langle A \rangle$ and depends on A. Thus we may write

$$\langle I(A) \rangle = I(\langle A \rangle) + \frac{1}{2} \left\langle \left. \frac{d^2 I}{dA^2} \right|_{A'(A)} (A - \langle A \rangle)^2 \right\rangle. \tag{44}$$

Now then

$$\frac{d^2 I}{dA^2} = A \exp(-A^2/2)/(2\pi)^{\frac{1}{2}} \tag{45}$$

which is zero or larger for $A \geqq 0$. Thus if we are averaging A over a probability distribution $P(A)$ for which $P(A < 0) = 0$, then $A'(A) \geqq 0$, and, consequently, the second term on the right-hand side will be zero or greater. Hence it follows that

$$\langle I(A) \rangle \geqq I(\langle A \rangle). \tag{46}$$

In Section VI this inequality is used to put a lower bound on the error rate for detecting digital signals.

REFERENCES

1. Boyle, W. S., private communication.
2. Shannon, Claude E., "Communication in the Presence of Noise," Proc. IRE, *37*, 1949, pp. 10–21.
3. Sangster, F. L. J., and Teer, K., "Bucket-Brigade Electronics—New Possibilities for Delay, Time-Axis Conversion, and Scanning," IEEE J. Solid-State Circuits, *SC-4*, June 1, 1969, pp. 131–136.
4. Sangster, F. L. J., presented at the 1970 Int. Solid-State Circuits Conf., Philadelphia, Pa., February 18–20, 1970.
5. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970) pp. 587–593.
6. Berglund, C. N., and Boll, H. J., presented at the 1970 Int. Electron Devices Meeting, Washington, D. C., October 28–30, 1970, "Performance Limitations of the IGFET Bucket-Brigade Shift Register," IEEE Trans. Electron Devices, *ED-19*, 1972, pp. 852–860.
7. Joyce, W. B., and Bertram, W. J., "Linearized Dispersion Relation and Green's Function for Discrete-Charge-Transfer Devices with Incomplete Transfer," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1741–1759.
8. Thornber, K. K., "Incomplete Charge Transfer in IGFET Bucket-Brigade Shift Registers," IEEE Trans. Electron Devices, *ED-18*, October 1971, pp. 941–950.
9. Berglund, C. N., "Analog Performance Limitations of Charge-Transfer Dynamic Shift Registers," IEEE J. Solid-State Circuits, *SC-6*, December 1971, pp. 391–394.
10. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards, 1965, pp. 374–8.
11. Thornber, K. K., "Error Rates of Digital Signals in Charge Transfer Devices," to be published in B.S.T.J., *52*, No. 10 (December 1973).
12. Boonstra, L., and Sangster, F. L. J., "Progress on Bucket-Brigade Charge-Transfer Devices," 1972 IEEE Solid-State Conf., Digest of Technical Papers, *15*, 1972, pp. 140–141.
13. Thornber, K. K., "Noise Suppression in Charge Transfer Devices," Proc. IEEE, *60*, September 1972, pp. 1113–1114.
14. Thornber, K. K., and Tompsett, M. F., "Spectral Density of Noise Generated in Charge Transfer Devices," IEEE Trans. Electron. Devices, *ED-20*, 1973, p. 456.

15. Tompsett, M. F., "Quantitative Effects of Interface States on the Performance of Charge-Coupled Devices," IEEE Trans. Electron Devices, *ED-20*, 1973, pp. 45–55.
16. Thornber, K. K., in preparation.
17. Carnes, J. E., and Kosonocky, W. F., "Noise Sources in Charge-Coupled Devices," RCA Rev., *33*, 1972, pp. 327–343.
18. van der Ziel, A., "Noise in Solid State Devices and Lasers," Proc. IEEE, *58*, 1970, pp. 1178–1206.
19. Thornber, K. K., in preparation.
20. Feynman, R. P., "Slow Electrons in a Polar Crystal," Phys. Rev., *97*, 1955, pp. 660–665.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

A Geometric Theory of Intersymbol Interference

Part I: Zero-Forcing and Decision-Feedback Equalization

By D. G. MESSERSCHMITT

(Manuscript received May 14, 1973)

A linear-space geometric theory of intersymbol interference is introduced in this paper. An equivalence between the structure of intersymbol interference and a wide-sense stationary discrete random process is demonstrated and exploited to demonstrate the equivalence of zero-forcing (decision-feedback) equalization to minimum mean-square error linear interpolation (prediction) of a random process. This equivalence is used to quickly derive the properties of these equalizers and give them additional geometric interpretation. Results from prediction theory are used to develop practical computational methods of determining the tap-gains of the infinite equalizers for both rational and nonrational channel power spectra. Finally, the theory of reproducing kernel Hilbert spaces is used to develop a theory of equalization for nonstationary channels with nonstationary noise.

I. INTRODUCTION

The analysis of digital communication systems from a geometrical viewpoint—the viewing of waveforms as points in a signal space and the identification of cross-correlation with the formation of an inner product—is by now well established. To a large extent, this approach has been popularized by the book of Wozencraft and Jacobs.[1] However, when it comes to analyzing systems with intersymbol interference, frequency-domain techniques have almost exclusively been relied upon. The purpose of this paper is to consider pulse-amplitude modulation (PAM) systems with intersymbol interference from a geometric standpoint, and more specifically to develop a geometric theory of equalization.

Consideration of the geometric structure of intersymbol interference leads immediately to the observation of a striking correspondence to the theory of minimum mean-square error (MMSE) linear estimation of a wide-sense stationary discrete-parameter random process. The fact that the latter subject is almost exclusively treated by geometric methods[2,3] is further impetus for this approach to equalization.

The theories of linear zero-forcing equalization and decision-feedback equalization are well established. The properties of linear equalization

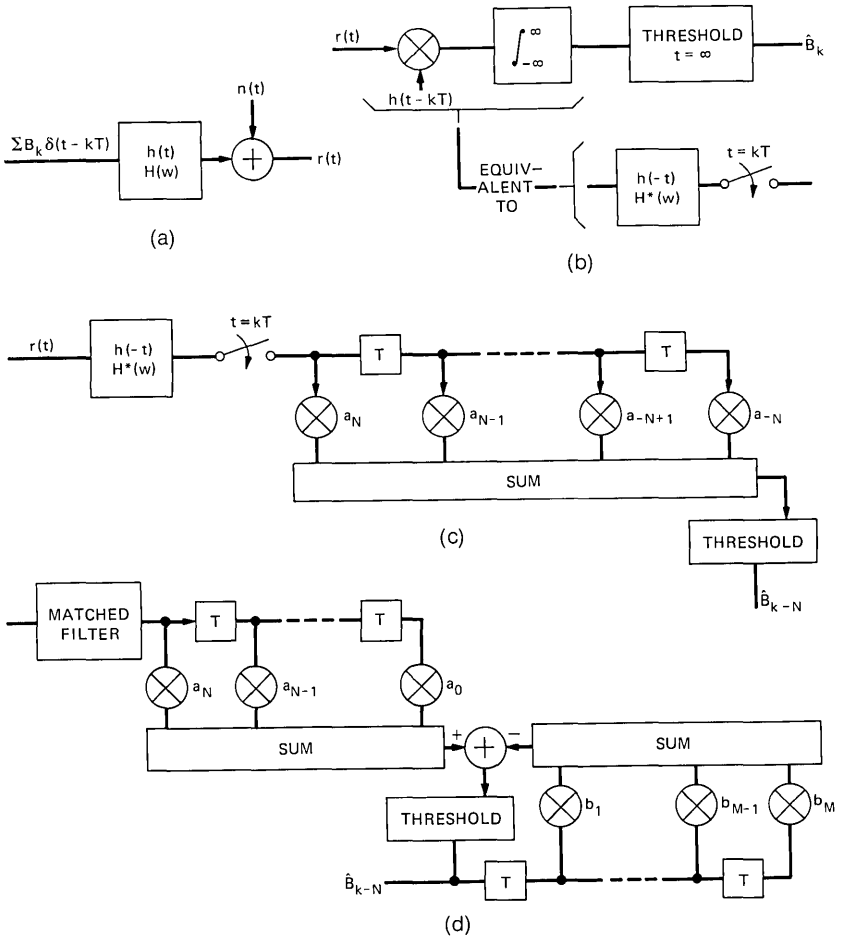


Fig. 1—(a) Communication system model. (b) Matched-filter receiver. (c) Zero-forcing equalizer. (d) Decision-feedback equalizer.

are summarized by Lucky, et al.,[4] while the present state of knowledge of decision-feedback equalization is summarized by Monsen[5] and Price.[6] The primary analysis tools which have been used are the calculus of variations in the case of linear equalization and Toeplitz forms in the case of decision-feedback equalization.

In this paper, the geometric approach enables us to treat the two types of equalization simultaneously using the same mathematical framework, in which the relationship between them becomes very clear and many of their known properties are given an additional geometric interpretation. Many of the results follow directly from the theory of MMSE estimation. In addition to the unification and reinterpretation of previously known results, the geometric approach leads to extensions of the theory in several directions. Among these are the derivation of an orthogonal expansion in Section 2.4 which is useful in many problems involving intersymbol interference, the development of practical iterative techniques for determining equalizer tap-gains (the infinite case) in Section 3.4, the extension of the theory of equalization to nonstationary noise and a time-varying channel in Section IV, and numerous results on the minimum distance problem associated with the performance analysis of the Viterbi algorithm maximum likelihood detector in a companion paper.[7]

This paper together with a companion one[7] expand upon an earlier talk.[8] Readers desiring a limited and short treatment of this subject may wish to refer there. The geometrical approach to intersymbol interference was also employed to a limited extent in the author's thesis.[9]

1.1 *Problem Statement*

We will consider the detection of a sequence of digital data digits, B_k, each assuming one of a finite and predetermined number of levels, from the reception

$$r(t) = \sum_{k=N_1}^{N_2} B_k h(t - kT) + n(t) \tag{1}$$

as determined from the communication system model of Fig. 1a. It will be assumed initially that $n(t)$ is white Gaussian noise (this assumption will be relaxed in Section IV).† A simple matched-filter receiver for the reception of $r(t)$ is shown in Fig. 1b. In the first of two equivalent formulations of this receiver, the reception is cross-correlated

† The assumption of Gaussian noise is not necessary for the majority of results to follow, and in particular those which involve only second-order statistics of the noise.

with $h(t - kT)$ and the decision on B_k made by applying a series of thresholds to the result; in the second formulation the cross-correlator is realized as a filter with impulse response $h(-t)$ (commonly called a matched filter) whose output is sampled at $t = kT$. The matched-filter receiver is optimum when there is no intersymbol interference, but in the presence of intersymbol interference the matched filter will respond to more than a single data digit and the performance of the receiver will be degraded.

When there is intersymbol interference, a common approach is to build a linear filter, called a zero-forcing equalizer (ZFE), which responds to only a single time-translate of $h(t)$ (this can only be approximated in practice). The most common form of this equalizer, shown in Fig. 1c, is a matched filter followed by transversal filter (MFTF). As $N \to \infty$ the tap-gains of the transversal filter can be chosen such that the threshold input is a function of only a single data digit. It is important to note for future reference that the MFTF can also be modeled in the manner of Fig. 1b as a cross-correlation of $r(t)$ with a linear sum of time translates of $h(t)$,

$$\sum_{m=-N}^{N} a_m h(t - mT).$$

The decision-feedback equalizer (DFE) embodies a slightly different philosophy in which the DFE forward filter is allowed to respond to past (but not future) translates at $h(t)$; the residual interference from past data digits is then subtracted out prior to the decision threshold using past decisions. A realization of the DFE using again the MFTF approach is shown in Fig. 1d. The tap coefficients are now chosen to null the response to future data digits; this can be accomplished as $N \to \infty$.

The shortcoming of both the ZFE and DFE is that their linear filters remove intersymbol interference without regard to the effect on the noise; the result is that in eliminating the intersymbol interference (or a portion thereof) they necessarily enhance the noise.[†] It seems clear intuitively that since the DFE eliminates interference from only future data digits, it has more degrees of freedom than the ZFE and should therefore be capable of less noise enhancement. A proof that this is always the case has been given by Price;[6] his method was to determine an explicit formula for the DFE S/N ratio using

[†] In addition, the DFE is susceptible to decision errors. The effect of errors will not receive consideration here.

Toeplitz form theory and compare it with the known S/N ratio of the ZFE.[4] Additional interpretation of this result will be given in Section 3.1.

A review of some requisite material on linear spaces and MMSE linear estimation is given in Sections 2.1 and 2.2. Readers familiar with this material are nevertheless urged to scan these sections for notation to be employed in the remainder of the paper. The ZFE and DFE are reformulated in Section 2.3. In Section 2.4 the relationship between intersymbol interference and MMSE estimation is discussed, and a useful orthogonal expansion arising out of this relationship is derived in Section 2.5.

Section III develops a geometric theory of the ZFE and DFE. Conditions necessary and sufficient for the existence of these equalizers are given in Section 3.1, their performance is discussed in Section 3.2, a useful property of the DFE with regard to its output noise sequence is interpreted in Section 3.3, methods of calculating the tap-gains are derived in Section 3.4, and the relationship between finite and infinite transversal filter equalizers receives consideration in Section 3.5.

Sections II and III are concerned with additive white noise exclusively. Section IV extends the theory to colored Gaussian noise, nonstationary Gaussian noise, and a time-varying channel using the theory of reproducing kernel Hilbert spaces (RKHS).

II. AN EQUIVALENCE TO DISCRETE RANDOM PROCESSES

The structure of the intersymbol interference in (1) will now be shown to have an equivalence to a wide-sense stationary random process. The starting point will be a quick review of linear spaces and of linear mean-square error (MMSE) estimation of a random process.

2.1 *Hilbert Space Notation*[10]

An inner product space $\mathfrak{L}$ consists of a linear space together with a defined inner product $\langle x, y\rangle$ between two elements x and y. All spaces in this paper are Hilbert spaces, which consist of an inner product space satisfying an additional closure property (specifically, the limits of Cauchy sequences must be in the space). The inner product induces a norm, or "length" of a vector,

$$\|x\| \triangleq \langle x,x\rangle \tag{2}$$

and the notion of the distance between two vectors, $\|x - y\|$. The geometrical interpretation of these quantities is illustrated in Fig. 2.

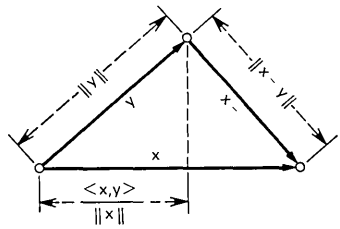


Fig. 2—Interpretation of inner product, norm, and distance.

A subspace of $\mathcal{L}$ is any set of vectors which itself constitutes a linear space. If x_k, $k \in I$ is a countable or finite sequence of vectors, then we denote by $M(x_k, k \in I)$ the closure of the subspace consisting of all finite linear combinations of elements of the set $\{x_k, k \in I\}$ and call this the subspace spanned by the x_k's. It is convenient to think of elements of $M(x_k, k \in I)$ as convergent (possibly) infinite sums of the form

$$\sum_{k \in I} a_k x_k$$

even though in some obscure cases not all elements can be expressed in this way.

In many minimization problems it is desired to find the element of some closed subspace M which is closest to a vector y; the resulting element is called the projection of y on M, is denoted by $P(y; M)$, and satisfies the orthogonality property

$$\langle y - P(y; M), x \rangle = 0 \tag{3}$$

for all $x \in M$.† The geometric interpretation of (3) is shown in Fig. 3 for a one-dimensional subspace spanned by x; for this case the projection must be a scalar times x and the validity of (3) is apparent.

2.2 *Review of Linear Mean-Square Interpolation and Prediction*[2,3]

We will now quickly review the theory of linear mean-square estimation of a random variable.

The set of random variables with zero mean and finite variance is a linear space, since the sum of any two such random variables itself has these properties. This set is also a Hilbert space with inner product

$$\langle X,Y \rangle = E(XY), \tag{4}$$

† When, as in (3), a vector is orthogonal to every vector in M, it is said to be orthogonal to M.

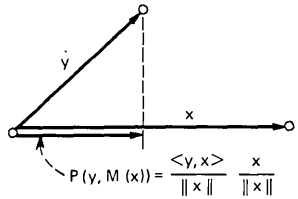


Fig. 3—Projection on subspace spanned by x.

where $E(\cdot)$ denotes expected value. It is standard to suppress the sample space dependence of a random variable as has been done in (4) because the geometric properties (inner product and norm) are determined by the value of the random variable on the whole sample space; that is, by its statistics in their entirety.

Consider now the following interpolation problem: Suppose that a sequence of zero-mean random variables X_k, $-\infty < k < \infty$, with finite variances are given and it is desired to estimate X_0 based on the observation of X_k, $k \neq 0$. If the estimate is further stipulated to be linear, it is the same as requiring that it be an element of $M(X_k, k \neq 0)$. Suppose that the estimate $\hat{X}_0$ is to be chosen in such a way that the mean-square error between X_0 and the estimate is minimized:

$$\min_{\hat{X}_0 \in M(X_k, k \neq 0)} E(X_0 - \hat{X}_0)^2. \tag{5}$$

From (4) and the previous section, the MMSE linear interpolator is

$$\hat{X}_0 = P[X_0, M(X_k, k \neq 0)], \tag{6}$$

the projection of X_0 on $M(X_k, k \neq 0)$.

A second estimation problem which will be of interest is the prediction of X_0 based only on X_k, $k > 0$ (an anticausal prediction). The MMSE linear predictor is the projection of X_0 on the subspace spanned by X_k, $k = 1, 2, \cdots$, denoted by $P[X_0, M(X_k, k > 0)]$.

2.3 *Zero-Forcing and Decision-Feedback Equalization*

We are now prepared to restate the problem of determining the ZFE and DFE filters in a linear space context. It will be assumed that the basic pulse $h(t)$ in (1) has finite energy (i.e., is square integrable),

$$\int_{-\infty}^{\infty} h^2(t)dt < \infty. \tag{7}$$

The set of waveforms which satisfies (7) is a linear space, which we

denote by L_2. L_2 is also a Hilbert space with inner product

$$\langle x, y \rangle = \int_{-\infty}^{\infty} x(t)y(t)dt \tag{8}$$

for any two L_2 waveforms $x(t)$ and $y(t)$. For the same reason that the sample space dependence of a random variable was suppressed in (4), the time dependence of the waveforms $x(t)$ and $y(t)$ has been suppressed on the left side of (8): it is the entire time waveform which determines the geometric properties.

The class of filters† which will be considered will be limited to those which can be modeled as an inner product (or cross-correlation) of the reception $r(t)$ with some L_2 waveform. A ZFE is a filter corresponding to a waveform $g_k(t)$ which does not respond to any translate of $h(t)$ except $h(t - kT)$,

$$\int_{-\infty}^{\infty} h(t - mT)g_k(t)dt = 0\,, \qquad m \neq k, \tag{9}$$

but does respond to $h(t - kT)$,

$$\int_{-\infty}^{\infty} h(t - kT)g_k(t)dt \neq 0\,, \tag{10}$$

in order that there be a signal on which to base the decision. It is evident that if $g_0(t)$ satisfies (9) and (10) for $k = 0$, then they are also satisfied by $g_k(t) = g_0(t - kT)$ for $k \neq 0$. Written in inner product notation, (9) and (10) become

$$\langle h_k, g_0 \rangle = 0, \qquad k \neq 0, \tag{11}$$

$$\langle h_0, g_0 \rangle \neq 0, \tag{12}$$

where we have written h_k for $h(t - kT)$. The analogous condition for a DFE forward filter is

$$\langle h_k, g_0 \rangle = 0, \qquad k > 0, \tag{13}$$

$$\langle h_0, g_0 \rangle \neq 0. \tag{14}$$

The forms of the ZFE and DFE in this symbolic notation are shown in Figs. 4a and b. The output of the linear filter is a function of B_k (a single data digit) for a ZFE and B_{k-m}, $m > 0$ (all past data digits) for a DFE. The tap-gains of the feedback transversal filter storing past decisions for the DFE are equal to the responses of g_0 to previous pulses, $\langle g_0, h_{-m} \rangle$, $m > 1$.

† In the case of the DFE, we refer only to the forward filter.

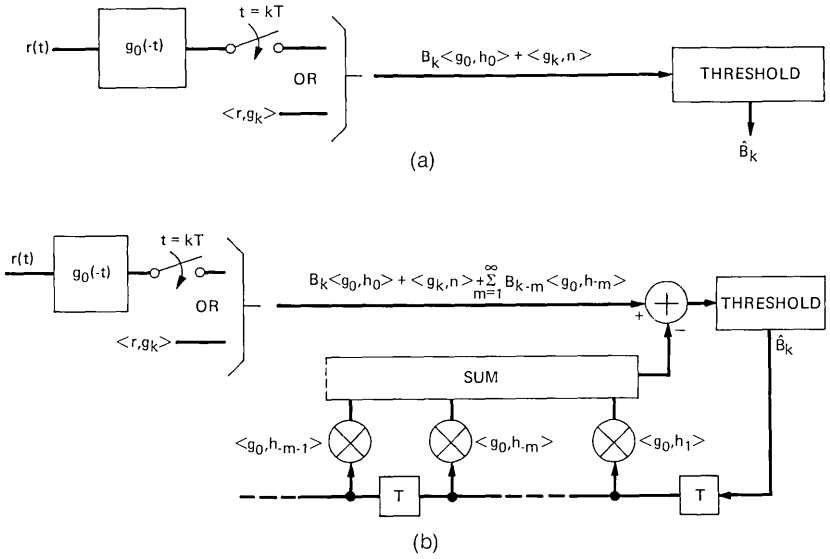


Fig. 4—Symbolic representations of the two equalizers: (a) zero-forcing equalizer; (b) decision-feedback equalizer.

2.4 *A Congruence Relationship*

Two Hilbert spaces which display an identical geometrical structure are said to be congruent[11] or unitarily equivalent.[10] Specifically, in order for two Hilbert spaces to be congruent, there must exist between them a one-to-one and onto linear mapping which preserves norms and inner products. Although the elements of two such spaces may be quite different entities, when considered as elements of their respective Hilbert spaces they have the same geometrical structure.

Define the autocorrelation function of the pulse sequence,

$$R_k = \langle h_m, h_{m+k} \rangle. \tag{15}$$

It follows from the inequality

$$0 \leqq \left|\left| \sum_{m=0}^{N} \alpha_m h_{k_m} \right|\right|^2 = \sum_{m=0}^{N} \sum_{n=0}^{N} \alpha_m \alpha_n R_{k_m - k_n}$$

that $\{R_k\}$ is a nonnegative definite function. Therefore, there exists a second-order discrete random process $\{X_k\}$ which has autocorrelation R_k,

$$\begin{aligned} \langle X_m, X_{m+k} \rangle &= E(X_m, X_{m+k}) \\ &= R_k. \end{aligned} \tag{16}$$

For the random process defined in (16), $M(h_k, k \in I)$ and $M(X_k, k \in I)$ are congruent through the obvious mapping

$$\phi\left[\sum_{m=1}^{N} \alpha_m h_{k_m}\right] = \sum_{m=1}^{N} \alpha_m X_{k_m} \tag{17}$$

which is a unitary linear transformation. To verify this, observe that the mapping is linear, preserves norms,

$$\begin{aligned} \left|\left|\phi\left(\sum_{m=1}^{N} \alpha_m h_{k_m}\right)\right|\right|^2 &= \left|\left|\sum_{m=1}^{N} \alpha_m X_{k_m}\right|\right|^2 \\ &= \sum_{m=1}^{N}\sum_{n=1}^{N} \alpha_m \alpha_n R_{k_m - k_n} \\ &= \left|\left|\sum_{m=1}^{N} \alpha_m h_{k_m}\right|\right|^2, \end{aligned} \tag{18}$$

and preserves inner products by an equally simple derivation.

The mapping of (17) is only defined for finite sums. When I is an infinite set, ϕ can be extended to all of $M(h_k, k \in I)$ by taking limits in the mean. For any $f \in M(h_k, k \in I)$ there exists a sequence $\{f_k\}$, each consisting of a finite sum of the form of (17), such that $f_k \to f$. Since $\phi(f_k)$ is a Cauchy sequence from (18) ,we define $\phi(f)$ as the limit of $\phi(f_k)$, which is in $M(X_k, k \in I)$ by completeness.

There is an additional congruence which is useful. From the definition of R_k in (15), we see that

$$\begin{aligned} R_k &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(\omega)|^2 e^{j\omega kT} d\omega \\ &= \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T} R(\omega) e^{j\omega kT}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} R(\omega) &\triangleq \sum_{m=-\infty}^{\infty} \left|H\left(\omega + m\frac{2\pi}{T}\right)\right|^2 \\ &= T\sum_{n=-\infty}^{\infty} R_n e^{jn\omega T}, \end{aligned} \tag{20}$$

where $R(\omega)$ is an equivalent power spectrum of the channel. From (16), $R(\omega)/T$ is the power spectrum of the random process $\{X_k\}$. Let $L_2(-\pi/T, \pi/T; R)$ denote the Hilbert space of all complex-valued Lebesque measurable functions $f(\omega)$ with domain $|\omega| < \pi/T$ which satisfy

$$\|f(\omega)\|^2 = \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T} |f(\omega)|^2 R(\omega) d\omega < \infty \tag{21}$$

with the obvious definition of the inner product. A frequently invoked congruence is between $M(X_k, -\infty < k < \infty)$ and $L_2(-\pi/T, \pi/T; R)$.[2] By implication, $L_2(-\pi/T, \pi/T; R)$ and $M(h_k, -\infty < k < \infty)$ are also congruent through the mapping

$$\psi\left(\sum_{m=1}^{N} \alpha_m h_{k_m}\right) = \sum_{m=1}^{N} \alpha_m e^{-j\omega k_m T} \tag{22}$$

as is readily verified.

In the remainder of this paper, the congruence demonstrated in this section will be exploited to demonstrate that many available results on MMSE interpolation and prediction theory are directly applicable to the equalization problems posed in Section 2.3.

2.5 *An Orthogonal Expansion*

The congruence relation of Section 2.4 will be used in this section to establish an orthogonal expansion in $M(h_k, -\infty < k < \infty)$ which will be particularly useful in the sequel.

Define the element

$$e_k^+ \triangleq h_k - P[h_k, M(h_m, m > k)] \tag{23}$$

which is the difference between a translate of $h(t)$, h_k, and its projection on the subspace of translates to its right. It will be shown later that this element is of particular significance to the DFE. For the moment, however, note that e_k^+ is equivalent to the MMSE prediction error of X_k based on X_m, $m > k$, since the projection is the optimum linear predictor. It is well known[3] that the successive prediction errors of a random process are uncorrelated random variables. The equivalent statement relating to e_k^+ is that

$$\langle e_m^+, e_n^+ \rangle = \|e_0^+\|^2 \delta_{m,n} \tag{24}$$

and it is an orthogonal sequence.† This is readily demonstrated directly by noting that e_m^+ is orthogonal to $M(h_k, k \geqq m)$, which contains e_n^+ for $n > m$. Hence, (24) follows for $n > m$ and by symmetry for $n < m$ also.

From (24) it follows that as long as

$$\|e_0^+\| > 0 \tag{25}$$

the sequence

$$w_n \triangleq e_n^+ / \|e_0^+\|, \qquad -\infty < n < \infty, \tag{26}$$

is an orthonormal set in L_2. The significance of (25) is that the equiv-

† The norm of e_k^+ is independent of k since e_k^+ is a time translate of e_0^+.

alent random process must not be linearly predictable with vanishing mean-square error (in the language of Ref. 3, p. 564, X_k must be "regular," or "nondeterministic").

Expanding h_n in a Fourier series in w_n,

$$\begin{aligned} h_n &= u_n + v_n \\ u_n &= \sum_{m=-\infty}^{\infty} c_m w_{n+m} \\ c_m &\triangleq \langle w_{n+m}, h_n \rangle = \langle w_m, h_0 \rangle \\ \langle v_n, w_m \rangle &= 0, \quad -\infty < n < \infty, \quad -\infty < m < \infty, \end{aligned} \tag{27}$$

where v_n is the remainder. Equation (27) can be simplified by observing that

$$\langle w_m, h_0 \rangle = 0, \qquad m < 0,$$

since $h_0 \in M(h_k, k \geqq 0)$ and w_m is orthogonal to $M(h_k, k \geqq m + 1)$, which contains $M(h_k, k \geqq 0)$ when $m < 0$. In addition, it can be shown (Ref. 3, pp. 571–575) that $v_n = 0$, since the spectrum under consideration here is absolutely continuous.† Thus, (27) reduces to

$$\begin{aligned} h_n &= \sum_{m=0}^{\infty} c_m w_{n+m} \\ c_m &= \langle h_0, w_m \rangle. \end{aligned} \tag{28}$$

The expansion of (28), which is used in the theory of linear prediction,[2,3] is similar in spirit to a straightforward Gram-Schmidt orthogonalization process, but is much more useful in that the coefficients of the expansion are independent of n. The main shortcoming of the expansion (28) is requirement (25).

The formula for c_m given in (27) is not very useful in explicitly evaluating the coefficients of (28). A more useful method of evaluation is to observe that it is a spectral factorization problem. Defining the bilateral z-transform‡ of the autocorrelation,

$$R^*(z) = \sum_{m=-\infty}^{\infty} R_m Z^m, \tag{29}$$

we claim that

$$R^*(z) = \sum_{n=0}^{\infty} c_n Z^n \sum_{n=0}^{\infty} c_n Z^{-n}, \tag{30}$$

† This is by virtue of the fact that integral (21) is in terms of $R(\omega)d\omega$; i.e., the underlying measure is presumed to be absolutely continuous with respect to Lebesque measure.

‡ Note that we define the z-transform in positive powers of z.

where the c_m are given by (28). To show (30), first calculate R_j from (15),

$$\begin{aligned} R_j &= \langle h_0, h_j \rangle \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} c_m c_n \langle w_m, w_{j+n} \rangle \\ &= \begin{cases} \sum_{n=0}^{\infty} c_n c_{n+j}, & j \geqq 0 \\ \sum_{n=-j}^{\infty} c_n c_{n+j}, & j < 0. \end{cases} \end{aligned} \tag{31}$$

Similarly, the right side of (30) can be manipulated,

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_n c_m Z^{n-m} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} c_n c_{n+m} Z^m + \sum_{m=1}^{\infty} \sum_{n=-m}^{\infty} c_n c_{n-m} Z^{-m}, \tag{32}$$

and comparing (31) and (32), (30) is established. The representation of (30) is not unique. However, Doob (Ref. 3, p. 160) shows that the coefficients of (27) uniquely satisfy (30) when the additional conditions

$$\sum_{n=0}^{\infty} c_n Z^n \neq 0, \qquad |Z| < 1, \tag{33}$$

$$\sum_{n=0}^{\infty} c_n^2 < \infty \tag{34}$$

are required.† The necessity of (34) is obvious from (27), while the reason why (33) is needed is that otherwise (30) could be satisfied on the unit circle by another sequence with a larger zeroth term, contradicting the fact that

$$c_0 = \|e_0^+\|. \tag{35}$$

Equation (35) follows from the observation that $M(h_k,\ k \geqq n) = M(w_k, k \geqq n)$ and therefore $P[h_n, M(h_k, k \geqq n+1)] = \sum_{m=1}^{\infty} c_m w_{n+m}$ or

$$e_n^+ = c_0 w_n. \tag{36}$$

A simple example will serve to illustrate (30). Suppose $h(t)$ has an exponential autocorrelation with

$$R_k = A^{|k|}, \qquad 0 < A < 1. \tag{37}$$

† Of course, condition (25) is also required.

Direct calculation of (29) reveals that

$$R^*(z) = \frac{1 - A^2}{(1 - AZ)(1 - A/Z)} \tag{38}$$

which is in the form of (30) with

$$c_n = \sqrt{1 - A^2}\, A^n. \tag{39}$$

The validity of (39) can be demonstrated directly for this simple example by noting that

$$e_k^+ = h_k - Ah_{k+1} \tag{40}$$

(as can be verified by showing that e_k^+ is orthogonal to h_m, $m \geqq k + 1$) and thus

$$\begin{aligned} w_n &= \frac{h_m - Ah_{m+1}}{\|h_m - Ah_{m+1}\|} \\ &= \frac{h_m - Ah_{m+1}}{\sqrt{1 - A^2}}. \end{aligned} \tag{41}$$

From (28),

$$\begin{aligned} c_m &= \langle h_0, w_n \rangle \\ &= \sqrt{1 - A^2}\, A^m \end{aligned} \tag{42}$$

agreeing with (39).

The procedure for higher-order rational spectra is equally simple. From (29) and the fact that R_m is real and even ($R_{-m} = R_m$), it follows that

$$R^*(z) = R^*\left(\frac{1}{z}\right). \tag{43}$$

Thus, for every zero a_i and pole b_i of $R^*(z)$, a_i^{-1} and b_i^{-1} are also a zero and a pole respectively. Thus, $R^*(z)$ can be written in the form

$$R^*(z) = K \frac{\prod_{i=1}^{m} (1 - a_i z)\left(1 - \frac{a_i}{z}\right)}{\prod_{i=1}^{n} (1 - b_i z)\left(1 - \frac{b_i}{z}\right)} \tag{44}$$

$$|a_i|, \qquad |b_i| < 1$$

so that from (30)

$$C(z) = \sum_{n=0}^{\infty} c_n Z^n = \sqrt{K} \frac{\sum_{i=1}^{m} (1 - a_i z)}{\sum_{i=1}^{n} (1 - b_i z)}, \tag{45}$$

where (33) has been insured by the choice of zeros in (45).

When $R^*(z)$ is not rational, a more general method of determining the coefficients of (28) is required. For this purpose, we use the equivalent power spectrum of (20). The first form in (20) is the one required for analytically determining $C(z)$, whereas the second form is the one which would usually be used in numerical calculations. The relationship of $R(\omega)$ to $R^*(z)$ is, of course,

$$R(\omega) = TR^*(e^{j\omega T}), \tag{46}$$

the evaluation of $R^*(z)$ on the unit circle. The equivalent of (30) for $R(\omega)$ is

$$\frac{R(\omega)}{T} = \left| \sum_{k=0}^{\infty} c_k e^{j\omega kT} \right|^2. \tag{47}$$

Intuitively, (47) requires the expansion of $\sqrt{R(\omega)/T}$, with an arbitrary phase characteristic, in a complex Fourier series with only positive frequencies. Following Doob (Ref. 3, p. 161), expand $\log \sqrt{R(\omega)/T}$ in a Fourier series,

$$\frac{1}{2} \log \frac{R(\omega)}{T} = \sum_k r_k e^{j\omega kT}. \tag{48}$$

This is always possible because, as will be demonstrated later, in order for (25) to be satisfied, it is necessary and sufficient that $\log R(\omega)$ be integrable. Define

$$g(z) = r_0 + 2 \sum_{k=1}^{\infty} r_k z^k \tag{49}$$

and note that

$$\operatorname{Re} g(e^{j\omega T}) = \frac{1}{2} \log \frac{R(\omega)}{T}. \tag{50}$$

We claim that

$$C(z) = e^{g(z)} \tag{51}$$

satisfies (47), since

$$|C(e^{j\omega T})| = \exp[\operatorname{Re} g\,(e^{j\omega T})] = \sqrt{\frac{R(\omega)}{T}}.$$

Equation (33) is also satisfied since $g(z)$ is analytic for $|z| < 1$.

Equation (51) is an analytic solution to the problem initially posed, but a practical means of applying it numerically is required. It is shown in Appendix A that the Fourier coefficients of (48) can be calculated efficiently and accurately using the fast Fourier transform (FFT) algorithm. The second difficulty is in determining $C(z)$ from

$g(z)$ in (51). This is easily resolved by noting that

$$\begin{aligned} c_m &= \frac{1}{m!}\frac{d^m}{dz^m} C(z)\Big|_{z=0} \qquad m \geqq 0 \\ r_m &= \frac{1}{2m!}\frac{d^m}{dz^m} g(z)\Big|_{z=0} \qquad m \geqq 1 \\ c_0 &= e^{r_0} \end{aligned} \tag{52}$$

and applying Leibniz's differentiation rule

$$\frac{d^n}{dz^n} uv = \sum_{m=0}^{n} \binom{n}{m} \frac{d^{n-m}u}{dz^{n-m}} \frac{d^m v}{dz^m}$$

to the product

$$\begin{aligned} \frac{d^n}{dz^n} C(z) &= \frac{d^{n-1}}{dz^{n-1}} \left(e^{g(z)} \frac{dg(z)}{dz} \right) \\ &= \sum_{m=0}^{n-1} \binom{n-1}{m} \frac{d^{n-m}g(z)}{dz^{n-m}} \frac{d^m C(z)}{dz^m} \end{aligned}$$

and, setting $z = 0$,

$$c_n = \frac{2}{n} \sum_{m=0}^{n-1} (n-m) r_{n-m} c_m , \quad n \geqq 1. \tag{53}$$

Equations (52)–(53) give us a practical recursive method of determining the coefficients of (28) when the channel spectrum is not rational.

III. GEOMETRIC THEORY OF THE ZERO-FORCING AND DECISION-FEEDBACK EQUALIZERS

The zero-forcing equalizer (ZFE) and decision-feedback equalizer (DFE) have been introduced in Sections 1.1 and 2.3. In this section, we will describe fully the characteristics of these equalizers in the context of the geometric structure developed in Section II.

3.1 *Conditions for the Existence of the ZFE and DFE*

The existence of a ZFE and DFE will now be related to the interpolation and prediction of the equivalent random process defined in Section 2.2. This relationship will then be used to obtain directly the known conditions for their existence.

The first observation is that the subspaces $M(h_k,\ k \neq 0)$ and $M(X_k,\ k \neq 0)$ are identical, as are the subspaces $M(h_k,\ k > 0)$ and $M(X_k,\ k > 0)$. The element

$$e_0 = h_0 - P[h_0, M(h_k, k \neq 0)] \tag{54}$$

is the same as the interpolation error vector defined in Section 2.2, $(X_0 - \hat{X}_0)$, while the prediction error vector is the same as

$$e_0^+ = h_0 - P[h_0, M(h_k, k > 0)]. \tag{55}$$

These two vectors are likely candidates for a ZFE and a DFE because they are orthogonal to the subspaces $M(h_k, k \neq 0)$ and $M(h_k, k > 0)$ respectively [see Section 2.1 and eq. (3)]. Hence, they satisfy (11) and (13) respectively. To verify that they are indeed a ZFE and a DFE, conditions (12) and (14) must be checked. Noting that e_0 is orthogonal to $M(h_k, k \neq 0)$, we have

$$\begin{aligned} \langle e_0, h_0 \rangle &= \langle e_0, h_0 - P[h_0, M(h_k, k \neq 0)] \rangle \\ &= \|e_0\|^2 \end{aligned} \tag{56}$$

by definition (54). Similarly, it follows that

$$\langle e_0^+, h_0 \rangle = \|e_0^+\|^2. \tag{57}$$

Thus, we see that a necessary and sufficient condition for e_0 (e_0^+) to be a ZFE (DFE) is that $\|e_0\| > 0$ $(\|e_0^+\| > 0)$. By definition, the projection of h_0 on a subspace is the element of that subspace which is at a minimum distance from h_0, and hence $\|e_0\|$ and $\|e_0^+\|$ are the minimum distances between h_0 and $M(h_k, k \neq 0)$ and $M(h_k, k > 0)$ respectively. Since $\|e_0\|$ can only vanish if $h_0 \in M(h_k, k \neq 0)$, and similarly for $\|e_0^+\|$, it follows that e_0 (e_0^+) is a ZFE (DFE) if and only if $h_0 \notin M(h_k, k \neq 0)$ $[h_0 \notin M(h_k, k > 0)]$. Physically, these conditions mean that $h(t)$ must not be representable as an infinite weighted sum of a subset of its own translates. Geometrically, it is evident in Fig. 5 that, as long as $\|e_0\| > 0$ (or $\|e_0^+\| > 0$), e_0 (or e_0^+) will have a component in the direction of h_0 and the equalizer will have a response to the desired signal.

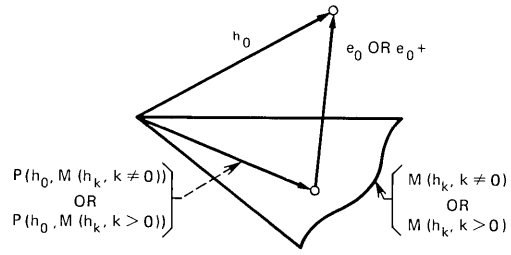


Fig. 5—Geometric interpretation of the zero-forcing equalizer and decision-feedback equalizer.

The weighting functions (54)–(55) can, under reasonable conditions,† be written in the form of a convergent linear sum of translates of h_0,

$$e_0 = h_0 - \sum_{k \neq 0} a_k h_k \tag{58}$$

$$e_0^+ = h_0 - \sum_{k>0} a_k^+ h_k \tag{59}$$

for some coefficients a_k^+. This demonstrates that these two elements are just the matched filter followed by transversal filter (MFTF) discussed in Section 1.1. It will be shown in the next section that the MFTF has particular significance, in that it maximizes the S/N ratio.

In general, there will be many ZFE's and DFE's other than (58)–(59). An example of a different ZFE is the element

$$h_0' - P[h_0', M(h_k, k \geqq 0)]$$

for any h_0' such that

$$\langle h_0', h_0 \rangle \neq 0$$
$$h_0' \notin M(h_k, k \neq 0).$$

An interesting question that arises is, then, whether there ever exists a ZFE and DFE when their corresponding MFTF's do not exist. To see that the answer is no for the ZFE (the proof for the DFE is identical), note that if $h_0 \in M(h_k, k \neq 0)$, then any g_0 orthogonal to $M(h_k, k \neq 0)$ is also necessarily orthogonal to h_0.‡ Thus, we have proven the following theorem:

Theorem 1: The following five statements are equivalent:

1. $h_0 \notin M(h_k, k \neq 0)$ $[h_0 \notin M(h_k, k > 0)]$.
2. $\|e_0\| > 0$ $[\|e_0^+\| > 0]$.
3. *There exists a ZFE* [*DFE*].
4. *There exists a ZFE* [*DFE*] *of the form of eq.* (*54*) [*eq.* (*55*)], *the MFTF.*
5. *The random process defined in* (*16*) *cannot be linearly interpolated* [*predicted*] *with vanishing mean-square error.*

The fifth condition of Theorem 1 follows from our earlier identification of e_0 and e_0^+ as the interpolation and prediction errors, respectively, of the equivalent random process. This observation also enables us to pull from the literature formulas for the norms of e_0 and e_0^+. The follow-

† This will be discussed fully in Section 3.5.
‡ We also make use of the trivial observation that any g_0 satisfying (11) is orthogonal to $M(h_k, k \neq 0)$.

ing corollary follows directly from the known formulas for the interpolation and prediction errors of a random process,[2,3]

$$\|e_0\|^2 = \left[\frac{T^2}{2\pi}\int_{-\pi/T}^{\pi/T} R^{-1}(\omega)d\omega\right]^{-1} \tag{60}$$

$$\|e_0^+\|^2 = \frac{1}{T}\exp\left[\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T}\log R(\omega)d\omega\right]. \tag{61}$$

Corollary 1: A ZFE [DFE] exists if and only if $R^{-1}(\omega)$ *[*$\log R(\omega)$*] is integrable.*

Both conditions relate to the fashion in which $R(\omega)$ vanishes. In particular, both require that $R(\omega)$ vanish on at most a set of measure zero. The relationship of (60) and (61) will be discussed more fully in the sequel.

It should be noted also that (61) follows directly from the orthogonal expansion of Section 2.5. From (35) we know that $\|e_0^+\|^2$ equals c_0^2, while (52) gives a relation for c_0. When the Fourier series of (48) is inverted and r_0 is substituted into (52), (61) results.

3.2 *Performance of the Equalizers*

It will now be shown that the MFTF among all ZFE's and DFE's maximizes the S/N ratio and minimizes the error probability in white Gaussian noise. The derivation will be a simple application of the Schwarz inequality.

Assume that the additive noise in (1) is white and Gaussian. Then the decision axis which is applied to a threshold is, for the ZFE,

$$\langle g_0, r\rangle = B_0\langle g_0, h_0\rangle + \langle g_0, n\rangle, \tag{62}$$

where $\langle g_0, n\rangle = n_0$ is a Gaussian random variable with mean zero and variance

$$En_0^2 = \frac{N_0}{2}\|g_0\|^2 \tag{63}$$

and $N_0/2$ is the two-sided spectral density of the noise. The minimum probability of error decision strategy is then to apply $\langle g_0, r\rangle$ to a series of $M-1$ thresholds, with the specific thresholds depending on the probability law on B_k. For any such law and series of thresholds the probability of error will be a monotone decreasing function of the S/N ratio, which is proportional to

$$\text{S/N} \propto \frac{\langle g_0, h_0\rangle^2}{\|g_0\|^2}, \tag{64}$$

since $\langle g_0, n \rangle$ is a zero-mean Gaussian random variable with variance proportional to $\|g_0\|^2$. Noting from (11) that g_0 is orthogonal to $P[h_0, M(h_k, k \neq 0)]$ whenever g_0 is a ZFE, (64) can be rewritten

$$\text{S/N} \propto \frac{\langle g_0, e_0 \rangle^2}{\|g_0\|^2} \leqq \|e_0\|^2 \tag{65}$$

by the Schwarz inequality, with equality if and only if g_0 equals e_0 (the MFTF) within a multiplicative constant. Thus, the MFTF, among all ZFE's, maximizes the S/N ratio. By the same method an identical result can be demonstrated for the DFE, if it is assumed that the decision-feedback mechanism correctly cancels the tails of earlier pulses.

The preceding derivation, which is a generalization of the Schwarz inequality derivation of the matched filter, has the geometric interpretation of Fig. 6. In writing (65), the maximization of (64) is restricted to those g_0 which lie in the hyperplane orthogonal to $P[h_0, M(h_k, k \neq 0)]$. Since every ZFE is also orthogonal to this vector, it follows that the hyperplane so described contains the set of all ZFE's. However, the maximization over elements of the hyperplane does not guarantee a result which is a ZFE. The vector in the hyperplane which has the greatest component in the direction of h_0 per unit length is evidently the one which lines up with e_0, as verified by (65). Fortunately, this vector also turns out to be a ZFE, so that the maximization is complete.

An additional observation relative to (65) is that the maximum S/N ratio is proportional to $\|e_0\|^2$ for the ZFE and $\|e_0^+\|^2$ for the DFE. The maximum S/N ratio is therefore directly proportional to the mean-square interpolation and prediction errors of the equivalent random process. Thus, the maximum S/N ratios of the ZFE and DFE are given by (60) and (61) respectively, while the factor by which the

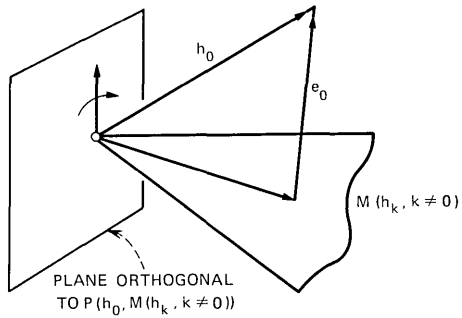


Fig. 6—S/N ratio maximized by the MFTF.

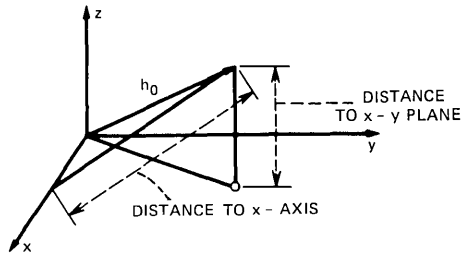


Fig. 7—Geometric interpretation of eq. (66).

S/N ratio is reduced relative to an isolated pulse with matched filter detection is obtained by dividing by R_0, the isolated pulse energy.

Price[6] derived (61) by a different method and used the geometric mean inequality for integrals to show from (60) and (61) that

$$\|e_0\|^2 \leqq \|e_0^+\|^2. \tag{66}$$

This important result implies that (*i*) the S/N ratio of the DFE MFTF always exceeds that of the ZFE MFTF,† and (*ii*) a DFE exists whenever a ZFE exists [the contrary is not true, as demonstrated by the important example of algebraic zeros in $R(\omega)$[6]]. Using the geometric method we have developed, two interpretations of (66) can be given. First, it is intuitively apparent that the mean-square interpolation error of a random process will be smaller than the mean-square prediction error, because an interpolation is based on more information; similarly, there will be some processes for which interpolation, but not prediction, with zero mean-square error is possible. Second, since $M(h_k,\ k \neq 0)$ contains $M(h_k,\ k > 0)$, the distance between h_0 and $M(h_k,\ k \neq 0)$ (equal to $\|e_0\|^2$) must be smaller than the distance between h_0 and $M(h_k,\ k > 0)$ (equal to $\|e_0^+\|^2$). This second interpretation is a rigorous way of establishing (66) by a method more direct than the integral inequality. It has the geometric interpretation of Fig. 7, where the distance between a vector h_0 and the larger subspace (the x-y plane) is less than between h_0 and the subspace it contains (the x axis).

The performance of the ZFE and DFE can be evaluated for any particular channel spectrum using (60)–(61). In particular, (60)–(61) can be evaluated in closed form for rational spectra. A different approach, which allows us to evaluate the tap-gains of the equalizers as well, will be pursued in Section 3.4.

† This result neglects the effect of decision errors on the DFE.

3.3 *On the DFE White Output Noise Property*

As observed by Price,[6] the DFE forward filter is identical to the "whitened matched filter" employed by Forney[12] as the first element of his maximum likelihood detector. The property of this filter which is essential to Forney's application is that the noise sequence at the filter output is uncorrelated. As with the other properties of this filter, this one has a simple explanation in terms of the relationship to linear prediction.

Identifying e_k^+ as $e_0^+(t - kT)$, the noise sequence at the DFE forward filter output is $\langle e_k^+, n\rangle$. Since $n(t)$ is white noise, this sequence will be uncorrelated if and only if

$$\langle e_m^+, e_n^+\rangle = 0, \qquad m \neq n. \tag{67}$$

The validity of (67) and an interpretation of this result in terms of the uncorrelated nature of the successive prediction errors of a random process has already been given in Section 2.5.

3.4 *Determination of Tap-Gains*

In this section, we will use the orthogonal expansion of Section 2.5 to derive methods of determining the tap-gains of the forward and feedback filters of the MFTF DFE. For comparison purposes the well-known relation for the tap-gains of the ZFE will also be briefly developed.

If we write the weighting response of the MFTF ZFE as

$$a_0 e_0 = \sum_{k=-\infty}^{\infty} a_k h_k, \tag{68}$$

where the tap-gains of the transversal filter are a_k, $-\infty < k < \infty$, condition (11)–(12) becomes

$$\begin{aligned}\langle e_0, h_m\rangle &= \|e_0\|^2 \delta_{m,0} \\ &= \frac{1}{a_0} \sum_k a_k R_{m-k}.\end{aligned} \tag{69}$$

Taking the bilateral z-transform of (69),

$$a_0 \|e_0\|^2 = A(z) R^*(z), \tag{70}$$

where $A(z)$ is the z-transform of the tap-gains

$$A(z) \triangleq \sum_k a_k z^k. \tag{71}$$

Thus, from (70),

$$A(z) = \frac{a_0\|r_0\|^2}{R^*(z)}. \tag{72}$$

This filter is illustrated in Fig. 8a. When $h(t)$ is applied to the input of a matched filter and the output sampled at a rate of $1/T$, the output has z-transform $R^*(z)$. The transversal filter weighting response has a z-transform proportional to $R^*(z)^{-1}$, so that the output is consistent with (69).

The S/N ratio of the ZFE, given by (60), is readily derived from (72). Writing the relation for tap-gain zero,

$$a_0 = \frac{1}{2\pi j}\oint \frac{A(z)}{z}\,dz = \frac{a_0\|e_0\|^2}{2\pi j}\oint \frac{dz}{zR^*(z)}, \tag{73}$$

and solving for $\|e_0\|^2$, we immediately get (60) using (46).

As an example, for the exponential autocorrelation of (37), (72) becomes

$$A(z) = a_0\|e_0\|^2\left[-\frac{A}{1-A^2}z^{-1} + \frac{1+A^2}{1-A^2} - \frac{A}{1-A^2}z\right] \tag{74}$$

from which we get

$$\begin{aligned} \|e_0\|^2 &= \frac{1-A^2}{1+A^2} \\ a_{-1} = a_1 &= -\frac{a_0 A}{1+A^2} \\ a_k &= 0, \qquad |k| > 1, \end{aligned} \tag{75}$$

a result derived by Tufts[13] by another method. This example points out that it is not ever necessary to actually evaluate (60) when the channel spectrum is rational, but rather the performance can be obtained by equating the zero-order tap-gains of (72) in the manner of (73).

The situation with the DFE is only slightly more complicated. In this case the DFE filter is

$$a_0^+ e_0^+ = \sum_{k=0}^{\infty} a_k h_k, \tag{76}$$

where only taps on one side are involved. Substituting from (28) and (36),

$$\begin{aligned} a_0^+ c_0 w_0 &= \sum_{k=0}^{\infty} a_k^+ \sum_{m=0}^{\infty} c_m w_{k+m} \\ &= \sum_{m=0}^{\infty} w_m \sum_{k=0}^{m} a_k^+ c_{m-k}, \end{aligned} \tag{77}$$

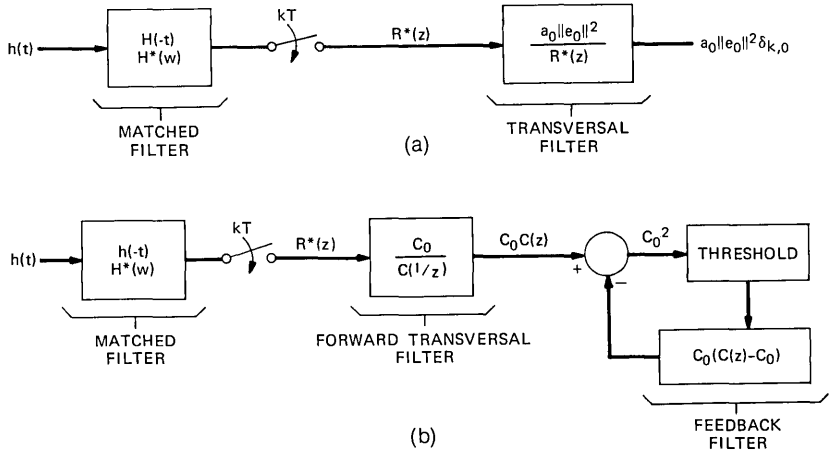


Fig. 8—Spectral representations of the MFTF zero-forcing equalizer (a) and decision-feedback equalizer (b).

and equating coefficients,

$$\sum_{k=0}^{m} a_k^+ c_{m-k} = \begin{cases} a_0^+ c_0, & m = 0 \\ 0, & m > 0. \end{cases} \tag{78}$$

From (78) we get a recursion relation for the tap coefficients which is useful for nonrational spectra,

$$a_m^+ = -\frac{1}{c_0} \sum_{k=0}^{m-1} a_k^+ c_{m-k}, \tag{79}$$

and a z-transform relation which is useful for rational spectra,

$$A^+(z) = \frac{a_0^+ c_0}{C(z)}, \tag{80}$$

where $A^+(z)$ is the z-transform of the tap-gains of (76). Performing (80) again for the autocorrelation of (37),

$$\begin{aligned} A^+(z) &= a_0^+(1 - Az) \\ \|e_0^+\|^2 &= c_0^2 = 1 - A^2 \end{aligned} \tag{81}$$

which is consistent with (40) and is larger than $\|e_0\|^2$ by a factor of $(1 + A^2)$. As with the ZFE, the performance of the DFE can be determined for rational spectra without the explicit evaluation of (61).

The comparison of (80) with (72) is interesting, in that they are identical except for the fact that in (80) $C(z)$ is substituted for $R^*(z)$

in (72). The annulus of convergence of $A(z)$ will always include the unit circle, since $R^*(z)$ converges in an annulus containing the unit circle. Similarly, $C(z)$ is analytic and nonzero in a region containing the unit disk, and hence $A^+(z)$ will have only positive powers of z and converge in a region containing the unit disk. Note that these properties of $A^+(z)$ are critically dependent on (33) being satisfied.

The spectral factorization method of determining the tap-gains of the DFE was given by Monsen[5] for rational spectra. Price[6] gave a formula valid for arbitrary spectra, but it is difficult to evaluate numerically. Since (79) is valid for arbitrary spectra, the method presented here represents a synthesis of the appeal and computational simplicity of the spectra factorization method with the generality of Price's Toeplitz form result.

We also need the tap-gains of the feedback filter for the DFE. From Fig. 4, the required feedback tap-gains are given by $\langle e_0^+, h_{-n}\rangle$, $1 \leqq n < \infty$. From (36) and (28),

$$b_n = \langle e_0^+, h_{-n}\rangle = c_0 \sum_{m=0}^{\infty} c_m \langle w_0, w_{m-n}\rangle$$

$$= c_0\, c_n. \tag{82}$$

Thus, the frequency response of the feedback filter is given by

$$\sum_{m=1}^{\infty} b_m z^m = c_0[C(z) - c_0]. \tag{83}$$

The z-transform representation of the DFE just derived is illustrated in Fig. 8b. When an isolated pulse $h(t)$ is applied to the matched filter, the sampled output has z-transform $R^*(z)$. The transversal filter multiplies by $A^+(1/z) = c_0/C(1/z)$, as can be verified from (76).† The z-transform of the forward transversal filter output is $c_0C(z)$ because of (30), which verifies the causal response which is characteristic of the DFE. The output of the feedback filter of (83) is then subtracted, to yield (hopefully) a delta function response c_0^2. The reader can verify that when the threshold is replaced by a gain of $1/c_0^2$ (the noise-free case) the response is as represented.

3.5 *Finite Transversal Filter Equalizers*

The previous sections have considered the rather idealized case of infinite transversal filter equalizers. Since only finite equalizers can

† This is because (76) is not in the form of a convolution sum. This distinction was not relevant to the ZFE due to the symmetry of that filter.

actually be implemented, the important question arises as to when and in what sense the infinite equalizer can be approximated by a finite one.

We have already seen in the example of the exponential autocorrelation that the infinite equalizer can degenerate into a finite transversal filter for some channel spectra. This will happen whenever $A(z)$ and $A^+(z)$ are finite polynomials in z. From (72) and (80) we see that this will occur whenever $R^*(z)$ is a rational function which has no zeros (only poles). When the spectrum is not rational, or is rational with zeros, it will be necessary to approximate the infinite MFTF.

It is straightforward to generalize the results of Sections 3.1 and 3.2 to subspaces spanned by a finite number of translates of h_0. In particular, if we replace the criteria of (11) and (13) by

$$\langle h_k, g_0 \rangle = 0 \qquad -N \leqq k \leqq N, \qquad k \neq 0 \tag{84}$$

for the ZFE and

$$\langle h_k, g_0 \rangle = 0 \qquad 1 \leqq k \leqq N \tag{85}$$

for the DFE, we are left with the consideration of the finite dimensional subspaces $M(h_k, -N \leqq k \leqq N, k \neq 0)$ and $M(h_k, 1 \leqq k \leqq N)$, which we will write as M_N and M_N^+ respectively. Then the MFTF equalizers which satisfy (84) and (85) are similar to (54) and (55),

$$e_0(N) \triangleq h_0 - P(h_0, M_N) \tag{86}$$

$$e_0^+(N) \triangleq h_0 - P(h_0, M_N^+). \tag{87}$$

It is straightforward to see that Theorem 1 can be replaced by the following version:

Theorem 2: The following four statements are equivalent:

1. $h_0 \notin M_N$ [$h_0 \notin M_N^+$].
2. $\|e_0(N)\| > 0$ [$\|e_0^+(N)\| > 0$].
3. *There exists a ZFE* [*DFE*] *in the restricted sense of* (84) [(85)].
4. *There exists an MFTF ZFE* [*DFE*] *in this restricted sense.*

The question of when it can be asserted that $\|e_0(N)\| > 0$ and $\|e_0^+(N)\| > 0$ deserves consideration. The condition that $h_0 \in M_N^+$ requires that coefficients $\{\alpha_m, 1 \leqq m \leqq N\}$ exist which satisfy

$$h_0 = \sum_{m=1}^{N} \alpha_m h_m. \tag{88}$$

This occurrence will be precluded if the set $\{h_m, -\infty < m < \infty\}$ is

linearly independent. Similarly, linear independence is sufficient for a ZFE to exist in the sense of (84). The following lemma, which is proven in Appendix B, establishes sufficient conditions for the linear independence of $\{h_m, -\infty < m < \infty\}$:

Lemma 1: The following two conditions are sufficient for the linear independence of $\{h_m, -\infty < m < \infty\}$:

1. $\|e_0\| > 0$ *or* $\|e_0^+\| > 0$.
2. *There exists an interval* $[a, b]$, $a < b$, *such that* $R(\omega) > 0, \omega \in [a, b]$.

The first condition of Lemma 1 satisfies our intuition that if an infinite MFTF ZFE or DFE exists then the finite MFTF version should also exist. The second condition assures us that the finite equalizers also exist under much weaker conditions.

The following theorem establishes a relationship between the finite and infinite equalizers, and is proven in Appendix B:

Theorem 3: As $N \to \infty$, $\|e_0(N)\|^2$ *is monotonically decreasing and approaches* $\|e_0\|^2$, *and likewise for* $e_0^+(N)$. *Furthermore,* $\|e_0(N) - e_0\|^2 \to 0$ *and* $\|e_0^+(N) - e_0^+\|^2 \to 0$.

The primary conclusion of Theorem 3 is that the infinite equalizer can be approximated with arbitrary accuracy (in the sense of L_2 convergence) by a finite equalizer. In addition, it asserts that the S/N ratio of this finite equalizer is greater than that of the infinite equalizer; however, this desirable property may be entirely or partially offset by any residual intersymbol interference.

Each member of the sequence of equalizers guaranteed by Theorem 3 has different tap-gains, because the projection on a different subspace is being taken with each N. A more aesthetically pleasing approximation results when (58) and (59) are valid, for then

$$\left|\left| h_0 - \sum_{k=-N}^{N} a_k h_k - e_0 \right|\right| \to 0, \tag{89}$$

$$\left|\left| h_0 - \sum_{k=1}^{N} a_k^+ h_k - e_0^+ \right|\right| \to 0, \tag{90}$$

by the definition of convergence of the infinite sums in (58)–(59). Each succeeding equalizer defined by (89)–(90) is obtained by adding an additional tap, without changing the other tap-gains. As observed by Doob (Ref. 3, p. 564), a convergent sum of the form of (58)–(59) does not always exist; the following theorem gives sufficient conditions

for the validity of (58)–(59) which are generally satisfied in practical problems:

Theorem 4†: *If there exist constants* K_1 *and* K_2, $0 < K_1 \leqq K_2$, *such that* $K_1 \leqq R(\omega) \leqq K_2$, $|\omega| < \pi/T$, *then convergent expansions of* e_0 *and* e_0^+ *of the form of (58)–(59) exist. Furthermore, the coefficients of the expansions are unique.*

This theorem is proven in Appendix B. The question of uniqueness of the tap-gains of the DFE is one which was not answered by Price.[6]

Finally, the white output noise property of the MFTF DFE also extends to a finite MFTF DFE in the following sense: If the reception of (1) extends from N_1 to N_2, where N_2 (but not necessarily N_1) is finite, then the DFE defined by

$$e_k^+ = h_k - P[h_k, M(h_m, k+1 \leqq m \leqq N_2)]$$

will have white output noise samples. This fact is easily verified from the same containment of subspaces that was used in the proof for the infinite case.

IV. EXTENSION TO NONSTATIONARY NOISE AND CHANNEL

The previous sections have considered only the case where the additive noise is white. The extension to colored Gaussian noise can be handled in a straightforward fashion with the addition of a whitening filter. In this section we will generalize the ZFE and DFE to the case of arbitrary nonstationary second-order Gaussian noise (which includes colored Gaussian noise as a special case) using the techniques of reproducing kernel Hilbert space (RKHS).[11] Although the cases for which the corresponding RKHS can be characterized explicitly correspond generally to those cases which can be handled by other techniques, the RKHS approach does allow us to treat all cases simultaneously and concisely. In addition, it enables us to generalize simultaneously to an arbitrary nonstationary channel (to be precise, a channel which is changing in time in a deterministic and known fashion) with no additional complications. Perhaps the most interesting outcome of this effort will be the observation that the DFE white output noise property (discussed in Section 3.3) remains valid in this general case. The result is an interesting generalization of Forney's whitened matched filter.[12]

---

† Theorem 4 remains valid under the weaker hypothesis that $0 <$ ess inf $R(\omega)$ and ess sup $R(\omega) < \infty$.

To this end, modify (1) to

$$r(t) = \sum_{m=N_1}^{N_2} B_m h_m(t) + n(t), \tag{91}$$

where, as before, N_1 and N_2 can be infinite. The noise will be assumed to be Gaussian with arbitrary autocorrelation

$$K(t, s) = E[n(t)n(s)]. \tag{92}$$

The subscript m on $h_m(t)$ indicates that the received pulses need not be translates of the same elementary waveform. The reception will be termed *channel stationary* when

$$h_m(t) = h(t - mT)$$

and *noise stationary* when

$$K(t, s) = K(t - s).$$

We denote by $L_2(n)$ the subspace of the Hilbert space of square integrable random variables spanned by $n(t)$, $-\infty < t < \infty$. This subspace is entirely analogous to $M(X_k, -\infty < k < \infty)$ defined earlier, except that the underlying parameter t is continuous. The following lemma is applicable:[11]

Lemma 2: Let $H(K)$ consist of all functions $g(\cdot)$ of the form

$$g(\cdot) = E[n(\cdot)U] \tag{93}$$

for some $U \in L_2(n)$. Then $H(K)$ is a Hilbert space with inner product

$$\langle g, g\rangle_{H(K)} = E|U|^2. \tag{94}$$

The mapping ψ: $L_2(n) \to H(K)$ defined by (93) is a congruence which maps $n(t)$ into $K(\cdot, t)$.

The Hilbert space $H(K)$ defined by Lemma 2 is known as the reproducing kernel Hilbert space with reproducing kernel K. It is straightforward to show from (93) and (94) that $H(K)$ has the properties

$$K(\cdot, t) \in H(K), \qquad -\infty < t < \infty, \tag{95}$$

$$\langle g(\cdot), K(\cdot, t)\rangle_{H(K)} = g(t), \qquad g \in H(K). \tag{96}$$

It can be shown[11] that for any symmetric positive-definite kernel K there exists a unique Hilbert space satisfying (95)–(96).

The inverse of $g(\cdot)$ under ψ is usually given the suggestive notation

$$\langle g, n\rangle_{H(K)} \triangleq \psi^{-1}(g) \tag{97}$$

even though $n \notin H(K)$ with probability one and therefore (97) cannot be given an interpretation as an inner product.

It will be assumed that $h_m(t) \in H(K)$, since otherwise the detection problem is singular.† [11] In nonstationary noise the space $H(K)$ takes the place of L_2 in the earlier white noise problem. Accordingly, we restrict the class of filters under consideration to $H(K)$ inner products with elements of $H(K)$. Thus, a filter can be written in the form

$$\langle g, r\rangle_{H(K)} = \sum_{m=-N_1}^{N_2} B_m \langle g, h_m\rangle_{H(K)} + \langle g, n\rangle_{H(K)}, \tag{98}$$

where the noise term in (98) assumes the special meaning of (97). Analogously to (15), we define the pulse autocorrelation

$$R(m, n) = \langle h_m, h_n\rangle_{H(K)}. \tag{99}$$

When the reception is noise and channel stationary, $R(m, n)$ is a function of the difference of its arguments, as in (15). In general, however, it is an arbitrary symmetric positive definite function defined for $N_1 \leqq m, n \leqq N_2$.‡

In the white noise case, we saw that the subspace of L_2 spanned by translates of $h(t)$ was congruent to the subspace of second-order random variables spanned by a wide-sense stationary random process. In the nonstationary noise case, the subspace of $H(K)$ spanned by h_m, $N_1 \leqq m \leqq N_2$, is congruent to the subspace of the second-order random variables spanned by a possibly nonstationary second-order random process. In the white noise case the theory of minimum mean-square error estimation of a wide-sense stationary random process was relevant; in the present case the random process becomes nonstationary. As before, the ZFE and DFE have interpretations as interpolation and prediction errors of the corresponding random process with autocorrelation $R(m, n)$. However, rather than pursue these correspondences further (in view of our results for the white noise case they are obvious), we will directly pursue the theory of the ZFE and DFE for the detection of B_m, $N_1 \leqq m \leqq N_2$, from $r(t)$ in (91).

---

† A singular detection problem is one in which a decision can be made which is correct with probability one.

‡ The positive definite property follows from the inequality

$$0 \leqq \left\| \sum_{m=1}^{N} \alpha_m h_{k_m} \right\|^2_{H(K)} = \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_m \alpha_n R(k_m, k_n).$$

The theory of Section 3.1 remains valid if the subspaces $M(h_m, m \in I)$ are considered as subspaces of $H(K)$ rather than L_2.† As before, the condition which is necessary and sufficient for the existence of a ZFE or DFE is that

$$h_k \notin M(h_m, m \in I).$$

The analogs of the MFTF versions of the DFE and ZFE are the elements given by (54) and (55), except that now we must work with e_k and e_k^+ instead of e_0 and e_0^+ (e_k is no longer necessarily simply a time translate of e_0, etc.). A derivation similar to that given in Section 3.3 establishes that e_k and e_k^+ maximize the S/N ratio as before. In particular, when the filter of (91) is restricted to be a ZFE, (91) becomes

$$\langle g, r\rangle_{H(K)} = B_k\langle g, h_k\rangle_{H(K)} + \langle g, n\rangle_{H(K)} \tag{100}$$

and the S/N ratio is proportional to

$$S/N \propto \frac{\langle g, h_k\rangle^2_{H(K)}}{\langle g, g\rangle_{H(K)}} \leqq \langle e_0, e_0\rangle_{H(K)} \tag{101}$$

since the variance of the noise term in (100) is, from (97),

$$\begin{aligned} E|\langle g, n\rangle_{H(K)}|^2 &\triangleq E|\psi^{-1}(g)|^2 \\ &= \langle g, g\rangle_{H(K)} \end{aligned}$$

through the congruence established in Lemma 2. Equation (101) demonstrates that the MFTF ZFE maximizes the S/N ratio, and the same result follows for the DFE by the same method.

A general equation can be given for the projection element required for the MFTF. This equation is entirely analogous to a result of Parzen[11] for stochastic estimation. To this end we require a lemma which is a restatement of Lemma 2:

Lemma 3: Let $H(R)$ consist of all functions $f(m)$, $m \in I$, of the form

$$f(m) = \langle h_m, F\rangle_{H(K)} m \in I \tag{102}$$

for some $F \in M(h_m, m \in I)$. Then $H(R)$ is the RKHS with reproducing kernel $R(m, n)$, $m,n \in I$, and has inner product

$$\langle f, f\rangle_{H(R)} = \langle F, F\rangle_{H(K)}. \tag{103}$$

The mapping ϕ: $M(h_m, m \in I) \to H(R)$ defined by (102) is a congruence which maps h_m into $R(\cdot, m)$.

---

† We use I as a set of indices to avoid repeating the equations twice. For the ZFE, $I = [N_1, k-1]U[k+1, N_2]$ and for the DFE $I = [k+1, N_2]$. For the infinite case, $N_2 = -N_1 = \infty$. The digit B_k is being detected.

The reader might find it instructive to verify from (102)–(103) that the RKHS properties hold for $H(R)$,

$$R(\cdot, n) \in H(R), \tag{104}$$

$$\langle f(\cdot), R(\cdot, n)\rangle_{H(R)} = f(n), \tag{105}$$

where $f(\cdot) \in H(R)$.

The problem we want to attack is finding the projection P of some vector Q on $M(h_m, m \in I)$ (later we will let $Q = h_k$). From (3) we have

$$\langle Q - P, h_m\rangle_{H(K)} = 0, \qquad m \in I \tag{106}$$

or

$$\langle P, h_m\rangle_{H(K)} = \rho_Q(m), \qquad m \in I, \tag{107}$$

where

$$\rho_Q(m) \triangleq \langle Q, h_m\rangle_{H(K)}, \qquad m \in I. \tag{108}$$

In (107), $\rho_Q(m)$ is a known function and P is to be determined. Assuming for the moment that $\rho_Q \in H(R)$, from Lemma 3 we see that ρ_Q is the image of P under the congruence ϕ, and hence

$$P = \phi^{-1}(\rho_Q), \tag{109}$$

which is the solution we desire. Using the congruence properties of ϕ, the length of $Q - P$ is

$$\begin{aligned}\|Q - \phi^{-1}(\rho_Q)\|^2_{H(K)} &= \|Q\|^2_{H(K)} - 2\langle Q, \phi^{-1}(\rho_Q)\rangle_{H(K)} + \|\phi^{-1}(\rho_Q)\|^2_{H(K)} \\ &= \|Q\|^2_{H(K)} - \|\rho_Q\|^2_{H(R)}.\end{aligned} \tag{110}$$

Establishing that in fact $\rho_Q \in H(R)$ is straightforward. Note that

$$\begin{aligned}\rho_Q(m) &= \langle Q, h_m\rangle_{H(K)} \\ &= \langle Q - P, h_m\rangle_{H(K)} + \langle P, h_m\rangle_{H(K)} \\ &= \langle P, h_m\rangle_{H(K)},\end{aligned} \tag{111}$$

which implies that $\rho_Q \in H(R)$ by Lemma 3 since $P \in M(h_m, m \in I)$.

Replacing Q by h_k in (109), we get the desired projection

$$P[h_k, M(h_m, m \in I)] = \phi^{-1}[R(k, \cdot)] \tag{112}$$

The ZFE and DFE are obtained by letting I equal the appropriate set. The S/N ratios of the receivers are proportional to, from (101) and (110),

$$\text{S/N} \propto \|h_k\|^2_{H(K)} - \|R(k, \cdot)\|^2_{H(R)}. \tag{113}$$

The RKHS approach has reduced the problem to that of finding

RKHS inner products. In some cases these inner products can be explicitly characterized, while in all others they can be determined by convergent iterative techniques.[11]

We can also quickly show that the DFE white output noise property discussed in Section 3.3 generalizes. From (98), the noise samples at the filter output are

$$\begin{aligned} n_k &= \langle e_k^+, n\rangle_{H(K)} \\ &= \psi^{-1}(e_k^+) \end{aligned} \tag{114}$$

by definition. From (114) and Lemma 2,

$$\begin{aligned} E(n_j n_k) &= E[\psi^{-1}(e_j^+)\psi^{-1}(e_k^+)] \\ &= \langle e_j^+, e_k^+\rangle_{H(K)} \\ &= 0, \qquad j \neq k \end{aligned} \tag{115}$$

by the same reasoning as before.

Finally, it is instructive to demonstrate that this RKHS formulation reduces to the whitening filter approach when the reception is noise and channel stationary. Assume that

$$K(t, s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j\omega(t-s)}N(\omega)d\omega\,, \tag{116}$$

where $N(\omega)$ is uniformly bounded and never vanishes. Under these conditions we claim that $H(K)$ consists of all integrable $g(t)$ with Fourier transforms $G(\omega)$ which satisfy

$$\|g\|_{H(K)}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\omega)|^2 \frac{1}{N(\omega)}\, d\omega\cdot \tag{117}$$

To verify this, properties (95)–(96) must be checked. Equation (95) is valid since $N(\omega)$ is integrable, while (96) follows from

$$\begin{aligned} \langle g(\cdot), K(\cdot, t)\rangle_{H(K)} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)[e^{-j\omega t}N(\omega)]^* \frac{1}{N(\omega)}\, d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)e^{j\omega t}\, d\omega \\ &= g(t), \end{aligned} \tag{118}$$

where (*) denotes complex conjugation. From (117), the $H(K)$ inner product consists of a filter with frequency response $N^{-1}(\omega)$ (which is the whitening filter) followed by an ordinary L_2 inner product, and is therefore consistent with the whitening filter formulation.

V. CONCLUSIONS

This paper has presented a unified and rather thorough treatment of the ZFE and DFE. In a companion paper,[7] the geometric model of intersymbol interference developed here will be used to study the minimum distance problem encountered in the performance analysis of the maximum likelihood detector[12] and in evaluating a lower bound on the performance of any receiver.[14] It is shown there that a canonical relationship exists between the minimum distance and the performance and tap-gains of the MFTF DFE.

No performance example comparing the DFE and ZFE on a channel of practical interest has been given in this paper in order that the maximum likelihood detector may enter into the comparison. In Ref. 7 the performance of three receivers is calculated for a channel whose loss in dB increases as the square-root of frequency. This channel is an excellent model of coaxial cable and some types of wire-pairs.

VI. ACKNOWLEDGMENTS

The author is indebted to R. Price for many valuable comments. In particular, it was he who suggested the extension to nonstationary noise using RKHS theory. The author also appreciates many valuable discussions with D. L. Duttweiler.

APPENDIX A

The purpose of this appendix is to derive an approximation to the Fourier coefficients of (48) in terms of discrete Fourier transform (DFT), which can be efficiently evaluated using the FFT algorithm.

Define a normalized function

$$F(\lambda) = \log \frac{R\left(\frac{2\pi}{T}\lambda\right)}{T} \tag{119}$$

so that

$$r_n = \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-jn2\pi\lambda} F(\lambda) d\lambda \cdot \tag{120}$$

Approximating the integral by a summation,

$$\begin{aligned} \hat{r}_n &\cong \frac{1}{2N} \sum_{k=0}^{N-1} F\left(\lambda_0 + \frac{k}{N} - \frac{1}{2}\right) e^{-jn2\pi(\lambda_0 + k/N - \frac{1}{2})} \\ &= \frac{1}{2} e^{-jn2\pi(\lambda_0 - \frac{1}{2})} \frac{1}{N} \sum_{k=0}^{N-1} F\left(\lambda_0 + \frac{k}{N} - \frac{1}{2}\right) e^{-j2\pi(kn/N)}, \end{aligned} \tag{121}$$

where the sum on the right is a discrete Fourier transform.

In order to determine the effect of this approximation, substitute

$$\tfrac{1}{2}F(\lambda) = \sum_k r_k e^{jk2\pi\lambda} \tag{122}$$

into the approximation equation (121) to yield

$$\begin{aligned}\hat{r}_n &= \sum_{m=-\infty}^{\infty} r_m \frac{1}{N} \sum_{k=0}^{N-1} e^{j(m-n)2\pi(\lambda_0+k/N-\frac{1}{2})} \\ &= r_n + \sum_{l\neq 0} e^{j2\pi lN\lambda_0}(-1)^l r_{n+lN}.\end{aligned} \tag{123}$$

Thus, the approximation of (121) yields the desired Fourier coefficient plus the sum of alias terms. N must be larger than the number of coefficients to be evaluated and large enough that the alias terms r_{n+lN} are small. In practice, $N \cong 5{,}000$ can be achieved with modest amounts of computer time using the FFT algorithm.

APPENDIX B

Proofs of Theorems

Proof of Lemma 1: Since $\|e_0^+\|^2 \geqq \|e_0\|^2$ it suffices to show that $\|e_0^+\| > 0$ implies that $\{h_m, -\infty < m < \infty\}$ is linearly independent set. To this end, assume that

$$\left\| \sum_{m=1}^{N} \alpha_m h_{k_m} \right\|^2 = 0, \qquad k_1 < k_2 < \cdots < k_N. \tag{124}$$

To show that $\alpha_1 = 0$, assume to the contrary that $\alpha_1 \neq 0$ and note that

$$0 = |\alpha_1|^2 \left\| h_{k_1} + \sum_{m=2}^{N} \frac{\alpha_m}{\alpha_1} h_{k_m} \right\|^2 \geqq |\alpha_1|^2 \|e_0^+\| > 0. \tag{125}$$

This contradiction establishes that $\alpha_1 = 0$. Continuing by induction in the same fashion, it can be shown that $\alpha_m = 0$, $1 \leqq m \leqq N$.

To show that the second condition of Lemma 1 implies linear independence, we use a proof similar to Tuft's.[13] By the congruence of (22), (124) is equivalent to

$$\int_{-\pi/T}^{\pi/T} \left| \sum_{m=1}^{N} \alpha_m e^{-j\omega k_m T} \right|^2 R(\omega) d\omega = 0,$$

which implies that the integrand is zero almost everywhere on $[a, b]$. This is impossible unless $\alpha_m = 0$, $1 \leqq m \leqq N$, since otherwise

$$\left| \sum_{m=1}^{N} \alpha_m e^{-j\omega k_m T} \right|^2$$

has at most a finite number of algebraic zeros on $[a, b]$ and $R(\omega)$ is strictly positive.

Proof of Theorem 3: We will prove the result for the ZFE; the proof for the DFE is identical. Since for $N \leqq M$

$$M(h_k, |k| \leqq N, k \neq 0) \subset M(h_k, |k| \leqq M, k \neq 0) \subset M(h_k, k \neq 0),$$

the inequality

$$\|e_0\| \leqq \|e_0(M)\| \leqq \|e_0(N)\|$$

follows. Hence $\|e_0(M)\|^2$ must approach a limit,

$$\lim_{N \to \infty} \|e_0(N)\| \geqq \|e_0\|.$$

Denote by the shortened notation P the projection of h_0 on $M(h_k, k \neq 0)$ (so that $e_0 \triangleq h_0 - P$). Since $P \in M(h_k, k \neq 0)$, there exists a sequence $\gamma_n \in M(h_k, |k| \leqq n, k \neq 0)$ such that $\gamma_n \to P$ and we have

$$\|h_0 - \gamma_n\|^2 = \|e_0\|^2 + \|P - \gamma_n\|^2.$$

For any $\epsilon > 0$, there exists an $N(\epsilon)$ such that

$$\|h_0 - \gamma_n\|^2 \leqq \|e_0\|^2 + \epsilon$$

for $n \geqq N(\epsilon)$, and since $\|e_0(n)\|^2 \leqq \|h_0 - \gamma_n\|^2$ we have

$$\|e_0\|^2 \leqq \|e_0(n)\|^2 \leqq \|e_0\|^2 + \epsilon,$$

which establishes that $\|e_0(n)\| \to \|e_0\|$. The remainder of the proof follows that of the projection theorem. By the parallelogram law,

$$\|e_0(N) - e_0\|^2 = 2\|e_0(N)\|^2 + 2\|e_0\|^2 - \|e_0(N) + e_0\|^2,$$

but defining $P(N) = P[h_0, M(h_k, |k| \leqq N, k \neq 0)]$

$$\begin{aligned} \|e_0(N) + e_0\|^2 &= \|h_0 - P(N) + h_0 - P\|^2 \\ &= 4\left\|h_0 - \frac{P(N) + P}{2}\right\|^2 \geqq 4\|e_0\|^2, \end{aligned}$$

we have

$$\|e_0(N) - e_0\|^2 \leqq 2[\|e_0(N)\|^2 - \|e_0\|^2] \to 0.$$

Proof of Theorem 4: From (22) we have

$$\left\| \sum_{m=1}^{N} \beta_m h_{k_m} \right\|^2 = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} \left| \sum_{m=1}^{N} \beta_m e^{-j\omega k_m T} \right|^2 R(\omega) d\omega.$$

A standard result of Toeplitz theory asserts that

$$\frac{1}{T}\left\{\sum_{m=1}^{N}|\beta_m|^2\right\}\operatorname{ess\,inf} R(\omega) \leqq \left|\left|\sum_{m=1}^{N}\beta_m h_k\right|\right|^2 \leqq \frac{1}{T}\left\{\sum_{m=1}^{N}|\beta_m|^2\right\}\operatorname{ess\,sup} R(\omega).$$

The conclusions of the theorem then follow from Theorem 5.17.18 of Ref. 10.

REFERENCES

1. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, New York: Wiley, 1967.
2. Grenander, U., and Rosenblatt, M., *Statistical Analysis of Stationary Time Series*, New York: Wiley, 1957.
3. Doob, J. L., *Stochastic Processes*, New York: Wiley, 1953.
4. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
5. Monsen, P., "Linear Equalization for Digital Transmission over Noisy Dispersive Channels," Ph.D. Dissertation, Columbia University, June 1970.
6. Price, R., "Nonlinearly Feedback-Equalized PAM vs Capacity for Noisy Filter Channels," 1972, Int. Conf. Commun., Philadelphia, June 1972.
7. Messerschmitt, D. G., "A Geometric Theory of Intersymbol Interference. Part II: Performance of the Maximum Likelihood Detector," B.S.T.J., this issue, pp. 1521–1539.
8. Messerschmitt, D. G., "A Unified Geometric Theory of Zero-Forcing and Decision-Feedback Equalization," 1973 Int. Conf. Commun., Seattle, June 1973.
9. Messerschmitt, D. G., "Digital Communications: Detectors and Estimators for the Time-Varying Channel with Intersymbol Interference," Ph.D. Dissertation, University of Michigan, December 1971.
10. Naylor, A. W., and Sell, G. R., *Linear Operator Theory in Science and Engineering*, New York: Holt, Rinehart and Winston, 1971.
11. Parzen, E., *Time Series Analysis Papers*, San Francisco: Holden-Day, 1967.
12. Forney, G. D., Jr., "Maximum Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-18*, May 1972, p. 363.
13. Tufts, D. W., "Nyquist's Problem—The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation," Proc. IEEE, *53*, March 1965.
14. Forney, G. D., Jr., "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," IEEE Trans. Commun., *COM-20*, February 1972, p. 76.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

A Geometric Theory of Intersymbol Interference

Part II: Performance of the Maximum Likelihood Detector

By D. G. MESSERSCHMITT

(Manuscript received May 24, 1973)

In a companion paper,[1] a geometric approach to the study of intersymbol interference was introduced. In the present paper this approach is applied to the performance analysis of the Viterbi algorithm maximum likelihood detector (MLD) of Forney.[2–4] It is shown that a canonical relationship exists between the minimum distance, which Forney has shown determines the performance of the MLD, and the performance and tap-gains of the decision-feedback equalizer (DFE). Upper and lower bounds on the minimum distance are derived, as is an iterative technique for computing it exactly.

The performances of the MLD, DFE, and zero-forcing equalizer (ZFE) are compared on the $\sqrt{f}$ channel representative of coaxial cables and some wire pairs. One important conclusion is that, previous statements notwithstanding,[2,4] even the MLD experiences a substantial penalty in S/N ratio relative to the isolated pulse bound on this channel of practical interest.

I. INTRODUCTION

Forney[2,3] has detailed the Viterbi algorithm version of the maximum likelihood detector (MLD) of digital sequences in the presence of intersymbol interference. He asserts that the probability of bit error of the MLD in additive white Gaussian noise can be bounded at high S/N ratios in the form

$$K_L Q\left(\frac{d_{\min}}{2\sigma}\right) \leqq P_e \leqq K_u Q\left(\frac{d_{\min}}{2\sigma}\right), \tag{1}$$

where K_L and K_u are constants, Q is the Gaussian distribution

function,

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2}\, dy\,, \tag{2}$$

$d_{\min}$ is the minimum distance between any two transmitted signals (it will be defined more fully in Section 2.2), and σ^2 is the noise variance. For comparison purposes, the probability of error for a matched filter receiver in the absence of intersymbol interference is

$$P_e = Q\left(\frac{\sqrt{R_0}}{2\sigma}\right), \tag{3}$$

where R_0 is the energy of an isolated pulse [(1) reduces to (3) in this case].

Forney also asserts that the lower bound of (1) is also a lower bound on the error probability of any receiver.[4] Thus, the MLD achieves, within the multiplicative constant K_u/K_L, the minimum probability of error attainable by any receiver at high S/N ratios, and, in a very fundamental sense, the quantity

$$d^2_{\min}/R_0$$

is a measure of the effective decrease in the S/N ratio (relative to the detection of an isolated pulse) resulting from intersymbol interference.

The determination of the quantity $d^2_{\min}$ (known as the "minimum distance problem") is therefore a very important one for, even if the implementation of the MLD is not contemplated on a particular channel, $d^2_{\min}$ is a measure of the potential performance which can be obtained using receivers of arbitrary complexity. Unfortunately, on channels with severe intersymbol interference, the exact analytical determination of $d^2_{\min}$ does not appear feasible because of the nonlinear nature of the problem.

The minimum distance can be determined numerically by the "brute force" technique of calculating a sequence of converging upper bounds. A shortcoming of this method is that it gives no assurance as to when convergence to the desired accuracy has occurred. In addition, it gives no insight into the nature of $d^2_{\min}$ and its relationship to the intersymbol interference or to the performances of other receivers.

In this paper, we attack the minimum distance problem using a geometric theory of intersymbol interference developed in companion papers.[1,5] A canonical relationship will be shown between $d^2_{\min}$ and the decision-feedback equalizer (DFE). This relationship will be exploited

to derive simple lower and upper bounds on $d^2_{\min}$ in terms of the tap-gains of the DFE transversal filter and the S/N ratio performance of the DFE. In addition, an iterative procedure will be derived for the calculation of $d^2_{\min}$ to any desired accuracy using a sequence of converging upper *and lower* bounds on $d^2_{\min}$. The lower bounds give us a measure of the degree of convergence and enable us to terminate the calculation when the desired accuracy is assured.

After consideration of the minimum distance problem in Section II, the performance of the zero-forcing equalizer (ZFE), DFE, and MLD is compared on a channel of practical interest in Section III.

II. PERFORMANCE OF THE MLD

The minimum distance problem will now receive consideration. The first step is to briefly review the notation of a companion paper.[1]

2.1 *Notation*

The reception from a PAM communication channel takes the form

$$r(t) = \sum_k B_k h(t - kT) + n(t), \tag{4}$$

where each B_k assumes one of a finite number of predetermined values (the data being transmitted), $h(t)$ is square-integrable (element of L_2),* and $n(t)$ is white Gaussian noise.

When we denote $h(t - kT)$ as an element of L_2 by h_k, $M(h_k, k \in I)$ is the smallest closed linear subspace of L_2 containing all finite linear combinations of elements of the set $\{h_k, k \in I\}$. The projection of a vector x on $M(h_k, k \in I)$ is denoted by $P[x, M(h_k, k \in I)]$. The forward matched-filter transversal-filter combination of the DFE corresponds to the L_2 inner product of the reception $r(t)$ with the element

$$e_k^+ \triangleq h_k - P[h_k, M(h_m, m > k)] \tag{5}$$

and is orthogonal to the subspace $M(h_m, m > k)$. The quantity

$$\frac{\|e_0^+\|^2}{R_0},$$

where

$$R_k \triangleq \langle h_m, h_{m+k} \rangle \tag{6}$$

* We denote by L_2 the space of square integrable waveforms with inner product

$$\langle x,y \rangle = \int_{-\infty}^{\infty} x(t)y(t)dt$$

and norm $\|x\|^2 = \langle x,x \rangle$.

is the effective decrease in S/N ratio relative to an isolated pulse for the DFE. Thus, $\|e_0^+\|^2$ plays the same role for the DFE as $d_{\min}^2$ plays for the MLD.

The sequence of vectors $\{w_k \triangleq e_k^+/\|e_k^+\|\}$ is an orthonormal sequence in L_2, and h_n has the orthogonal expansion

$$h_n = \sum_{m=0}^{\infty} C_m w_{m+n}, \tag{7}$$

where the coefficients $\{C_m\}$ can be determined by the method of Ref. 1 for channels with either a rational or nonrational spectrum. In particular, we have

$$C_0 = \|e_0^+\|. \tag{8}$$

Of course, it is apparent that (7) is valid only as long as $\|e_0^+\| > 0$, which is true if and only if a DFE exists.

2.2 *Interpretation of the Minimum Distance*

The MLD described by Forney[2] consists of a combination of a matched filter followed by a causal or anticausal transversal filter, the combination of which he calls a "whitened matched filter," followed by a dynamic programming algorithm known as the Viterbi algorithm.[3] The whitened matched filter forms a sequence of sufficient statistics for the detection of the data digits and has independent noise samples at the output. As pointed out by Price,[6] the anticausal whitened matched filter is identical to the forward linear filter portion of the DFE.

The signal at the output of the whitened matched filter (or DFE forward filter) is[1]

$$r_k = C_0^2 B_k + \sum_{m=1}^{\infty} C_0 C_m B_{k-m} + n_k, \tag{9}$$

where n_k is a noise sample. The DFE forms the quantity

$$r_k' = r_k - \sum_{m=1}^{\infty} C_0 C_m \hat{B}_{k-m} \tag{10}$$

and applies it to a decision threshold to determine the estimated digit $\hat{B}_k$. The MLD detector, on the other hand, assumes that the sum in (9) is truncated to M terms and determines the sequence $\{\hat{B}_k\}$ so as to minimize

$$\sum_{k=1}^{N} \left\{ r_k - \sum_{m=0}^{M} C_0 C_m \hat{B}_{k-m} \right\}^2. \tag{11}$$

Thus, the two receivers perform similar functions on the same sufficient statistics r_n, the major differences being the greater complexity of the MLD and the susceptibility of the DFE to decision errors. We will now demonstrate the less obvious conclusion that the *performance* of the MLD is closely related to the DFE as well.

The minimum distance, $d^2_{\min}$, is defined as[2]

$$d^2_{\min} \triangleq \inf_{\epsilon_0 \neq 0} \left|\left| \sum_{n=0}^{N} \epsilon_n h_n \right|\right|^2, \tag{12}$$

where the infimum is over all error sequences $(\epsilon_0, \cdots, \epsilon_N)$ and all N.* Each ϵ_k assumes the value $+1$, -1, or zero (for simplicity, the binary case with $B_k = 1$ or 0 is considered). Thus, $d_{\min}$ is the minimum distance in L_2 between two signals in the signal set. It is apparent that

$$d^2_{\min} \leqq R_0, \tag{13}$$

since R_0 corresponds to $\epsilon_n = 0$, $n > 0$. Thus, $d^2_{\min}/R_0$, which is the S/N ratio penalty, is a number between zero and unity as it should be.

It is apparent in (12) that without loss of generality we can choose $\epsilon_0 = 1$ and write

$$d^2_{\min} = \inf \left|\left| h_0 + \sum_{n=1}^{N} \epsilon_n h_n \right|\right|^2. \tag{14}$$

The sum in (14) is an element of $M(h_k, k \geqq 1)$, and the minimization in (14) is an attempt to find the element of $M(h_k, k \geqq 1)$ with manifold coefficients $(+1, -1, 0)$ which is closest (in $\mathcal{L}_2$ metric) to h_0. We know that the closest element without the restriction in coefficients is the projection of h_0 on $M(h_k, k \geqq 1)$, $P[h_0, M(h_k, k \geqq 1)]$. Thus, intuitively, $d^2_{\min}$ is determined by how closely the projection can be approximated by an element with restricted manifold coefficients. To formalize this intuition, add and subtract the projection from (14) and utilize (5),

$$\begin{aligned} d^2_{\min} &= \inf \left|\left| e_0^+ + P[h_0, M(h_k, k \geqq 1)] + \sum_{n=1}^{N} \epsilon_n h_n \right|\right|^2 \\ &= \|e_0^+\|^2 + \inf \left|\left| P[h_0, M(h_k, k \geqq 1)] + \sum_{n=1}^{N} \epsilon_n h_n \right|\right|^2, \end{aligned} \tag{15}$$

where the fact that e_0^+ is orthogonal to $M(h_k, k > 0)$ has been used to eliminate the cross-product in (15). The most immediate consequence of (15) is that

$$d^2_{\min} \geqq \|e_0^+\|^2. \tag{16}$$

* In most cases of interest, the infimum will be achieved for finite N.

We have thus succeeded in proving formally what should be obvious from considerations of the relative complexity of the two receivers: The effective S/N ratio of the MLD always exceeds that of the DFE (and hence ZFE[1]).* The second consequence of (15) is the formalization of our intuition through the assertion that the amount by which the S/N ratio of the MLD exceeds that of the DFE is governed by the coarseness of the best approximation to the projection by the element with restricted coefficients: The poorer the approximation, the better the S/N ratio of the MLD.

Writing the projection in the form

$$P[h_0, M(h_k, k > 0)] = -\sum_{m=1}^{\infty} a_m^+ h_m, \tag{17}$$

we note that the a_m^+ are the tap-gains of the DFE forward transversal filter, and rewrite (15) as†

$$d_{\min}^2 = \|e_0^+\|^2 + \inf \left|\left| \sum_{n=1}^{\infty} (\epsilon_n - a_n^+) h_n \right|\right|^2. \tag{18}$$

Equation (18) shows the fundamental relationship between the minimum distance, the effective S/N ratio of the DFE (in the form of $\|e_0^+\|^2$), and the tap-gains of the DFE transversal filter. In particular, we can assert that $d_{\min}^2 = \|e_0^+\|^2$ if and only if the tap-gains are all $+1$, -1, or zero.

2.3 *Bounds on the Minimum Distance*

Equation (18) can be used to derive bounds on $d_{\min}^2$ in terms of the DFE tap-gains. From the identity‡

$$\left|\left| \sum_{n=1}^{N} (\epsilon_n - a_n^+) h_n \right|\right|^2 = (\epsilon_k - a_k^+)^2 \left|\left| h_k + \sum_{\substack{n=1 \\ n \neq k}}^{N} \frac{\epsilon_n - a_n^+}{\epsilon_k - a_k^+} h_n \right|\right|^2, \tag{19}$$

we immediately get the bounds

$$\left|\left| \sum_{n=1}^{N} (\epsilon_n - a_n^+) h_n \right|\right|^2 \geqq \begin{cases} (\epsilon_1 - a_1^+)^2 \|e_0^+\|^2, & k = 1 \\ (\epsilon_k - a_k^+)^2 \|e_0\|^2, & k > 1, \end{cases} \tag{20}$$

---

* We are tempted to argue that (16) is implied by the assertion in Ref. 2 that the MLD achieves the lowest effective S/N ratio of any receiver. However, that is not the case, because of the effect of decision errors on the DFE. The effective S/N ratio of the DFE could be higher than that of the MLD, and yet the DFE could have at the same time a higher error probability because of error propagation.

† We have taken the liberty of writing a sum over infinite error sequences, where it is understood that the infimum is only over error sequences with a finite number of nonzero terms.

‡ In (19) it is assumed that $(\epsilon_k - a_k^+) \neq 0$. When $\epsilon_k - a_k^+ = 0$, (20) is trivially satisfied.

since

$$\sum_{\substack{n=1 \\ n \neq k}}^{N} \frac{(\epsilon_n - a_n^+)}{(\epsilon_k - a_k^+)} h_n$$

is an element of $M(h_m, m \neq k)$. In (20), e_0 is the ZFE filter defined in Ref. 1,

$$e_0 \triangleq h_0 - P[h_0, M(h_k, k \neq 0)]. \tag{21}$$

In addition, if we define $\lambda_{\min}(N)$ and $\lambda_{\max}(N)$ as the minimum and maximum eigenvalues of the correlation matrix

$$\mathbf{R}_N \triangleq [R_{m-n}] \qquad 1 \leqq m,n \leqq N,$$

then we can assert that

$$\lambda_{\min}(N) \sum_{n=1}^{N} (\epsilon_n - a_n^+)^2 \leqq \left|\left| \sum_{n=1}^{N} (\epsilon_n - a_n^+) h_n \right|\right|^2 \leqq \lambda_{\max}(N) \sum_{n=1}^{N} (\epsilon_n - a_n^+)^2. \tag{22}$$

A standard Toeplitz form result[7] asserts that*

$$\lim_{N \to \infty} \lambda_{\min}(N) = \frac{1}{T} \text{ ess inf } R(\omega)$$

$$\lim_{N \to \infty} \lambda_{\max}(N) = \frac{1}{T} \text{ ess sup } R(\omega).$$

Applying (18), (20), and (22), we get three lower and one upper bound on $d_{\min}^2$ in terms of the tap coefficients of the DFE,

$$d_{\min}^2 \geqq \|e_0^+\|^2 + \begin{cases} \|e_0^+\|^2 \min\limits_{\epsilon_1} (\epsilon_1 - a_1^+)^2 \\ \|e_0\|^2 \min\limits_{\epsilon_k} (\epsilon_k - a_k^+)^2, \qquad k > 1 \\ \dfrac{1}{T} \{\text{ess inf } R(\omega)\} \lim\limits_{N \to \infty} \min\limits_{\epsilon_1, \cdots, \epsilon_N} \sum\limits_{n=1}^{N} (\epsilon_n - a_n^+)^2 \end{cases}$$

$$d_{\min}^2 \leqq \|e_0^+\|^2 + \frac{1}{T} \{\text{ess sup } R(\omega)\} \lim_{N \to \infty} \min_{\epsilon_1, \cdots, \epsilon_N} \sum_{n=1}^{N} (\epsilon_n - a_n^+)^2. \tag{23}$$

In addition, an upper bound can be obtained by substituting any error sequence into (18); a reasonable choice is

$$\epsilon_k = \begin{cases} +1, & a_k^+ < -\frac{1}{2} \\ 0, & -\frac{1}{2} < a_k^+ < \frac{1}{2}. \\ -1, & a_k^+ > \frac{1}{2} \end{cases} \tag{24}$$

* For all practical purposes, "ess inf" and "ess sup" can be replaced by "min" and "max," respectively.

These five bounds can be useful in estimating the penalty in S/N ratio for the MLD. They all require the existence of a DFE and require that the projection can be written as the convergent sum of (17).* The second and third bounds of (23) are an improvement on (16) only when the increasingly stringent requirements that a ZFE exist ($\|e_0\| > 0$) and $R(\omega)$ be uniformly bounded away from zero (almost everywhere) are imposed. The requirement of the upper bound of (23) that $R(\omega)$ be uniformly upper bounded (almost everywhere) will generally be satisfied in practice. All the bounds require a pointwise minimization over error sequences, a task much simpler than minimizing (12) directly.

As a simple application of these bounds, consider the exponential autocorrelation

$$R_k = A^{|k|}, \qquad 0 < A < 1. \tag{25}$$

Then we have[1,2]

$$\begin{aligned} d^2_{\min} &= \begin{cases} 1, & 0 < A \leqq \frac{1}{2} \\ 2(1 - A), & \frac{1}{2} < A < 1 \end{cases} \\ \|e_0\|^2 &= (1 - A^2)/(1 + A^2) \\ \|e_0^+\|^2 &= 1 - A^2 \\ a_1^+ &= -A, \qquad a_k^+ = 0, \qquad k > 1. \end{aligned} \tag{26}$$

The first and third bounds of (23) become

$$d^2_{\min} \geqq \begin{cases} 1 - A^4, & 0 < A \leqq \frac{1}{2} \\ (1 - A^2)(2 + A^2 - 2A), & \frac{1}{2} < A < 1 \end{cases} \tag{27}$$

$$d^2_{\min} \geqq \begin{cases} 1 - 2A^3/(1 + A), & 0 < A \leqq \frac{1}{2} \\ 2(1 - A)(1 + A^2)/(1 + A), & \frac{1}{2} < A < 1 \end{cases} \tag{28}$$

and the upper bound of (23) becomes

$$d^2_{\min} \leqq \begin{cases} 1 + \dfrac{2A^3}{1 - A}, & 0 < A \leqq \frac{1}{2} \\ 2(1 - A^2), & \frac{1}{2} < A < 1. \end{cases} \tag{29}$$

These bounds are plotted in Fig. 1. The upper bound of (24) is equal to $d^2_{\min}$ and is not plotted.

* If the projection of h_0 on $P(h_k,\ k \geqq 1)$ cannot be written in the form of (17), the bounds of (22) to (24) can be fixed up by considering the projection on $P(h_k,\ 1 \leqq k \leqq N)$ and taking limits as $N \to \infty$. The tap-gains will then be a function of N, and the process will be more difficult.

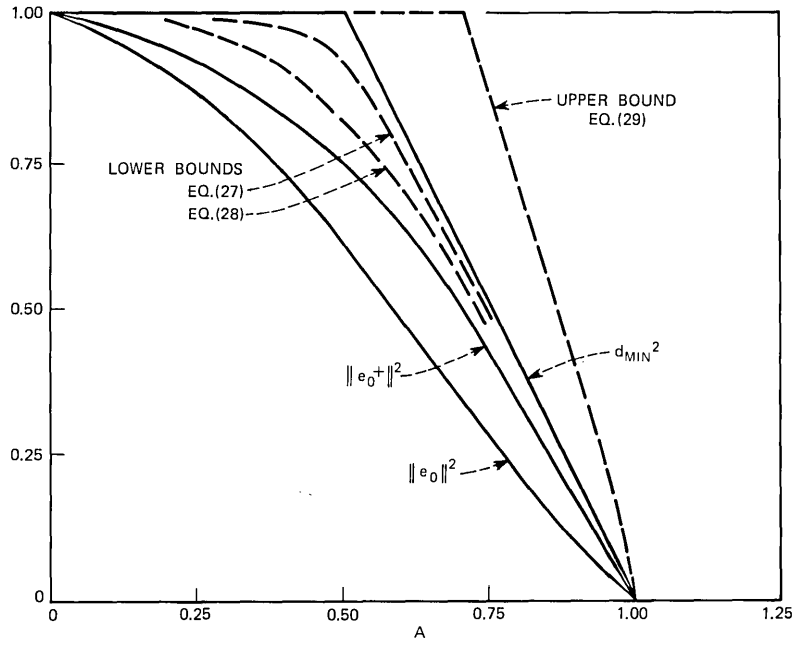


Fig. 1—Bounds on $d_{\min}^2$ for an exponential autocorrelation.

The bounds just determined have the disadvantages that (*i*) they require calculation of the DFE tap coefficients and (*ii*) they do not give precise results on $d_{\min}^2$. The exact value of $d_{\min}^2$ can be determined numerically by the direct minimization of (12); by letting $N \to \infty$ while exhaustively minimizing over error sequences, we get a sequence of upper bounds on $d_{\min}^2$ which approach $d_{\min}^2$ monotonically. The obvious difficulty with this method is that the number of error sequences which must be checked grows as 3^N, and the computational effort soon becomes unreasonable. What happens in practice is that the true minimum is achieved for a finite (and small) N. However, unless we have some method of determining when the true minimum is reached, there must always remain a degree of uncertainty as to whether the true minimum has been reached.

Our approach to this computational problem will be to derive a sequence of *lower* bounds on $d_{\min}^2$ which also approach $d_{\min}^2$ monotonically. We can then halt the process at a value of N where the upper and lower bounds are close enough to ensure knowledge of $d_{\min}^2$ within the desired accuracy. To this end, we will utilize the orthogonal expansion

of (7). Substituting (7) into the sum of (12),

$$\sum_{n=0}^{\infty} \epsilon_n h_n = \sum_{n=0}^{\infty} \epsilon_n \sum_{m=0}^{\infty} C_m w_{n+m}$$

$$= \sum_{m=0}^{\infty} \beta_n w_n, \tag{30}$$

where

$$\beta_m = \sum_{k=0}^{m} \epsilon_k C_{m-k}. \tag{31}$$

Then, because the $\{w_n\}$ are orthonormal,

$$\left|\left| \sum_{n=0}^{\infty} \epsilon_n h_n \right|\right|^2 = \sum_{n=0}^{\infty} \beta_n^2. \tag{32}$$

It appears that we may have made life more difficult for ourselves, because even when we substitute a finite sum on the left of (32) we must still evaluate an infinite sum on the right. However, note that since the terms in the sum are positive,

$$\left|\left| \sum_{n=0}^{\infty} \epsilon_n h_n \right|\right|^2 \geqq \sum_{n=0}^{N} \beta_n^2, \tag{33}$$

where the sum on the right is always finite and is in terms of a finite length error sequence $(\epsilon_0, \cdots, \epsilon_N)$. Hence,

$$d_{\min}^2 \geqq \min_{\substack{\epsilon_1, \cdots, \epsilon_N \\ \epsilon_0 = 1}} \sum_{n=0}^{N} \beta_n^2 \tag{34}$$

and, furthermore, the right side of (34) approaches the left side monotonically as $N \to \infty$.

The minimization of (34) is no more or less difficult to perform than that of the direct minimization of (12). It does require the existence of a DFE and evaluation of the coefficients $\{C_m\}$. A reasonable procedure is, at each stage of N, to minimize the right side of (34) to obtain a lower bound on $d_{\min}^2$ and substitute the minimizing sequence into (12) to obtain the upper bound* on $d_{\min}^2$. When the lower and upper bounds are sufficiently close, the process can be terminated.

* Note that any sequence substituted into (12) yields an upper bound on $d_{\min}^2$, and the one which minimizes (34) is as good as any. On the other hand, only the sequence which minimizes (34) yields a valid lower bound, so it must be minimized.

The minimization of (34) can be assisted slightly by dynamic programming. Defining

$$f_{N-m}(\epsilon_1, \cdots, \epsilon_{m-1}) = \min_{\epsilon_m, \cdots, \epsilon_N} \sum_{n=m}^{N} \beta_n^2, \tag{35}$$

we note that

$$\min_{\epsilon_1, \cdots, \epsilon_N} \sum_{n=1}^{N} \beta_n^2 = \min_{\epsilon_1} [f_{N-2}(\epsilon_1) + \beta_1^2] \tag{36}$$

with a recursion relation for $f_{N-m}(\epsilon_1, \cdots, \epsilon_{m-1})$,

$$\begin{aligned} f_{N-m+1}(\epsilon_1, \cdots, \epsilon_{m-2}) &= \min_{\epsilon_{m-1}, \cdots, \epsilon_N} \sum_{n=m-1}^{N} \beta_n^2 \\ &= \min_{\epsilon_{m-1}} \left[\min_{\epsilon m \cdots \epsilon N} \sum_{n=m}^{N} \beta_n^2 + \beta_{m-1}^2 \right] \\ &= \min_{\epsilon_{m-1}} [f_{N-m}(\epsilon_1 \cdots \epsilon_{m-1}) + \beta_{m-1}^2]. \end{aligned} \tag{37}$$

Because there is no possibility of using forward dynamic programming in this case, the savings in computation for this method is not too spectacular. Each β_n must still be evaluated for 3^N error sequences; the savings is in eliminating the need for summing β_n^2 for most of the combinations of 3^N error sequences.

We note in passing that using the FFT algorithm to reduce the computational effort in the convolutional sum of (31) is a possibility. However, the 3^N sequences for which it must be evaluated becomes a limiting factor long before the savings of that method becomes substantial.

In the foregoing discussion, the existence of a DFE has been required [that is, $\|e_0^+\| > 0$ or equivalently $\log R(\omega)$ is integrable, where $R(\omega)$ is the equivalent power spectrum of the channel[1]]. When $\log R(\omega)$ is not integrable (as when it vanishes on an interval), there does not appear to exist an expansion of the type (31) to (32). What can be done is to use the Gram-Schmidt expansion of the form

$$h_m = \sum_{k=0}^{m} \langle h_m, w_k \rangle w_k, \tag{38}$$

where w_k is the orthonormal sequence obtained from $\{h_k\}$ by the usual Gram-Schmidt orthonogalization procedure. This expansion merely requires that $\{h_k\}$ be linearly independent, which is guaranteed by the existence of an interval on which $R(\omega)$ does not vanish.[1] From (38), it

follows that

$$\sum_{m=0}^{\infty} \epsilon_m h_m = \sum_{m=0}^{\infty} \epsilon_m \sum_{k=0}^{m} \langle h_m, w_k \rangle w_k$$

$$= \sum_{k=0}^{\infty} \beta_k w_k \tag{39}$$

$$\beta_k = \sum_{m=k}^{\infty} \epsilon_m \langle h_m, w_k \rangle. \tag{40}$$

The key point is that the summation in (40) is infinite, so that evaluation of the lower bound of (34) is now necessarily over infinite error sequences. The finite sum in (31) results from the form of the expansion (7) in which h_n is expanded in terms of all future w_k's, and this expansion is in turn dependent on h_n not being an element of $M(h_k, k > n)$. Thus, when a DFE does not exist there appears to be no alternative to evaluating a sequence of upper bounds to $d^2_{\min}$ obtained by a finite sum approximation without the benefit of lower bounds to measure the degree of convergence.

III. THE PERFORMANCE OF THREE RECEIVERS ON THE $\sqrt{f}$ CHANNEL

Results of a calculation of the performance of the MLD, DFE, and ZFE will now be reported for the $\sqrt{f}$ channel, for which the attenuation in decibels increases as the square root of frequency. The $\sqrt{f}$ channel is a good approximation to coaxial cable, as well as to some cables consisting of wire pairs, and for this reason it is of great practical interest.

Many present high-speed digital transmission systems use some form of linear equalization, and their performance will be reasonably well approximated by that of the ZFE. Thus, the comparison between the ZFE and the MLD gives us an indication of the size of the gap in performance between common transmission systems in use today and what could theoretically be achieved by much more complex receiver designs.* The comparison with the DFE is much less interesting, because the susceptibility of the DFE to decision errors is not included in the present analysis and, as will be shown shortly, is of such a magnitude on the $\sqrt{f}$ channel as to essentially invalidate the performance estimate we calculate.

* This comparison is, of course, very idealized. The only impairment we consider is additive Gaussian noise.

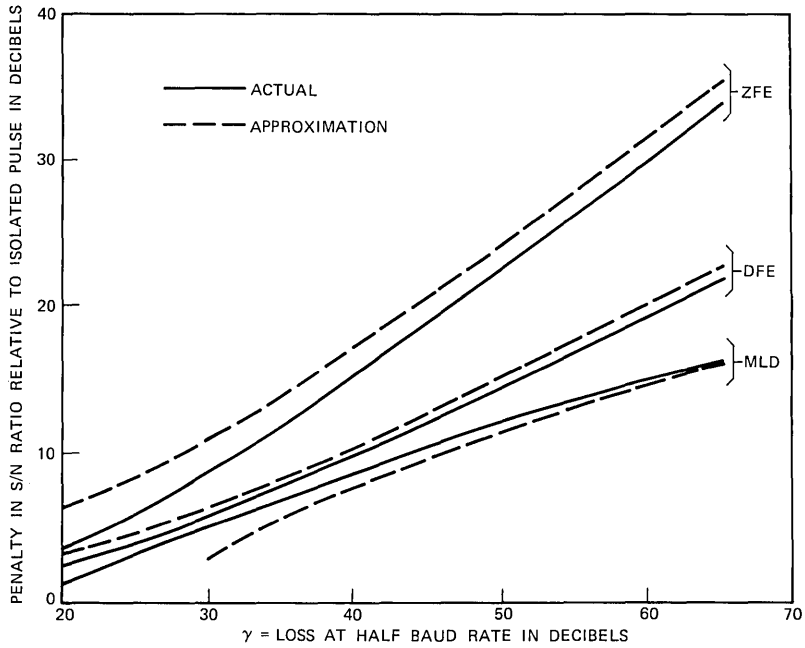


Fig. 2—Performance of three receivers on the $\sqrt{f}$ channel.

The power spectrum of the $\sqrt{f}$ channel is given by

$$|H(\omega)|^2 = 2\pi K^2 R_0 e^{-2K\sqrt{\omega}}, \tag{41}$$

where $H(\omega)$ is the frequency response of the channel and K is a parameter proportional to the line length. The usual convention is to designate the loss at the half-baud rate ($\omega = \pi/T$),

$$\gamma = -10 \log \frac{\left|H\left(\frac{\pi}{T}\right)\right|^2}{|H(0)|^2} \text{ (dB)}, \tag{42}$$

in which case

$$K = \sqrt{\frac{T}{\pi}} \frac{\gamma}{20 \log e}. \tag{43}$$

The effective penalties in S/N ratio relative to the isolated pulse bound can be calculated for the ZFE and DFE using the methods of Ref. 1, and for the MLD using the methods developed in Section II. The result is shown in Fig. 2 for the range of γ of practical interest. Most high-speed transmission systems in use today have a γ less than

about 65 dB because of limitations in the maximum gain which can be incorporated into a repeater without excessive coupling of the output back into the input.

One interesting feature of Fig. 2 is that even the MLD has a substantial S/N ratio penalty (15 dB) on the $\sqrt{f}$ channel. Thus, Forney's statement[3] that on most channels intersymbol interference does not have to lead to a significant degradation in performance does not apply to channels with very severe intersymbol interference, such as are commonly used in high-speed transmission systems.

The value of $d^2_{\min}$, valid for Fig. 2, as well as many other examples considered by this author and Forney,[4] is

$$d^2_{\min} = 2(R_0 - R_1), \tag{44}$$

where R_k is the autocorrelation of the received pulse.* An approximation to (44) valid for large γ is derived in Appendix A and plotted in Fig. 2 as a dotted line. Approximations to the S/N ratio penalty of the ZFE and DFE are also derived in Appendix A and plotted in Fig. 2. An intuitive interpretation of eq. (44) is given in Appendix B.

As an illustration of the speed of convergence of (34), the sequence of upper and lower bounds is illustrated in Fig. 3 for a $\sqrt{f}$ channel with $\gamma = 60$ dB. These bounds are within 1 dB for $N = 1$ and 0.5 dB for $N = 3$. Thus, convergence is very rapid, even for severe intersymbol interference.

A word of caution is in order with respect to the curve for the DFE in Fig. 2. This curve does not take into account the effect of decision errors on the performance of the receiver. The DFE subtracts, prior to the decision threshold on data digit B_k, the quantity

$$\sum_{m=1}^{\infty} b_m \hat{B}_{k-m}, \tag{45}$$

where $\hat{B}_{k-m}$ is the receiver's previous decision on $\hat{B}_{k-m}$ and b_m is the tap-gain of the DFE feedback filter. The resulting quantity which is applied to the threshold is[1]

$$b_0 B_k + \sum_{m=1}^{\infty} b_m (B_{k-m} - \hat{B}_{k-m}) + n_k, \tag{46}$$

where n_k is a noise sample. Whenever the b_m's are large with respect to b_0, a single decision error will likely cause many more errors. The

* This corresponds to the error sequence $(1, -1, 0, 0, \cdots)$ or, in the notation of Forney, $(1 - D)$.

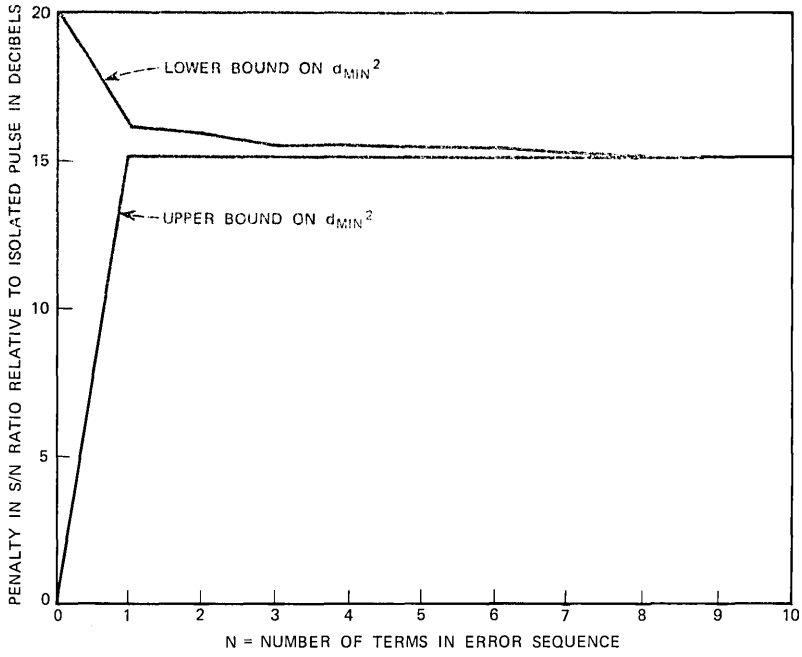


Fig. 3—Convergence of lower and upper bounds on $d_{\min}^2$ ($\sqrt{f}$ channel with $\gamma = 60$ dB).

coefficients of (46), given by (9), are tabulated in Table I for several values of γ.

Needless to say, the situation is hopeless for the large γ; the effect of a single decision error will be major and will last for a long time. Even for $\gamma = 20$, the reduction in noise margin resulting from a pre-

TABLE I—COEFFICIENTS OF THE DFE FEEDBACK FILTER (b_m)

m	b_m		
	$\gamma = 20$	$\gamma = 40$	$\gamma = 60$
0	1	1	1
1	0.61	1.4	2.2
2	0.36	1.3	2.8
3	0.25	1.1	2.9
4	0.18	0.94	2.9
5	0.14	0.80	2.8
10	0.06	0.42	1.9
47	0.006	0.06	0.38
174	0.001	0.009	0.06

vious decision error will be significant for five or ten subsequent decisions. We must conclude, then, that Fig. 2 will not be representative of the true performance of the DFE, and further that the DFE may not be a suitable receiver for the $\sqrt{f}$ channel*.

In terms of repeater spacing and baud rate, Fig. 1 can be interpreted in two ways. If the ZFE is replaced by an MLD, the same level of performance can be maintained while either increasing the repeater spacing with a constant baud rate or increasing the baud rate with the same repeater spacing. To illustrate this, consider the example of a ZFE operating at a given level of performance on a $\sqrt{f}$ channel with $\gamma = 40$ dB. Then γ can be increased to 60 dB at the same effective S/N ratio. This corresponds to a 50-percent increase in repeater spacing at a constant baud rate (since γ goes up linearly with the repeater spacing). However, since the repeater spacing has increased, the transmitted power must also be increased by 3.5 dB to maintain a constant isolated pulse energy at the receiver.†

If the repeater spacing is held constant, an increase in baud rate by a factor of $(1.5)^2$, or 125 percent, will also result in a 50-percent increase in γ. Here too, the average (but not peak) transmitted power is increased by 3.5 dB.

The conclusion of these results is that there is a fairly large gap between the performance of linear equalizers and the theoretical limit on the $\sqrt{f}$ channel. It is probably fair to say, however, that practical constraints on repeater complexity, speed of operation, and gain makes the attainment of a substantial portion of this potential improvement on high-speed transmission systems very difficult, at least for the present. Such is not the case for low-speed applications, such as voiceband data, where the implementation of the MLD can be contemplated on the basis of existing technology.

IV. CONCLUSIONS

In this paper, the minimum distance measure has been interpreted geometrically, related to equalization (the decision-feedback equalizer in particular), and bounded in several ways. A practical numerical technique has been developed for calculating the minimum distance without considering unnecessarily long error sequences.

---

* Tomlinson[8] has invented a method of avoiding the error propagation problem by subtracting out interference from past data digits in the transmitter.

† The received pulse energy is proportional to γ^{-2}, so that the peak and average transmitted power must be increased by $20 \log(60/40) = 3.5$ dB.

Numerical results for the $\sqrt{f}$ channel reveal that the penalty in S/N ratio relative to the isolated pulse bound for the MLD can be substantial for this channel, and that the gap in performance between the MLD and linear equalization is also substantial. The latter suggests that further attempts at finding receivers without the complexity of the Viterbi algorithm MLD but which nevertheless improve on the performance of linear equalization might well be fruitful. The decision-feedback equalizer does not appear to fit this bill because of its serious error propagation problem when confronted with intersymbol interference as severe as that found on the $\sqrt{f}$ channel.

APPENDIX A

Autocorrelation of the $\sqrt{f}$ Channel

From (41), the autocorrelation is

$$R_k = \frac{1}{\pi}\int_0^\infty |H(\omega)|^2 \cos(\omega kT)d\omega$$

$$= \frac{4K^2R_0}{T}\int_0^\infty x \exp\left(-\frac{2K}{\sqrt{T}}x\right)\cos(kx^2)dx. \tag{47}$$

Integrating by parts with $u = \exp\left(-\frac{2K}{\sqrt{T}}x\right)$ and $dv = x\cos(kx^2)dx$, we get

$$R_k = \frac{4K^3R_0}{(kT)^{\frac{3}{2}}}\int_0^\infty \exp\left(-\frac{2K}{\sqrt{kT}}x\right)\sin x^2 dx,$$

which is given in terms of the Fresnel Integral,[9]

$$R_k = \sqrt{\frac{\pi}{2}}\frac{4K^3R_0}{(kT)^{\frac{3}{2}}}\left\{\left[\frac{1}{2} - C\left(\frac{K}{\sqrt{kT}}\sqrt{\frac{2}{\pi}}\right)\right]\cos\left(\frac{K^2}{kT}\right)\right.$$

$$\left. + \left[\frac{1}{2} - S\left(\frac{K}{kT}\sqrt{\frac{2}{\pi}}\right)\right]\sin\left(\frac{K^2}{kT}\right)\right\}, \tag{48}$$

where

$$C(x) = \int_0^x \cos\left(\frac{\pi}{2}y^2\right)dy$$

$$S(x) = \int_0^x \sin\left(\frac{\pi}{2}y^2\right)dy.$$

An accurate approximation to R_1 valid for large γ is easily obtained from (47) by substituting the first two terms of a Taylor series for

$\cos x^2$,

$$R_1 \cong \beta^2 R_0 \int_0^\infty x \left(1 - \frac{x^4}{2}\right) e^{-\beta x} dx$$

$$= R_0 \left(1 - \frac{60}{\beta^4}\right), \tag{49}$$

where

$$\beta = \frac{2K}{\sqrt{T}}.$$

Hence

$$2(R_0 - R_1) \cong \frac{120}{\beta^4}$$

and

$$-10 \log \frac{2(R_0 - R_1)}{R_0} \cong 40 \log \gamma - 56.2. \tag{50}$$

Approximations to $\|e_0\|^2$ and $\|e_0^+\|^2$ can also be derived by assuming that $H(\omega) = 0$, $|\omega| > \pi/T$, or equivalently that $|H(\omega)|^2 = R(\omega)$. The resulting S/N ratio penalties are

$$-10 \log \|e_0\|^2/R_0 \cong \gamma + 25.15 - 30 \log \gamma \tag{51}$$

$$-10 \log \|e_0^+\|^2/R_0 \cong \tfrac{2}{3}\gamma + 15.76 - 20 \log \gamma\ . \tag{52}$$

Equations (50) to (52) are plotted in Fig. 2 as dotted lines.

APPENDIX B

Interpretation of Equation (44)

It is straightforward to show that whenever

$$\frac{R_1}{R_0} \geqq 0.5 \tag{53}$$

we have

$$d_{\min}^2 \leqq 2(R_0 - R_1) \leqq R_0. \tag{54}$$

Noting that

$$R_1 = \langle h_0, h_1 \rangle = \|h_0\| \, \|h_1\| \cos \theta$$
$$= R_0 \cos \theta,$$

where θ is the angle between h_0 and h_1, eq. (53) becomes

$$\theta \leqq 60°. \tag{55}$$

The geometric interpretation of (55) is shown in Fig. 4, where it is seen that (54) is satisfied until $\theta = 60°$, when the triangles become

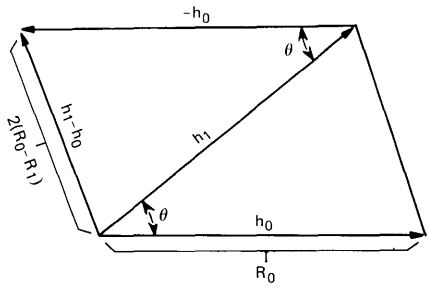


Fig. 4—Geometric interpretation of eq. (44).

equilateral. As long as (55) is satisfied, $h_0 - h_1$ is a shorter vector than h_0.

In the case of the $\sqrt{f}$ channel, R_1/R_0 is very close to unity. Thus, $h_0 - h_1$ is a very short vector. Although it will certainly not always be the case, a plausible explanation for the fact that longer error sequences do not yet yield a shorter vector is that the addition of other translates of h_k (such as $\pm h_2$) adds further components in other dimensions. Presuming that it does not reduce the component in the $h_0 - h_1$ plane, it can then only increase the length of the total vector.

REFERENCES

1. Messerschmitt, D. G., "A Geometric Theory of Intersymbol Interference. Part I: Zero-Forcing and Decision-Feedback Equalization," B.S.T.J., this issue, pp. 1483–1519.
2. Forney, G. D., Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-18,* May 1972, p. 363.
3. Forney, G. D., Jr., "The Viterbi Algorithm," Proc. IEEE, *61,* March 1973, p. 268.
4. Forney, G. D., Jr., "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," IEEE Trans. Commun., *COM-20,* February 1972, p. 76.
5. Messerschmitt, D. G., "A Unified Geometric Theory of Zero-Forcing and Decision-Feedback Equalization," 1973 Int. Conf. Commun., Seattle, June 1973.
6. Price, R., "Nonlinearly Feedback-Equalized PAM vs. Capacity for Noisy Filter Channels," 1972 Int. Conf. Commun., Philadelphia, June 1972.
7. Grenander, U., and Szego, G., *Toeplitz Forms and Their Applications,* Berkeley: University of California Press, 1958.
8. Tomlinson, M., "New Automatic Equalizer Employing Moduluo Arithmetic," Elec. Letters, *7,* March 1971, p. 138.
9. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions,* New York: Dover, 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Adaptive Channel Memory Truncation for Maximum Likelihood Sequence Estimation

By D. D. FALCONER and F. R. MAGEE, JR.

(Manuscript received May 15, 1973)

Maximum likelihood data sequence estimation, implemented by a dynamic programming algorithm known as the Viterbi algorithm (VA), is of considerable interest for data transmission in the presence of severe intersymbol interference and additive Gaussian noise. Unfortunately, the required number of receiver operations per data symbol is an exponential function of the duration of the channel impulse response, resulting in unacceptably large receiver complexity for high-speed PAM data transmission on many channels.

We propose a linear prefilter to force the overall impulse response of the channel/prefilter combination to approximate a desired truncated impulse response (DIR) of acceptably short duration. Given the duration of the DIR, the prefilter parameters and the DIR itself can be optimized adaptively to minimize the mean-square error between the output of the prefilter and the desired prefilter output, while constraining the energy in the DIR to be fixed.

In this work we show that the minimum mean-square error can be expressed as the minimum eigenvalue of a certain channel-dependent matrix, and that the corresponding eigenvector represents the optimum DIR. An adaptive algorithm is developed and successfully tested. The simulations also show that the prefiltering scheme, used together with the VA for two different channel models, compares favorably in performance with another recently proposed prefiltering scheme. Limiting results for the case where the prefilter is considered to be of infinite length are obtained; it is shown that the optimum DIR of length two must be one of two possible impulse responses related to the duobinary impulse response. Finally we obtain limiting results for the case where the transmitting filter is optimized.

I. INTRODUCTION

Forney[1] has recently proposed a receiver structure for a communication system operating over a known time-dispersive channel with little loss in performance due to intersymbol interference by using maximum likelihood sequence estimation, or the Viterbi algorithm (VA).[2] This has resulted in much attention being given to practical methods of applying his results. Magee and Proakis[3] proposed the use of the VA directly in conjunction with a channel estimator. This approach can result in a receiver too complex for practical use because the complexity of the VA depends exponentially on the duration of the channel impulse response.

In particular, if the impulse response of the channel has an effective duration of τ seconds and if an L-level PAM system transmits $1/T$ data symbols per second, the number of operations per received symbol is proportional to $L^{\tau/T}$. For channels such as voiceband telephone channels, the bandwidth of which is used efficiently, typical values of τ/T may be between about 20 and 200, making direct application of the VA infeasible.

Thus, it seems clear that effective practical application of the VA or of related techniques involves a compromise between optimum performance and receiver complexity. The complexity-limiting approach we take here is to use a linear prefilter at the receiver to "condition" the overall sampled impulse response seen by the VA so that it is significantly different from zero over only a small number of samples, and any remaining intersymbol interference is considered to be noise. Additional joint optimization of the transmitting filter is also treated, but would be much harder to implement in a real system.

The simplest example of a prefilter is a linear equalizer, which yields an approximate overall impulse response of just one sample. Another example of prefiltering for a different purpose is the linear portion of a decision-feedback equalizer; in that case the initial sample of the desired overall impulse response is required to be large relative to the additive noise.

In any application of prefiltering to approximate a desired impulse response (DIR), the DIR itself and the prefilter should be chosen to minimize the error due to noise and to the difference between the DIR and the actual impulse response that is achieved. The latter error results from intersymbol interference components outside the interval accounted for by the DIR samples as well as from errors in approximating the DIR inside the time-limited interval. This error could be

eliminated by using a zero-forcing criterion at the cost of additive noise.

Qureshi and Newhall[4] have recently proposed a receiver incorporating prefiltering with the VA. They use a mean-square error (MSE) criterion to force the overall response of the channel plus the linear equalizer to approximate a truncated version of the channel pulse response. In order to decode, the VA assumes this truncated response, resulting in much simplified processing. There is no effort made in Ref. 4 to optimize the desired truncated response. It is the purpose of this paper to see how this desired response can be chosen to minimize MSE and to show that this receiver structure can be made adaptive.

In Section II we formulate the MSE-minimization problem, assuming a fixed number of samples in the DIR and in the impulse response of the prefilter. The minimum achievable MSE is the minimum eigenvalue of a certain channel-dependent matrix. In Section III we indicate how the prefilter tap coefficients and the samples of the DIR can be determined adaptively by a gradient algorithm based on the MSE minimization. Section IV is a study of the limiting situation in which the tapped delay line prefilter consists of an infinite number of taps and it is preceded by a matched filter. Compact expressions for the prefilter impulse response, DIR, and minimum MSE are derived, which lend further insight. Section V describes the results of computer simulations of an adaptive prefilter/VA receiver structure, including comparison of the receiver with that described in Ref. 4 and with performance lower bounds. Section VI presents performance calculations for the prefilter/VA system, a decision-feedback equalizer, and a linear equalizer for a particular channel. Plots of minimum MSE versus bit rate for each of the three types of receiver structures are shown. Section VII considers asymptotic transmitter optimization.

II. OPTIMIZATION OF THE RECEIVER

The channel model and the preliminary receiver processing are shown in Fig. 1. The channel is modeled as a linear continuous filter with additive white Gaussian noise. It has been shown by Forney[1] that the channel can then be followed by a matched filter, symbol rate sampler, and noise whitening filter with no loss of information. Alternatively, the reader may assume that the channel is band-limited and symbol rate sampling can be used with no information loss.

Due to these considerations, the discrete-time model of Fig. 2 was adopted with the additional assumption that the channel pulse re-

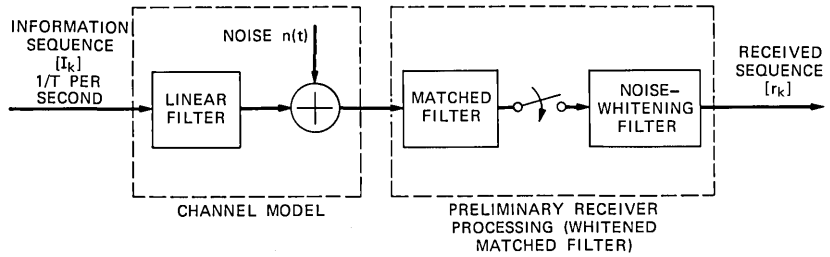


Fig. 1—Channel model and preliminary receiver processing.

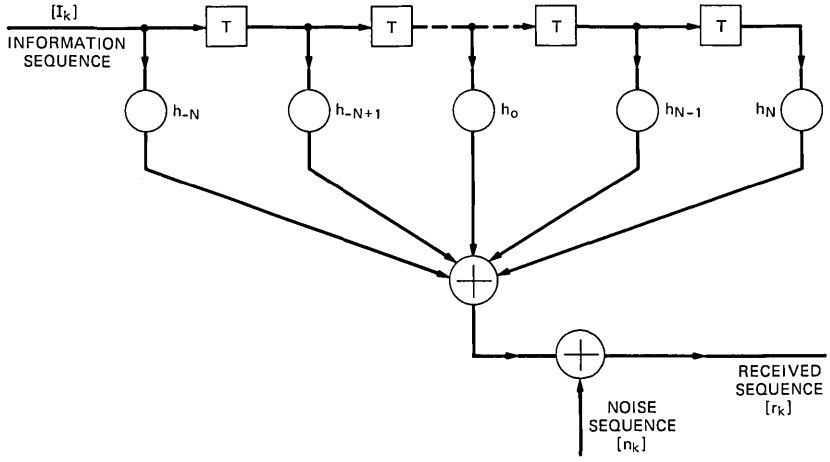


Fig. 2—Discrete-time channel model.

sponse is time-limited. The noise sequence $\{n_k\}$ is additive, uncorrelated, and Gaussian with variance σ^2. Note that a discrete-time model with uncorrelated noise samples also results from the commonly used but nonoptimum expedient of passing the received signal through a flat Nyquist band-limiting filter prior to sampling.*

The proposed receiver structure is shown in Fig. 3. The received sequence feeds a linear tapped delay line filter whose function is to shorten the overall impulse response length. The filter has $L(=2M+1)$ taps which are chosen in the manner to be described later. The output of this filter feeds the Viterbi algorithm which detects the information sequence.

* Although a white noise model was used throughout, the correlated noise case can be considered in a similar manner.

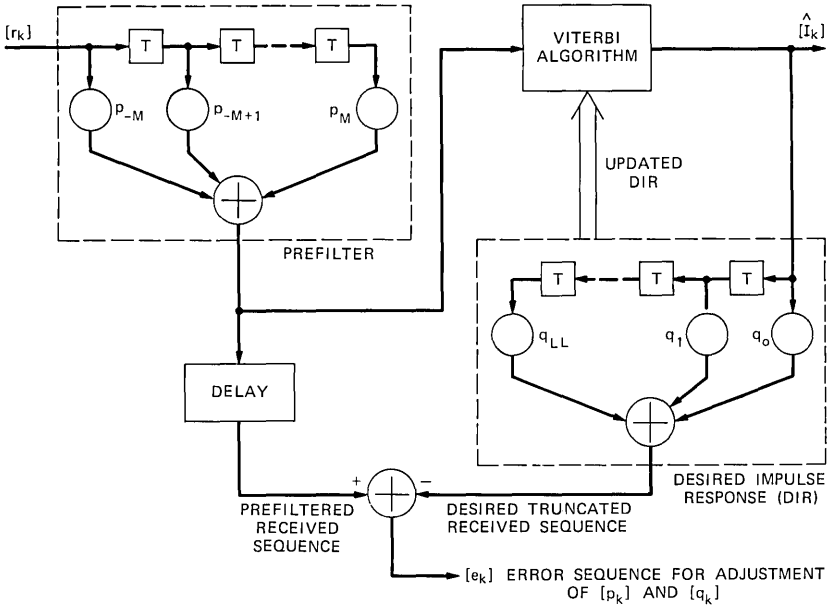


Fig. 3—Receiver structure.

The Viterbi algorithm makes decisions on the assumption that the DIR $\{q_k\}_{k=0}^{LL}$ is the actual overall channel response. The value of LL (the length of the DIR) is much less than $2N + 1$ (the length of the actual channel response). LL is chosen to make acceptable the complexity of the Viterbi algorithm while taking a small noise penalty in the linear preprocessing.* An error signal is formed by feeding the information sequence estimate through the tapped delay line representing the desired channel response. This forms the desired truncated channel received sequence which is then compared with a delayed version of the actual linear prefilter output to form an error sequence. It is this error which is to be minimized since it represents a sum of the additive noise, and the difference between the desired and actual overall impulse responses.

If the sampled channel impulse response, sequence of information symbols, and sequence of uncorrelated noise samples are represented respectively by $\{h_l\}_{l=-\infty}^{\infty}$, $\{I_l\}_{l=-\infty}^{\infty}$, and $\{n_l\}_{l=-\infty}^{\infty}$, then the kth dis-

* Obviously if LL is allowed to be very large, the DIR can closely approximate a delayed version of the original channel impulse response, and there is no significant noise penalty, since the prefilter simply approximates a delay tine.

crete channel output is

$$r_k = \sum_l h_l I_{k-l} + n_k. \tag{1}$$

Then if the vectors $\mathbf{P}^+ \equiv (p_{-M}, \cdots, p_0, \cdots p_M)$ where $+$ indicates transpose and $\mathbf{Q}^+ \equiv (q_o, \cdots, q_{LL})$ represent the tap coefficients of the prefilter and the DIR respectively, the error in the kth interval is

$$e_k = \sum_{l=-M}^{M} p_l r_{k-l} - \sum_{l=o}^{LL} q_l I_{k-l}. \tag{2}$$

In order to simplify the following equations, it is assumed that the information sequence is uncorrelated ($\overline{I_k I_j} = \delta_{kj}$) and that the information sequence estimate equals the information sequence. Substituting (1) into (2) and averaging e_k^2 we get

$$\overline{e_k^2} \equiv \overline{e^2} = \mathbf{P}^+ A \mathbf{P} + \mathbf{Q}^+ \mathbf{Q} - 2\mathbf{P}^+ H \mathbf{Q}, \tag{3}$$

where

$$H = \begin{bmatrix} h_M & \cdots h_{M+LL} \\ \vdots & \vdots \\ h_0 & \cdots h_{LL} \\ \vdots & \vdots \\ h_{-M} \cdots & h_{-M+LL} \end{bmatrix} \tag{4}$$

is an $L \times (LL + 1)$ matrix and A is an $(L \times L)$ channel covariance matrix with elements $a_{ij} = \overline{r_i r_j}$.

First, the error is minimized with respect to the prefilter by taking the gradient with respect to the taps $\{p_i\}$ and setting it equal to zero. The taps $\{q_i\}$ are constrained to be nonzero.

$$\frac{\partial \overline{e^2}}{\partial \mathbf{P}} = 2A\mathbf{P} - 2H\mathbf{Q} = \mathbf{0} \tag{5}$$

and therefore

$$\mathbf{P}_{\text{opt}} = A^{-1} H \mathbf{Q}. \tag{6}$$

The interpretation—thus far—is that some desired (and truncated) channel response is chosen; and the linear prefilter taps are chosen to force the overall response to this with a minimum MSE. The question of what this desired response should be naturally arises. If a fixed length is assumed for this desired response, the desired response can be optimized in the sense of minimizing MSE. Substituting (6) into (3), one obtains

$$\overline{e^2} = \mathbf{Q}^+ [I - H^+ A^{-1} H] \mathbf{Q}, \tag{7}$$

where I is the identity matrix.

Since $\overline{e^2} \geqq 0$, this is a positive definite quadratic form in $\mathbf{Q}$ which depends only upon $\mathbf{Q}$ and the channel characteristics. This can be minimized by choosing $\mathbf{Q}$ to be the eigenvector with the minimum eigenvalue of the matrix $[I - H^+A^{-1}H]$. The constraint

$$\mathbf{Q}^+\mathbf{Q} = 1 \tag{8}$$

is necessary to avoid the trivial case of no MSE. The trivial case corresponds, of course, to no transmission through the channel.

It should be noted that when the MSE is minimized a reasonable definition of the signal-to-noise ratio (SNR) seen by the Viterbi algorithm is maximized. This is true because $\{q_i\}$ is considered to be the effective channel pulse response, constrained to unit energy; the additive noise plus any residual intersymbol interference is the effective noise seen by the algorithm. Since this noise is equal to the MSE which has been minimized, the SNR has been maximized.

In summary, to minimize the MSE and thus maximize the SNR seen by the VA receiver, choose

$$\mathbf{Q}_{\text{opt}} = \text{eigenvector of } [I - H^+A^{-1}H] \text{ corresponding to its minimum eigenvalue,} \tag{9}$$

$$\mathbf{P}_{\text{opt}} = A^{-1}H\mathbf{Q}_{\text{opt}}, \tag{10}$$

and then

$$\overline{e^2}\ \min = \min \text{ eigenvalue of } [I - H^+A^{-1}H]. \tag{11}$$

III. AN ADAPTIVE ALGORITHM FOR OPTIMUM RECEPTION

In order to make the receiver structure practical, the procedure of choosing the $\{p_i\}$ and the $\{q_i\}$ must be made adaptive since the channel pulse response will not usually be known prior to the start of transmission. An algorithm to choose the taps adaptively will now be described.

Consider the conditions for the optimum operating point of this receiver to be reached. The condition that the gradient with respect to $\mathbf{P}$ be equal to zero is easily implemented by using the products of sampled values of quantities in the receiver as noisy estimates of the required cross correlations, assuming the data sequence is known or has been correctly estimated by the receiver. Thus,

$$\mathbf{P}^{(r+1)} = \mathbf{P}^{(r)} - \Delta_1 e_r \mathbf{R}^{(r)}, \tag{12}$$

where $\mathbf{P}^{(r)}$ is the set of tap values at the rth iteration, Δ_1 is an adjustment parameter which controls accuracy and speed of convergence,

$\mathbf{R}^{(r)}$ is a vector of the received samples contained in the linear preprocessing filter, and e_r is the error in (2). $\mathbf{P}^{(r+1)}$ is thus the new estimate of the $\{p_i\}$ taps, and when a steady state is reached a noisy unbiased estimate of these taps is obtained. Note that the value of $\mathbf{P}$ implicitly depends on the value of $\mathbf{Q}$ through the e_r terms.

The algorithm to obtain the $\mathbf{Q}$ taps is not so easily obtained. Consider the unconstrained gradient with respect to the $\mathbf{Q}$ vector. If a noisy estimate of the required cross correlation is used, then the recursion for the unconstrained gradient algorithm is

$$\mathbf{Q}^{(r+1)} = \mathbf{Q}^{(r)} + \Delta_2 e_r \mathbf{I}^{(r)}, \tag{13}$$

where $\mathbf{I}^{(r)}$ is a vector of the information symbols contained in the channel reference filter. If (12) and (13) are followed without a constraint at each iteration, then the trivial solution results. The algorithm is therefore modified so that the $\mathbf{Q}$ vector is renormalized at each step. That is,

$$\tilde{\mathbf{Q}}^{(r+1)} = \mathbf{Q}^{(r)} + \Delta_2 e_r \mathbf{I}^{(r)} \tag{14}$$

$$\mathbf{Q}^{(r+1)} = \frac{\tilde{\mathbf{Q}}^{(r+1)}}{(\tilde{\mathbf{Q}}^{(r+1)})^{+}(\tilde{\mathbf{Q}}^{(r+1)})}. \tag{15}$$

By following the combined algorithm of (12), (14), and (15), a stationary point in $\mathbf{P}$ will be reached, and the energy in $\mathbf{Q}$ will be constrained to one.

Now consider the noiseless unconstrained gradient of $\overline{e^2}$ with respect to $\mathbf{Q}$. Then

$$\frac{\partial \overline{e^2}}{\partial \mathbf{Q}} = 2\mathbf{Q} - 2H^{+}\mathbf{P}. \tag{16}$$

Consider $\mathbf{P}$ to be in the neighborhood of the correct solution (6) with respect to $\mathbf{Q}$ (that is, $\mathbf{P}$ is adjusted more quickly than $\mathbf{Q}$). Then (16) becomes

$$\frac{\partial \overline{e^2}}{\partial \mathbf{Q}} = 2\mathbf{Q} - 2H^{+}A^{-1}H\mathbf{Q}. \tag{17}$$

Thus the gradient algorithm, in terms of the actual matrix quantities, becomes

$$\begin{aligned}\tilde{\mathbf{Q}}^{(r+1)} &= \mathbf{Q}^{(r)} - \tfrac{1}{2}\Delta_2 \frac{\partial \overline{e^2}}{\partial \mathbf{Q}^{(r)}} \\ &= \mathbf{Q}^{(r)} - \tfrac{1}{2}\Delta_2(2\mathbf{Q}^{(r)} - 2H^{+}A^{-1}H\mathbf{Q}^{(r)}) \\ &= \Delta_2 H^{+}A^{-1}H\mathbf{Q}^{(r)} + \mathbf{Q}^{(r)}(1 - \Delta_2)\end{aligned} \tag{18}$$

and then $\tilde{\mathbf{Q}}^{(r+1)}$ is renormalized to form $\mathbf{Q}^{(r+1)}$. Now note that if

$\Delta_2 = 1$ this corresponds exactly to the method of Vianello and Stodola[5] for determining the maximum eigenvalue and corresponding eigenvector of $H^+A^{-1}H$. Since the maximum eigenvalue of $H^+A^{-1}H$ corresponds to the minimum eigenvalue of $(I - H^+A^{-1}H)$ this technique will converge to the minimum MSE. This method will fail only when the starting vector $\mathbf{Q}^{(1)}$ is exactly orthogonal to the desired solution. Since the algorithm actually used (14)–(15) deals with noisy estimates rather than the exact expressions, the noise will prevent the case of the algorithm becoming stuck on a vector orthogonal to the solution.

In the practical case it is not possible to choose Δ_2 to be one because when the noisy estimates are used the algorithm will amplify the noise and diverge. Actually, Δ_2 will be much smaller than one. Again looking at (18), one can see that a steady state is reached when $\mathbf{Q}$ becomes nonrotating with respect to the transformation. This occurs when $\mathbf{Q}$ is the maximum eigenvalue of $H^+A^{-1}H$ (i.e., the maximum eigenvalue will dominate as in the method of Vianello and Stodola). Thus, the unique solution for $\mathbf{Q}$ has been obtained.

IV. LIMITING RESULTS

We now study the limiting situation where the prefilter is allowed to be any general linear filter with impulse response $p(t)$, while the desired impulse response $\{q_m\}_{m=0}^{LL}$ is still finite. In addition we assume that the additive noise on the channel is white, with double-sided power spectral density $N_o/2$.

In this case we wish to minimize the mean square of the sampled error

$$e_k = \int_{-\infty}^{\infty} p(\tau)r(kT - \tau)d\tau - \sum_{l=o}^{LL} q_l I_{k-l}\,, \tag{19}$$

where

$$r(kT - \tau) = \sum_{l=-\infty}^{\infty} h(kT - lT - \tau)I_l + n(kT - \tau) \tag{20}$$

is the received signal.

The MSE is then

$$\overline{e^2} = \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\tau_1)p(\tau_2)h(lT - \tau_1)h(lT - \tau_2)d\tau_1 d\tau_2 + \frac{N_o}{2}\int_{-\infty}^{\infty} p(\tau)^2 d\tau - 2\sum_{l=o}^{LL} q_l \int_{-\infty}^{\infty} p(\tau)h(lT - \tau)d\tau + \sum_{l=o}^{LL} q_l^2 \cdot \tag{21}$$

Using a simple calculus of variations argument to minimize MSE with respect to the prefilter impulse response $p(t)$, we get the following integral equation defining the optimum $p(t)$.

$$\frac{N_o}{2} p(t) = \sum_{l=o}^{LL} q_l h(lT - t) - \sum_{l=-\infty}^{\infty} S_l h(lT - t) , \tag{22}$$

where

$$S_l = \int_{-\infty}^{\infty} p(\tau) h(lT - \tau) d\tau \tag{23}$$

is the overall sampled impulse response of the channel and prefilter. Note that we would hope for $0 \leqq l \leqq LL$, $S_l \approx q_l$ and for $l < 0$ and $l > LL$, $S_l \approx 0$.

Equation (22) tells us that the optimum prefilter structure is a matched filter with impulse response $h(-t)$, followed by an infinite-length tapped delay line whose tap gains $\{p_l\}$ are given by (22) and (23).

$$\frac{N_o}{2} p_l = q_l - \sum_{l=-\infty}^{\infty} p_l \phi_{m-l} , \tag{24a}$$

where

$$\phi_m = \int_{-\infty}^{\infty} h(mT - \tau) h(-\tau) d\tau = \phi_{-m} \tag{24b}$$

is the channel's sampled covariance function, and where we later require that $\{q_l\}$ is nonzero only for $0 \leqq l \leqq LL$. We remark that the development so far is analogous to that of Berger and Tufts[6] for the case $LL = 0$. Equation (24) may be solved in terms of z-transforms. Defining

$$q(z) \equiv \sum_{m=-\infty}^{\infty} q_m z^m$$

$$p(z) = \sum_{m=-\infty}^{\infty} p_m z^m$$

$$\phi(z) = \sum_{m=-\infty}^{\infty} \phi_m z^m$$

we can take z-transforms of both sides of (24) and solve for $p(z)$.

$$p(z) = \frac{q(z)}{\phi(z) + \frac{N_o}{2}} , \tag{25}$$

where we have used the fact that $\phi(z) = \phi(z^{-1})$ since the sequence $\{\phi_l\}$ is symmetric about $l = o$.

Using (25) we get the z-transform of the autocovariance sequence of the $\{e_k\}$, when the tap coefficients $\{p_n\}_{-\infty}^{\infty}$ are chosen to minimize the mean-squared error. Defining $E_m = \overline{e_k e_{k+m}}$ and $E(z) = \sum_{m=-\infty}^{\infty} E_m z^m$ we have

$$E(z) = \frac{N_o}{2} \frac{q(z)q(z^{-1})}{\phi(z) + \frac{N_o}{2}}. \tag{26}$$

We now minimize $\overline{e^2} = E_o$ with respect to the desired impulse response samples $\{q_n\}_{n=o}^{LL}$, under an appropriate energy constraint. Taking the inverse transform of $E(z)$ we have

$$\begin{aligned} E_o &= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} E(e^{-j\omega T}) d\omega \\ &= \frac{N_o T}{4\pi} \int_{-\pi/T}^{\pi/T} \frac{|q(e^{j\omega T})|^2}{\phi(e^{-j\omega T}) + \frac{N_o}{2}} d\omega , \end{aligned} \tag{27}$$

where

$$q(e^{j\omega T}) = q_o + q_1 e^{j\omega T} + \cdots + q_{LL} e^{j\omega LLT}.$$

Defining the $LL + 1$ dimensional vector $\mathbf{Q}^+ = (q_o, q_1, \cdots q_{LL})$ we can rewrite (27) as a quadratic form

$$E_o = \mathbf{Q}^+ R \mathbf{Q}, \tag{28}$$

where R is a square matrix of dimension $(LL + 1)$ whose i–jth element is

$$r_{ij} = \frac{N_o T}{4\pi} \int_{-\pi/T}^{\pi/T} \frac{e^{j\omega(i-j)T}}{\phi(e^{j\omega T}) + \frac{N_o}{2}} d\omega. \tag{29}$$

Note that $\phi(e^{j\omega T})$ is the discrete Fourier transform of an autocovariance sequence, and hence is an even, real, positive function of ω. Thus $r_{ij} = r_{ji}$ is a real function of $|i - j|$, and so R is a positive definite symmetric Toeplitz matrix.

Minimization of E_o under the energy constraint $|\mathbf{Q}|^2 = 1$ is then accomplished by making $\mathbf{Q}$ that normalized eigenvector of R corresponding to its minimum eigenvalue. The matrix R is evidently the limiting case of the matrix $I - H^+A^{-1}H$ for the finite-tap receiver [displayed in expressions (7) through (11)]. Then

$$\overline{e^2}_{\min} = \lambda_{\min}(R). \tag{30}$$

To recapitulate, the minimum is taken over the set of tap-coefficients

$$\{p_n\}_{n=-\infty}^{\infty} \text{ and } \{q_n\}_{n=o}^{LL} \text{ under the constraint } \sum_{n=o}^{LL} q_n^2 = 1,$$

which can be expressed as

$$\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} |q(e^{j\omega T})|^2 d\omega = 1.$$

Thus, from (27) we have the lower bound

$$\lambda_{\min}(R) = E_o = \overline{e}_{\min}^2 \geqq \frac{N_o}{2} \frac{1}{\sup\limits_{-\pi/T \leqq \omega \leqq \pi/T} \left(\phi(e^{j\omega T}) + \frac{N_o}{2}\right)}. \tag{31}$$

Now $\phi(e^{j\omega T})$ is the discrete Fourier transform of the sequence $\{\phi_n\}$ defined in terms of the channel's impulse response by (24b). Thus, if the channel's transfer function is denoted by

$$H(\omega) = \int_{\infty}^{\infty} h(t)e^{-j\omega T}dt\,,$$

$$\phi(e^{j\omega T}) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \left| H\left(\omega + \frac{2n\pi}{T}\right)\right|^2. \tag{32}$$

The term $\phi(e^{j\omega T})$ can be interpreted as the channel's "folded" power spectrum.[7]

When $LL + 1$, the number of components in the desired impulse response $\{q_n\}_{n=o}^{LL}$, is relatively small, say less than 10, the minimum eigenvalue and corresponding eigenvector of R can be evaluated without difficulty. For much longer values of LL, the lower bound (31) which is easily computed using (32) may be quite tight. A particular case of interest is where $LL = 1$. Then R has the form

$$R = \begin{bmatrix} r_o & r_1 \\ r_1 & r_o \end{bmatrix}$$

and

$$\overline{e}_{\min}^2 = \lambda_{\min}(R) = \min(r_o + r_1, r_o - r_1),$$

where

$$r_o + r_1 = \frac{N_o T}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{\cos^2\frac{\omega T}{2}}{\phi(e^{j\omega T}) + \frac{N_o}{2}}\, d\omega \tag{33}$$

and

$$r_o - r_1 = \frac{N_o T}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{\sin^2\frac{\omega T}{2}}{\phi(e^{j\omega T}) + \frac{N_o}{2}}\, d\omega. \tag{34}$$

The normalized eigenvectors (optimum (q_o,q_1)) corresponding to the eigenvalues $r_o + r_1$ and $r_o - r_1$ are respectively $(1/\sqrt{2}, 1/\sqrt{2})$ and $(1/\sqrt{2}, -1/\sqrt{2})$.

Thus, we have the curious result that the optimum desired impulse response of length two is one of only two possible forms, depending only on whether the channel's folded power spectrum is such that (33) or (34) is smaller. For example, if the channel's folded power spectrum has a single minimum near the band edge, $\omega = \pi/T$, the best choice for (q_o,q_1) would be $(1/\sqrt{2}, 1/\sqrt{2})$ since $\cos^2(\omega T)/2$ has a zero at the band edge. However, if the channel's folded power spectrum has a single minimum near zero frequency, the best choice for (q_o,q_1) would be $(1/\sqrt{2}, -1/\sqrt{2})$, since $\sin^2(\omega T)/2$ is zero at $\omega = o$. These two cases are illustrated in Fig. 4.

It is interesting to point out that the two possible optimum desired impulse responses $(1/\sqrt{2}, 1/\sqrt{2})$ and $(1/\sqrt{2}, -1/\sqrt{2})$ are reminiscent of duobinary and partial response impulse responses.[8]

V. PERFORMANCE OVER SIMULATED CHANNELS

In order to observe performance obtainable from this receiver structure, the arbitrary discrete time channels shown in Fig. 5 were used. Figure 6 shows the results of the simulations performed with the receiver developed here and that of Qureshi and Newhall on these channels. Underlined in Fig. 5 are the desired response used for the

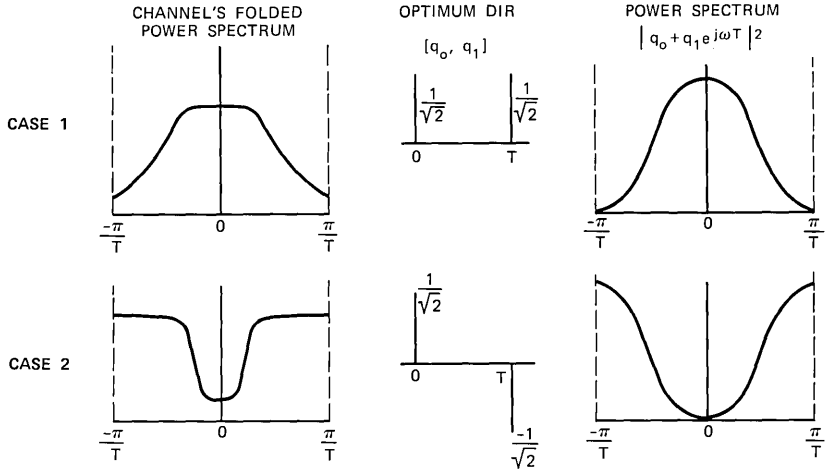


Fig. 4—Optimum desired impulse response of length two.

-0.1, -0.2, -0.1, 0.3, -0.3, 0.33, 0.67, 1.0, 0.67, 0.33, 0.3, -0.5, -0.1, 0.1, 0.1

CHANNEL A

-0.1, -0.1, -0.3, -0.15, -0.33, 0.16, 0.33, 1.0, 0.33, 0.16, -0.33, -0.3, -0.15, -0.1, -0.1

CHANNEL B

Fig. 5—Sampled channel impulse responses used in the simulation.

Qureshi and Newhall receiver. As can be seen from the performance curves, the Qureshi and Newhall receiver performs about as well as our receiver for Channel A and much worse for Channel B. The difference in performance is presumably due to the use of different criteria to choose the desired response. In the case of Channel B, the channel passes virtually no dc, yet the DIR from truncating the channel response does pass dc. This causes considerable noise enhancement by the Qureshi and Newhall linear prefilter.

Figure 6 also shows the matched filter lower bound, the lower bound on performance derived by Forney,[9] and a lower estimate which is used to predict actual optimum reference receiver performance. This lower estimate is obtained by computing the MSE and minimum coding distance of the DIR. Thus, it is assumed that the MSE is uncorrelated and Gaussian in this approximation.

$$P(e) \lesssim K \operatorname{erfc}\left(\frac{1}{2}\sqrt{\frac{d_{\min}^2}{2\ \mathrm{MSE}}}\right), \tag{35}$$

where K is a constant depending on the error structure of the channel, and $d_{\min}$ is the minimum Euclidian distance between all possible pairs of noiseless sequences with differing first information symbols emerging from the prefilter.[9] This lower estimate is found without considering the fact that the noise is correlated. If a more accurate estimate of performance is desired, the results of Qureshi and Newhall[4] can be used to consider the effects of noise correlation.

The simulations were run with a 31-tap prefilter whose taps were adjusted with $\Delta = 0.001$, and a 5-tap desired overall response length with the adjustment parameter equal to 0.01. In the case of the Qureshi and Newhall receiver, the prefilter was adjusted with $\Delta = 0.001$ and the channel was estimated with a filter with adjustment parameter equal to 0.01.* As the curves show, the receiver structure given here

* The performance loss due to the adjustment parameters has not been evaluated; however, simulation results indicate that this loss is very small.

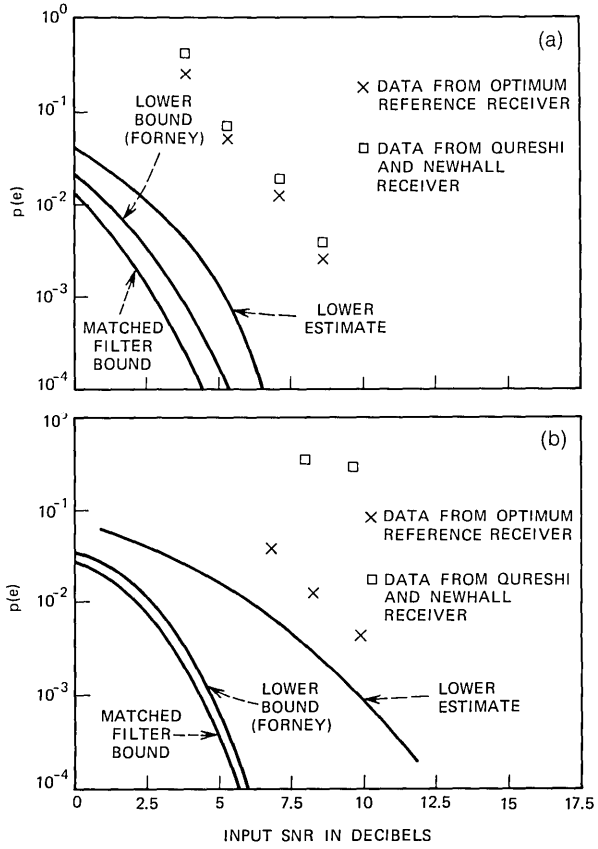


Fig. 6—Simulation results: (a) channel A; (b) channel B.

assures that a good choice of a DIR is made which is not always the same as the truncated channel impulse response.

VI. COMPARISON WITH OTHER SYSTEMS—AN EXAMPLE

Based on the results in Section II, performance calculations were made for baseband PAM transmission on the channel whose frequency response is shown in Fig. 7. The results shown in Fig. 8 were made with the following assumptions:

(*i*) A matched filter preceded the receiver.
(*ii*) There was a 31-tap prefilter.
(*iii*) There was a 5-tap desired impulse response.

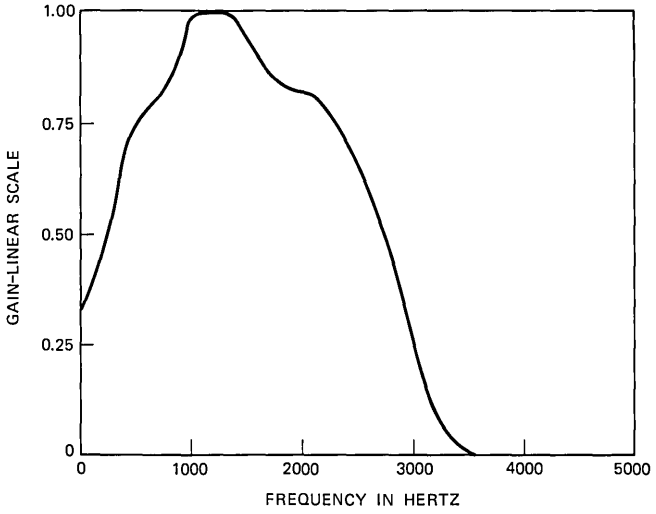


Fig. 7—Channel amplitude characteristic.

(*iv*) Although the noise may be correlated, at the input to the Viterbi algorithm, it has a negligible effect on performance.

Of the assumptions made, only the one about the noise correlation might not be realistic. Work is currently being done to deal with the correlated noise problem. In any case, it is not expected that it would affect performance more than a few dB and it clearly would not affect the place in the performance curve at which the performance begins to degrade seriously.

The curves representing the linear and decision-feedback equalizers, provided by J. Salz, show the MSE versus rate for additive white Gaussian noise with $N_o/2 = 0.0001$. In the linear and decision-feedback cases the MSE may be roughly related to performance in terms of probability of error.* [6,10] The curve for the prefilter/VA combination, labeled "VA equalizer," is a plot of $(\text{MSE}/d^2_{\text{min}})$ versus rate, where d_{min} is the minimum distance for the DIR. This is done because the attainable system performance is not determined by MSE alone, but rather by $\text{MSE}/d^2_{\text{min}}$ as in expression (35). Direct minimization of this ratio by analytical or numerical means has not been accomplished. Note however that the minimum value d_{min} can attain (over all

* The analysis for decision-feedback equalization ignores the effect of decision errors on the MSE.

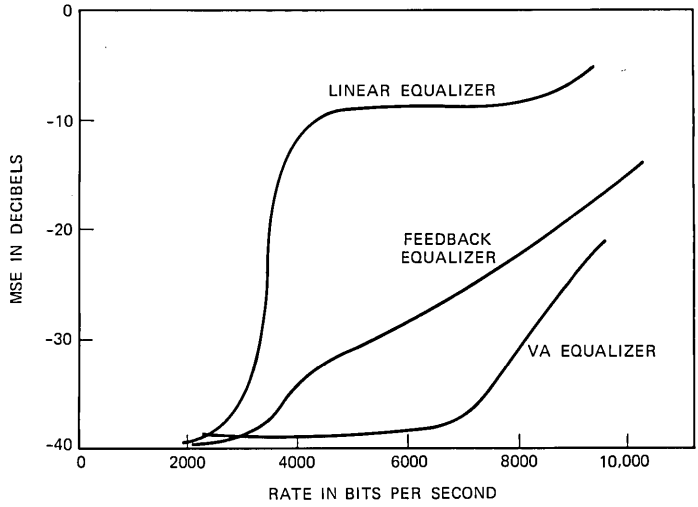


Fig. 8—Indication of attainable performance for three receiver structures.

channels with equal-energy impulse responses) is limited by the duration of the DIR which is chosen.[11]

As can be seen from the curves, this receiver structure can be expected to perform well while using only binary signaling over a much greater range of transmitted data rates than the linear and decision-feedback receivers. This result occurred despite the fact that the linear and decision-feedback computations were made for infinite filters while the prefilter was finite. It is the more relaxed criterion for our system compared to the decision-feedback criterion which results in lower MSE and thus better performance. Nevertheless the results are considered preliminary until a better understanding of the effect of noise correlation is achieved.

VII. TRANSMITTER OPTIMIZATION

The "channel's" frequency response $H(\omega)$ actually includes the transmitting filter, i.e.,

$$H(\omega) = C(\omega)G(\omega), \tag{36}$$

where $C(\omega)$ is the frequency response of the transmission channel alone, and $G(\omega)$ is the frequency response of the transmitting filter, which we have hitherto assumed fixed. In a practical data communication system, a "reasonable" transmitting filter would likely be fixed to avoid having to provide an extra feedback channel for adjust-

ing the transmitter parameters, and because of the complexity of the transmitter optimization argument itself for general channels.[6]

Nevertheless, the performance attainable with transmitter optimization is of theoretical interest. In this section we obtain expressions for the optimum transmitter filter $G(\omega)$ and the resulting minimum MSE under a transmitted power constraint. For simplicity, we assume a "well behaved" channel $C(\omega)$ for which $|C(\omega)|$ is monotone decreasing, and for which $|C(\omega)|/N_o$ is sufficiently large in the range $\{-\pi/T, \pi/T\}$ that the optimum transmitter uses the entire Nyquist band $|\omega| \leqq \pi/T$. Treatment of more general channel characteristics is more complicated, but can be carried out as in Ref. 6.

The minimum MSE for a fixed transmitting filter $G(\omega)$ and DIR $\{q_l\}$ is given by eq. (27) and by the channel's folded power spectrum, which from (32) and (36) can be written

$$\phi(e^{j\omega T}) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \left| C\left(\omega - \frac{2\pi n}{T}\right) \right|^2 \left| G\left(\omega - \frac{2\pi n}{T}\right) \right|^2. \tag{37}$$

The constraint that the transmitted power be fixed at P_T can be written

$$\frac{1}{T} \int_{-\pi/T}^{\pi/T} \sum_n \left| G\left(\omega - \frac{2\pi n}{T}\right) \right|^2 d\omega = P_T. \tag{38}$$

A necessary condition for minimizing the MSE, given by (27), subject to the power constraint (38), is obtained using a simple variational argument: for $-\pi/T \leqq \omega \leqq \pi/T$ and every integer m, either $G(\omega - 2\pi m/T) = 0$ or $G(\omega - 2\pi m/T) \neq 0$ and

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} \left| C\left(\omega - \frac{2\pi n}{T}\right) \right|^2 \left| G\left(\omega - \frac{2\pi n}{T}\right) \right|^2 + \frac{N_o}{2} = \lambda |q(e^{j\omega T})| \left| C\left(\omega - \frac{2\pi m}{T}\right) \right|, \tag{39}$$

where λ is a Lagrange multiplier whose value will be determined from the constraint (38). Furthermore, for any ω such that $C(\omega) = 0$, $G(\omega) = 0$.

For any frequency ω, there will be only one integer m for which $G(\omega - 2\pi m/T) \neq 0$, since the left-hand side of (39) does not depend on m and the right-hand side does. Indeed, if $|C(\omega)|$ is monotone decreasing, then best use is made of the transmitter power if $G(\omega) = 0$ for $|\omega| > \pi/T$. Thus we can rewrite (39) as

$$\frac{1}{T} |C(\omega)|^2 |G(\omega)|^2 + \frac{N_o}{2} = \lambda |q(e^{j\omega T})| |C(\omega)| \quad \text{for } C(\omega) \neq 0 \text{ and } |\omega| \leqq \pi. \tag{40}$$

For simplicity, we assume that P_T and the ratio $|C(\omega)|/N_o$ are sufficiently large that (40) can be satisfied for all $|\omega| < \pi/T$. [Otherwise $G(\omega)$ would be zero[6] beyond a certain frequency $\omega_0 < \pi/T$.] Then the amplitude frequency response of the optimum transmitting filter is given by

$$\frac{1}{T}|G_{\text{opt}}(\omega)|^2 = \frac{\lambda|q(e^{j\pi T})|}{|C(\omega)|} - \frac{N_o}{2|C(\omega)|^2} \qquad \text{for } |\omega| \leqq \frac{\pi}{T}$$

$$= 0 \qquad \text{for } \omega > \frac{\pi}{T}. \tag{41}$$

The Lagrange multiplier λ is determined by (41) and the power constraint (38). Substitution of the expression for $|G_{\text{opt}}|$ into expression (27) for the MSE gives

$$\text{MSE} = \frac{N_o T}{4\pi\lambda} \int_{-\pi/T}^{\pi/T} \frac{|q(e^{j\omega T})|}{|C(\omega)|}\, d\omega\,, \tag{41a}$$

where

$$\lambda = \frac{P_T + \dfrac{N_o}{2}\displaystyle\int_{-\pi/T}^{\pi/T} \frac{1}{|C(\omega)|^2}\, d\omega}{\displaystyle\int_{-\pi/T}^{\pi/T} \frac{|q(e^{j\omega T})|}{|C(\omega)|}}. \tag{41b}$$

It is interesting to look now at the frequency response of the optimum receiver prefilter

$$\frac{1}{T}\,|P_R(\omega)|^2 \equiv \frac{1}{T}\,|C(\omega)|^2\,|G_{\text{opt}}(\omega)|^2\,|p(e^{j\omega T})|^2\,, \tag{42}$$

corresponding to the optimum transmitter filter. The left side of (42) follows from the cascade of the channel and transmitter and the appropriate matched filter, followed by the discrete filter $\{p_l\}$. Substitution of expressions (41) for $|G_{\text{opt}}(\omega)|$ and (25) for $p(e^{j\omega T})$ results in

$$\frac{1}{T}\,|P_R(\omega)|^2 = \frac{1}{\lambda^2 T}\left[\frac{\lambda|q(e^{j\omega T})|}{|C(\omega)|} - \frac{N_o}{2|C(\omega)|^2}\right] \qquad \text{for } |\omega| \leqq \frac{\pi}{T}. \tag{43}$$

Thus the transmitting filter and receiving prefilter frequency responses are identical in the Nyquist band except for constant factors (clearly, the transmitting and receiving filters' phase characteristics can be chosen arbitrarily). This equal sharing of the filtering load between the transmitter and receiver is a well-known result for optimum *linear* communication systems (see pp. 118–121 of Ref. 7).

It is also of interest to evaluate the power spectrum of the error sequence that the Viterbi algorithm assumes to be additive uncorre-

lated Gaussian noise samples. From expressions (26) and (40), this is given by

$$E(e^{j\omega T}) = \frac{N_o}{2}\frac{|q(e^{j\omega T})|^2}{\lambda|q(e^{j\omega T})||C(\omega)|}$$
$$= \frac{N_o}{2\lambda}\frac{|q(e^{j\omega T})}{|C(\omega)|} \qquad \text{for } |\omega| \leqq \frac{\pi}{T}. \tag{44}$$

Thus, the extent that the amplitude frequency response of the chosen DIR approximates that of the channel in the Nyquist band determines how close the power spectrum $E(e^{j\omega T})$ is to being flat, and hence, to what extent successive errors are uncorrelated.

From (41a) and (41b) we obtain an expression for the MSE for a given DIR after the transmitting and receiving filters have been optimized.

$$\text{MSE} = \frac{\frac{N_o T}{4\pi}\alpha^2(\mathbf{Q})}{P_T + \frac{N_o}{2}\int_{-\pi/T}^{\pi/T}\frac{1}{|C(\omega)|^2}\,d\omega}, \tag{45a}$$

where

$$\alpha(\mathbf{Q}) = \int_{-\pi/T}^{\pi/T}\frac{|q(e^{j\omega T})|}{|C(\omega)|}\,d\omega$$
$$= \int_{-\pi/T}^{\pi/T}\frac{\left[\sum_{l=0}^{LL}\sum_{m=0}^{LL} q_l q_m e^{j(m-l)\omega T}\right]^{\frac{1}{2}}}{|C(\omega)|}\,d\omega. \tag{45b}$$

Minimization of the MSE expression with respect to the DIR $\mathbf{Q}$ under the constraint $|\mathbf{Q}|^2 = 1$ is then equivalent to minimization of $\alpha(\mathbf{Q})$ under this constraint. Necessary conditions for the optimum $\mathbf{Q}^+ = (q_0, \cdots q_{LL})$ are then

$$\mu q_l = \sum_{m=0}^{LL} q_m \rho_{l-m}(\mathbf{Q}), \qquad l = 0, \cdots LL, \tag{46a}$$

where

$$\rho_l(\mathbf{Q}) = \int_{-\pi/T}^{\pi/T}\frac{e^{jl\omega T}}{|C(\omega)|\left[\sum_{i=0}^{LL}\sum_{k=0}^{LL} q_i q_k^{j(i-k)\omega T}\right]^{\frac{1}{2}}}\,d\omega \tag{46b}$$

and where μ is a Lagrange multiplier. Again the optimum DIR $\mathbf{Q}$ is the solution of an eigenvalue problem, this time nonlinear. It is easy to verify that the optimum DIR of length two is again either of the two "duobinary" impulse responses shown in Fig. 4. In that case the

minimum achievable MSE is given by (45a) and (45b) as

$$\text{MSE} = \frac{\frac{N_o T}{4\pi}\left[\int_{-\pi/T}^{\pi/T} \frac{[1 \pm \cos \omega T]^{\frac{1}{2}}}{|C(\omega)|} d\omega\right]^2}{P_T + \frac{N_o}{2}\int_{-\pi/T}^{\pi/T} \frac{1}{|C(\omega)|^2} d\omega}, \tag{47}$$

the + or − being chosen to minimize (47).

VIII. CONCLUSIONS

We have presented a scheme of linear prefiltering to optimally "condition" the impulse response of a channel to approximate an impulse response of limited duration for which maximum likelihood estimation of the data sequence is implementable in practice. This scheme in conjunction with the VA can be adaptive, to deal with unknown or slowly time-varying channels. In the simulations its performance compared favorably with the similarly motivated scheme of Ref. 4.

The optimization criterion we used—minimization of the MSE with respect to the prefilter taps and the DIR, with the energy, duration, and relative delay of the DIR being fixed—is admittedly somewhat *ad hoc*. If the sequence of errors $\{e_i\}$ emerging from the prefilter is still assumed to be stationary Gaussian, with zero mean and covariance $\{E_m\}$, then it can be shown that the error rate of a VA which assumes correlated noise is minimized if a certain weighted minimum distance is maximized, namely

$$\min_{\mathbf{d} \in s'} \mathbf{d}^+ \Lambda^{-1} \mathbf{d},$$

where the s' is the set of all possible vectors representing error events and Λ is a covariance matrix whose dimension equals that of $\mathbf{d}\{\Lambda_{ij} = E_{|i-j|}\}$. The above quantity is clearly difficult to maximize, and even if it could be done, the non-Gaussianness of the error sequence would render the solution suspect.

Nevertheless, the performance estimates for the sample channel reported in Section VI make the use of the VA is conjunction with prefiltering appear attractive for high-speed data transmission relative to other schemes. Further studies should be done on the correlatedness of the error sequence and the minimum distance properties of the desired impulse responses.*

---

* S. Fredricsson presented a paper dealing with this subject at the International Symposium on Information Theory, Israel, June 1973.

IX. ACKNOWLEDGMENTS

We are grateful to R. D. Gitlin for helpful suggestions in connection with the adaptive algorithm. We are also grateful to R. R. Anderson and J. Salz for providing the data and performance curves of comparative systems in Section VI.

REFERENCES

1. Forney, G. D., 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.
2. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," IEEE Trans. Inform. Theory, *IT-13*, April 1967, pp. 260–269.
3. Magee, F. R., Jr., and Proakis, J. G., "Adaptive Maximum-Likelihood Sequence Estimation for Digital Signaling in the Presence of Intersymbol Interference," IEEE Trans. Inform. Theory (Corresp.), *IT-19*, January 1973, pp. 120–124.
4. Qureshi, S., and Newhall, E., "An Adaptive Receiver for Data Transmission Over Time-Dispersive Channels," IEEE Trans. Inform. Theory, *IT-19*, July 1973, pp. 448–457.
5. Hildebrand, F. B., *Method of Applied Mathematics*, second edition, Englewood Cliffs, N. J.: Prentice-Hall, 1965, pp. 62–65.
6. Berger, T., and Tufts, D., "Optimum Pulse Amplitude Modulation Part I: Transmitter-Receiver Design and Bounds from Information Theory," IEEE Trans. Inform. Theory, *IT-13*, April 1967, pp. 196–208.
7. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968, pp. 45–51.
8. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. Commun. Tech., *COM-14*, February 1966, pp. 67–68.
9. Forney, G. D., Jr., "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," IEEE Trans. Commun. Tech. (Corresp.), *COM-20*, February 1972, pp. 276–277.
10. Salz, J., "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–1373.
11. Magee, F. R., Jr., and Proakis, J. G., "An Estimate of the Upper Bound on Error Probability for Maximum Likelihood Sequence Estimation on Channels having a Finite Duration Pulse Response," IEEE Trans. Inform. Theory (Corresp.), *IT-19*, September 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Multimode Theory of Graded-Core Fibers

By D. GLOGE and E. A. J. MARCATILI

(Manuscript received March 29, 1973)

New technologies of fiber manufacture and a demand for unusual fiber qualities in communication systems have intensified the interest in a comprehensive theory of multimode fibers with nonuniform index distributions. This paper deals with a general class of circular symmetric profiles which comprise the parabolic distribution and the abrupt core-cladding index step as special cases. We obtain general results of useful simplicity for the impulse response, the mode volume, and the near- and far-field power distributions. We suggest a modified parabolic distribution for best equalization of mode delay differences. The effective width of the resulting impulse is more than four times smaller than that produced by the parabolic profile. Of course, practical manufacturing tolerances are likely to influence this distribution. A relation is derived between the maximum index error and the impulse response.

I. INTRODUCTION

Conventional optical fibers consist of a high-index core surrounded by a cladding of lower index. The index step between core and cladding contains the light inside the core and isolates it from the outer fiber surface, whose quality is usually difficult to control. In a more general way, inside guidance can be accomplished by any index profile which decreases from a maximum inside the fiber to a lower (cladding) value. The specific shape of the profile has an effect on the distribution of the guided optical power in the fiber and on the overall loss encountered, but, more importantly, the profile profoundly influences the velocities of the various propagating modes. A good example is the parabolic index distribution which was predicted to nearly equalize the group velocities of the propagating modes.[1,2] The Selfoc fiber which closely approximates these conditions has indeed since exhibited an extremely narrow impulse response.[3,4]

These effects greatly enhance the chances of multimode fibers to be used in optical communication systems. On the other hand, a theory

of the interrelations between index profile, impulse response, and power distribution is presently only available for the two special cases of the uniform and the parabolic core index. This paper provides a more general theory and studies a broad class of index profiles potentially useful in communication applications. The uniform and the parabolic profile are special cases within this class.

Our concern with multimode fibers for communication applications allows us to make four simplifying assumptions:

(*i*) The index profile is circular symmetric.
(*ii*) The core diameter measures hundred wavelengths or more and, hence, a great number of modes can propagate.
(*iii*) The total index change within the guiding core region is only a few hundredths, so the propagating modes can be considered essentially as transverse electromagnetic.[5]
(*iv*) Index variations within the distance of a wavelength are negligible, and the conditions of geometrical optics (or the zeroth order of the WKB method) apply.

Except for these four restrictions and the requirement of guidance, the index profile can be of the most general form. It can, for example, have an index depression in the center and one or several ring-shaped index maxima.[6]

For the sake of clarity, this paper is restricted to the simpler type of profile illustrated in Fig. 1. We assume the index profile will decrease monotonically from the center and converge into a flat cladding region which guarantees isolation from the outside surface.

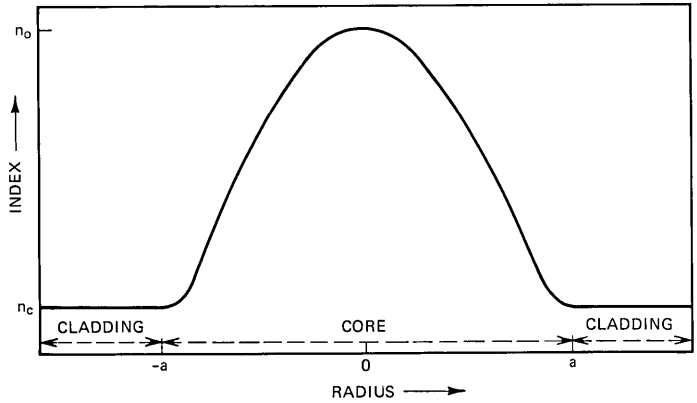


Fig. 1—Cross-sectional sketch of circular symmetric index profile in multimode fiber.

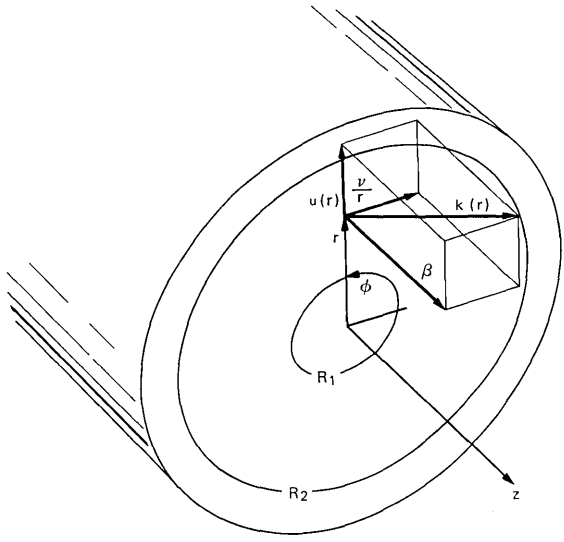


Fig. 2—Wave vector diagram in the propagating region of a multimode fiber.

Apart from the index profile, there are, of course, other influences which affect the impulse response and the optical power distribution inside the fiber. Mode excitation, loss differences in the process of propagation, and coupling among the modes play a part. To isolate the effect of the index profile, we assume here the ideal case of uniform loss, absence of coupling, and equal and simultaneous excitation of all propagating modes at the input. For the computation of the impulse response, the input is assumed to be an infinitely narrow pulse of unit energy.

II. MODE DESIGNATION AND MODE COUNT

All guided modes are essentially transverse electromagnetic and, with some proviso, can be decomposed into linearly polarized pairs.[5,7] Because of the circular symmetry of the index n, the modes have a circular periodicity and can be identified in the conventional way by an azimuthal order number ν. To characterize the radial field distribution, we need an additional mode number μ. The propagation constant β of a particular mode (μ, ν) can then be approximately determined by the WKB method.[6,8] Figures 2 and 3 give a physical description of these relationships. In Fig. 2, the local wave number

$$k(r) = 2\pi n(r)/\lambda \tag{1}$$

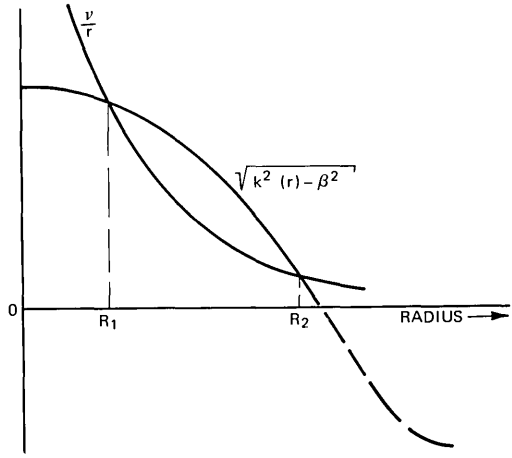


Fig. 3—Sketch defining regions of periodic and aperiodic field characteristics of a mode of azimuthal order ν.

is decomposed into its components in a cylindrical coordinate system (r, ϕ, z). The unknown radial component becomes

$$u(r) = [k^2(r) - \beta^2 - \nu^2/r^2]^{\frac{1}{2}}. \tag{2}$$

Given β and ν, we can find two radii R_1 and R_2, at which $u(r)$ vanishes (see Fig. 3). These radii define a ring-shaped region within which u is real, causing a radial periodicity of the mode field. Outside this region, the field is aperiodic.

Radially decreasing (or evanescent) field conditions obtain outside, when the phase inside (approximately) adds up to an integer number of half periods between R_1 and R_2. Consequently, if μ designates this number of half periods,

$$\mu\pi = \int_{R_1}^{R_2} u(r)dr = \int_{R_1}^{R_2} [k^2(r) - \beta^2 - \nu^2/r^2]^{\frac{1}{2}}dr. \tag{3}$$

We would have obtained the same result by way of the WKB method, with the only difference that μ and ν^2 would be replaced by $\mu + \frac{1}{4}$ and $\nu^2 + \frac{1}{4}$. These corrections are important in the case of small μ or ν, and particularly for the fundamental mode which has $\mu = \nu = 0$. On the other hand, to obtain a general view of the mode structure, we can ignore the $\frac{1}{4}$-terms as long as we refrain from discussing individual low-order modes.

For the purpose of a total mode count, let us consider the limits of μ, ν, and β. The requirement of evanescent field conditions in the

cladding (index n_c in Fig. 1) limit β to a minimum value

$$\beta_c = 2\pi n_c/\lambda. \tag{4}$$

Modes with smaller β find propagating conditions in the cladding and are no longer bounded by the core profile. Condition (4) defines mode cutoff. The largest value for ν results for $\beta = \beta_c$ and $\mu = 0$, and alternatively μ is largest for $\beta = \beta_c$ and $\nu = 0$. We obtain the total number of modes M from a summation of (3) over all ν from 0 to $\nu_{\max}$. If $\nu_{\max}$ is a large number, we may consider ν a continuous variable and replace the sum by an integral. In this case,

$$M = \frac{4}{\pi}\int_0^{\nu_{\max}}\int_{R_1(\nu)}^{R_2(\nu)} [k^2(r) - \beta_c^2 - \nu^2/r^2]^{\frac{1}{2}} dr d\nu . \tag{5}$$

The factor 4 in front of the expression allows for the fact that each combination μ, ν designates a (degenerate) group of four modes of different polarization or orientation.[5] Figure 4 illustrates the area of the double integration indicated in (5). A change of order in the integration leads to

$$M = \frac{4}{\pi}\int_0^{a}\int_0^{r(k^2-\beta_c^2)^{\frac{1}{2}}} (k^2 - \beta_c^2 - \nu^2/r^2)^{\frac{1}{2}} d\nu dr , \tag{6}$$

where a is the radius at which the index $n(r)$ reaches the cladding value n_c. Integrating (6) with respect to ν yields

$$M = \int_0^a [k^2(r) - \beta_c^2]^{\frac{1}{2}} r dr = \left(\frac{2\pi}{\lambda}\right)^2 \int_0^a [n^2(r) - n_c^2]^{\frac{1}{2}} r dr . \tag{7}$$

For small index differences, the integral represents the volume under the (circular symmetric) profile plot. It may be worth noting, though, that the substance of this relation is not limited to circular symmetry.

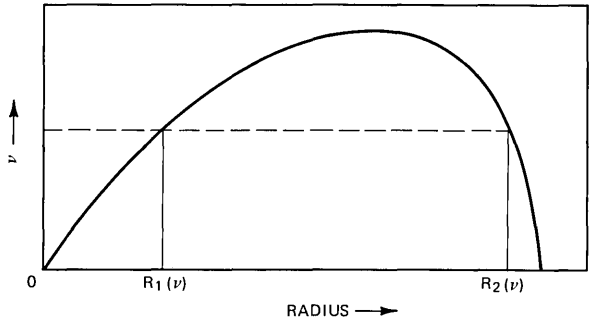


Fig. 4—Region of double integration in eq. 5.

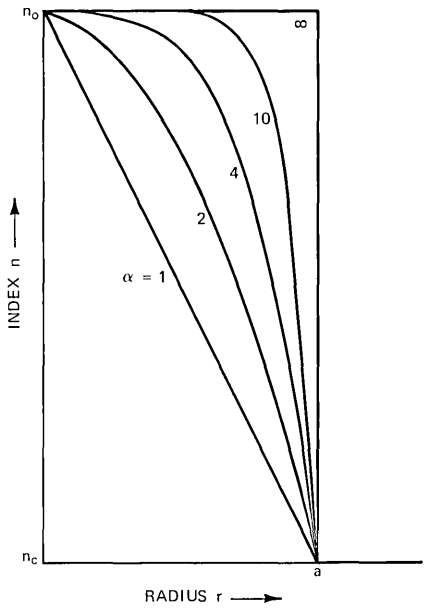


Fig. 5—A few of the index profiles defined by $n = n_0[1 - 2\Delta(r/a)^\alpha]^{\frac{1}{2}}$ for small Δ.

For later use we write (7) in the somewhat different form

$$m(\beta) = \int_0^{R_2(0)} [k^2(r) - \beta^2]^{\frac{1}{2}} r dr , \tag{8}$$

where $m(\beta)$ denotes the number of modes having a propagation constant larger than β. The upper limit $R_2(0)$ of the integration is the radius at which $k(r) = \beta$.

Let us now consider a particular class of profiles defined by

$$n(r) = \begin{cases} n_o[1 - 2\Delta(r/a)^\alpha]^{\frac{1}{2}} & \text{for } r < a \\ n_o[1 - 2\Delta]^{\frac{1}{2}} & \text{for } r > a , \end{cases} \tag{9}$$

where α is a parameter between 1 and ∞. Figure 5 illustrates the cases $\alpha = 1, 2, 4, 10$, and ∞. All profiles reach a constant cladding value at $r = a$. The core profile has a cone shape for $\alpha = 1$, becomes nearly parabolic for $\alpha = 2$, and converges to the case of the step profile for $\alpha = \infty$. Using (1) we introduce (9) into (8) and obtain

$$m(\beta) = a^2 \Delta k_o^2 \frac{\alpha}{\alpha + 2} \left(\frac{k_o^2 - \beta^2}{2\Delta k_o^2} \right)^{(2/\alpha)+1} , \tag{10}$$

where

$$k_o = 2\pi n_o/\lambda. \tag{11}$$

For $\beta = \beta_c$ from (4), the total mode number becomes

$$M = \frac{\alpha}{\alpha + 2} a^2 k_o^2 \Delta . \tag{12}$$

It is proportional to the index difference and the core cross section. The uniform profile accepts twice as many modes as the parabolic one and three times more than the cone-shaped one.

III. IMPULSE RESPONSE

Consider all modes to be excited by the same narrow pulse at the input. Each mode transports an equal amount of energy to the fiber end. The individual pulses are expected to suffer a certain distortion, depending on the β–ω characteristic of each mode and dispersion in the dielectric. We assume, however, that the resultant broadening is small, or at least not much larger than the group delay differences between adjacent modes. Because of this effect and other limitations in the system response, the pulses from individual modes are likely to fuse into one continuous output pulse called the impulse response. Since all modes carry the same energy, the power profile of the impulse response is equal to the mode density per unit time interval. In the following theory, the continuity of the impulse response results not from the broadening of the individual mode responses, but from the assumption that μ and ν are continuous functions.

The straightforward method of computing the impulse response starts from (3) to find the propagation constant β for each pair, μ, ν. The group delay in a fiber of length L is then

$$\tau(\mu, \nu) = \frac{Ln_o}{c} \frac{d\beta(\mu, \nu)}{dk_o} , \tag{13}$$

where c is the vacuum velocity of light. A simplification of this approach for the purpose of numerical computations is indicated in the appendix. Once $\tau(\mu, \nu)$ is known, the impulse response results from a count of the combinations μ, ν which arrive between τ and $\tau + d\tau$. This number plotted versus τ then constitutes the impulse response.

For the special class of profiles defined by (9), group delay and impulse response can be computed in a much simpler way. First we postulate that, in this case, the relation between τ and β according to (13) is independent of μ and ν. If this holds—and we shall prove it

later with the help of eq. (16)—we can replace β by τ in (3) and still perform the same integration over ν which led to (8) and, more specifically, to (10). Solving the result of this integration for τ yields

$$\tau = \frac{Ln_o}{c}\frac{d}{dk_o}\left[k_o^2 - \left(\frac{2m}{a^2}\frac{\alpha+2}{\alpha}\right)^{\alpha/(\alpha+2)}(2\Delta k_o^2)^{2/(\alpha+2)}\right]^{\frac{1}{2}}. \tag{14}$$

This result can easily be verified by solving (10) for β and introducing it into (13). With the help of (10) and the abbreviation

$$\delta = \tfrac{1}{2}(1 - \beta^2/k_o^2)\,, \tag{15}$$

eq. (14) takes the form

$$\tau = \frac{Ln_o}{c}\frac{1 - 4\delta/(\alpha+2)}{(1-2\delta)^{\frac{1}{2}}}. \tag{16}$$

This expression proves indeed to depend on β alone (and not explicitly on m), thus justifying the approach chosen.

To obtain the impulse response, we can now introduce (16) into (10) and differentiate with respect to τ. Although this is not difficult to do, it leads to rather unwieldy expressions. We shall therefore merely consider some special cases of interest. To normalize the impulse response for total unit energy, we divide (10) by (12) and obtain

$$\frac{m}{M} = \left(\frac{\delta}{\Delta}\right)^{(2/\alpha)+1}. \tag{17}$$

Furthermore, since δ can at most assume the value Δ (for $\beta = \beta_c$) and is therefore small compared to unity within the scope of our theory, we develop (16) into a power series in terms of δ and obtain

$$\tau = \frac{Ln_o}{c}\left(1 + \frac{\alpha-2}{\alpha+2}\delta + \frac{3\alpha-2}{\alpha+2}\frac{\delta^2}{2}\right). \tag{18}$$

We relate τ to the total propagation time Ln_o/c and introduce a new time reference, which ignores the delay common to all modes. Hence,

$$t = \frac{\tau c}{Ln_o} - 1 = \frac{\alpha-2}{\alpha+2}\delta + \frac{3\alpha-2}{\alpha+2}\frac{\delta^2}{2}. \tag{19}$$

In this time frame, the fundamental mode arrives at $t = 0$.

As long as α is not too close to 2, the linear term in (19) dominates. Therefore,

$$\delta = \begin{cases} \dfrac{\alpha+2}{\alpha-2}t & \text{except for } \alpha \approx 2 \\ \sqrt{2t} & \text{for } \alpha = 2. \end{cases} \tag{20}$$

Insert this into (17) and differentiate with respect to t to obtain the impulse response

$$\frac{1}{M}\frac{dm}{dt} = \begin{cases} \dfrac{\alpha+2}{\alpha}\left|\dfrac{\alpha+2}{\alpha-2}\dfrac{1}{\Delta}\right|^{(2/\alpha)+1}|t|^{2/\alpha} & \text{except for } \alpha \approx 2 \\ \dfrac{2}{\Delta^2} & \text{for } \alpha = 2. \end{cases} \tag{21}$$

As δ varies from 0 to Δ, the time t changes from 0 to

$$T = \begin{cases} \dfrac{\alpha-2}{\alpha+2}\Delta & \text{except for } \alpha \approx 2 \\ \dfrac{\Delta^2}{2} & \text{for } \alpha = 2. \end{cases} \tag{22}$$

Outside of this time interval, the impulse response is zero. Figure 6 shows plots of (21) for the profiles sketched in Fig. 5. A change from $\alpha = \infty$ to $\alpha = 10$, which implies a relatively small change in the profile, narrows the impulse response by $\frac{1}{3}$. The response becomes

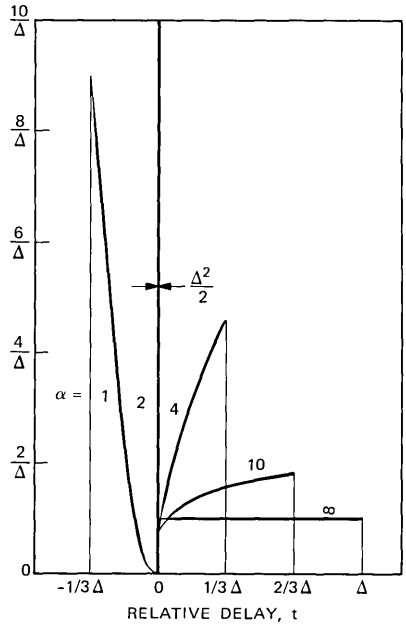


Fig. 6—Impulse response of multimode fibers having the profiles of Fig. 5.

extremely narrow for $\alpha \approx 2$, then broadens again, as α decreases further. For $\alpha < 2$, the high-order modes overtake the fundamental and arrive earlier.

In the vicinity of $\alpha = 2$, where both terms of (19) contribute, the impulse response is a rather complicated function. The most interesting of these cases is the one for which the impulse response has the narrowest possible width. This optimum condition arises for

$$\alpha_{\text{opt}} = 2 - 2\Delta , \tag{23}$$

which yields

$$t = \tfrac{1}{2}(\delta^2 - \Delta\delta). \tag{24}$$

In this case, the modes of highest and lowest order both arrive at the same time $t = 0$; all other modes are faster, the fastest one being determined by $\delta = \Delta/2$. It arrives at

$$t = -\frac{\Delta^2}{8}. \tag{25}$$

Equation (24) has two solutions for δ. Hence, (17) yields two values for the same t, indicating that two mode groups, a high and a lower order, contribute to the impulse response at every particular instant in time. By introducing δ into (17), differentiating with respect to t, and then adding the two contributions, we find the impulse response

$$\frac{4}{\Delta^2}\left(1 + \frac{8t}{\Delta^2}\right)^{-\frac{1}{2}}. \tag{26}$$

This function is plotted in Fig. 7. It peaks at $t = -\Delta^2/8$ and decreases towards $t = 0$. Because of the normalization introduced in (19), the absolute temporal width is

$$\frac{Ln_o}{c}\frac{\Delta^2}{8}. \tag{27}$$

The time slot in which a pulse of this kind can be transmitted is narrower than that, because 70 percent of the power is concentrated in the first half of the interval (27).

A practical implementation must, of course, allow for a certain tolerance or error in the profile, as a result of which the total width of the impulse response is likely to exceed (27). To obtain some indication of the pulse broadening as a result of this index deviation, we assume that the erroneous profile is still of the type (9), but has

$$\alpha = \alpha_{\text{opt}} + d\alpha. \tag{28}$$

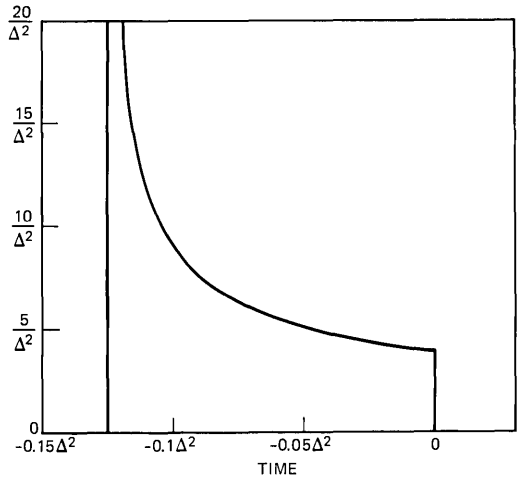


Fig. 7—Impulse response in the case of optimal profile shape.

The maximum index deviation from the optimum profile then appears approximately at

$$r = ae^{-\frac{1}{2}} \tag{29}$$

and has the value

$$dn_{\max} = d\alpha \frac{n_o \Delta}{2e}, \tag{30}$$

where e is the base of the natural logarithm. As a result of this profile error, the normalized width of the impulse response becomes

$$\tfrac{1}{8}(\Delta + \tfrac{1}{2}|d\alpha|)^2 \tag{31}$$

or, in absolute terms,

$$\frac{Ln_o}{8c}\left(\Delta + \frac{e}{n_o\Delta}|dn_{\max}|\right)^2. \tag{32}$$

Consider a guide with a maximum index $n_o = 1.5$ and an index variation $\Delta = 2$ percent. If the profile is optimal, mode delay should produce an effective broadening of only 0.25 ns/km. An index deviation of 10^{-4} from the optimal profile increases the broadening to 0.53 ns/km.

IV. NEAR- AND FAR-FIELD POWER DISTRIBUTION

We take again into account the fact that the core cross section measures many wavelengths in diameter. If this cross section is illuminated by an incoherent source (exciting all modes uniformly), the power

incident per unit solid angle at any point in the cross section is constant. To compute the power accepted by the fiber, we merely have to know the solid angle of acceptance at any point. We find this angle from the wave vector diagram of Fig. 2, which yields

$$\cos \theta(r) = \beta/k(r). \tag{33}$$

The maximum angle θ_c results for $\beta = \beta_c$; hence,

$$\cos \theta_c(r) = \frac{\beta_c}{k(r)} = \frac{n_c}{n(r)}. \tag{34}$$

Using this relation, we can define a local numerical aperture at the fiber front face

$$A(r) = n(r) \sin \theta_c(r) = [n^2(r) - n_c^2]^{\frac{1}{2}}. \tag{35}$$

The power accepted at r is then

$$p(r) = p(0) \frac{A^2(r)}{A^2(0)} = p(0) \frac{n^2(r) - n_c^2}{n^2(0) - n_c^2}. \tag{36}$$

If all modes propagate equally attenuated and without coupling, the

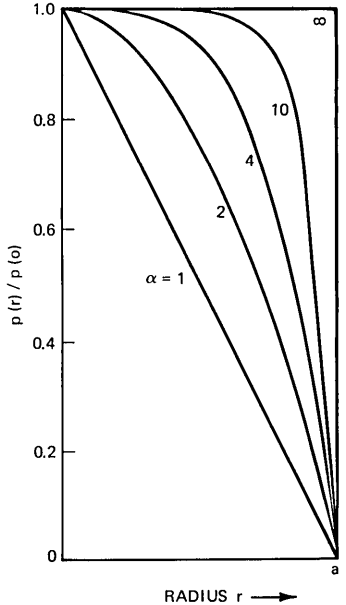


Fig. 8—Power distribution in the core of multimode fibers having the profiles of Fig. 5.

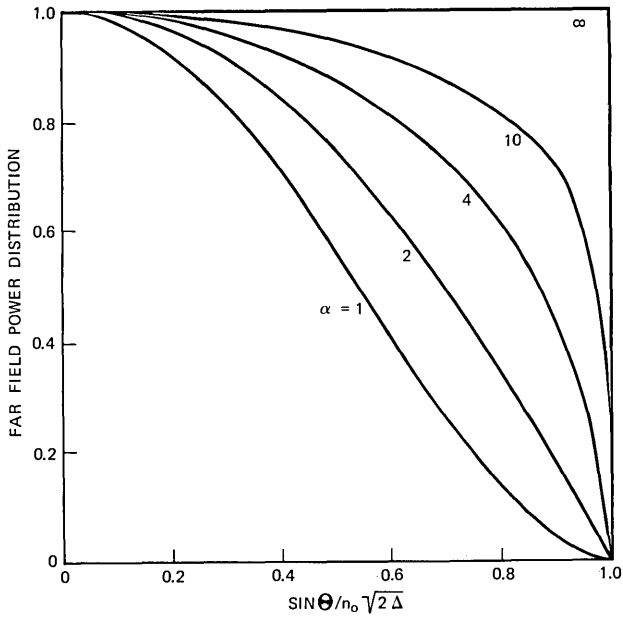


Fig. 9—Power distribution in the far field of multimode fibers having the profiles of Fig. 5.

same power distribution should hold for the fiber end face. The class of profiles described by (9) has

$$A(r) = n_o(2\Delta)^{\frac{1}{2}}[1 - (r/a)^\alpha]^{\frac{1}{2}} \tag{37}$$

and

$$p(r) = p(0)[1 - (r/a)^\alpha]. \tag{38}$$

The agreement between the profile plots (Fig. 5) and the near-field power plots (Fig. 8) is not a coincidence, but holds in general as long as the total index variation is small.

Under the conditions assumed here, every incremental area of the core cross section at the fiber end uniformly illuminates its cone of acceptance. For this reason, all those areas that have a numerical aperture

$$A(r) \geqq \sin\theta \tag{39}$$

contribute equally to the far-field power at θ. For the class of profiles described by (9), the areas contributing to θ are within a circle whose radius is obtained by solving (37) for r. Consequently,

$$P(\theta) = P(0)\left(1 - \frac{\sin^2\theta}{2n_o^2\Delta}\right)^{2/\alpha} \tag{40}$$

is the far-field power distribution. Figure 9 shows a plot of (40) for the profiles of Fig. 5. The uniform illumination for $\alpha = \infty$ changes to a parabolic distribution for $\alpha = 2$. All plots must be understood as the average power expected under the idealized conditions mentioned earlier. Monochromatic mode excitation results in mode interference phenomena and a local fine structure, which can greatly modify the average distribution considered here.

V. CONCLUSIONS

By assuming somewhat idealized conditions for mode excitation, coupling, and loss in a multimode fiber, we can isolate the influence of the index profile upon mode volume, near- and far-field power distribution, group delay, and impulse response. Surprisingly simple relations exist for a special class of profiles which comprises most multimode fibers of interest. Particular attention is given to a near-parabolic profile which accomplishes optimal delay equalization of all modes. If the (relative) index difference between center and periphery of this profile is Δ, mode delay broadens the impulse response by a fraction $\Delta^2/8$ of the total propagation time. This amounts to about 0.25 ns/km for $\Delta = 2$ percent. On the other hand, an index deviation of 10^{-4} from the optimal profile increases the broadening to 0.53 ns/km.

APPENDIX

Some Further Relations for the Group Velocity

The numerical evaluation of β as a function of μ, ν, and k_o and its subsequent differentiation to obtain τ are usually tedious and time-consuming. A substantial simplification results from a direct computation of τ by applying the operation

$$\tau = \frac{Ln_o}{c} \frac{\partial \mu / \partial k_o}{\partial \mu / \partial \beta} \tag{41}$$

to (3). The result is

$$\tau = \frac{L}{c} \frac{\int_{R_1}^{R_2} k(r) n(r) dr / u(r)}{\int_{R_1}^{R_2} \beta dr / u(r)}. \tag{42}$$

To understand the physical significance of this relation, consider a ray propagating along the fiber core of Fig. 2 in such a way that it has the direction of $k(r)$ at r. A line element along this ray is

$$ds = (dr^2 + r^2 d\phi^2 + dz^2)^{\frac{1}{2}} \tag{43}$$

and therefore

$$\frac{ds}{dr} = \frac{k(r)}{u(r)} \qquad \text{and} \qquad \frac{dz}{dr} = \frac{\beta}{u(r)}. \tag{44}$$

The condition $u(r) = dr = 0$ at R_1 and R_2 indicates a reflection (turn-around) of the ray. The ray performs periodic undulations between R_1 and R_2, simultaneously moving sideways in a helical fashion. By introducing (44) into (42), we obtain

$$\tau = L \frac{\oint n(r)ds/c}{\oint dz}, \tag{45}$$

where $\oint$ denotes integration over a full period of the ray. The denominator describes the axial length of one ray period, and the numerator the propagation time along the ray within this length. Multiplied by the fiber length, this ratio yields the total group delay. This result emphasizes the equivalence between ray theory and the zeroth-order WKB approach followed in this paper.

Within this order of approximation, the only quantities that depend on the wavelength are the mode numbers. Normalization of these numbers and subsequent transition to continuous variables eliminates the wavelength entirely; group velocity and impulse response are then independent of wavelength. More specifically, if we write

$$\rho = \mu/ak_o \qquad \text{and} \qquad \sigma = \nu/ak_o \tag{46}$$

and

$$n = n_o[1 - 2d(r)]^{\frac{1}{2}}, \tag{47}$$

eq. (3) assumes the form

$$\rho = \frac{1}{\pi a} \int_{R_1}^{R_2} [2\delta - 2d - (\sigma a/r)^2]^{-\frac{1}{2}} dr, \tag{48}$$

and (42) becomes

$$\tau = \frac{Ln_o}{c} (1 - 2\delta)^{-\frac{1}{2}} \frac{\int_{R_1}^{R_2} (1 - 2d)dr/[2\delta - 2d - (\sigma a/r)^2]^{\frac{1}{2}}}{\int_{R_1}^{R_2} dr/[2\delta - 2d - (\sigma a/r)^2]^{\frac{1}{2}}}. \tag{49}$$

These two equations are sufficient to calculate group velocity and impulse response in the case of large mode numbers.

REFERENCES

1. Miller, S. E., "Light Propagation in Generalized Lens-Like Media," B.S.T.J., *44*, No. 9 (November 1965), pp. 2017–2064.
2. Kawakami, S., and Nishizawa, T., "An Optical Waveguide with the Optimum Distribution of the Refractive Index with Reference to Waveform Distortion," IEEE Trans. Microwave Theory and Tech., *MTT-16* (October 1968), pp. 814–818.
3. Uchida, M., Furukawa, M., Kitano, I., Koisumi, K., and Matsumura, H., "A Light-Focusing Fibre Guide," IEEE J. Quan. Elec. (Digest of Technical Papers), *QE-5* (June 1969), p. 331.
4. Gloge, D., Chinnock, E. L., and Koizumi, K., "Study of Pulse Distortion in Selfoc Fibers," Elec. Letters, *8*, 21 (October 19, 1972), pp. 526–627.
5. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, 10, pp. 2252–2258.
6. Gloge, D., and Marcatili, E. A. J., "Impulse Response of Fibers with Ring-Shaped Parabolic Index Distribution," B.S.T.J., *52*, No. 7 (September 1973), pp. 1161–1168.
7. Matsuhara, M., "Analysis of Electromagnetic-Wave Modes in Lens-Like Media," J. Opt. Soc. Am., *63*, 2 (February 1973), pp. 135–138.
8. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics*, New York: McGraw-Hill, 1953, p. 1092.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Optical Fiber End Preparation for Low-Loss Splices

By D. GLOGE, P. W. SMITH, D. L. BISBEE, and E. L. CHINNOCK

(Manuscript received May 8, 1973)

Cables made from brittle materials like glass require new techniques of end preparation for the purpose of splicing, especially if such splices are to be made in the field. We report here on a method of breaking fibers in a way which invariably produces flat and perpendicular end faces. We explain the underlying theory and derive optimal parameters that permit the design of a simple breaking tool. Experiments with a tool of this kind show that the tolerances for successful fracture are not critical. Laboratory splices of multimode fibers prepared by this method exhibited losses of less than 1 percent (0.04 dB) when joined in index-matching fluid.

I. INTRODUCTION

With installation and maintenance consuming an ever-larger share of system costs, simple and inexpensive splicing techniques have become a prerequisite for competitive communication systems. One bottleneck in optical fiber cable splicing is the fiber end preparation, as conventional grinding and polishing techniques turn out to be time-consuming and costly, especially in the field. It is well known that glass fibers sometimes break with flat and perpendicular end faces if they are previously scored,[1] and it has thus become common practice in the laboratory to obtain good ends in this way by trial and error. Besides being faster and simpler, this technique has the added advantage of producing perfectly clean surfaces uncontaminated by lossy residues. Such ends were recently used in fiber joining experiments to determine eventual splice losses.[2–5] The lowest losses obtained were about 10 percent for single-mode fibers[4,5] and 3 percent for multimode fibers.[2]

For such laboratory practice to become useful technology, absolute control of the breaking process and utmost reliability in obtaining a

successful result are required. We report here on an approach which guarantees this reliability through control of the stress distribution in the fracture zone. The break is initiated by lightly scoring the fiber periphery at the correct point. We explain the underlying theory which allows us to predict the character of the break from the initial stress distribution. By modifying a previous design,[6] we obtained a simple tool that permits us to vary the amount and distribution of stress in the fracture zone. All 130 breaks we have made with this tool have produced the predicted fracture surface. The range within which perfectly flat and perpendicular end faces were obtained was found to be so wide that the eventual construction of a simple hand tool for this purpose should present no problem. The quality of the surfaces obtained makes this method the most promising of all the techniques investigated so far.[7–9] This notion is supported by some fiber-joining experiments which we describe in Section IV of this paper. Low-loss multimode silica glass fibers were prepared by our breaking technique and then joined in an index-matching liquid. With proper alignment, the splice losses were always less than 1 percent. Results on alignment tolerances for multimode fiber splices are also given in Section IV.

II. BRITTLE FRACTURE OF GLASS RODS AND FIBERS

It has been well documented that glass rods tend to break in such a way that the fracture face comprises three regions known as the mirror, the mist, and the hackle zones.[10,11] The mirror zone is an optically

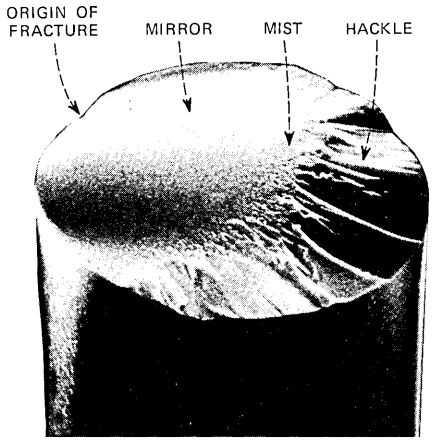


Fig. 1—A typical glass fiber fracture.

smooth surface adjacent to the fracture origin, the hackle zone corresponds to an area where the fracture has forked and the specimen is separated into three or more pieces, and the mist zone is a transition region between these two zones. Such behavior is also observed with glass fibers. Figure 1 shows the fractured end of a 125-μm glass fiber which clearly exhibits these three regions.

It has been experimentally demonstrated[11] that the distance from the origin of fracture to a point on the boundary between the mirror and mist zones, r, is given by

$$Z\sqrt{r} = K\,, \tag{1}$$

where Z is the local stress at the point in question and K is a constant for a given material.

A theoretical justification for eq. (1) can be given. Anderson[12] gives the energy balance equation for a crack of length $2c$ propagating in a brittle isotropic material subject to a plane stress, Z, as

$$\frac{d}{dc}\left(-\pi\frac{Z^2c^2}{E} + \frac{1}{2}k\rho\dot{c}^2\frac{c^2Z^2}{E^2} + 4\gamma c\right) = 0\,. \tag{2}$$

Here E is Young's modulus, ρ is the density, and γ is the surface tension of the material. The parameter k is a geometrical factor which depends on the shape of the crack. The three terms in eq. (2) represent, respectively, the released strain energy, the kinetic energy associated with the moving crack, and the surface energy of the newly created surfaces. As the crack propagates, more and more strain energy is converted into kinetic energy until the crack reaches a limiting velocity, $\dot{c} = v_f$, where v_f is roughly $\frac{1}{3}$ the longitudinal sound velocity for the material (see, for example, Reference 12). At this point the excess energy begins to be taken up by the creation of subsurface cracks (the mist zone). When the released strain energy is sufficient to create four new surfaces, a hackle zone is created. Thus, at the boundary of the mirror and mist zones,

$$\frac{d}{dc}\left(-\pi\frac{Z^2c^2}{E} + \frac{1}{2}k\rho v_f^2\frac{c^2Z^2}{E^2} + 4\gamma c\right) = 0\,. \tag{3}$$

By differentiating, we find

$$Z^2c = \frac{4\gamma E}{2\pi - k\rho v_f^2/E} = \text{a constant}\,, \tag{4}$$

which is of the same form as eq. (1). A similar derivation is given in Reference 11. The value of the constant K in eq. (1) is found experi-

mentally to have the value 6.1 kg/mm$^{\frac{3}{2}}$ for soda-lime-silicate glass and 7.5 kg/mm$^{\frac{3}{2}}$ for fused silica, in reasonably good agreement with the value found[11] from the evaluation of the constant from eq. (4).

In order to break an optical fiber in such a way that the mirror zone extends across the entire fiber, it is necessary to have the stress at all points within the fiber low enough so that $Z\sqrt{r} < K$. The required value of Z at the origin of the fracture depends on the size of the crack or flaw from which the fracture originates.[12] The value of Z cannot be allowed to become zero or negative at any point across the fiber, or the crack will cease to propagate or propagate in a direction which is not perpendicular to the axis of the fiber. Under these conditions, a lip is formed on one fiber end. We see, then, that, to make a reliable clean mirror zone fracture, the stress distribution across the fiber must be suitably adjusted.

III. THE FIBER BREAKING MACHINE

In the preceding section, we have given the conditions necessary to create a mirror zone fracture across an entire fiber end. To determine experimentally the range of stress distributions over which clean mirror zone fractures can be obtained, an apparatus was constructed which could simultaneously bend the fiber and place it under tension. In this way, the stress distribution across the fiber can be varied, as shown in Fig. 2. For a given average tension (force per unit area), T, the stress distribution across the fiber depends on the radius, R, of the form over which the fiber is bent. (We assume no shear friction between the fiber and the form.) In fact, the stress across the fiber, $Z(x)$, is given by

$$Z(x) = T + \frac{E(a - x)}{R}, \tag{5}$$

where T is the average tension on the fiber, E is Young's modulus, and a is the radius of the fiber.

If $R = \infty$, the maximum diameter, d_M, of fiber that can be fractured with a mirror zone across the entire surface is given by

$$Z'\sqrt{d_M} = K, \tag{6}$$

where Z' is the stress necessary to initiate the break.

In the experiments to be described later using a diamond or carbide scorer to initiate the break, we find $Z' \approx 25$ kg/mm^2. Thus, for fused silica fibers we find $d_M \approx 100$ μm and for $R = \infty$, when fracturing fused silica fibers with diameters $\gtrsim 100$ μm, we expect hackle to appear.

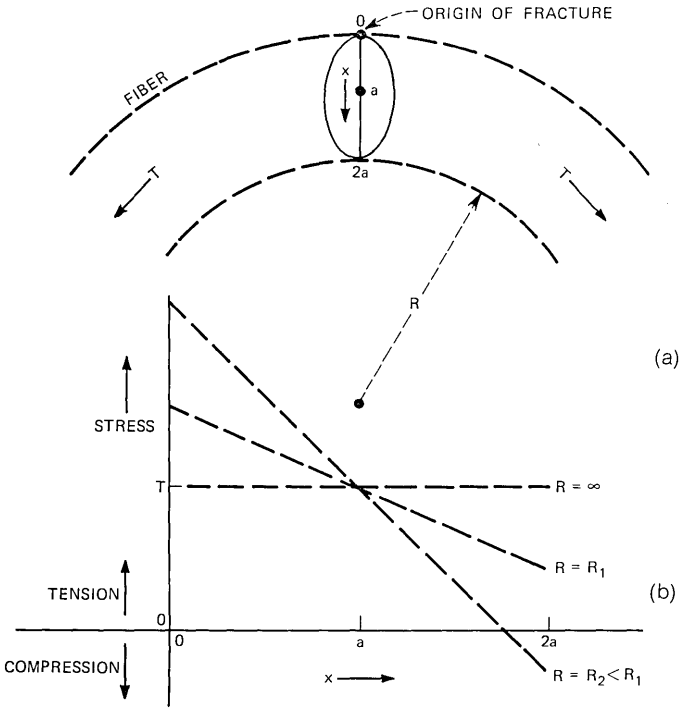


Fig. 2—(a) A glass fiber bent over a form of radius R and subjected to a tension T. (b) The stress as a function of position in the fiber for various bending radii R.

Our experiments showed this to be the case. If we compute the stress from eq. (5), assume T is adjusted so that the stress at $x = 0$ is Z', and select R ($=R_0$) so that $Z = 0$ at $x = 2a$, we find that the maximum value of $Z\sqrt{r}$ occurs on the surface of the fiber at the position where $r = (\sqrt{4/5})a$, and if we require this product to be $<K$, we find $d_M(R = R_0) = 3.50\ d_M\ (R = \infty)$. Thus, using this technique, fused silica fibers of up to $\sim$350 μm in diameter can be fractured with clean, mirror zone ends.

The fibers used for the experiments reported here were multimode silica glass fibers with an outer diameter of 125 μm and a core diameter of 80 μm. R_0 can be found from eq. (5), letting $Z = 0$ at $x = 125$ μm, assuming the stress necessary to initiate the break, Z', to be equal to the experimentally determined value of 25 kg/mm^2, and using the values $E = 7.2 \times 10^3$ kg/mm^2; $K = 7.5$ kg/mm$^{\frac{3}{2}}$ appropriate for silica glasses. We obtain $R_0 = 3.7$ cm.

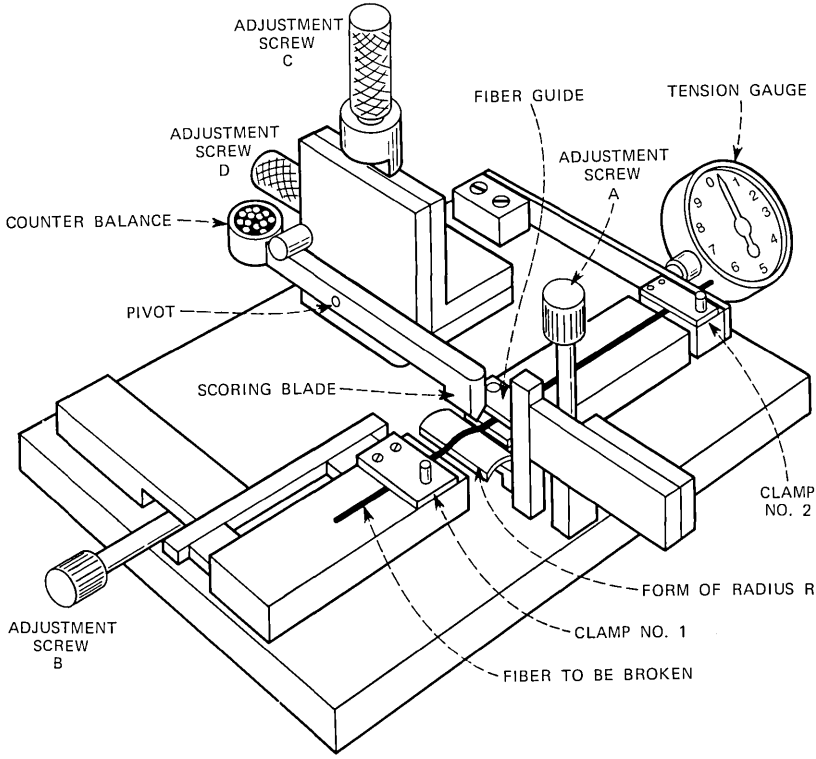


Fig. 3—A semi-schematic view of the fiber breaking machine.

Figure 3 shows a semi-schematic view of the device used to investigate the fracture properties of optical fibers. The fiber to be broken is clamped by clamps No. 1 and No. 2 and slides freely under the Teflon*-coated fiber guide. A Teflon-coated form of suitable radius R can be raised to cause the fiber to conform to the form by adjustment screw A. The tension on the fiber is measured by a tension gauge, which measures the mechanical displacement of a stiff steel bar on which clamp No. 2 is mounted. The tension can be adjusted with the adjustment screw B. A scoring blade can be lowered onto the fiber by adjustment screw C and pulled across the fiber by adjustment screw D. The pressure on the scorer blade can be adjusted by changing the weight in the counterbalance.

* Registered trademark of Dupont Co.

IV. EXPERIMENTAL RESULTS

Breaks were made on samples of a low-loss multimode silica glass fiber having a core diameter of 80 μm and a cladding thickness of 22 μm. A wide range of breaking tensions and fiber-bending radii was studied using the fiber-breaking machine described above. We used a variety of scoring techniques and attempted breaks in atmospheres of various relative humidities. The results can be summarized as follows: If the radius of curvature of the form was less than about 2 cm, a lip would be formed. When fractures were made without using a form, i.e., $R = \infty$ or negative, a hackle region was produced. "Good," clean fractures were obtained when a 5.7-cm radius of curvature form was used. These results are illustrated in Fig. 4.

Using the 5.7-cm radius of curvature form, clean fractures with no visible hackle or lip were always produced using breaking tensions in the range of 125 to 175 g and scorer pressures ranging from 1.5 to 7.5 g. The smallest scores were produced when a sharp diamond scorer* was lowered onto the fiber after the tension had been applied. We found no effect on the fracture characteristics when the relative humidity was varied from 7 to 100 percent, or even when water was applied to the point of fracture. In all, a total of 33 fractures were made within this range of conditions. *In no case was there any visible evidence of any hackle or lip.* In the worst case the disturbed region associated with the score extended over a distance of $\sim$22 μm. As the cladding thickness on this optical fiber was 20 μm, this means that in all cases a perfect mirror zone fracture occurred over essentially the entire core region of the fiber.

To establish the minimum splice loss in joining such fiber ends, we used the setup shown in Fig. 5. The joints were made from ends obtained from the same fracture, but rotated with respect to the original fracture position. In this way, the time between fracture and joining was kept at a minimum in order to avoid contamination of the ends. Moreover, utmost accuracy was achieved by comparing the losses immediately before fracture and immediately after joining. This time was typically 10 minutes, while instabilities in the setup caused a power drift at the detector of not more than $\frac{1}{4}$ percent in 30 minutes. Joining adjacent ends, of course, eliminated the possibility of diameter discrepancies which would be encountered in practical splices.

---

* The diamond scorer was supplied by Victory Diamond Tool Co. of East Hanover, N. J.

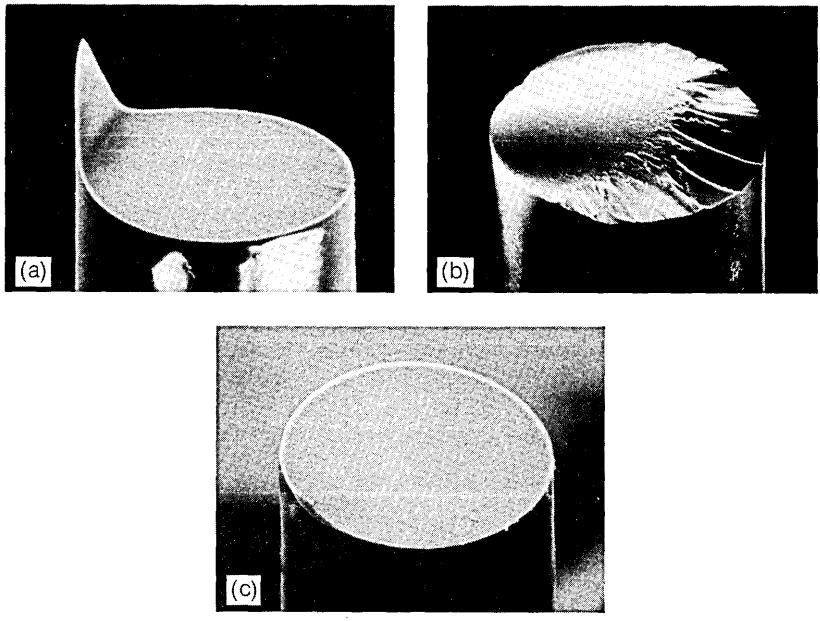


Fig. 4—Electron microscope photographs of 125-μm diameter silicate glass fibers broken using various form radii (R): (a) $R = 0.75$ cm; (b) $R = \infty$ or negative; (c) $R = 5.7$ cm.

Fibers of the type that were used for the splice loss measurements reach a steady-state power distribution after a certain distance independent of the injection conditions. This distribution was measured for the fiber in question at the end of a 1.2-km length. The power distribution in the splice should preferably be the steady-state distribution. Since a sufficient fiber length to achieve such conditions was not available for our measurements, we approximated as well as

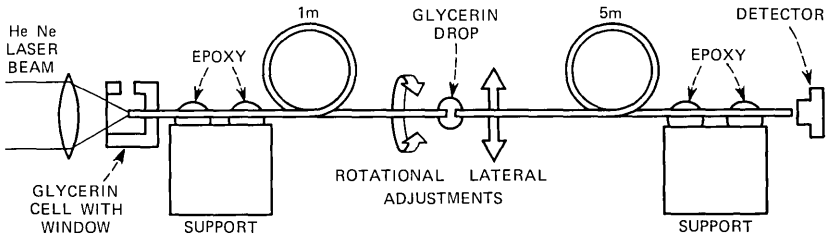


Fig. 5—Schematic of apparatus used to measure laboratory splice loss.

possible the steady-state distribution at the input by properly focusing the input beam onto the fiber front face. Specifically, we made the $1/e^2$ width of the Gaussian field distribution in the input cone equal to the $1/e^2$ width of the steady-state far-field distribution measured at the end of the long sample (0.14 rad half-width). The distance between input and splice was 1 m.

If the splice disturbs the power distribution substantially, a sufficient fiber length must be provided after the splice to allow the distribution to settle, a process which generally is associated with some excess loss. To study the magnitude of this effect, we measured several joints with fiber lengths from 1 to 5 m after the joint. We did not find a consistent increase in loss as the length was increased, although 5 m is admittedly not a sufficient length to reach the steady state. Further study is necessary to estimate the error involved.

To make a good joint, the ends were aligned using a microscope to within a fraction of a degree in angle and within 1 μm laterally, but kept apart by at least 10 μm to avoid damage of the ends by mutual abrasion. A drop of glycerin was then added, which was held between the ends by surface tension. The refractive index of glycerin is 1.473 and almost coincides with that of silica glass ($n = 1.458$). This procedure invariably produced a splice with a loss of less than 1 percent (typically, 0.5 percent). This result was unaffected by rotating one end with respect to the other. Note that no information on the transmitted optical signal was required to achieve this optimal alignment.

To establish the order of magnitude of alignment tolerances permissible in practical splices, we measured the increase in loss as a result of longitudinal or lateral misalignments. The fiber ends could be parted axially by 100 μm (one core diameter) before the losses increased by 1 percent. Lateral displacements were more critical. The loss increased by 1 percent for a 5-μm displacement and by 4 percent for a 10-μm displacement (10 percent of the core diameter).

V. SUMMARY AND CONCLUSIONS

We have presented a theory of glass fiber fracture which allows us to design a machine for reliably producing clean breaks which leave the fiber ends in a suitable condition for splicing. We have built such a machine and demonstrated that, with 125-μm silica glass multimode optical fibers, such breaks are consistently obtained. Laboratory splicing experiments using fibers broken with this machine always produced splices with losses of less than 1 percent (0.04 dB).

VI. ACKNOWLEDGMENTS

We would like to thank R. D. Standley and F. A. Braun for the electron microscope photographs shown in Figs. 1 and 4.

REFERENCES

1. Bisbee, D. L., "Optical Fiber Joining Technique," B.S.T.J., *50*, No. 10 (December 1971), pp. 3153–3158.
2. Bisbee, D. L., "Measurements of Loss Due to Offsets and End Separation of Optical Fibers," B.S.T.J., *50*, No. 10 (December 1971), pp. 3159–3168.
3. Dyott, R. B., Stern, J. R., and Stewart, J. H., "Fusion Junctions for Glass Fiber Waveguides," Elec. Letters, *8*, 11 (June 1, 1972), pp. 290–292.
4. Krumpholz, O., "Detachable Connector for Monomode Glass Fiber Waveguides," Archiv Elektronik Ubertragungstechnik, *26* (1972) pp. 288–289.
5. Someda, C. G., "Simple, Low-Loss Joints Between Single-Mode Optical Fibers," B.S.T.J., *52*, No. 4 (April 1973), pp. 583–596.
6. McCormick, A. R., "Fiber Breaking Technique," unpublished memorandum.
7. Cherin, A. H., Eichenbaum, B. R., and Schwartz, M. I., unpublished memorandum.
8. Saunders, M. J., "Results for the Quality of the Edges of Fibers Cut with a Razor Blade," unpublished memorandum.
9. Gandrud, W. B., "On the Possibility of Two-Dimensional Fiber Splicing," unpublished memorandum.
10. Andrews, A. H., "Stress Waves and Fracture Surfaces," J. Appl. Phys. *30* (May 1959), pp. 740–743.
11. Johnson, J. W., and Holloway, D. G., "On the Shape and Size of the Fracture Zones on Glass Fracture Surfaces," Phil. Mag., *14* (1966), pp. 731–743.
12. Anderson, O. L., "The Griffith Criterion for Glass Fracture," in *Fracture*, B. L. Averbach, et al., eds., New York: Wiley, 1960.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Overload Model of Telephone Network Operation

By R. L. FRANKS and R. W. RISHEL

(Manuscript received June 5, 1973)

An analytic model for the steady-state behavior of an overloaded telephone network is given. The model includes trunk and machine congestion, retrials, "don't answer and busy," and some network management controls. It is significantly cheaper to use than Monte Carlo simulations for moderate size networks. It compares well with Monte Carlo simulation calculations of point-to-point completion probabilities and the expected number of messages in progress. It compares less well for sender attachment delay and probability of time-out calculations in switching machines.

I. INTRODUCTION TO THE PROBLEM

The purpose of this paper is to develop an analytic model of a telephone network which displays the major steady-state behavior of the network under overload conditions and which is computationally tractable. Besides being used to predict steady-state network operation in the presence of overload, such a model should help in the development of insight into network operation. Also, with an analytic model available, optimization theory can be brought to bear on various problems in network management.

Analytic network modeling seems to have been aimed at the trunking network design problem in the past. The usual approach was to assume that switching machines had enough capacity to have no effect on the traffic through them. Under these conditions, the stream of call attempts on a trunk group may have its distribution changed in two ways: the calls have previously been offered to a different trunk group and overflowed to the present one, or some of the calls in the stream were removed as a result of blocking on a trunk group in series with the present one. These cases have been handled by Wilkinson's Equivalent Random Method[1] and Katz's Carried Equivalent Method.[2]

When a network is overloaded, the effect of machine congestion is not negligible and must be taken into account. Early work on toll

machine congestion was done by Helly,[3] who considered a homogeneous group of identical machines connected by infinite trunk groups. His approach suggested the way we treat sender holding time in the switching machine model. Recently Szybicki[4] gave a model for an overloaded local switching machine.

Monte Carlo call-by-call simulations have been used to study network behavior. Recent examples at Bell Laboratories are simulations by J. A. Kohut[5] and J. M. McCormick. These simulations have the advantage of great flexibility. They also give transient response as well as the steady-state response of the network. Call-by-call simulations may require many runs, or long runs, to obtain reliable statistics for a process under study. They tend to be more expensive to run than analytic models.

II. INTRODUCTION TO A TELEPHONE NETWORK

From a traffic point of view, the network consists of end offices, switching machines, and trunks. The end offices serve as sources and destinations of calls. The trunks are message paths through the network. The switching machines are nodes in the network at which the choice is made of the path to be taken.

To illustrate the important effects in network operation, let us trace the progress of a typical call through the simple network shown in Fig. 1. The call enters the network through end office 1. It finds a free circuit on the trunk group from 1 to 2, attaches to it, and simultaneously bids for a sender in switching machine 2. After a short wait, it is accepted into machine 2. It finds the trunk group from 2 to 4 full, and attempts to attach to the trunk group from 2 to 3. There is a free circuit on that trunk group, so the call attaches to it and bids for a sender in machine 3. This process continues until the call enters end office 5. If the destination telephone is not busy, it rings. If it is answered, the attempted call is successful and becomes a message.

This typical call went through three switching machines. The block

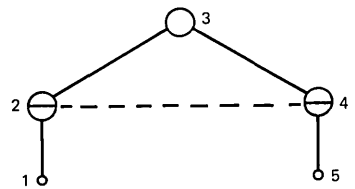


Fig. 1—Network used to show progress of typical call.

diagram in Fig. 2 shows the sequence of operations in a Bell System No. 4A-ETS switching machine.

A call coming into a switching machine enters a queue to wait for a vacant sender. When a call gets a sender, its destination information is impulsed to the sender, and the sender it had in the previous machine is released.

The call then queues for the remaining common control equipment, indicated here as a decoder-marker combination. This decoder-marker decides on the machine the call should be routed to next, tests for a vacant trunk, and sets up the connection, if possible. If there are no vacant trunks to appropriate subsequent machines, a no-circuit announcement is given. After a no-circuit announcement, the trunks on all the links over which the call had progressed are released. If the call is routed to a subsequent machine, it enters a queue for a sender in that machine.

The sender in the current machine is occupied by the call from the time it begins processing the call until the call has transmitted its destination information to the sender which processes the call in the subsequent machine. If a call waits longer than a fixed time to get a sender in the subsequent office, it is timed out. If it is timed out, the call is sent back to the marker-decoder which then connects it to a no-circuit announcement.

Under normal network operation, very few calls receive a no-circuit announcement, and fewer still time out while waiting for a sender. By far the most important causes of a call failing to become a message are for the called telephone to be busy when the call arrives and for the called customer to fail to answer the phone when it rings. When the network is overloaded, the number of no-circuit announcements increases, and time-outs become more frequent. Not only does the percentage of failures increase as the network becomes overloaded, but the number of successful attempts may actually decrease.

The factors underlying the decrease in the number of calls carried by the telephone network as it becomes highly overloaded were already understood in early work, such as Reference 3. As a call is being set up, it uses equipment in one switching machine until the next switching machine on its route accepts the call and receives the destination of the call from the previous machine. If a switching machine becomes overloaded, machines adjacent to it will have to wait longer to have their calls accepted and the destinations passed on. This causes an increase in the service time for putting a call through these machines. This in turn may cause the adjacent machines to

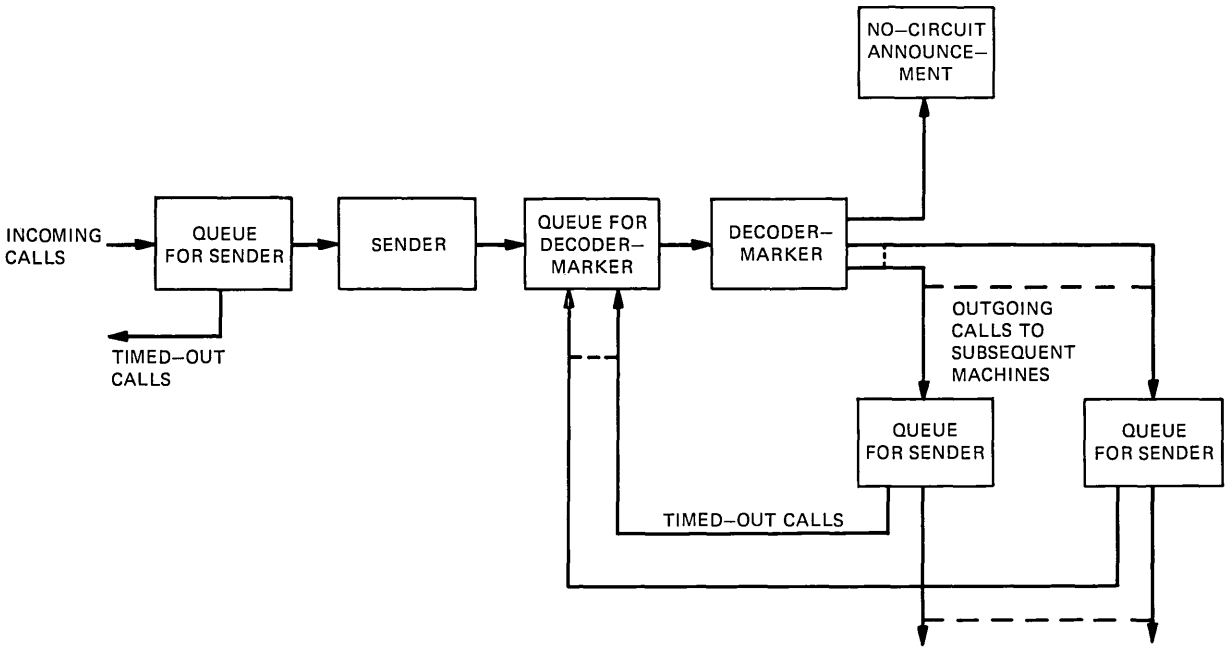


Fig. 2—Switching machine block diagram.

become congested. The time-out mechanism helps to relieve this congestive phenomenon. However, even with time-outs, switching machine congestion can back up throughout the network. Calls being set up occupy trunks on the partial route over which they have progressed. If a large number of calls are attempting on routes which are blocked, a portion of the capacity of certain links could be used by these ineffective attempts trying to set up. This would use capacity that could be utilized by talking calls. These ineffective attempts also use switching capacity in machines preceding the blockage. Most blocked attempts try again; these retrials increase the congestion.

The model to be set up will incorporate the features mentioned above. Based on those observations, the model must take into account both trunking congestion and switching machine congestion. Any call which enters the network will take trunk and machine capacity, even if it fails. For this reason, its effect on the network depends not only on whether a call succeeded or failed, but also on how far it progressed and over which particular route. This information must also be included in the model.

III. INTRODUCTION TO THE MODEL

Our view of modeling the network is probabilistic. We assume that for each trunk group there is a probability that an attempt on it will find a free circuit and for each switching machine there is a probability that an attempt on it will be accepted before timing out. We further assume that these acceptance probabilities are independent of the past history of the attempt.

The model has two conceptually distinct parts. First, the global problem is, given these acceptance probabilities, to find the various quantities of interest such as the expected number of messages in progress between each source-destination pair, the attempt rate on each trunk group and on each switching machine, and the point-to-point completion probability. Second, the local problem is, given those quantities, to find the acceptance probabilities for each switching machine and each trunk group. The local and global problems together give a large number of coupled nonlinear equations which describe the steady-state behavior of the network. These equations form the model.

For the model to be computationally tractable, the number of equations involved must be as small as possible. For this reason we have assumed that each stochastic process in the model can be described by a single parameter. (For example, the call origination process

between a given source-destination pair is assumed to be Poisson.) Even with this assumption, the number of equations involved is very large for any reasonable size network.

It must be emphasized that this is a first-generation model for an overloaded network. While the model does very well at predicting certain statistics, it is relatively poor at predicting others. A model using two parameters to describe each stochastic process could be more accurate than this one. Within the structure of this model, the switching machine treatment could be improved.

The network model given here is in several ways similar to the model used in the optimization problem of Reference 6.

IV. GLOBAL ASPECTS OF THE MODEL

This section deals with global, or network, effects caused by local phenomena. An example of such a global statistic is the mean number of messages in progress between a source-destination pair, which depends on various local effects such as the probability a call offered to each trunk group will be accepted onto it.

Before proceeding further, some definitions are required.

A complete route, $R = (a, b_1, b_2 \cdots b_n, c)$, is a list of the switching machines through which a call may pass in going from the end office connected to a machine a to the end office connected to machine c. For example, in Fig. 1 there are two complete routes, (2, 4) and (2, 3, 4), from end office 1 to end office 5.

A partial route, $r = (a, b_1, b_2 \cdots b_m; c)$ of a complete route R describes the route occupied by a call in the process of being set up and its destination. For example, a call on $r = (a, b_1, b_2, b_3; c)$ started in the end office attached to switching machine a, passed through machines a, b_1, and b_2, has entered (or is waiting to enter) machine b_3, and has as its destination the end office connected to machine c. Define

x_R = Expected rate that calls on complete route $R = (a, \alpha_1 \cdots \alpha_t, b)$ attach to the trunk group from switching machine b to its associated end office.

x_r = Expected rate that calls on partial route $r = (a, \alpha_1 \cdots \alpha_t; b)$ attach to the trunk group from α_{t-1} to α_t.

z_r = Expected rate that calls on partial route r are connected to senders in switching machine α_t.

t_r = Expected rate that calls on partial route r time out while waiting for senders in switching machine α_t.

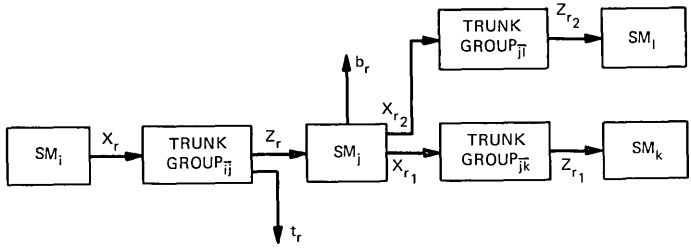


Fig. 3—Relationship of x_r, z_r, t_r, and b_r.

b_r = Expected rate that calls on partial route r which attach to senders in switching machine α_ℓ are blocked because of a lack of outgoing circuits.

SM_i = the ith switching machine.

To make the meaning of these variables more obvious, consider Fig. 3. Calls on partial route $r = (a, i, j; b)$ attach to trunk group $\overline{ij}$, and therefore bid for senders in SM_j at a rate x_r. Some of them are accepted, at a rate z_r, into SM_j and the remaining ones time out at a rate t_r.

Those calls accepted into SM_j then attempt to reach end office b by attaching to link $\overline{jk}$, at rate x_{r_1}, where $r_1 = (a, i, j, k; b)$.

If $x_{r_1} < z_r$, some of the calls alternate route over link $\overline{j\ell}$, with rate x_{r_2}, where $r_2 = (a, i, j, \ell; b)$. If $x_{r_1} + x_{r_2} < z_r$ and there is no further alternate route available, then some of the calls are blocked with rate

$$b_r \triangleq z_r - (x_{r_1} + x_{r_2}).$$

The global aspect of the network relates the x_r's, z_r's, b_r's, and t_r's for all partial routes in the network. The following assumptions are made:

(*i*) The model assumes that every type of call attempting to enter switching machine i has the same probability, P_i, of being accepted. Therefore

$$z_r = P_i x_r \tag{1}$$

for each partial route r entering SM_i.

(*ii*) It further assumes that every type of call attempting to attach to trunk group, $\overline{ij}$, has the same probability $P_{\overline{ij}}$ of attachment. The resulting equations are given below as (2) for the case with no network management control. Appendix A contains the equations that result when network management controls are included.

If a call from a to b in SM_i has a preferred route over link $\overline{ij}$ and a second most preferred route over link $\overline{ik}$, then

$$\begin{aligned} x_{r_1} &= P_{\overline{ij}} z_r \\ x_{r_2} &= P_{\overline{ik}}(z_r - x_{r_1}), \end{aligned} \tag{2}$$

etc., where

$$\begin{aligned} r &= (a,i;b) \\ r_1 &= (a,i,j;b) \\ r_2 &= (a,i,k;b). \end{aligned}$$

(*iii*) It is assumed that any call successfully attached to the trunk group leading to its destination end office, b, has a probability PA_b of being answered and therefore becoming a complete call. This probability depends on the destination. Therefore,

$$x_R = PA_b P_{\overline{bb}} z_r, \tag{3}$$

where

$$\begin{aligned} R &= (a,i,j,k,b) \\ r &= (a,i,j,k,b;b). \end{aligned}$$

(*iv*) The stream of original calls attempting to go from a to b is Poisson with mean $\lambda_{a,b}$. Calls that do not complete retry with probability PR. The expected long run attempt rate for calls from a to b is

$$A_{a,b} = \lambda_{a,b} + PR(A_{a,b} - C_{a,b}),$$

where $C_{a,b} = \sum_{R \text{ from } a \text{ to } b} x_R$, is the completion rate for all calls from a to b. That is,

$$A_{a,b} = \frac{\lambda_{a,b} - PR \sum_{R\ a \text{ to } b} x_R}{1 - PR}. \tag{4}$$

On the link from end office a to its switching machine, SM_a,

$$x_{(a;b)} = P_{\overline{aa}} A_{a,b}. \tag{5}$$

Clearly, with assumptions (*i*)–(*iv*), all the rates involved can be found from the P's. A method for finding the P's in terms of the x_r's and z_r's is discussed in the next section.

V. LOCAL ASPECTS OF THE MODEL

The assumptions made in Section IV for the global portion of the model place few constraints on the local part, requiring only certain acceptance probabilities, P_i, to satisfy those assumptions.

5.1 *Toll Machine Model*

The toll machines to be modeled are No. 4A-ETS switching machines. The statistics of interest in the model are the acceptance probabilities, P, mentioned in the last section, and the expected waiting time, T_w, for a call to get a sender. For relatively light loads, these statistics behave as though the machines selected the next call to be served on a first-come, first-served basis.[7] When the machines are overloaded, they behave as though calls were selected at random for service.[8]

We calculate the expected waiting time based on the former when loads are light and on the latter when they are heavy. The switch from the first to the second method is made where the curves of waiting time versus load cross.

In the light-load case the assumptions are

(*i*) The stream of calls attempting to enter each switching machine is a Poisson stream with mean λ_i.
(*ii*) The time a sender is held in SM_i by a call is an exponentially distributed random variable with mean $1/\mu_i$.
(*iii*) All toll machines have the same time-out interval, T.
(*iv*) The queuing discipline is first-come, first-served.

The problem of finding the probability, P_i, of acceptance of calls into SM_i under these assumptions has been solved.[9] The result is

$$P_i = 1 - \left\{ e^{T(N_{s_i}\mu_i - \lambda_i)} \left[\frac{1}{B(N_s, \lambda_i \mu_i^{-1})} - \frac{\lambda}{\lambda_i - N_{s_i}\mu_i} \right] + \frac{\lambda_i}{\lambda_i - N_{s_i}\mu_i} \right\}^{-1}, \quad \text{for} \quad \lambda_i \neq N_{s_i}\mu_i \tag{6}$$

$$P_i = 1 - \left\{ \frac{1}{B(N_{s_i}, N_{s_i})} + \lambda_i T \right\}^{-1}, \quad \text{for} \quad \lambda_i = N_{s_i}\mu_i$$

where

N_{s_i} = the number of senders in machine i

$B(n, a) = \dfrac{a^n/n!}{\sum_{j=0}^{n} a^j/j!}$, the Erlang B function.

Under these assumptions Reference 9 also gives the expected waiting time to get a server. The result is

$$T_{w_i}^r = (1 - P_i)\left\{\frac{N_{s_i}\mu_i e^{T(N_{s_i}\mu_i - \lambda_i)} - [N_{s_i}\mu_i + \lambda_i T(N_{s_i}\mu_i - \lambda_i)]}{(N_{s_i}\mu_i - \lambda_i)^2}\right\}, \quad \text{for} \quad \lambda_i \neq N_{s_i}\mu_i \tag{7}$$

$$T_{w_i}^r = (1 - P_i)T\left(1 + \frac{\lambda_i T}{2}\right), \qquad \text{for} \quad \lambda_i = N_{s_i}\mu_i.$$

In the heavy load case, assumption (*iv*) is replaced with

(*iv*)′ The queuing discipline is random.

An asymptotic expression for waiting time under this assumption is

$$T_{w_i}^r = T(1 - P_i/2), \qquad \lambda_i > N_{s_i}\mu_i. \tag{8}$$

Fitting eqs. (7) and (8) together

$$T_{w_i} = \begin{cases} T_{w_i}^r, & \lambda_i \leqq N_{s_i}\mu_i \\ \min\{T_{w_i}^r, T_{w_i}^r\}, & \lambda_i > N_{s_i}\mu_i. \end{cases} \tag{9}$$

To find P_i and T_{w_i}, all that is needed are T, λ_i, and μ_i^{-1}. T is given and

$$\lambda_i \equiv \sum_{r \in I_i} x_r \tag{10}$$

where

$$I_i = \{r \mid r = (a,\alpha_1 \cdots i; b) \text{ for some } a, b \text{ and } \alpha_1 \cdots\}.$$

To find μ_i^{-1}, consider Fig. 4. Calls on partial route r bid for a sender and connector in SM_i at rate x_r. All calls waiting to enter SM_i have an expected waiting time, T_{W_i}. Calls on partial route r are accepted into SM_i at rate z_r. It takes T_C seconds to connect an incoming trunk to a sender.

Once a sender is connected, it requires T_P seconds to pulse the digits into that sender. It then waits T_{WM} seconds for a translation device. The time to translate the digits, look for an available trunk on an acceptable route, and connect to that trunk is taken as a constant, T_M seconds. If no circuit is available, a call is attached to a no-circuit announcement. Otherwise it is attached to a trunk connected to some switching machine, say SM_j. The call, and the sender in SM_i, wait for a sender and connector in SM_j. If the call is accepted into SM_j, it takes T_C seconds to connect to the next sender and another T_P to pulse its digits into that sender. If the call hasn't been accepted by T seconds, it times out, and must return to the common control equipment to be connected to a no-circuit announcement.

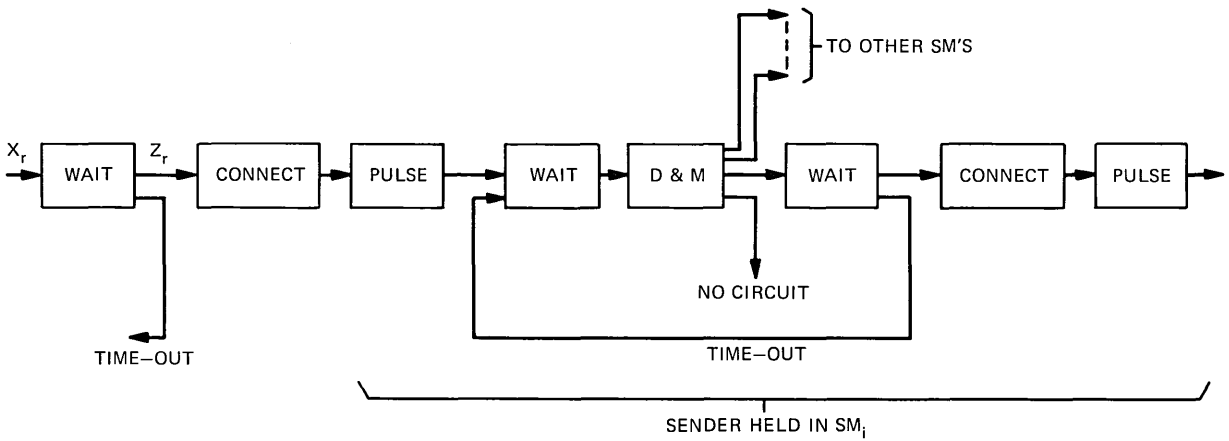


Fig. 4—Flow diagram for expected sender holding time.

The expected time a sender is held by any call accepted into SM_i is

$$\begin{aligned}\mu_i^{-1} = {} & (T_P + T_{WM} + T_M) \\ & + (\text{Expected waiting time for calls leaving } SM_i) \\ & + (T_{WM} + T_M)(\text{Probability a call leaving } SM_i \text{ times out}) \\ & + (T_P + T_C)(\text{Probability a call leaving } SM_i \text{ is accepted into the next machine}).\end{aligned}$$

More explicitly,

$$\begin{aligned}\mu_i^{-1} = {} & [T_P + T_{WM} + T_M] + \left[\frac{\sum_{r\in 0_i} T_{W_{j_r}} x_r}{\sum_{r\in I_i} z_r}\right] \\ & + [T_{WM} + T_M]\left[\frac{1 - \sum_{r\in 0_i} z_r + \sum_{R\in \bar{0}_i} x_R}{\sum_{r\in I_i} z_r}\right] \\ & + [T_P + T_C]\left[\frac{\sum_{r\in 0_i} z_r + \sum_{R\in \bar{0}_i} x_R}{\sum_{r\in I_i} z_r}\right], \quad (11)\end{aligned}$$

where

$$j_r = \alpha_t$$

when

$$r = (a,\alpha_1 \cdots \alpha_t; b)$$

and 0_i and $\bar{0}_i$ are sets of partial and complete routes, respectively, defined by

$$\begin{aligned}0_i &= \{r \mid r = (a,\alpha_1 \cdots i,\alpha_t; b) \text{ for some } a, b, \alpha_1 \cdots\} \\ \bar{0}_i &= \{R \mid R = (a,\alpha_1 \cdots i,b) \text{ for some } a, b, \alpha_1 \cdots\}.\end{aligned}$$

The only symbol in (11) remaining to be explained is T_{WM}, the expected time spent waiting for a decoder-marker. The decoder-markers are modeled as a finite source queue. The assumptions are

(*i*) There are N_m exponential servers.
(*ii*) The queuing discipline is first-come, first-served.
(*iii*) Each N_s sender in the switching machine either is waiting for or receiving marker service or is generating its next marker bid with an exponential interarrival time of mean $1/\gamma$.

This finite source queuing model has been analyzed in Reference 10. The result, in a convenient computational form thanks to D. Jagerman, is

$$T_{WM} = \frac{T_M}{N_m} \frac{N_s - M_m - A[1 - B(N_s - N_m - 1, A)]}{1 + B(N_s - N_m - 1, A) \sum_{i=1}^{N_m} \frac{N_m^{(i)} (N_s - N_m)^{(-i)}}{(\gamma T_M)^i}}, \quad (12)$$

where the subscripts corresponding to the switching machine have been suppressed and

$$A = \frac{N_m}{\gamma T_m}$$

$$N^{(i)} = N(N-1)\cdots(N-i+1)$$

$$N^{(-i)} = \frac{1}{N(N+1)\cdots(N+i-1)}.$$

To compute T_{WM}, it is necessary to know γ, which depends on the mean rate, m, at which calls arrive at the marker-decoder queue

$$m = \sum_{r \in I_i} z_r + \sum_{R \in O_i} t_r \tag{13}$$

Making use of Little's Theorem[11] and the definitions of m and γ,

$$\gamma = \frac{m}{N_s - m(T_M + T_{WM})}. \tag{14}$$

Equations (13) and (14) have a unique solution for all rates, m, which can be handled by the machines.

5.2 *Comments on the Switching Machine Model*

The interaction of switching machines in the real network is known to be an important cause of congestion. That interaction is included in this model by the waiting time of senders in one machine affecting the holding time of senders in adjacent machines.

This model has some features that appear to be ad hoc. The assumptions, however, are computationally convenient and give results similar to gross machine behavior. The network model has been structured to accept expanded machine models if they are required.

Section VII, on validation, includes a discussion of the accuracy of this switching machine model.

5.3 *Trunk Group Model*

The probability that an attempt will be accepted on a trunk group is found by assuming that the arrival process is Poisson. The required result is the Erlang B function, which depends only on the mean number of calls which would be on the trunk group if it were infinite and the actual number of trunks.

To find $P_{\overline{ij}}$ on the trunk group between i and j, we need the following definitions:

$N_{\overline{ij}}$ = number of trunks between SM_i and SM_j.
$E_{\overline{ij}}$ = expected number of calls on the trunk group between i and j.

$F_{\overline{ij}}$ = expected number of calls that would be on the trunk group if $N_{\overline{ij}}$ were infinite.

n_R = expected number of messages on complete route R.

S_r = expected number of calls on partial route r being processed in SM_ℓ where $r = (a \cdots \ell; b)$.

W_r = expected number of calls on partial route r waiting to enter SM_ℓ where $r = (a \cdots \ell; b)$.

$\frac{1}{\nu}$ = mean holding time for a message.

Then

$$\begin{aligned} E_{\overline{ij}} &= \sum_{R \text{ over } \overline{ij}} n_R + \sum_{r \text{ over } \overline{ij}} S_r + \sum_{r \text{ over } \overline{ij}} W_r \\ &= \frac{1}{\nu} \sum_{R \text{ over } \overline{ij}} x_R \\ &\quad + \sum_{\text{all } TC_\ell} \{T_{W\ell} \sum_{\substack{r \in I_\ell \\ r \text{ over } \overline{ij}}} x_r + (T_C + T_P + T_{WM} + T_M) \sum_{\substack{r \in I_\ell \\ r \text{ over } \overline{ij}}} z_r\} \\ &\quad + (T_{WM} + T_M) \sum_{\substack{r \text{ over } \overline{ij} \\ r \notin I_i \bigcup I_j}} t_r. \end{aligned} \tag{15}$$

To find $F_{\overline{ij}}$, simply notice that $P_{\overline{ij}}$ is a linear factor of every x_r, z_r, and t_r in eq. (15). Given all the P's except $P_{\overline{ij}}$, choose any positive value for $P_{\overline{ij}}$, say $\hat{P}_{\overline{ij}}$, use eqs. (1) through (15) to find $\hat{E}_{\overline{ij}}$, the value of $E_{\overline{ij}}$ corresponding to $\hat{P}_{\overline{ij}}$. Then

$$F_{\overline{ij}} = \frac{\hat{E}_{\overline{ij}}}{\hat{P}_{\overline{ij}}}. \tag{16}$$

Finally

$$P_{\overline{ij}} = 1 - \mathrm{B}(N_{\overline{ij}}, F_{\overline{ij}}), \tag{17}$$

where B is the Erlang B function.

It is unlikely that the arrival process at each trunk group in the network is Poisson. In fact, much of the recent trunking analysis has been directed toward non-Poisson processes. However, the Poisson assumption is a reasonable one to make in a mean value model. The accuracy of the overall model will be discussed in Section VII.

VI. SOLVING THE EQUATIONS

In the last two sections the model was given as a set of equations, (1) through (17). Most reasonable uses of the model require simultaneous solution of the equations and then computation of quantities of interest from this solution.

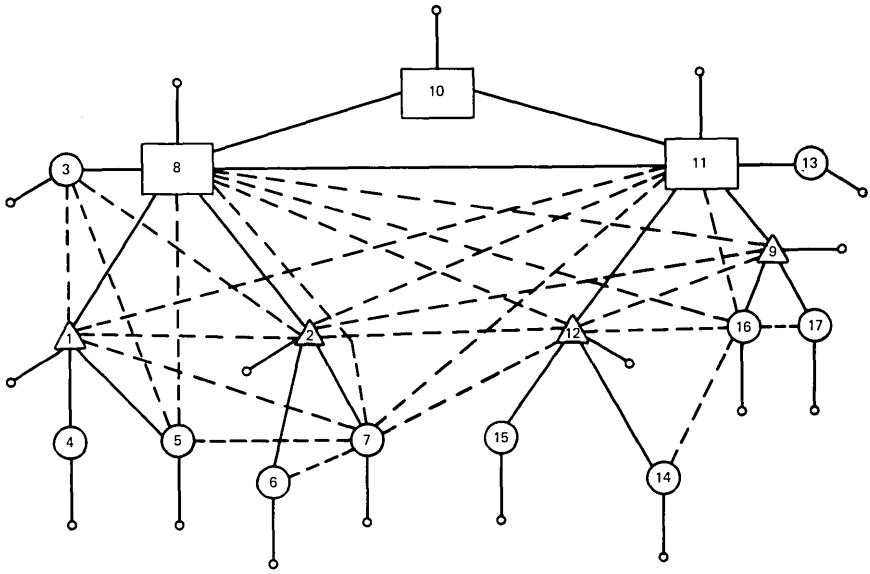


Fig. 5—Network used in point-to-point comparison run.

Solving these equations, at reasonable cost, is essential to the usefulness of the model. As given in Sections IV and V, there are a very large number of equations, both nonlinear and coupled. In the network shown in Fig. 5 are about 50,000 equations and variables. Fortunately, the P's and T_w's calculated in Section V form a fundamental set of variables from which everything else can be calculated. There are 91 of these variables for the network in Fig. 5. In another network of interest there are 240 of these variables.

The approach taken to solving the equations is iterative. Given a set of P's and T_w's, all other variables of Section IV are calculated. Then a new set of P's and T_w's are calculated. If the new set and the old set are the same, a solution has been found. More precisely, let $\mathbf{y}$ be a vector whose components are the P's and T_w's, then the equations of the model specify a function, $F(\mathbf{y})$, which gives the new value of P's and T_w's. In this framework solving the model equation is equivalent to solving

$$F(\mathbf{y}) = \mathbf{y}. \tag{18}$$

To make the solution of (18) easier, the components were normalized to the interval [0, 1]. The components corresponding to P's are necessarily in this interval. The components corresponding to T_w's were forced to be in the interval by replacing T_{w_i} by T_{w_i}/T. The re-

sulting function F maps $[0, 1]^n$ into itself, where n is the number of trunk groups plus twice the number of toll machines. The function is continuous so (18) has a solution by Brouwer's Fixed Point Theorem.[12] The question of uniqueness of the solution will be discussed later.

For this model to be a useful tool for network analysis, it is necessary to solve (18) inexpensively. There are two basic problems to be overcome. First, F is so complicated that, for reasonable size networks, its derivative is unavailable. This means that any method which requires F' cannot be used. Second, in these same networks a single evaluation of F costs on the order of \$0.50. The cost of estimating F' by n evaluations of F depends on n, the dimension of y. For the network in Fig. 5 it would cost about \$45 for a single estimate of F'.

In order to solve (18) economically, an algorithm which doesn't require F' or estimates of it was devised. The algorithm adapts the step size on the basis of the last ten evaluations of F.

The algorithm is as follows:

(*i*) Initialize $\mathbf{y}^0 \epsilon [0, 1]^n$ and set $i = 0$.

(*ii*) If $\|F\mathbf{y}^i - \mathbf{y}^i\| < \epsilon$, stop.

(*iii*) Otherwise,

$$\alpha_i = \min_{j \in I_i} \frac{1}{1 + \delta_j}$$

$$\mathbf{y}^{i+1} = \mathbf{y}^i + \alpha_i(F\mathbf{y}^i - \mathbf{y}^i),$$

where

$$I_i = \{j \mid j \text{ an integer} \geqq 0,\ i \geqq j \geqq i - 10\}$$

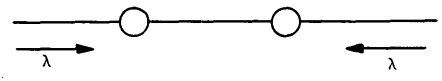
$$\delta_0 = 1$$

$$\delta_j = \frac{\|(F\mathbf{y}^j - \mathbf{y}^j) - (F\mathbf{y}^{j-1} - \mathbf{y}^{j-1})\|}{\|\mathbf{y}^j - \mathbf{y}^{j-1}\|}, \qquad j > 0.$$

(*iv*) Repeat (*ii*) with $i = i + 1$.

A computer program was written to evaluate $F(\mathbf{y})$, i.e., eqs. (1) through (17), and to implement the algorithm for solving (18). Our experience with the program has been that the algorithm usually reaches a satisfactory solution in less than $n/2$ steps. The cost of the program depends on the network size and on the value of various calling parameters. However, the cost for the network shown in Fig. 5 was usually around \$10, with some runs as high as \$40. For a larger network with 240 variables, the cost was usually around \$20.

An important point to mention is that F is not a contraction mapping. This means that the existence of a unique solution cannot be



N_S = 14 λ = 1 CALL/SEC

T_P = 2.5 SEC PR = 0

T = 30 SEC PA = 1

Fig. 6—Example network which has two quasi-stationary solutions.

guaranteed. In fact, in one example, two solutions to (18) were found. The question immediately arises, "Are both of these solutions physically meaningful?" The answer is "Apparently, yes." The existence of two stable operating regimes for the toll network has been suggested before.[3] The argument for their existence is as follows: If large queues form before switching machines, other machines will have their holding times greatly increased and will be able to switch only a fraction of their usual capacity. All other calls will time out. That means, if somehow the queues become large, they might stay large and the network completion rate would be very low, while under the same system parameters if the queues ever became small, they would stay small and the network completion rate would be much larger.

To test this argument, J. A. Kohut's Monte Carlo Simulation[5] of the network shown in Fig. 6 was run. A brief description of this simulation is given in Appendix B. The system started empty and was run for one hour of simulated time, reaching a quasi-stationary condition within the first 10 minutes. Then the offered load was doubled for 10 minutes, causing large queues to form. After that, the simulation was run for one more hour at the original traffic level. Again it reached a quasi-stationary state within 10 minutes. The results are given in Table I. The results show conclusively, for this simulated network, that two quasi-stationary operating regimes exist. For comparison, the results from the analytic model are also included in Table I.

It is possible for this network to be operating in the uncongested regime, receive an unusually large number of calls during some short time interval, and go into the congested regime. In the congested regime, if an unusually small number of calls arrived for a period of time, the system could go into the uncongested regime. It seems reasonable, and the simulation run helps confirm it, that the mean time before spontaneously leaving one of the regimes is quite long. This is the reason for using the term quasi-stationary.

TABLE I—EXAMPLE SHOWING TWO QUASI-STATIONARY SOLUTIONS

Model	Mean Attempts (per 5 min.)	Mean Time-outs (per 5 min.)	Mean Waiting Time (in sec.)
Simulation			
10 to 60 min.	602 ± 9	0	0.17 ± 0.02
80 to 130 min.	605 ± 8	454 ± 11	21.5 ± 0.2
Analytic			
Solution 1	600	10^{-9}	0.22
Solution 2	600	459	22.7

The existence of two quasi-stationary operating regimes apparently has implications for network management. If the network is congested, it may not be due to high calling rates but only due to high sender queues. A control which clears out these queues may be enough to decongest the network. Short sender timing which is currently used in the network is such a control.

VII. MODEL VALIDATION

The model discussed in previous sections contains many important network features. The machine model includes the stochastic arrival of attempts, office work times depending on waiting for adjacent offices, and time-outs. The trunk model includes stochastic arrivals and holding times, with the mean holding times dependent on how far the calls progressed toward becoming messages.

In the final analysis, the model stands or falls by how well it predicts the operation of a real network under overload. While from a validation viewpoint this could best be done by comparing the model with a real situation, there are two good reasons not to do so. First, the data collection problem would be extremely difficult and prohibitively expensive. Second, getting meaningful comparisons would require allowing the network to operate in an unacceptable mode. In addition, any real network would include things not modeled here, such as other types of switching machines and additional network management controls.

An alternative is to compare the model with a Monte Carlo simulation which is currently being used to evaluate network controls. While this comparison cannot evaluate the modeling of effects treated similarly in both models, it does help evaluate the modeling of effects treated differently. We compare our model with Kohut's simulation.

It also contains a simplified machine model, but does not include our simplifying assumptions on how senders wait for senders, that all arrival processes are Poisson, or that machine holding times are exponentially distributed. These assumptions are perhaps the most suspect in our model.

Two kinds of comparison between the two models were carried out. The first compares gross behavior over a very large range of offered loads. The second looks at more detailed statistics under a reasonable overload. In both cases, the results are similar.

The first type of comparison was carried out on the network in Fig. 7. The two models were given exactly the same data on machine sizes, pulsing times, trunk group sizes, etc. A series of runs was made with the only change between runs being the calling rates. The first run used a nominal set of calling rates, the second used twice the nominal calling rates, the third used three times the nominal calling rates, etc. For each run, the Monte Carlo simulation was run until, on the basis of the retrial rates, it appeared to be in steady state. It was then run for another one or two simulated hours to estimate the expected number of messages in progress in the network in steady state. The sample variance was used to estimate the 68 percent confidence interval for the mean. The analytical model was then used to find the expected number of messages in progress for each offered load. Figure 8 shows the curve generated by the analytic model as well as the Monte Carlo simulation's estimates of the corresponding means and confidence intervals.

In Fig. 9, exactly the same runs were made as in Fig. 8, except that switching machines timed out after 5 seconds in Fig. 8 and after 30 seconds in Fig. 9. Monte Carlo runs for more than double the nominal calling rates were not made in the 30-second case, since the simulation was not intended to handle the very large queues that would develop.

From Figs. 8 and 9, it appears that the models predict the same behavior for the expected number of messages in progress over a very large range of calling rates. The numerical values given by the two

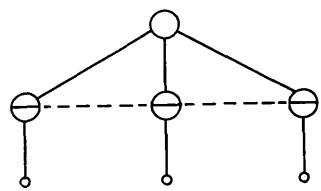


Fig. 7—Network used in massive overload comparison runs.

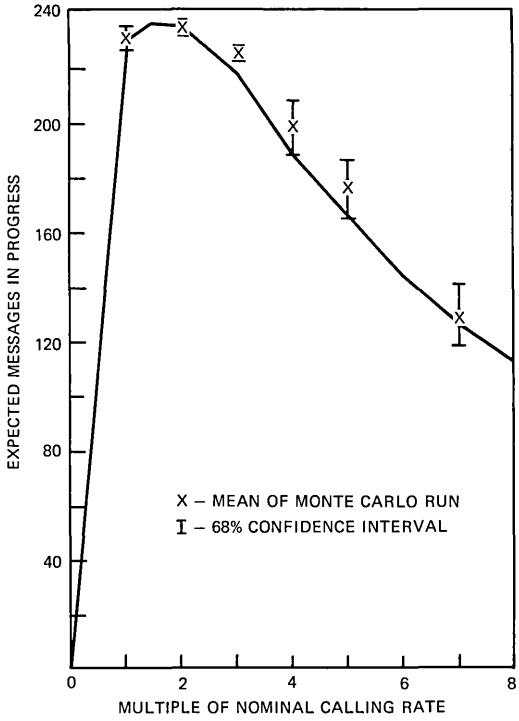


Fig. 8—Carried vs offered load, time-out = 5 seconds.

models also seem to agree well. At three times the nominal calling rate, the two means differ by less than 4 percent in Fig. 8.

The previous examples were generated for the network in Fig. 7. The second type of comparison was made between the Monte Carlo simulation and the analytic model on the network in Fig. 5. This network configuration was used in early network management simulation studies. While it is similar in structure to the toll network, it has one less level of hierarchy.

In order to get reliable statistics, the Monte Carlo simulation was run for three simulated hours. The network appeared to have reached equilibrium by the end of the first hour. Statistics were printed out at 10-minute intervals for the next two hours, and these were used to estimate completion probabilities, expected sender attachment delay, and the probability a call would time out in each switching machine.

Table II shows the comparison of the expected sender attachment delay and probability of time-out given by each model for each of the

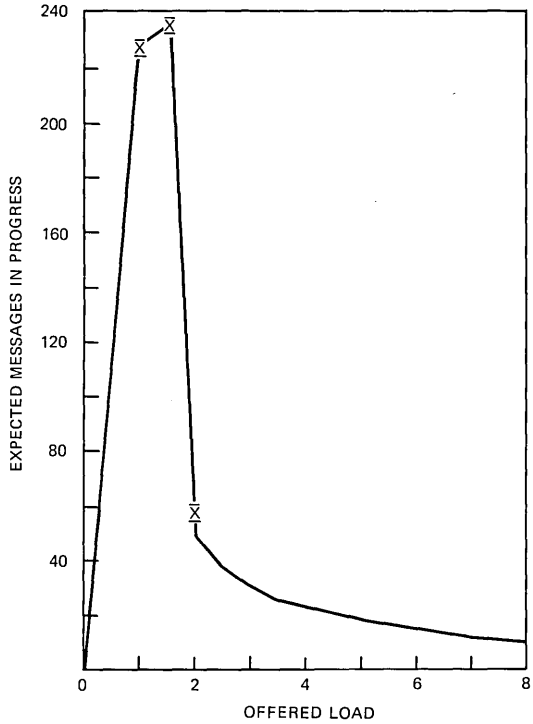


Fig. 9—Carried vs offered load, time-out = 30 seconds.

17 switching machines. For the Monte Carlo run, estimates of the standard deviation in the estimates are also given. It can be seen that the analytic model tends to give larger numbers for both quantities than does the Monte Carlo simulation.

To estimate the completion ratio for each point-to-point pair, the total number of attempts and completions were recorded for each pair for the last two simulated hours of the Monte Carlo run. The estimate of the completion ratio from source i to destination j is

$$\hat{C}R_{ij} = \frac{C_{ij}}{A_{ij}}$$

C_{ij} = number of completions from i to j

A_{ij} = number of attempts from i to j.

This estimate of the completion ratio was used since it has a smaller variance than would result if the completion ratio was calculated for each 10-minute interval and then averaged. The corresponding com-

TABLE II—COMPARISON OF ANALYTIC AND MONTE CARLO MODEL RESULTS FOR SWITCHING MACHINES

Switching Machine	Attachment Delay (in sec.)		Probability of Time-out	
	Analytic	Monte Carlo	Analytic	Monte Carlo
1	0.33	0.19 ± 0.03	0.008	0.008 ± 0.002
2	0.12	0.08 ± 0.01	0.002	0.002 ± 0.001
3	0.30	0.20 ± 0.03	0.016	0.012 ± 0.002
4	0.25	0.16 ± 0.03	0.013	0.007 ± 0.002
5	0.29	0.21 ± 0.03	0.015	0.012 ± 0.002
6	0.18	0.12 ± 0.02	0.008	0.007 ± 0.002
7	0.10	0.04 ± 0.02	0.004	0.001 ± 0.001
8	0.79	0.31 ± 0.03	0.011	0.009 ± 0.002
9	0.49	0.19 ± 0.03	0.016	0.007 ± 0.002
10	0.15	0.07 ± 0.01	0.005	0.003 ± 0.001
11	3.04	2.03 ± 0.07	0.219	0.163 ± 0.009
12	1.83	0.90 ± 0.06	0.085	0.054 ± 0.006
13	0.46	0.32 ± 0.04	0.027	0.020 ± 0.003
14	0.17	0.07 ± 0.02	0.008	0.002 ± 0.001
15	0.18	0.12 ± 0.02	0.008	0.007 ± 0.002
16	0.19	0.08 ± 0.01	0.009	0.003 ± 0.001
17	0.12	0.04 ± 0.01	0.005	0.002 ± 0.001

pletion ratio calculated by the analytic model will be denoted CR_{ij}. To compare $\hat{C}R_{ij}$ with CR_{ij}, it is necessary to have an estimate of the standard deviation of $\hat{C}R_{ij}$. This estimate was made as follows: If the actual completion ratio really is CR_{ij} and if the probabilities of completion are independent for successive ij attempts, then given the number of ij attempts, the number of ij completions is a binomial random variable. Therefore, $\hat{C}R_{ij}$ has mean and standard deviation

$$E[\hat{C}R_{ij}] = CR_{ij}$$

$$\sigma_{ij} = \sqrt{\frac{CR_{ij}(1 - CR_{ij})}{A_{ij}}},$$

respectively. The assumption that successive completion probabilities are independent is not unreasonable, since in this network the trunk groups between a typical ij pair will have an average of 10 to 50 message completions between successive ij attempts.

To conveniently compare CR_{ij} and $\hat{C}R_{ij}$, consider the standardized random variable

$$\xi_{ij} \triangleq \frac{\hat{C}R_{ij} - CR_{ij}}{\sigma_{ij}}. \tag{19}$$

Figure 10 has a histogram of the ξ_{ij}'s for 72 arbitrarily chosen ij pairs.

In the pairs plotted, 68 percent of the ξ_{ij}'s were in $[-1, 1]$, 92 percent were in $[-2, 2]$, and 100 percent were in $[-3, 3]$. This is consistent with the above assumptions. If they were correct, the expected percentages would be approximately 68, 95, and 99.7 percent, respectively.

To get a quantitative estimate of the difference in the completion ratios given by the two models, we require an estimate of $E[\xi_{ij}]$. To get such an estimate, treat the ξ_{ij}'s as independent, identically distributed random variables. Then, using the 72-pair sample to estimate $E[\xi_{ij}]$ and the standard deviation of that estimate gives

$$E[\xi_{ij}] = 0.026 \pm 0.143. \tag{20}$$

This is consistent with $E[\xi_{ij}] = 0$. However, using the estimated mean allows us to estimate the relative error, ϵ_{ij}, between the two models.

$$\epsilon_{ij} \triangleq E\left[\frac{\hat{C}R_{ij} - CR_{ij}}{CR_{ij}}\right]. \tag{21}$$

From (19),

$$\epsilon_{ij} = \frac{\sigma_{ij}E[\xi_{ij}]}{CR_{ij}}. \tag{22}$$

Using $E[\xi_{ij}] = 0.026$, ϵ_{ij} was calculated for all 72 point-to-point pairs. All the calculated values were in (0.0017, 0.0051) and the average was 0.0033. This gives the estimated relative difference in the completion ratios calculated from the two models as 0.33 percent. This error is negligible for practical purposes.

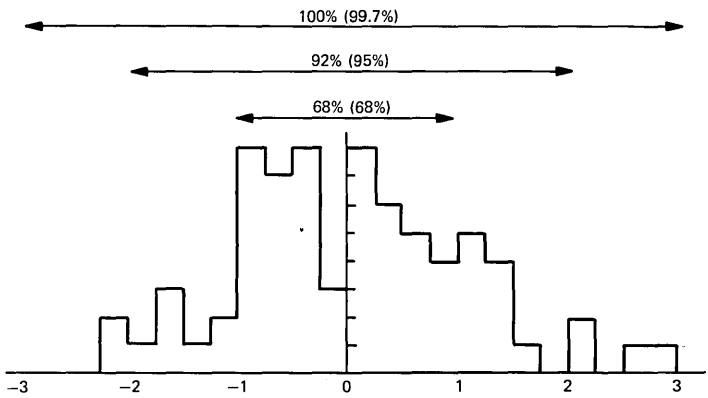


Fig. 10—Relative frequency of ξ_{ij}'s for 72 point-to-point pairs.

From the examples given in this section, it seems reasonable to conclude that

(*i*) The two models agree well when computing network quantities such as point-to-point completion ratios and the expected number of messages in progress.

(*ii*) They show somewhat less agreement in local phenomena such as the sender attachment delay and probability of time-out in individual machines.

VIII. CONCLUDING REMARKS

This paper presented an analytic model of the response of a telephone network to overloads. The model agrees well with a currently used simulation for network quantities such as point-to-point completion ratios and the expected number of calls in progress in the network. It is much cheaper to use, with typical costs being $20 versus $200. The model only gives quasi-stationary results. To get the transient response, a Monte Carlo simulation must be used.

The overall model structure permits changes in the model of individual components. Expanding the No. 4A switching machine model, including other types of switching machines, and including additional network management controls might be useful.

IX. ACKNOWLEDGMENTS

Many people have contributed ideas to this work. We wish particularly to acknowlecge contributions by Jack Holtzman, Sheldon Horing, Edwin Messerli, Irvin Yavelberg, Pat Spagon, Dave Jagerman, and George Hallas. We also take this opportunity to thank Elizabeth Murphy for early programming assistance, and Ellen Hill for producing the final program. Finally, we wish to thank John Kohut for help in connection with his Monte Carlo simulation model of the network.

APPENDIX A

Introducing Some Network Management Controls into the Model

For ease of exposition, eqs. (2) omitted network management controls. These equations must be modified to include network management controls corresponding to switching machine code blocking on the basis of destination, skip routing, cancellation of alternate routing from a trunk group, and cancellation of alternate

routing to a trunk group. To see how these controls are included, let

$$r = (a,i;d)$$
$$r_1 = (a,i,j;d)$$
$$r_2 = (a,i,k;d).$$

Partial route r enters switching machine i. Partial route r_1 is the preferred route out of machine i toward the destination d. Partial route r_2 is the alternative to route r_1. Also let

NCB_{id} = one minus the fraction of calls code blocked in switching machine i because they have destination d

NCF_{ij} = one minus the fraction of calls cancelled because they attempt to alternate route from trunk group ij

NCT_{ik} = one minus the fraction of calls cancelled because they attempt to alternate route to trunk group ik

NSK_{ik} = one minus the fraction of calls which skip over trunk group ik when alternate routing.

In terms of these symbols, the replacements for eqs. (2) are

$$\begin{aligned} x_{r_1} &= P_{ij} z_r NCB_{id} \\ x_{r_2} &= P_{ik} NSK_{ik} \{NCT_{ik} NCF_{ij} (z_r NCB_{id} - x_{r_1})\}. \end{aligned} \tag{2'}$$

The interpretation of the equations is as follows: The calls entering SM_i on route r which are not code blocked are offered to trunk group ij. They are accepted with probability P_{ij}. The quantity in parentheses corresponds to calls which are neither code blocked nor accepted into route r_1. The quantity in brackets corresponds to calls which are not cancelled because of alternate routing controls. Of those calls, the ones which do not skip trunk group ik and do find a free trunk enter route r_2. Those calls which do skip trunk group ik or find it full will be offered to the next alternate route, if one exists.

APPENDIX B

A Brief Description of the Monte Carlo Simulation in Reference 5

The Monte Carlo simulation in Reference 5 is a call-by-call simulation in the sense that it generates calls individually and processes them through the simulated network as individual entities. That is, each run of the simulation of Reference 5 produces a realization of the underlying stochastic process as opposed to the model presented here which analytically arrives at statistics for that process. The remainder of this

appendix describes the assumptions and treatments used in the simulation.

The underlying traffic between each source-destination pair is a Poisson stream. Any attempt which reaches its destination end office has a fixed probability of failing because of a "don't answer" or "busy" condition. Any attempt which fails to become a message, for any reason, will retry with a fixed probability. If a failed attempt will retry, the time until retrial is chosen from an exponential distribution. The conversation length for each successful attempt is also chosen from an exponential distribution.

An attempt which arrives at a trunk group can seize a trunk only if one is free at that time. Once a trunk is seized, it is held while the attempt progresses through the network. If the attempt fails, the trunk is released at the time of failure. If the attempt becomes a message, the trunk is also held for the duration of the conversation.

The simulation contains several switching machine models, only one of which was used in the comparisons in this paper. It consists of two groups of parallel servers: the first models the senders, while the second models the common control responsible for translation, trunk testing, and switching. We will refer to the first group as the senders and the second as the markers.

An attempt bids for a sender, if one is seized, then a constant delay is introduced to represent receiving digits. If a sender is not seized within a specified time, the attempt abandons the queue. After a sender has received the digits, a bid is made for a marker. If one is available, it is seized and held for a constant holding time. If one is not available, the sender will wait for a marker. During the marker operation, a test is made for a free trunk. If no trunk is available, the call is immediately blocked and the sender and all prior seized trunks are released. In the simulation, it is assumed that announcements and reorder tone do not extend the holding time of blocked attempts.

After the marker holding time, the sender bids for an attachment to a sender at a distant machine. This bid will result in either an attachment of a sender or an intersender time-out. In the former case, the sender is held for an additional constant length of time which simulates outpulsing the digits. In the latter case, no out-pulsing occurs, but an additional marker usage is required to route the attempt to an announcement.

The queuing discipline for senders and markers is random. When a piece of common control becomes free, a bid is selected at random from the bids waiting. The simulated switch of an attempt through a

switching machine may encounter delay in four different ways. An attempt will be delayed during the sender and marker service times and may be delayed by waiting for these pieces of equipment if they are not available at the time of the bid. The service time delays are fixed, so these delays are equal for all attempts. However, the delays caused by queuing are random and are dependent upon how long an attempt must wait for equipment to become free.

REFERENCES

1. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
2. Katz, S. S., "Statistical Performance Evaluation of a Switched Communications Network," Fifth International Teletraffic Conference, London (June, 1967), pp. 566–575.
3. Helly, W., "Two Stochastic Traffic Systems Whose Service Times Increase With Occupancy," Operations Research, *12* (1964), pp. 951–963.
4. Szybicki, E., "Approximate Method for Determination of Overload Ability in Local Telephone Systems," Seventh International Teletraffic Conference, Stockholm (June, 1973).
5. Kohut, J. A., unpublished work (see Appendix B above).
6. Franks, R. L. and Rishel, R. W., "Optimum Network Call Carrying Capacity," B.S.T.J., *52*, No. 7 (Sept. 1973), pp. 1195–1214.
7. Cardwell, R. H., unpublished work.
8. Hallas, G. A., unpublished work.
9. Gnedenko, B. V. and Kovalenko, I. N., *Introduction to Queueing Theory*, Israel Program for Scientific Translation (1968), pp. 33–39.
10. Descloux, A., *Delay Tables for Finite- and Infinite-source Systems*, New York: McGraw–Hill (1962).
11. Little, J. D. C., "A Proof of the Queueing Formula: $L = \lambda W$," Operations Research, *9* (1961), pp. 383–387.
12. Dunford, M., and Schwartz, J. T., *Linear Operators, Part 1*, London: Interscience (1958), p. 453.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Peakedness of Traffic Carried by a Finite Trunk Group With Renewal Input

By H. HEFFES and J. M. HOLTZMAN

(Manuscript received May 17, 1973)

In trunking theory, peakedness is defined conventionally as the variance-to-mean ratio of a traffic load when carried on an infinite trunk group. For analysis of switching machine delays, it has proven useful to define a peakedness measure associated with the Carried Arrival Process (CAP), the stream of call arrivals carried on an incoming trunk group. The peakedness of the CAP is defined to be the conventional peakedness of a fictitious traffic-load process generated by associating with each carried arrival an independent exponentially distributed holding time with mean equal to the mean of calls actually carried on the trunk group.

The problem considered is the effect of trunk group congestion on the peakedness of the CAP for traffic consisting of renewal inputs offered on a blocked-calls-cleared basis to a finite trunk group with exponential holding times. The CAP is characterized as a semi-Markov process. This model leads to the determination of the peakedness of the CAP. Numerical results illustrate the reduction of peakedness, or smoothing, introduced by the congestion.

I. INTRODUCTION

This paper is concerned with characterizing the traffic offered to a switching machine, taking into account both the alternate routing that the traffic may have undergone and the smoothing of the traffic resulting from congestion on the trunk group incoming to the machine. In trunking theory, peakedness is defined conventionally as the variance-to-mean ratio of a traffic load carried on an infinite trunk group. It is well known that trunk group blocking of peaked traffic, such as overflow traffic, can be substantially larger than the blocking seen by Poisson traffic with the same intensity. Similarly, switching machine* delay and capacity can be quite sensitive to the peakedness

* Throughout this paper, when we refer to a switching machine we mean the common control devices in a switching machine.

of the incoming traffic.[1] To determine the peakedness of the traffic offered to a switching machine, we must take into account the smoothing effect of the incoming trunk groups. To this end, we consider the process of arrivals offered to a trunk group which are carried by that trunk group. We call this process the Carried Arrival Process, or CAP.

To illustrate the CAP, consider the alternate routing network shown in Fig. 1. Here traffic overflowing trunk group AB is then offered to trunk group AC [Fig. 1(c)]. Those calls finding free circuits on AC then appear at node C as requests for service. The CAP is illustrated in Fig. 1(d).

The basic model used in the analysis is shown in Fig. 2 where a renewal process is offered to a group of N trunks on a blocked-calls-cleared (BCC) basis. The renewal input allows us to consider overflow traffic offered to an incoming trunk group. The holding times on the trunks are assumed to be independent, identically distributed, exponential random variables with service rate μ.

For analysis of machine performance it has proven useful to define a measure of peakedness for the CAP, z_c, equal to the variance-to-mean ratio of the traffic load carried (number of trunks occupied) on an infinite trunk group to which the Carried Arrival Process has been offered. By definition, the holding times on the infinite trunk group

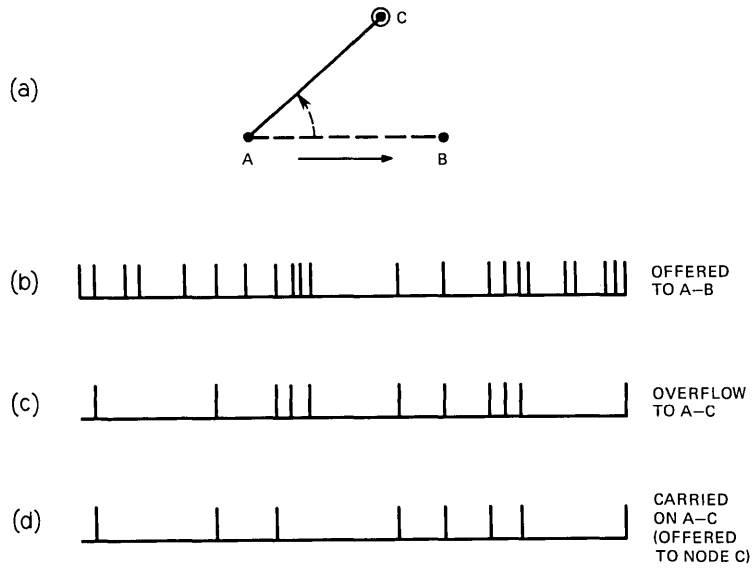


Fig. 1—Carried Arrival Process.

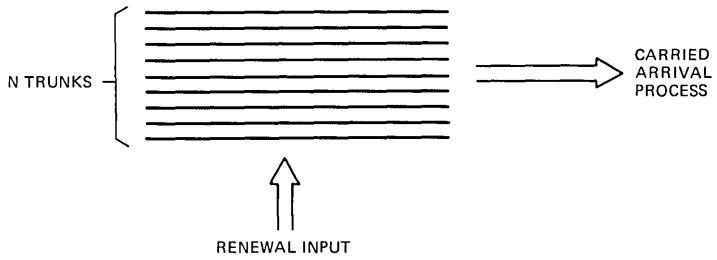


Fig. 2—Carried arrival model.

are independent, identically distributed, exponential random variables (with service rate μ) which are *independent* of the holding times on the incoming trunks.*

The peakedness of the CAP should be distinguished from the variance-to-mean ratio of busy trunks on the incoming group, a quantity which is discussed, for example, in Section 8.4 of Reference 2 for the case of Poisson input. This distinction can be made clear by considering an example of Poisson traffic of intensity λ(calls/second) offered to N trunks. As λ approaches zero, the two measures approach one. Clearly, as λ gets large, the variance of busy circuits in the trunk group goes to zero and the mean goes to N, giving a variance-to-mean ratio of zero. On the other hand, as λ gets large, the time differences between successive carried calls approach independent, exponential random variables with rate $N\mu$ (i.e., a Poisson stream) and the peakedness z_c approaches unity. This is illustrated graphically in Fig. 3 which plots z_c and $(v/m)_{B.S.}$ (variance-to-mean ratio of busy servers on the N trunks) as a function of offered load for $N = 10$. The example is a special case of the general results derived in this paper for arbitrary renewal input to the trunk group.

By modeling the CAP as a semi-Markov process (SMP), it becomes possible to calculate peakedness z_c as a function of the peakedness of the traffic offered to the incoming trunk group and of the congestion encountered on the group. The resulting z_c may then be used in the determination of machine performance.[1] Numerical results illustrate the reduction in peakedness, or smoothing, introduced by the trunk group congestion. In the course of determining the peakedness, the transform of the distribution of the time between carried calls is derived.

* That is, although carried arrivals are accepted simultaneously on the finite and infinite trunk groups, the departure times are different.

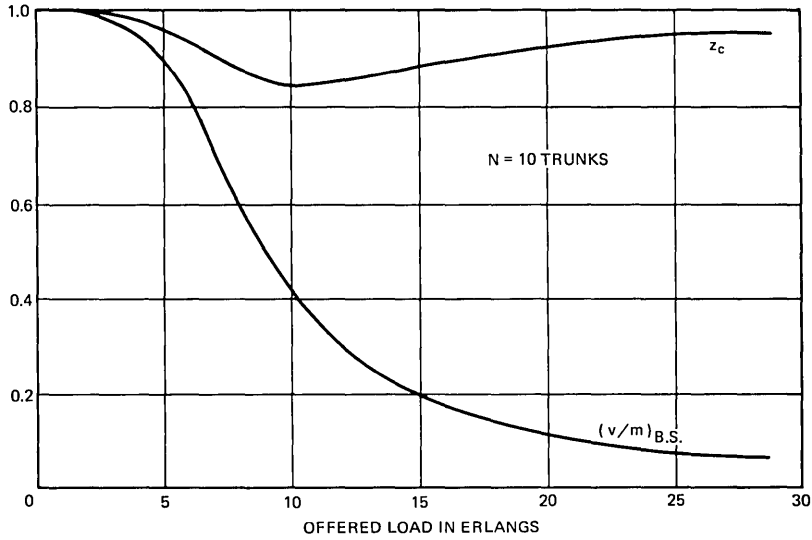


Fig. 3—Distinction between peakedness and variance-to-mean ratio of busy servers.

II. OUTLINE OF RESULTS

In this section, we give an informal overview of the results. In particular, we give the key equations that a user can employ to determine the peakedness, z_c, of a Carried Arrival Process. The equation numbers are the same as will be used in the derivation. Stationarity is assumed throughout. In all that follows, we assume unity holding time (or the time unit is the mean holding time).

First of all,

$$z_c = M^+ - \frac{1}{\beta}, \tag{11}$$

where M^+ is the mean number of calls up on the infinite trunk just after the time that a call is accepted onto the finite (and infinite) trunk group. $(1/\beta)$ is the mean number of carried calls (on both the finite and infinite trunk group).*

M^+ is determined from the following equation:

$$M^+ = \frac{1}{1 - \phi(1)} - M_N^+ P_N^+ \left[\frac{\phi(N+1)}{1 - \phi(N+1)} \right], \tag{32}$$

where $\phi(s)$ is the Laplace-Stieltjes transform of the interarrival time

* Note that, since we are assuming unity holding time, M^+ and $(1/\beta)$ are in erlangs.

distribution. M_j^+ is the mean number of calls up on the infinite trunk group immediately after arrival of a carried call, given j calls in progress on the finite trunk group immediately following the arrival of the carried call. P_j ($j = 0, 1, \cdots, N$) is the probability that an arrival finds j trunks busy on the finite access group, and

$$P_j^+ = \frac{P_{j-1}}{1 - P_N} \tag{24}^*$$

is the probability of j calls up on the finite trunk group immediately following a carried call arrival.

The only quantity left to be determined in (32) is M_N^+ which is calculated by solving the linear equations

$$[M_m^+ - 1]P_m^+ = \sum_{l=\max(1,m-1)}^{N} C_{lm}M_l^+, \qquad m = 1, 2, \cdots, N, \tag{33}$$

where

$$C_{lm} = P_l^+ \binom{l}{m-1} \sum_{\eta=0}^{l-m+1} \binom{l-m+1}{\eta} (-1)^\eta \phi(\eta + m) \qquad m - 1 \leqq l \leqq N - 1, \quad 1 \leqq m \leqq N, \tag{34}^\dagger$$

$$C_{Nm} = \frac{P_N^+ \binom{N}{m-1} \sum_{\eta=0}^{N-m+1} \binom{N-m+1}{\eta} (-1)^\eta \phi(\eta + m)}{1 - \phi(N + 1)}. \tag{35}^\dagger$$

A simple method of solving (33) is discussed at the end of Section VI. In the course of deriving the expressions which ultimately determine the peakedness of the CAP, the Laplace-Stieltjes transform of the distribution of time between carried call arrivals is obtained. This is given by

$$\phi_c(s) = \phi(s) - \frac{[1 - \phi(s)]\phi(s + N)}{1 - \phi(s + N)} P_N^+. \tag{23}$$

Note that the CAP is not completely characterized by (23), since it is not generally a renewal process.

Examples are given in Section VII.

---

* One method of computing P_j is via the equations given on p. 179 of Reference 3. Alternate methods which may avoid some of the numerical difficulties inherent in this approach will be discussed in Appendix C.

† Alternate expressions for special cases, more suitable for computation, are given in Appendix C.

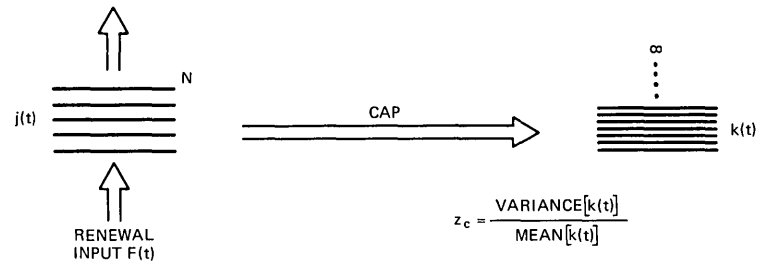


Fig. 4—Peakedness of the Carried Arrival Process.

III. CHARACTERIZATION OF THE CARRIED ARRIVAL PROCESS AS A SEMI-MARKOV PROCESS

Consider a renewal process, with a nonlattice interarrival time distribution $F(t)$, as an input to N trunks with mutually independent exponential holding times each with mean unity (or the time unit is the mean holding time). $F(0^+)$ is assumed to be zero. Blocked calls are cleared. Such a system is analyzed in Chapter 4 of Reference 3.

Let $\{\tau_i, i = 1, 2, \cdots\}$ denote the sequence of times at which calls are accepted by the N servers (this is, of course, a subset of the times at which calls are offered). Let $j(t)$ be the number of servers busy at time t. Then $j(\tau_n^+)$ is the number of servers busy just after the nth carry (nth carried call). Note that $P\{j(\tau_n^+) = 0\} = 0$. It is clear that the $j(t)$ process held fixed at $j(\tau_n^+)$ for $\tau_n < t < \tau_{n+1}$ is an SMP.* The transition probabilities for the embedded Markov chain are derived as follows.

Since there is a death process on the finite trunk group between carried calls, we have for $m = 1, 2, \cdots, N$ and for $l = m - 1, \cdots, N - 1$ (in which case, the next arrival is the next carry)

$$\begin{aligned} P\{j(\tau_n^+) = m \mid j(\tau_{n-1}^+) = l\} &= \int_0^\infty \binom{l}{m-1} e^{-(m-1)t}[1 - e^{-t}]^{l-m+1} dF(t) \\ &= \binom{l}{m-1} \sum_{\eta=0}^{l-m+1} \binom{l-m+1}{\eta} (-1)^\eta \phi(\eta + m - 1), \quad (1) \end{aligned}$$

where $\phi(s)$ is the Laplace-Stieltjes transform of $F(t)$,

$$\phi(s) = \int_0^\infty e^{-st} dF(t). \quad (2)$$

* For an introduction to semi-Markov processes, see Reference 4, Chapter 5.

When $j(\tau_n^+) = N$, note that the next arrival need not be the next carry. With s.c. denoting service completion and $F_i(t)$ the ith-fold convolution of $F(t)$, we have

$$P\{j(\tau_n^+) = m \mid j(\tau_{n-1}^+) = N\}$$

$$= \sum_{i=1}^{\infty} P\left\{\begin{array}{l}\text{no s.c. before the first } (i-1) \\ \text{arrivals after } \tau_{n-1},\ (N-m+1) \text{ s.c.'s} \\ \text{before } i\text{th arrival after } \tau_{n-1}\end{array} \middle| \; j(\tau_{n-1}^+) = N\right\}$$

$$= \sum_{i=1}^{\infty} \int_0^{\infty} \int_0^t e^{-Ns} \binom{N}{m-1} e^{-(m-1)(t-s)}[1 - e^{-(t-s)}]^{N-m+1} \times dF_{i-1}(s)dF(t-s)$$

$$= \frac{\binom{N}{m-1} \sum_{\eta=0}^{N-m+1} \binom{N-m+1}{\eta} (-1)^{\eta}\phi(\eta + m - 1)}{1 - \phi(N)}. \tag{3}$$

Letting $F_{lm}(t)$ be the conditional probability that a transition will take place within a time t, given that the process has just entered l and will next enter m, we have

$$dF_{lm}(t) = \begin{cases} \dfrac{\binom{l}{m-1} e^{-(m-1)t}[1 - e^{-t}]^{l-m+1} dF(t)}{P\{j(\tau_n^+) = m \mid j(\tau_{n-1}^+) = l\}}, \\ \qquad\qquad l = m-1, \cdots, N-1, \\ \sum_{i=1}^{\infty} \int_0^t e^{-Ns} \binom{N}{m-1} e^{-(m-1)(t-s)} \\ \quad \dfrac{\times[1-e^{-(t-s)}]^{N-m+1} dF_{i-1}(s)dF(t-s)}{P\{j(\tau_n^+) = m \mid j(\tau_{n-1}^+) = N\}}, \qquad l = N. \end{cases} \tag{4}$$

Although the SMP characterization of the CAP is of general interest, it is particularly useful in determining the peakedness of the CAP as we shall see in Section IV.

IV. PEAKEDNESS OF THE CARRIED ARRIVAL PROCESS

Recall that the peakedness of a process is the variance-to-mean ratio of the number of calls up on an infinite trunk group when that process is offered to the infinite trunk group. To determine the peakedness of the CAP, consider the situation shown in Fig. 4. In this figure, each time a call is carried on the finite group, a call is put up on the infinite group with an exponentially distributed holding time with the same mean as on the N-trunk group but independent of the N-trunk

group holding time.* As before, $j(t)$ is the number of busy servers on the N-trunk group at time t. Let $k(t)$ be the number of busy servers on the infinite trunk group at time t. We use the following familiar result (given for renewal input in Reference 3 and for semi-Markov input in Reference 5, which is the form applicable to our problem). The following limits

$$\bar{P}_i = \lim_{n\to\infty} P\{k(\tau_n^-) = i\} \tag{5}$$

and

$$\bar{P}_i^* = \lim_{t\to\infty} P\{k(t) = i\} \tag{6}$$

exist and satisfy

$$\bar{P}_i^* = \frac{\bar{P}_{i-1}}{i\beta}, \tag{7}$$

where β is the mean time between transitions of the SMP (i.e., mean time between carried calls). From (7) we obtain

$$\lim_{t\to\infty} E\{k^2(t)\} = \frac{1}{\beta}\left[\lim_{n\to\infty} M(\tau_n^-) + 1\right] = \frac{1}{\beta}\lim_{n\to\infty} M(\tau_n^+), \tag{8}$$

where we have defined

$$M(t) = E\{k(t)\}. \tag{9}$$

Defining

$$M^+ = \lim_{n\to\infty} M(\tau_n^+), \tag{10}$$

the peakedness of the CAP (denoted z_c) is given by

$$z_c = M^+ - \frac{1}{\beta}. \tag{11}^\dagger$$

Note that $1/\beta$ corresponds to the mean of the carried load (recall that we are assuming unity mean holding time).

We are thus left with the problem of determining M^+. This determination will be in terms of the distribution of time between carried calls, to which the next section is devoted.

V. DISTRIBUTION OF TIME BETWEEN CARRIED CALLS

Consider an arrival at τ_n which finds a free circuit [i.e., $j(\tau_n^-) < N$]. Let $F_c(t)$ be the distribution of time until the next carry (carried call);

---

* It is this independence which distinguishes the peakedness from the variance-to-mean ratio of busy trunks on the N-trunk group (as discussed in Section I).

† This is given in Reference 6 for the case of renewal input and weaker assumptions on service times.

i.e.,

$$F_c(t) = P\{ict \leqq t | \text{carry at } \tau_n\}, \tag{12}$$

where *ict* denotes inter-carry time. Denoting

$$F_c[t | j(\tau_n^+) < N] = P\{ict \leqq t | j(\tau_n^+) < N\}$$

and

$$\bar{F}_c(t) = P\{ict \leqq t | j(\tau_n^+) = N\} \tag{13}$$

and recognizing that

$$F_c[t | j(\tau_n^+) < N] = F(t) \tag{14}$$

yields

$$F_c(t) = F(t)\left[1 - \frac{P_{N-1}}{1 - P_N}\right] + \bar{F}_c(t)\left[\frac{P_{N-1}}{1 - P_N}\right], \tag{15}$$

where

$$P_j = P\{j \text{ trunks busy on the finite group just before a call arrival}\} \tag{16}$$

is the stationary call congestion probability given on p. 179 in Chapter 4 of Reference 3. In particular, P_N is the blocking probability.

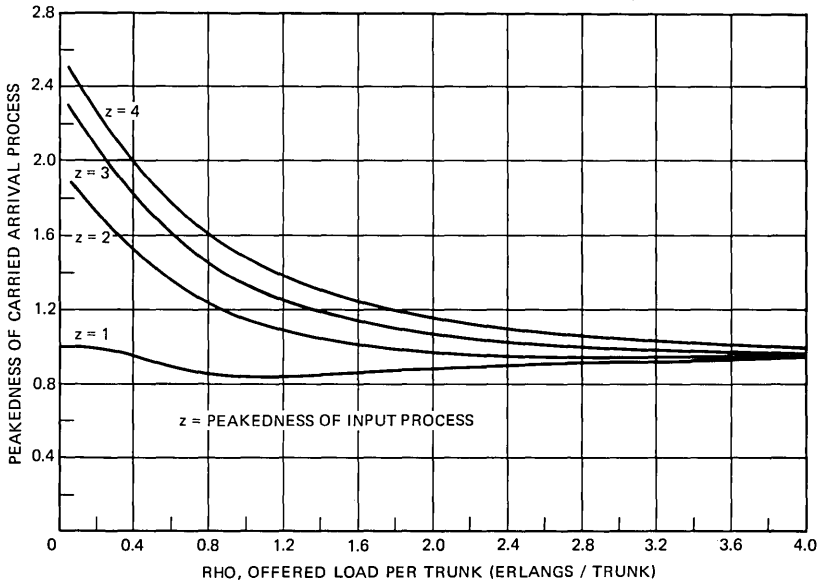


Fig. 5—Peakedness of CAP (5 trunks).

Let $\phi_c(s)$ and $\bar{\phi}_c(s)$ be the Laplace-Stieltjes transforms of $F_c(t)$ and $\bar{F}_c(t)$, respectively. Transforming (15) gives

$$\phi_c(s) = \phi(s)\left[1 - \frac{P_{N-1}}{1 - P_N}\right] + \bar{\phi}_c(s)\left[\frac{P_{N-1}}{1 - P_N}\right]. \tag{17}$$

The function $\bar{\phi}_c(s)$ can be obtained from the solution to the Type I counter problem given on p. 207 of Reference 3 with renewal input transform $\phi(s)$:

$$\bar{\phi}_c(s) = [1 - \phi(s)] \int_0^\infty e^{-sy} H(y) dm(y) , \tag{18}$$

where, for our problem,

$$H(t) = 1 - e^{-Nt} \tag{19}$$

and

$$m(t) = E\{\text{number of arrivals in } (0, t)\}. \tag{20}$$

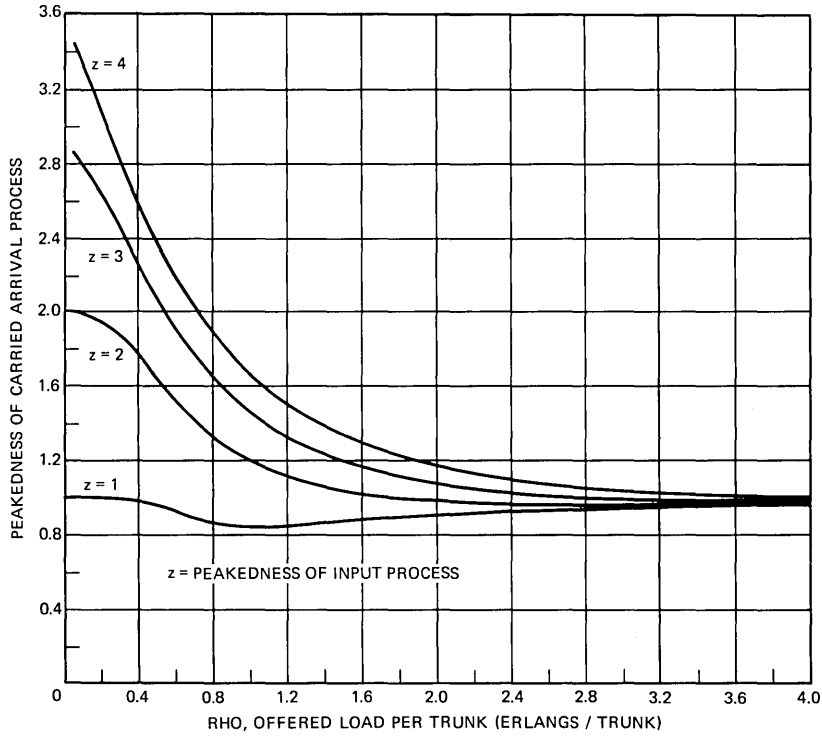


Fig. 6.—Peakedness of CAP (10 trunks).

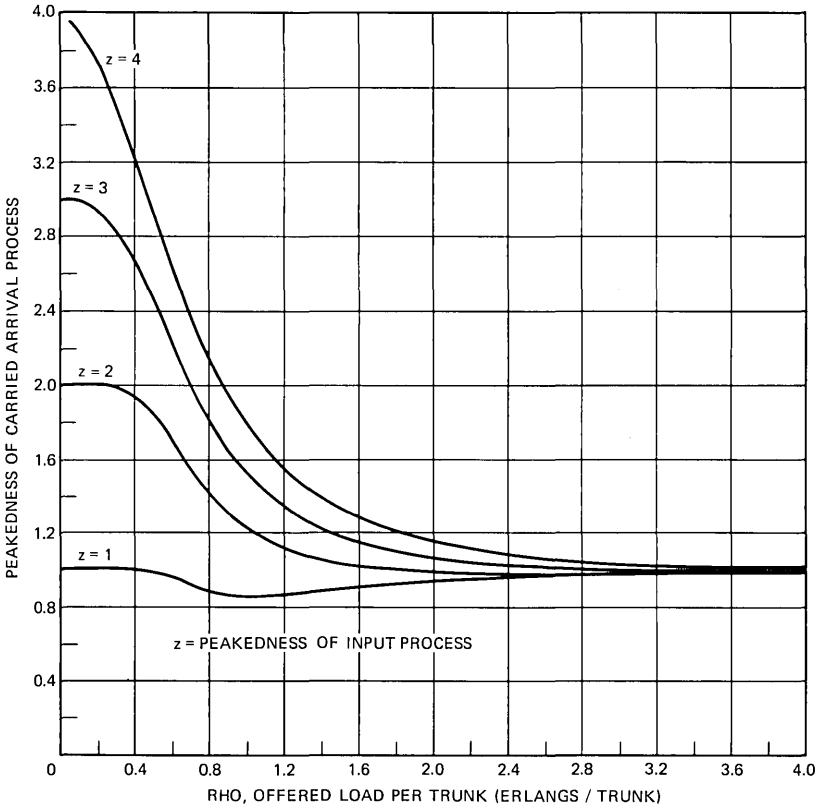


Fig. 7—Peakedness of CAP (20 trunks).

From (18) we have

$$\tilde{\phi}_c(s) = [1 - \phi(s)]\left[\frac{\phi(s)}{1 - \phi(s)} - \frac{\phi(s + N)}{1 - \phi(s + N)}\right], \tag{21}$$

where we have used

$$\int_0^\infty e^{-st} dm(t) = \frac{\phi(s)}{1 - \phi(s)}. \tag{22}$$

Combining (17) and (21) gives the transform of the intercarry time distribution:

$$\phi_c(s) = \phi(s) - \frac{[1 - \phi(s)]\phi(s + N)}{1 - \phi(s + N)}\left[\frac{P_{N-1}}{1 - P_N}\right]. \tag{23}$$

We are now in position to determine M^+, defined by (10), and subsequently the peakedness of the CAP.

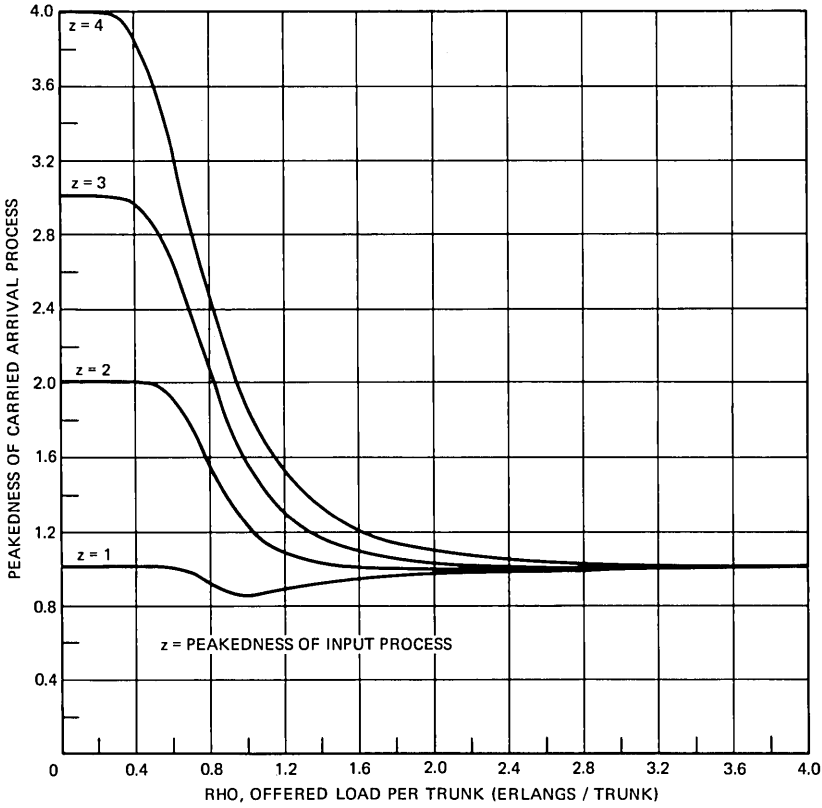


Fig. 8—Peakedness of CAP (50 trunks).

VI. DETERMINATION OF M^+

In order to determine the peakedness of the CAP [eq. (11)], we need to evaluate M^+ defined in (10). This will be done by characterizing the conditional mean of trunks up on the infinite trunk group, given j calls up on the finite trunk group.

Recall that we are considering an arrival at τ_n that finds a free circuit on the finite trunk group (i.e., $j(\tau_n^-) < N$). The state distribution on the finite trunk group at τ_n^+ is thus given by

$$\begin{aligned} P_j^+ &= \Pr\{j(\tau_n^+) = j\} = \frac{P_{j-1}}{1 - P_N}, \qquad 1 \leqq j \leqq N, \\ P_0^+ &= 0, \end{aligned} \tag{24}$$

where the P_j's are the call congestion probabilities defined in (16).

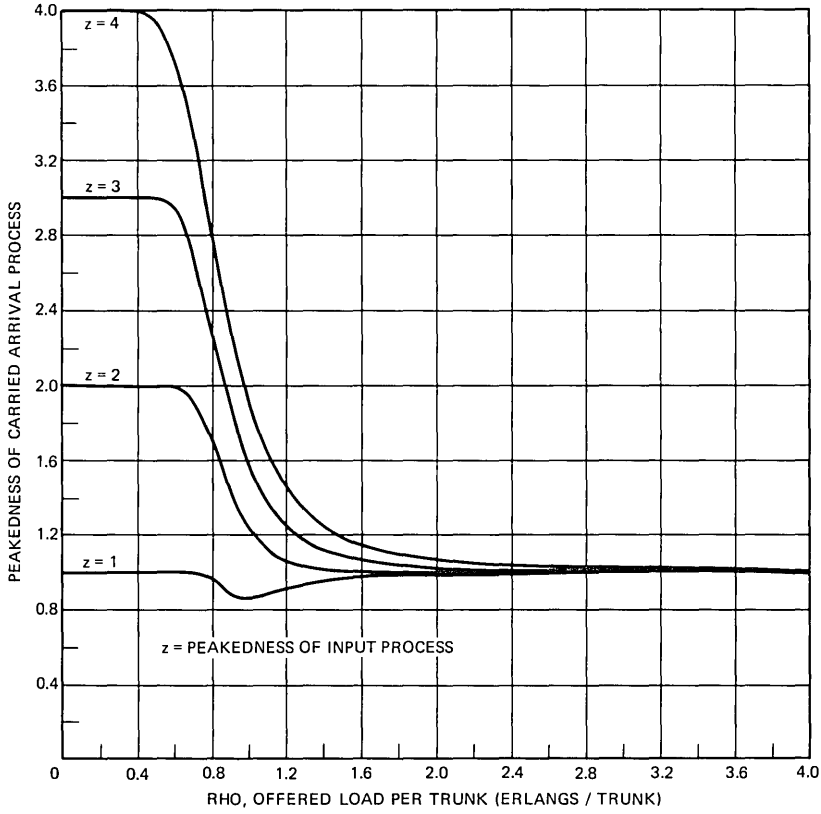


Fig. 9—Peakedness of CAP (100 trunks).

We define

$$P^{+}_{<N} = \Pr\{j(\tau_n^{+}) < N\}, \tag{25}$$

$$M^{-} = E\{k(\tau_n^{-})\} = E\{k(\tau_n^{+})\} - 1 = M^{+} - 1, \tag{26}$$

$$M_l^{+} = E\{k(\tau_n^{+}) \mid j(\tau_n^{+}) = l\}, \tag{27}$$

and

$$M^{+}_{<N} = E\{k(\tau_n^{+}) \mid j(\tau_n^{+}) < N\}, \tag{28}$$

where k corresponds to the infinite trunk group and j corresponds to the finite trunk group. In terms of these quantities, we have

$$M^{+} = M^{+}_{<N} P^{+}_{<N} + M^{+}_{N} P^{+}_{N}. \tag{29}$$

Recall that, if $j(\tau_n^{+}) < N$, the distribution of time until the next

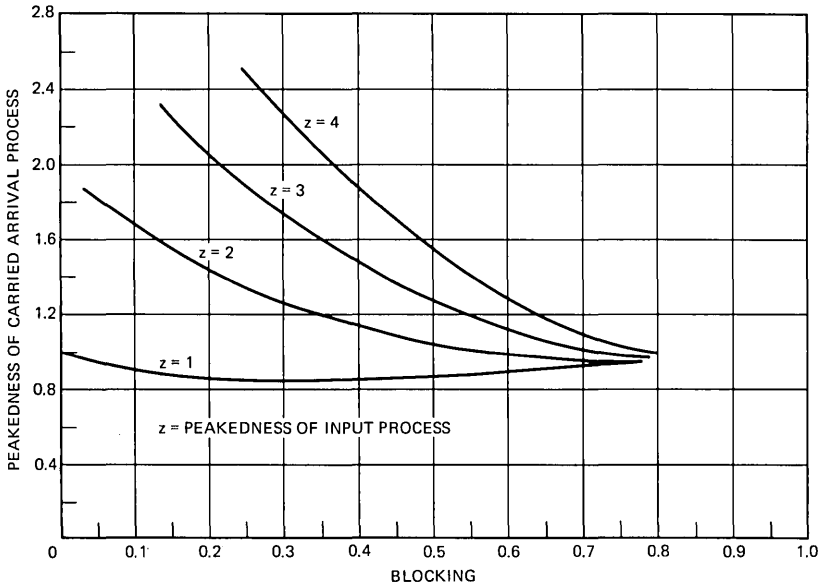


Fig. 10—Peakedness of CAP versus blocking (5 trunks).

carry is $F(t)$ and if $j(\tau_n^+) = N$ the distribution of time until the next carry is $\bar{F}_c(t)$. Using this together with some conditioning arguments, the following relationship is obtained (see Appendix A):

$$M^- = M^+ - 1 = P_{<N}^+ M_{<N}^+ \phi(1) + P_N^+ M_N^+ \tilde{\phi}_c(1). \tag{30}$$

From (29) and (30) we obtain

$$M^+ = \frac{1}{1 - \phi(1)} + M_N^+ P_N^+ \left[\frac{\tilde{\phi}_c(1) - \phi(1)}{1 - \phi(1)} \right]. \tag{31}$$

Use of (21) simplifies (31) to

$$M^+ = \frac{1}{1 - \phi(1)} - \frac{M_N^+ P_N^+ \phi(N+1)}{1 - \phi(N+1)}. \tag{32}$$

It should be noted that the first term in (32) corresponds to the value M^+ would assume if the renewal input process was offered directly to the infinite trunk group. The second term corresponds to the reduction in M^+ as a result of blocking on the finite trunk group. We are now left with the problem of determining M_N^+.

It is shown in Appendix B that M_m^+ for $m = 1, 2, \cdots, N$ satisfies

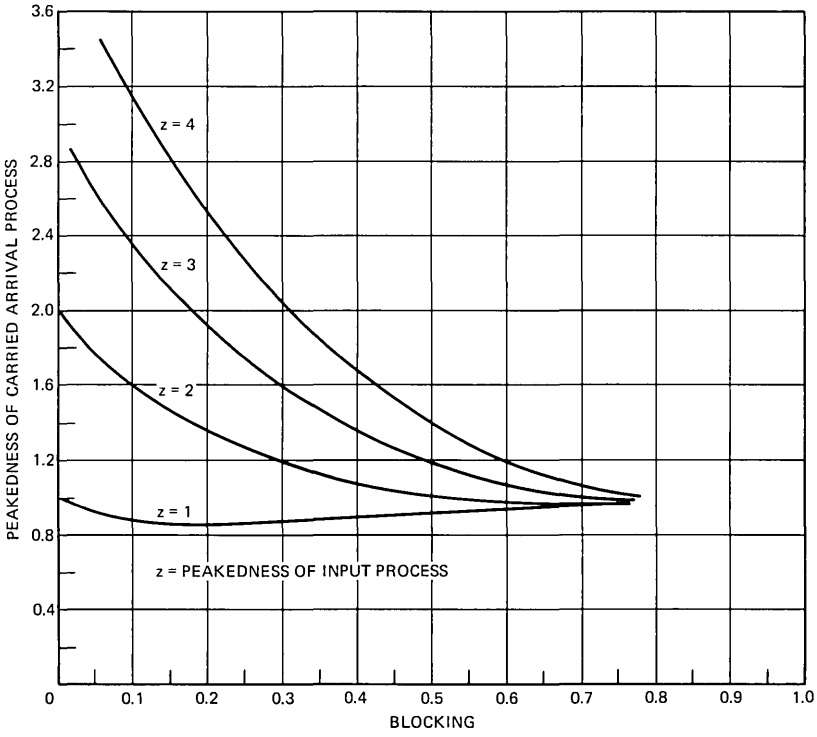


Fig. 11—Peakedness of CAP versus blocking (10 trunks).

the following set of equations:

$$[M_m^+ - 1]P_m^+ = \sum_{l=\max(1,m-1)}^{N} C_{lm}M_l^+, \qquad m = 1, 2, \cdots, N, \quad (33)$$

where for $m - 1 \leqq l \leqq N - 1$, $1 \leqq m \leqq N$,

$$C_{lm} = P_l^+ \binom{l}{m-1} \sum_{\eta=0}^{l-m+1} \binom{l-m+1}{\eta} (-1)^\eta \phi(\eta + m). \quad (34)$$

Further, for $l = N$ we have

$$C_{Nm} = \frac{P_N^+ \binom{N}{m-1} \sum_{\eta=0}^{N-m+1} \binom{N-m+1}{\eta} (-1)^\eta \phi(\eta + m)}{1 - \phi(N+1)}. \quad (35)$$

It should be noted that the above set of equations is in a form which is amenable to solution for the desired quantity M_N^+. Written in matrix-vector form, the matrix in question is triangular with additional entries below the diagonal. Transforming the matrix associated with

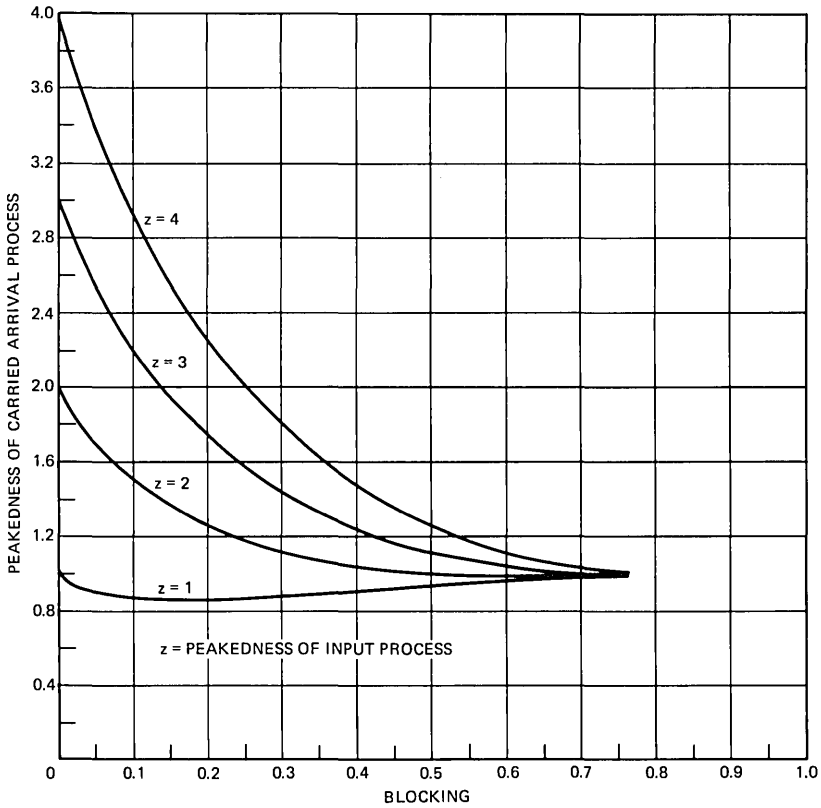


Fig. 12—Peakedness of CAP versus blocking (20 trunks).

(33) into triangular form* leads directly to the quantity of interest M_N^+ for use in (32), which is subsequently used to determine the peakedness of the CAP [eq. (11)].

VII. EXAMPLES

We ran some examples with a 2-moment match† interrupted Poisson process (Reference 7) as the renewal input to the finite trunk group (the computational aspects are discussed in Appendix C). Figures 5 to 9 show z_c, the peakedness of the CAP, as a function of ρ, the offered load per trunk for $N = 5$, 10, 20, 50, and 100 trunks, respectively.

* Details are in Reference 9.

† The blocking experienced by an overflow process is less than the blocking seen by a 2-moment match interrupted Poisson process and more than that seen by the 3-moment match process (all with the same mean and peakedness).

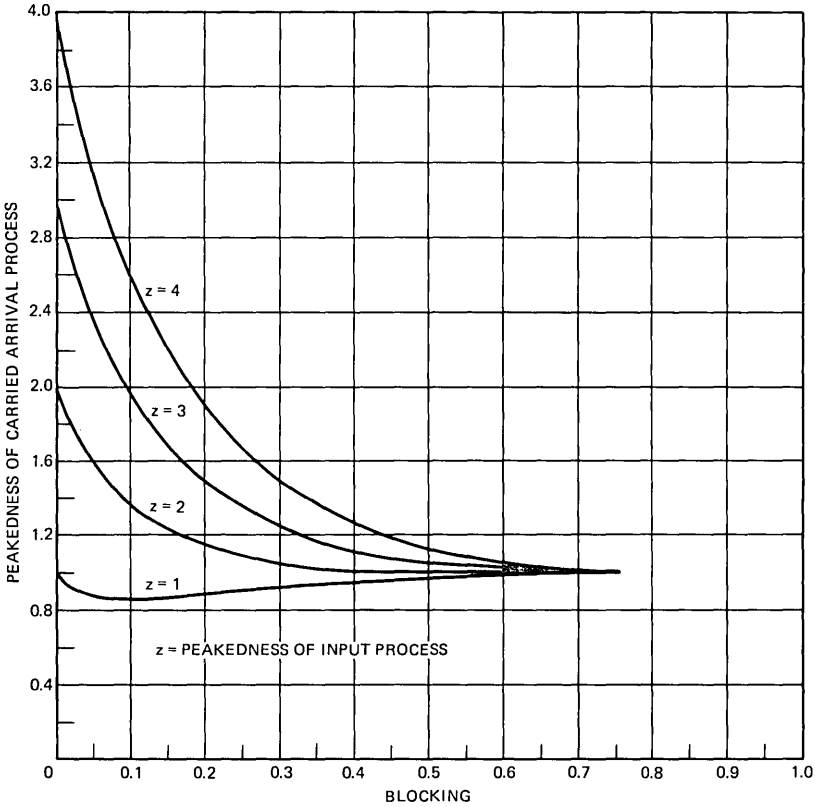


Fig. 13—Peakedness of CAP versus blocking (50 trunks).

On each figure are plots for offered z's of 1, 2, 3, and 4. It is seen that, for fixed N, as z increases, the smoothing effect (reduction of peakedness) becomes appreciable at lower ρ's. This is because of blocking remaining negligible for larger ρ's as z is decreased. If we fix z, we see that the smoothing effect becomes appreciable at lower ρ's as N is decreased. This is again explainable from the point of view of blocking, i.e., blocking is larger on the less efficient small trunk groups.

Since blocking is an important parameter, Figs. 10 to 14 show the peakedness of the CAP versus the blocking for the same cases as shown in Figs. 5 to 9. Note that, for final trunk groups which are normally operated with blockings of 0.01, the smoothing effect is very small, while for high-usage trunk groups which may reach blockings of a few tenths, the smoothing is substantial. Also, note that z_c, in all the cases, approaches unity as the load (and blocking) increases which

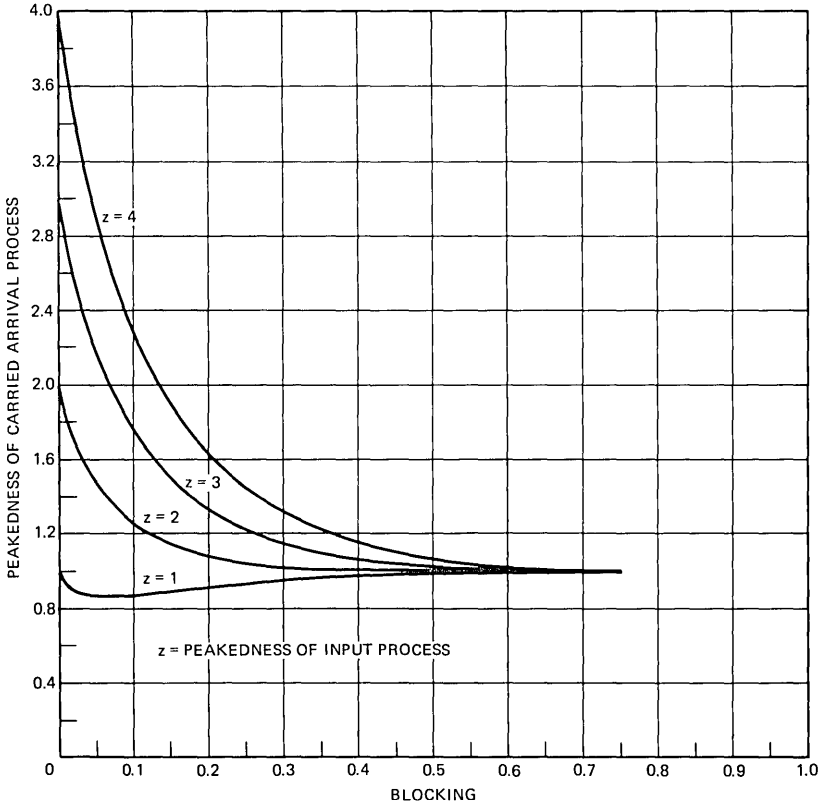


Fig. 14—Peakedness of CAP versus blocking (100 trunks).

is consistent with the explanation in Section I as to how peakedness differs from the variance-to-mean ratio of busy servers (which approaches zero).

It was observed in Reference 10 that, when $z > 1$, the blocking probability is bounded away from zero no matter how small the input mean is. This is evident in Fig. 10.

VIII. CONCLUSION

We have shown how to determine the peakedness of a CAP. The use of mean and peakedness to characterize a CAP is attractive from the point of view of simplicity and is consistent with the use of the equivalent random method (Reference 2) in trunking analyses. To approximately calculate the delays at a switching machine, we could replace the CAP (or, more usually, a superposition of CAP's) with an

interrupted Poisson process with the same mean and variance and proceed as in Reference 1. This is being investigated.

We mention, in passing, that we also tried out a renewal approximation for z_c. That is, although the CAP is a semi-Markov process and not generally a renewal process, we tried to approximate z_c with

$$\frac{1}{1-\phi_c(1)} - \frac{1}{\beta},$$

which is the peakedness that the CAP would have if it were a renewal process. This approximation did not compare well enough with the true z_c to recommend its use. That is, although one might use a renewal approximation to a superposition of CAP's *after* the peakedness is determined (see the last paragraph), we do not recommend using a renewal assumption to determine the peakedness.

In the course of determining the CAP peakedness, we have more fully characterized the CAP as a semi-Markov process. Any queuing results available for semi-Markov inputs could be used with the CAP semi-Markov characterization given in Section III.

IX. ACKNOWLEDGMENT

The excellent programming of Mary Zeitler is gratefully acknowledged.

APPENDIX A

Derivation of Equation (30)

Consider the system in equilibrium with times of carried calls $\{\tau_n\}$. From (26) we have

$$M^- = E\{k(\tau_n^-)\}, \tag{36}$$

which can be expanded as

$$\begin{aligned} M^- = {} & E\{k(\tau_n^-) \mid j(\tau_{n-1}^+) = N\}P\{j(\tau_{n-1}^+) = N\} \\ & + E\{k(\tau_n^-) \mid j(\tau_{n-1}^+) < N\}P\{j(\tau_{n-1}^+) < N\}. \end{aligned} \tag{37}$$

This can be written as

$$\begin{aligned} M^- = {} & P_N^+ \sum_{n=0}^{\infty} \sum_{i=1}^{n} iP\{k(\tau_n^-) = i \mid k(\tau_{n-1}^+) = n,\ j(\tau_{n-1}^+) = N\} \\ & \times P\{k(\tau_{n-1}^+) = n \mid j(\tau_{n-1}^+) = N\} \\ & + P_{<N}^+ \sum_{n=0}^{\infty} \sum_{i=1}^{n} iP\{k(\tau_n^-) = i \mid k(\tau_{n-1}^+) = n,\ j(\tau_{n-1}^+) < N\} \\ & \times P\{k(\tau_{n-1}^+) = n \mid j(\tau_{n-1}^+) < N\}. \end{aligned} \tag{38}$$

Writing

$$P\{k(\tau_n^-) = i \mid k(\tau_{n-1}^+) = n,\ j(\tau_{n-1}^+) = N\}$$
$$= \int_0^\infty P\{k(\tau_n^-) = i \mid k(\tau_{n-1}^+) = n,\ j(\tau_{n-1}^+) = N,\ \tau_n - \tau_{n-1} = t\}\, d\bar{F}_c(t)$$
$$= \int_0^\infty \binom{n}{i} e^{-it}(1 - e^{-t})^{n-i} d\bar{F}_c(t) \tag{39}$$

and

$$P\{k(\tau_n^-) = i \mid k(\tau_{n-1}^+) = n,\ j(\tau_{n-1}^+) < N\} = \int_0^\infty \binom{n}{i} e^{-it}(1 - e^{-t})^{n-i} dF(t) \tag{40}$$

and observing that

$$\sum_{i=1}^{n} i \binom{n}{i} e^{-it}(1 - e^{-t})^{n-i} = ne^{-t} \tag{41}$$

leads to the desired result

$$M^- = P^+_{<N} M^+_{<N} \phi(1) + P^+_N M^+_N \bar{\phi}_c(1). \tag{42}$$

APPENDIX B

Derivation of Equations (33) through (35) for M_m^+

Consider the events

$$A_{lm} = \{j(\tau_{n-1}^+) = l,\ (l - m + 1) \text{ s.c.'s before next arrival after } \tau_{n-1}\} \tag{43}$$

$$B_{km} = \{j(\tau_{n-1}^+) = N, \text{ no s.c. before } (k-1)\text{st arrival after } \tau_{n-1},\ (N - m + 1) \text{ s.c.'s before } k\text{th arrival after } \tau_{n-1}\}, \tag{44}$$

where s.c. denotes service completion on the finite trunk group. From these definitions we obtain

$$P\{j(\tau_n^+) = m\} = \sum_{l=m-1}^{N-1} P\{A_{lm}\} + \sum_{k=1}^{\infty} P\{B_{km}\}. \tag{45}$$

The conditional mean of interest is thus given by

$$M_m^- = E\{k(\tau_n^-) \mid j(\tau_n^+) = m\}$$
$$= \frac{\sum_{l=m-1}^{N-1} E\{k(\tau_n^-) \mid A_{lm}\} P\{A_{lm}\} + \sum_{k=1}^{\infty} E\{k(\tau_n^-) \mid B_{km}\} P\{B_{km}\}}{P\{j(\tau_n^+) = m\}} \tag{46}*$$

Defining the events

$$A_l = \{j(\tau_{n-1}^+) = l\} \tag{47}$$

* When $m = 1$, $P\{A_{0m}\} = 0$, because $P\{j(\tau_{n-1}^+) = 0\} = 0$. In that case, sum from $l = 1$ to $l = N - 1$.

and

$$A_{2lm} = \{(l - m + 1) \text{ s.c.,'s before next arrival after } \tau_{n-1}\} \quad (48)$$

gives

$$E\{k(\tau_n^-) \mid A_{lm}\} P\{A_{lm}\}$$

$$= \sum_{i=1}^{\infty} iP\{k(\tau_n^-) = i, A_{2lm} \mid A_l\} P_l^+ \quad (49)$$

$$= P_l^+ \sum_{r=0}^{\infty} P\{k(\tau_{n-1}^+) = r \mid A_l\} \sum_{i=1}^{r} iP\{k(\tau_n^-) = i, A_{2lm} \mid A_l, k(\tau_{n-1}^+) = r\}. \quad (50)$$

Letting $l < N$ and using

$$P\{k(\tau_n^-) = i, A_{2lm} \mid A_l, k(\tau_{n-1}^+) = r\} = \int_0^{\infty} \binom{r}{i} e^{-it}(1 - e^{-t})^{r-i} \times \binom{l}{m-1} e^{-(m-1)t}(1 - e^{-t})^{l-m+1} dF(t) \quad (51)$$

and

$$\sum_{i=1}^{r} i \binom{r}{i} e^{-it}(1 - e^{-t})^{r-i} = re^{-t} \quad (52)$$

yields

$$E\{k(\tau_n^-) \mid A_{lm}\} P\{A_{lm}\} = P_l^+ M_l^+ \int_0^{\infty} \binom{l}{m-1} \times e^{-mt}[1 - e^{-t}]^{l-m+1} dF(t). \quad (53)$$

From (1) we have

$$E\{k(\tau_n^-) \mid A_{lm}\} P\{A_{lm}\} = P_l^+ M_l^+ \binom{l}{m-1} \times \sum_{\eta=0}^{l-m+1} \binom{l - m + 1}{\eta} (-1)^{\eta} \phi(\eta + m) \quad (54)$$

$$= C_{lm} M_l^+, \quad (55)$$

where we are using (54) to define C_{lm} in (55). This yields (34).

It now remains to show the desired relationship for C_{Nm}. We consider the second summation of (46)

$$E\{k(\tau_n^-) \mid B_{km}\} P\{B_{km}\} = \sum_{i=1}^{\infty} iP\{k(\tau_n^-) = i, B_1, B_{2km}\}, \quad (56)$$

where

$$B_1 = \{j(\tau_{n-1}^+) = N\}$$

$$B_{2km} = \{\text{no s.c. before } (k-1)\text{st arrival after } \tau_{n-1}, (N - m + 1) \text{ s.c.'s before } k\text{th arrival after } \tau_{n-1}\}. \quad (57)$$

Now

$$E\{k(\tau_n^-)\,|\,B_{km}\}P\{B_{km}\}$$

$$= \sum_{n=0}^{\infty}\sum_{i=1}^{n} iP\{k(\tau_n^-) = i, B_{2km}|B_1, k(\tau_{n-1}^+) = n\} \times P\{k(\tau_{n-1}^+) = n|B_1\}P_N^+ \quad (58)$$

$$= P_N^+ \sum_{n=0}^{\infty} nP\{k(\tau_{n-1}^+) = n|B_1\} \int_0^{\infty}\int_0^t \sum_{i=1}^{n}\binom{n-1}{i-1} e^{-it} \times (1-e^{-t})^{(n-i)}e^{-Ns}\binom{N}{m-1}e^{-(m-1)(t-s)} \times (1-e^{-(t-s)})^{N-m+1}dF_{k-1}(s)dF(t-s). \quad (59)$$

Upon performing the summation over i, (59) simplifies to

$$E\{k(\tau_n^-)\,|\,B_{km}\}P\{B_{km}\}$$

$$= P_N^+M_N^+\binom{N}{m-1}\int_0^{\infty}\int_0^t e^{-(N+1)s}e^{-m(t-s)} \times \sum_{\eta=0}^{N-m+1}\binom{N-m+1}{\eta}(-1)^{\eta}e^{-\eta(t-s)}dF_{k-1}(s)dF(t-s)$$

$$= P_N^+M_N^+\binom{N}{m-1}\sum_{\eta=0}^{N-m+1}(-1)^{\eta}\binom{N-m+1}{\eta} \times \phi^{k-1}(N+1)\phi(\eta+m). \quad (60)$$

Summing over k gives

$$\sum_{k=1}^{\infty} E\{k(\tau_n^-)\,|\,B_{km}\}P\{B_{km}\} = M_N^+\frac{P_N^+\binom{N}{m-1}\sum_{\eta=0}^{N-m+1}(-1)^{\eta}\binom{N-m+1}{\eta}\phi(\eta+m)}{1-\phi(N+1)}, \quad (61)$$

which is the desired result for (35).

APPENDIX C

Computational Considerations

In this appendix we briefly* discuss some computational problems experienced in the numerical solution of the carried process problem and point out some possible approaches to circumvent them.† In

* A more detailed description of our computational experience is in Reference 9.
† We first discuss the approaches and then the numerical experience we have had.

particular, we are concerned with the computation of the state probabilities, P_j, $j = 0, \cdots, N$, defined in (16) and the computation of the quantities C_{lm} defined in (34) and (35).

We also specialize and subsequently simplify the above results for the case where the interarrival time distribution is a sum of exponentials. In particular, we consider the interrupted Poisson process[7] as the renewal input. This specialization appears to have eliminated the numerical problems associated with the required calculations for large trunk groups.

The problem of determining the call congestion state probabilities for a renewal input to a BCC system with N independent exponential servers is considered on p. 179 of Reference 3. The results are as follows: Let

$$C_{j+1} = \left(\frac{\phi(j+1)}{1 - \phi(j+1)} \right) C_j , \qquad j = 0, \cdots, N-1 , \tag{62}$$

with $C_0 = 1$. Then B_r, the rth binomial moment of the P_j's, is given by

$$B_r = C_r \frac{\sum_{j=r}^{N} \binom{N}{j} \frac{1}{C_j}}{\sum_{j=0}^{N} \binom{N}{j} \frac{1}{C_j}} . \tag{63}$$

The B_r satisfies the backward recursion

$$B_r = \left(\frac{1 - \phi(r+1)}{\phi(r+1)} \right) B_{r+1} + \binom{N}{r} B_N , \tag{64}$$

with

$$\frac{1}{B_N} = \frac{1}{P_N} = \sum_{j=0}^{N} \binom{N}{j} \frac{1}{C_j} \tag{65}$$

(Reference 11, p. 93). The P_j's are given by

$$P_j = \sum_{r=j}^{N} (-1)^{r-j} \binom{r}{j} B_r . \tag{66}$$

The computation of the C_j coefficients and the binomial moments B_r is fairly straightforward and does not pose much of a numerical problem. It is the computation of the P_j's, using (66), that is sensitive to numerical errors. The alternating sign in (66), together with the facts that $\binom{r}{j} B_r$ can be quite large and the summation in (66) is between zero and unity, lead to a numerical problem.*

* Actually, for an $N = 10$ case seven significant decimal digits were lost in one subtraction and the resultant probability was computed to be zero. This plays havoc with the solution to (33). Details are in Reference 9.

An alternate approach* to the computation of the state probabilities is by way of the equations

$$P_k = \sum_{j=k-1}^{N} p_{jk} P_j, \tag{67}$$

where the transition probabilities p_{jk} represents the probability of going from state j prior to one arrival to state k prior to the next arrival. The p_{jk}'s are given by

$$p_{jk} = \binom{j+1}{k} \int_0^\infty e^{-kt}(1 - e^{-t})^{j+1-k} dF(t), \qquad j < N, \tag{68}$$

and

$$p_{N,k} = p_{N-1,k}. \tag{69}$$

From (67) and (68), we obtain the backward relation

$$P_{k-1} = \frac{P_k}{\phi(k)} - \frac{1}{\phi(k)} \sum_{j=k}^{N} p_{jk} P_j. \tag{70}$$

Expanding (68) gives

$$p_{jk} = \binom{j+1}{k} \sum_{\eta=0}^{j+1-k} \binom{j+1-k}{\eta} (-1)^\eta \phi(\eta + k). \tag{71}$$

The computational procedure is outlined as follows: Compute P_N from (65)† and use (70) to compute P_j for $j < N$. It should be noted that, although the terms in (71) alternate in sign, ϕ never exceeds unity and is monotonically decreasing. Also note the relation between p_{jk} given by (71) and C_{lm} given by (34) and (35). At this point, the accuracy in computing the P_j's should be comparable to the accuracy in computing the C_{lm}'s.

For the case where $F(t)$ is the sum of exponentials (e.g., interrupted Poisson process) we can further simplify (and more accurately compute) the P_j's and C_{lm}. We go to the integrals from which the sums (with alternating signs) appearing in (34), (35), and (71) were derived. Note that we have

$$\binom{l}{m-1} \sum_{\eta=0}^{l-m+1} \binom{l-m+1}{\eta} (-1)^\eta \phi(\eta + m) = \int_0^\infty \binom{l}{m-1} e^{-mt}(1 - e^{-t})^{l-m+1} dF(t) \tag{72}$$

---

* Motivation for investigating this approach stems from remarks made by P. J. Burke. (In recent unpublished work, Burke showed a more accurate approach for the case where the renewal interarrival time distribution is a sum of exponentials.)

† Note that each term in the sum of (65) is positive.

[see (53) and (54)]. Let

$$F(t) = \sum_{i=1}^{S} k_i(1 - e^{-r_i t}); \tag{73}$$

then the integral in (72) can be identified as a beta function.* Repeated integration by parts in (72) gives

$$\int_0^\infty \binom{l}{m-1} e^{-mt}(1 - e^{-t})^{l-m+1} dF(t) = \sum_{i=1}^{S} f_{lm}(k_i, r_i), \tag{74}$$

where

$$f_{lm}(k_i, r_i) = \left(\frac{k_i r_i}{l + r_i + 1}\right)\left(\frac{l}{l + r_i}\right)\left(\frac{l-1}{l + r_i - 1}\right) \cdots \left(\frac{l - (l-m)}{l + r_i - (l-m)}\right) \quad \text{for} \quad l > m - 1. \tag{75}$$

For $l = m - 1$, we obtain from (72)

$$f_{m-1,m}(k_i, r_i) = \frac{k_i r_i}{m + r_i}. \tag{76}$$

Note that f_{lm} can be computed recursively from

$$f_{l+1,m}(k_i, r_i) = \left(\frac{l+1}{l + r_i + 2}\right) f_{l,m}(k_i, r_i) \tag{77}$$

with initialization from (76).

The direct calculation of the integral has thus led to a computationally tractable method of computing the C_{lm}'s and the P_j's. C_{lm} is calculated from

$$C_{lm} = P_l^+ \left(\sum_{i=1}^{S} f_{lm}(k_i, r_i)\right), \qquad m - 1 \leqq l \leqq N - 1, \tag{78}$$†

[see (53)] and

$$C_{Nm} = \frac{P_N^+}{1 - \phi(N+1)} \left(\sum_{i=1}^{S} f_{N,m}(k_i, r_i)\right) \tag{79}$$†

[see (35) and (72)].

The P_j's are calculated from (70) where (68) is simplified as above (using integration by parts) and then used to compute the transition probabilities. Note that for an interrupted Poisson process,[7] $S = 2$, and k_1, k_2, r_1, and r_2 are given in Reference 7 in terms of the switch

* Identification made by D. L. Jagerman.

† Note that this procedure does not involve the calculation of either binomial moments or binomial coefficients.

parameters. An alternate method of computing the P_j's for the interrupted Poisson process is via the use of birth and death equations and conditioning the results on the switch being closed. A computer program for doing this was available (Reference 8).

At this point, it is of interest to discuss our computational experience using some of the aforementioned procedures. The first method considered was to calculate the P_j's by first obtaining the binomial moments [eqs. (62) to (66)] and to compute the C_{lm}'s from (34) and (35). Using single precision arithmetic the procedure worked for $N = 2$, but failed for $N = 10$. The problem was traced to inaccurate calculation of the probabilities from binomial moments. Double precision arithmetic extended the range of N (worked for $N = 10$). The method failed at $N = 20$. The failure was traced to the same cause as above. At this point we used the birth and death equation approach to calculate the P_j's,[8] which assumes an interrupted Poisson input. This extended the range to $N = 20$. For $N = 30$ we ran into problems computing the C_{lm}'s from (34) and (35).* Modification of the C_{lm} computation using (78) and (79) significantly extended the useful range on N. The results presented in Section VII were computed using this method of calculation.

REFERENCES

1. Heffes, H., "Analysis of First-Come First-Served Queuing Systems With Peaked Inputs," B.S.T.J., *52*, No. 7 (September 1973), pp. 1215–1228.
2. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
3. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
4. Ross, S. M., *Applied Probability Models with Optimization Applications*, San Francisco: Holden-Day, 1970.
5. Franken, P., "Erlang Formulas for Semi-Markovian Input," Elektron Informationsverarbeit Kybernetik, *4* (1968), pp. 197–204.
6. Descloux, A., "On Markovian Servers with Recurrent Input," Proc. Sixth International Teletraffic Congress, 1970, pp. 331/1–331/6.
7. Kuczura, A., "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *52*, No. 3 (March 1973), pp. 437–448.
8. Marzec, R. P., unpublished work.
9. Zeitler, M. J., unpublished work.
10. Holtzman, J. M., "The Accuracy of the Equivalent Random Method With Renewal Inputs," B.S.T.J., this issue, pp. 1673–1679.
11. Riordan, J., *Stochastic Service Systems*, New York: Wiley, 1962.

* This implicitly indicates the range of accuracy for the computation of the P_j's from (70) and (71).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

Model Approximations to Visual Spatio-Temporal Sine-Wave Threshold Data

By Z. L. BUDRIKIS

(Manuscript received May 8, 1973)

Experimental data on visual spatio-temporal sine-wave thresholds obtained by Robson and Kelly are considered. In seeking model approximations to the data it is assumed that the subject's visual threshold to modulation at different spatial and temporal frequencies gives the image of his filter function to within a multiplicative constant. It is further assumed that the data can be approximated by a system with a spatially uniform, isotropic, and temporally invariant response which consists of the difference between an excitatory and an inhibitory term, and that each term is separable into a product of a spatial and a temporal function.

I. INTRODUCTION

Tests of vision with sine-wave flicker go back at least fifty years to H. E. Ives.[1] He determined flicker fusion frequencies with a number of wave shapes, including sinusoids. Spatial sinusoid test stimuli are more recent. The first to use them was probably Schade[2] in the fifties. Soon after that Kelly[3] suggested a stimulus which would simultaneously test the spatial and the temporal sine-wave response of vision. Such tests were implemented by Robson,[4] Kelly,[5,6] and others.

The special interest in the sine wave as a test stimulus stems from the ease with which one can extrapolate from its results. Provided a system is linear and time-invariant, Fourier analysis can be used to predict the system response to any input from its response to sinusoidal inputs. However, the visual system is neither linear nor time-invariant. Nevertheless, given a sufficiently constant adaptation state and input variations that result in small output variations,† linear theory can be used.

† It is often incorrectly stated or implied that, for linearity, the input needs to be small. But consider the situation where a flickering light appears fused visually. The input may then swing between zero and many times the average luminance, yet the behavior is linear.

Our interest in the visual system is related to visual communications. When visual messages are transmitted digitally then there are potentially very many different ways—some more advantageous than others—in which the messages might be coded and still give acceptable fidelity at the receiver. Clearly, it would be good if the likely subjective effect of given quantizing procedures could already be predicted at the computer simulation stage without involving repeated subjective tests. Such predictions will probably be possible soon.[7] However, there is still need for complete specifications both of the linear behavior of the visual system and of the nonlinear effects of background masking. We will concern ourselves here only with the linear characteristics. To this end we will examine several alternative mathematical models to see whether they could be used to represent published experimental data on spatio-temporal sine-wave thresholds.

The data that we will use were reported by Kelly[6] and Robson.[4] In both cases threshold values of m were determined in a target described by

$$L = L_o(1 + m \cos 2\pi u_o x \cdot \cos 2\pi f_o t) , \qquad (1)$$

where L_o is the average luminance, u_o the spatial frequency, and f_o the temporal frequency.

Kelly's measurements† were made at four different values of L_o. The entire target area, a circular 7-degree CRT face, filled with the flickering grating, was viewed monocularly through a 2.3-mm artificial pupil. Robson made all measurements at a single L_o value. The target had a 2.5-degree × 2.5-degree grating in the center of a 10-degree × 10-degree screen which had a luminance equal to L_o, and it was viewed binocularly without artificial pupils.

In both cases the subject's threshold was measured by the method of adjustment. The subject judged whether he could see the signal or not. He did not attempt to distinguish between seeing flicker and seeing the bar pattern. During each session of Kelly's experiment, the subject made 5 settings at each of 12 frequencies, with the 60 presentations given to him at random. Robson made his measurements in orderly sequences. Their results are shown as log-log plots of $(1/m)$ against frequency in Figs. 1–5.

Kelly's measurements obtained for vision with an artificial pupil are converted to equivalent luminances viewed through a natural pupil. To calculate the equivalent luminance one needs to take into account changes in the size of the natural pupil and the Stiles-Crawsford effect. From data tabulated by LeGrand[8] it can be inferred that,

† D. H. Kelly kindly supplied a listing of his measurements and standard deviations.

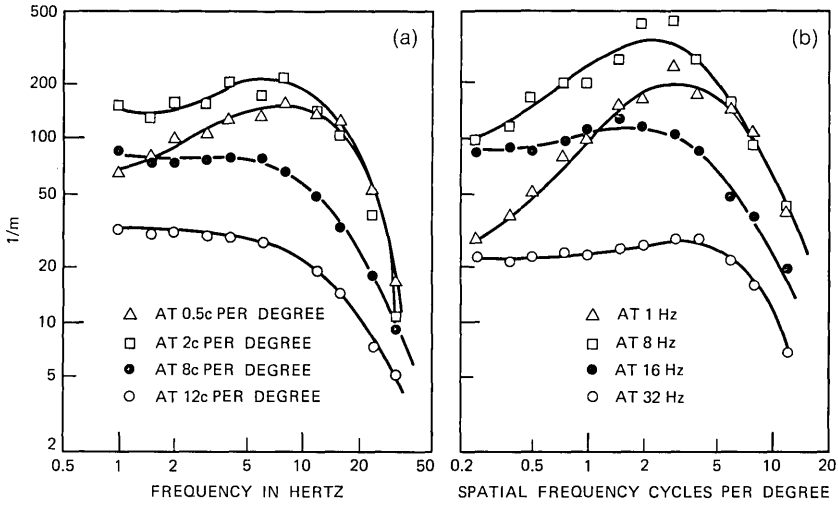


Fig. 1—Kelly's data at 62.8 mL. (a) Temporal frequency response. (b) Spatial frequency response.

given an illuminance I, in trolands, the corresponding luminance L in mL is

$$L = 1.142 \times 10^{-2} I^{1.223}; \qquad 10 < I < 2000 \text{ td.} \tag{2}$$

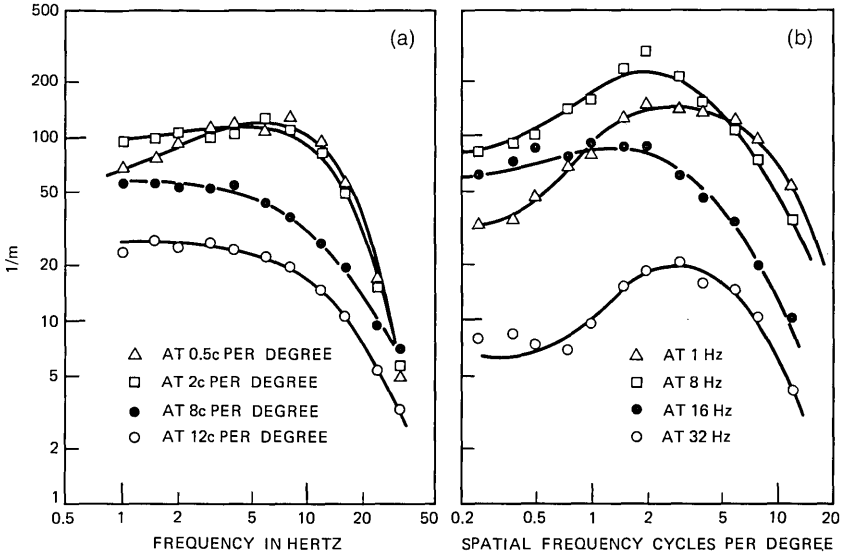


Fig. 2—Kelly's data at 15.2 mL. (a) Temporal frequency response. (b) Spatial frequency response.

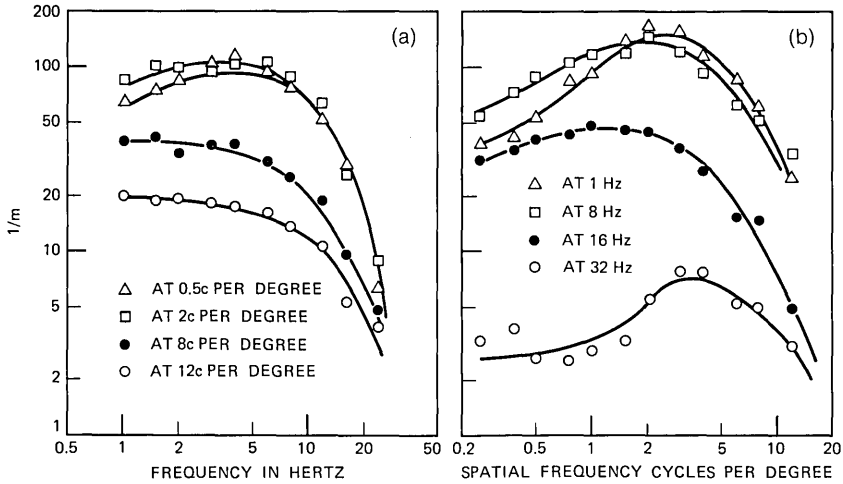


Fig. 3—Kelly's data at 3.7 mL. (a) Temporal frequency response. (b) Spatial frequency response.

We will consider six different, though similar, mathematical models as possible candidates for representing the data of Figs. 1–5. There is a similarity between the models in that: (*i*) they all consist of an algebraic difference between an excitatory and inhibitory term, (*ii*) these terms are in all models separable functions of spatial and tem-

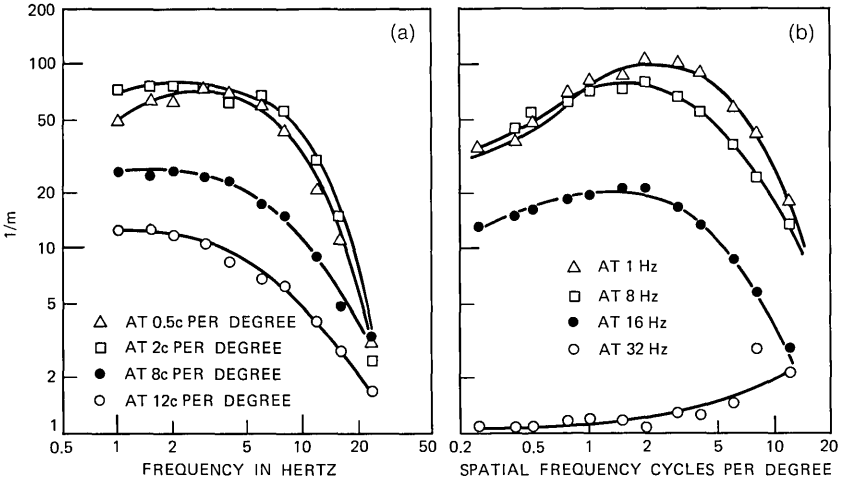


Fig. 4—Kelly's data at 0.91 mL. (a) Temporal frequency response. (b) Spatial frequency response.

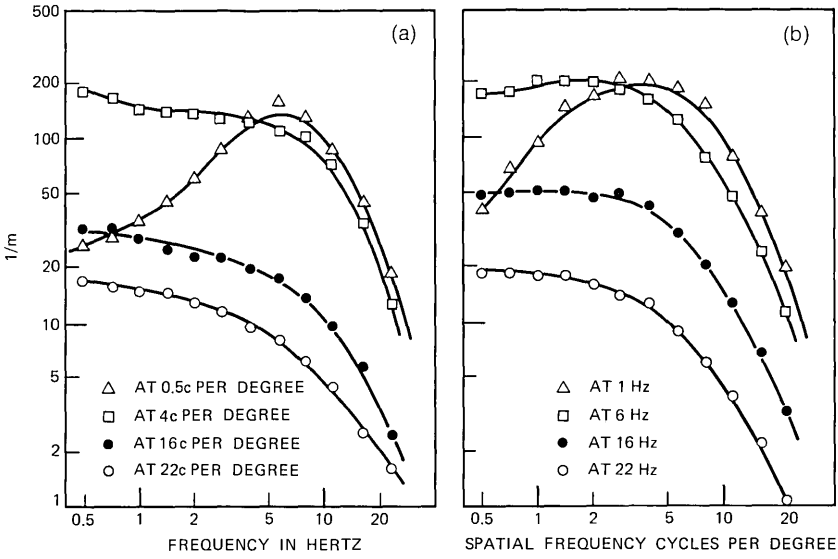


Fig. 5—Robson's data at 6.3 mL. (a) Temporal frequency response. (b) Spatial frequency response.

poral frequencies, and (*iii*) in each model there will be six undetermined parameters. We find values for the parameters by digitally searching for the smallest weighted mean-square deviation of experimental points from the models. The fit of none of the models is compellingly good, but in several cases the degree of fit is useful. We find the best all-round fit with a model with diffusion-like temporal response of excitation, a Gaussian function for the temporal response of inhibition, and Cauchy functions for the spatial response. From the point of view of economy in computer simulation, a model with simple exponential time responses and Gaussian spatial responses would be preferable. However, the mean-square departure from the model is somewhat larger than the best.

II. THE MODELS

2.1 *The Framework*

By the nature of things, the retinal image is a somewhat blurred version of the light distribution in object space. Over isoplanatic patches,[9] or areas A which are large compared to the size of a blurred point and small compared to inhomogeneities of the image-forming properties of the eye, we can model the formation of the image by a

convolution integral with a fixed point-spread function:

$$I(x, y) = \iint_A W(x - \xi, y - \eta)L(\xi, \eta)d\xi\, d\eta\,, \tag{3}$$

where W is the point spread function and L and I are the object and image distributions.

There is virtually no time lag in forming the retinal image. If L were switched on at some instant of time then the retinal image $I(x, y)$ would be formed at that instant.

The object-space to image-plane spatial frequency response of the given isoplanatic patch, or its modulation transfer function, is the Fourier transform of W:

$$H(u, v) = \int\int W(x, y)e^{-2\pi j(ux+vy)}dx\, dy\,, \tag{4}$$

where u and v are the spatial frequencies in the x and the y directions. Point spread in the space domain becomes filtering when transformed into the frequency domain. The point spread function W is necessarily positive and, with a normal pupillary aperture, has a maximum at the center and decreases monotonically.[10] Consequently $H\ (u, v)$ is a low-pass function.

It is natural to think of perception being based on an "image" at some deeper location beyond the retina. This "image" is physiologically mediated and must suffer appreciable time lags. Hence, the response at the deeper location will be time-dependent. There will also be further spatial filtering as a result of lateral physiological interactions.[11]

Say we designate the resulting point response function by $R(x, y, t)$ and the internal "image" distribution by $C(x, y, t)$. At least for a restricted class of object-space luminance functions, $L(x, y, t)$, C can be obtained by superposition, so that

$$C(x, y, t) = \int_A \int\int_o^{+\infty} R(x - \xi, y - \eta, t - \tau)L(\xi, \eta, \tau)d\tau\, d\eta\, d\xi. \tag{5}$$

The three-dimensional Fourier transform of R is the spatio-temporal frequency response function

$$S(u, v, f) = \int\int\int R(x, y, t)e^{-2\pi j(ux+xy+ft)}dx\, dy\, dt. \tag{6}$$

The integration is over all x, y, and t. f is the temporal frequency.

We may assume that the response function R is even in x and y, i.e., $R(x, y, t) = R(-x, y, t) = R(x, -y, t) = R(-x, -y, t)$. This

means that S is even in u and v, i.e., $S(u, v, f) = S(-u, v, f) = S(-u, -v, f)$. No symmetry can be assumed for R in t, and hence, for S in f. Indeed, $R(x, y, t) = 0$ for $t \leqq 0$, and hence, $S(u, v, -f) \neq S(u, v, f)$.

If the input to the system is the L of eq. (1), then the internal "image" is

$$C(x, y, t) = S(0, 0, 0)L_o + |S(u_o, 0, f_o)|L_o m \cos(2\pi u_o x) \times \cos(2\pi f_o t + \phi), \quad (7)$$

where

$$|S(u_o, 0, f_o)| = \{S(u_o, 0, f_o) \cdot S^*(u_o, 0, f_o)\}^{\frac{1}{2}}$$

and

$$\phi = \tan^{-1}\{\mathrm{Im}[S(u_o, 0, f_o)]/\mathrm{Re}[S(u_o, 0, f_o)]\},$$

The * designates the complex conjugate and Im and Re the imaginary and real parts.

Now we ask: What size must m be before the flickering grating is seen with a given level of certainty? We assume thresholds correspond to fixed differences, i.e., the flickering grating is seen with probability p if

$$|S(u, 0, f)|L_o m = T(p), \quad (8)$$

where T is a monotonically increasing function of p, but is independent of all other variables. We may assume that subjects adjusted m so that it always resulted in the same probability of seeing. Therefore, the values of $1/m$, as plotted in Figs. 1–5, are regarded as experimental determinations of $|S(u, 0, f)|$ [to within the multiplier $T(p)/L_o$ which is a constant when the criterion T and the average luminance L_o are fixed].

If the visual system were truly linear it would have the same response functions irrespective of luminance level L_o. But all evidence, including that contained in Figs. 1–4, shows that the system adapts. It does so somewhat ponderously, much faster with rising L_o than in reverse, but still quite effectively, changing gain, spatial spread, and temporal lag. There is just one aspect of S which Kelly[12] found unchanging over more than four decades of luminance, L_o. In large-area flicker threshold determinations, using an artificial pupil, he found that at different L_o values plots of $(1/mL_o)$ approached a common asymptote for large values of f. However, in other parts of the functional domain, different S functions hold for different adaptation luminances.[6]

In searching for suitable mathematical expressions for R or S it would be convenient if these functions were isotropic, and even more

so, if they were also separable into spatial and temporal factors. Isotropism would mean that the space variables x and y would reduce to a single distance ρ and the frequencies u and v to a direction-independent spatial frequency ν. Then

$$S(\nu, f) = S(u = \nu, 0, f) = S(0, v = \nu, f)$$
$$= \int_0^\infty \left[\int_0^\infty R(\rho, t) 2\pi\rho J_o(2\pi\rho\nu) d\rho \right] e^{-j2\pi ft} dt , \tag{9}$$

where J_o is the Bessel function of order zero.

Man's vision is not isotropic. It is astigmatic, having better resolution in the horizontal and vertical directions than at other angles. But, to a first order of approximation, we may assume isotropism.

Separability of R would mean that we could write it as

$$R(\rho, t) = U(\rho)V(t) \tag{10}$$

and then S would also be separable:

$$S(\nu, f) = G(\nu)H(f) , \tag{11}$$

where

$$G(\nu) = \int_0^\infty U(\rho) 2\pi\rho J_o(2\pi\rho\nu) d\rho \tag{12}$$

and

$$H(f) = \int_0^\infty V(t) e^{-j2\pi ft} dt . \tag{13}$$

Moreover, because $U(\rho)$ is symmetrical, and hence $G(\nu)$ is a real-valued function, it would follow that

$$|S(\nu, f)| = G(\nu)|H(f)| . \tag{14}$$

However, even a superficial look at the families of experimental curves in Figs. 1–5 will convince one that $|S(\nu, f)|$ is not separable. If it were, then curves of $|S(\nu, f)|$, as functions of f at different values of ν, would differ from each other only by constant multipliers. Plotted against a logarithmic ordinate this would result in fixed vertical shifts. The same result would hold for plots of $|S(\nu, f)|$ versus ν at different values of f. But neither of these outcomes are found to be true. This is particularly evident when looking at Figs. 5a and b. The curves at high values of ν or f are low-pass in shape, while for low values of the parameters they are bandpass. Figure 6 shows a linearly scaled perspective view of a surface[13] to which the measured values of Fig. 5 approximate. Measurements apply only to positive frequencies, while

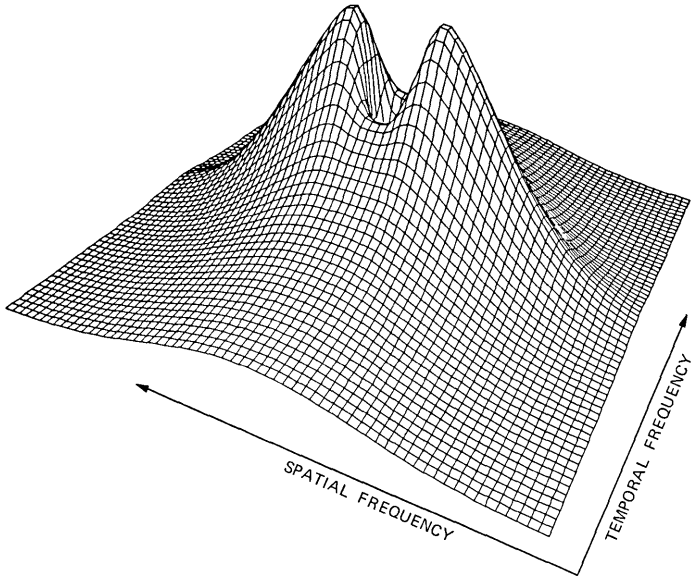


Fig. 6—Perspective view of spatio-temporal frequency response.

the surface has been drawn over all four quadrants making use of symmetry. It suggests a volcano with a deep central crater.

It is customary[7] to think of the response as being brought about by an interplay of excitation and inhibition, with inhibition responsible for the crater. Looked at in this way the measurements suggest, at least to a first approximation, that excitatory and inhibitory responses in themselves may be separable and that the effects of inhibition simply subtract from the effects of excitation. These assumptions will be made. The response functions can then be formally broken down:

$$\begin{aligned} R(\rho, t) &= R_e(\rho, t) - R_i(\rho, t) \\ &= U_e(\rho)V_e(t) - U_i(\rho)V_i(t) \end{aligned} \tag{15}$$

$$\begin{aligned} S(\nu, f) &= S_e(\nu, f) - S_i(\nu, f) \\ &= G_e(\nu)H_e(f) - G_i(\nu)H_i(f) \end{aligned} \tag{16}$$

with

$$G_e(\nu) = \int_0^\infty U_e(\rho)2\pi\rho J_o(2\pi\rho\nu)d\rho,$$

$$H_e(f) = \int_0^\infty V_e(t)e^{-j2\pi ft}dt,$$

and similarly for the inhibitory functions.

2.2 *Choice of Functions*

To satisfy physical considerations, all the component functions should be low-pass in character. Of the immense number of possibilities we consider just several. The Gaussian function comes readily to mind particularly for spatial spreads.

If

$$U_e(\rho) = \frac{1}{2\pi a^2} e^{-\rho^2/2a^2}, \tag{17}$$

then of course the Fourier transform is also Gaussian:

$$G_e(\nu) = e^{-2\pi^2 a^2 \nu^2}. \tag{18}$$

The function has another property which can be especially useful in computations, namely that as a function of two variables x and y, i.e., $\rho^2 = x^2 + y^2$, it is separable:

$$U_e(x, y) = \left(\frac{1}{\sqrt{2\pi a^2}} e^{-x^2/2a^2}\right)\left(\frac{1}{\sqrt{2\pi a^2}} e^{-y^2/2a^2}\right). \tag{19}$$

None of the other functions of interest to us has this property.

Another possible candidate for point spreads is the exponential

$$U_e(\rho) = e^{-2\pi b\rho}; \qquad p > 0 \tag{20}$$

and then

$$G_e(\nu) = \frac{b}{2\pi(b^2 + \nu^2)^{\frac{3}{2}}}. \tag{21}$$

At high values of ν, $\nu \gg b$, the function decreases as $(1/\nu)^3$ which corresponds to a fall-off of 18 dB/octave.

On the other hand, if the spatial frequency response function were an exponential then there would be no straight-line asymptote on a log-log plot, but rather a response which would be

$$U_e(\rho) = \frac{c}{2\pi(c^2 + \rho^2)^{\frac{3}{2}}} \tag{22}$$

with

$$G_e(\nu) = e^{-2\pi c|\nu|}. \tag{23}$$

This is often called the Cauchy response.

Since the temporal frequency responses are similar to the spatial frequency responses similar functions can be used to model these. The important differences are that the function $V(t)$ is one-sided and that eq. (13), instead of (12), is used to obtain the Fourier transform.

The Gaussian function can be used in an approximate way by shifting it a distance t_o to the right along t and deleting it leftward of $t = 0$:

$$V_e(t) = \frac{1}{\sqrt{2\pi}\tau} e^{-(t-t_o)^2/2\tau^2}; \qquad t \geqq 0$$
$$= 0; \qquad t < 0. \tag{24}$$

When t_o/τ is greater than three, say, then there is negligible error in assuming that $V_e(t)$ is the Gaussian function for all t, $t < 0$ included. Then

$$H_e(f) = e^{-2\pi^2\tau^2 f^2 - j2\pi f t_o}. \tag{25}$$

In computer simulation a simple exponential time response, often known as the Poissonian, would be the easiest because it can be effected by recursion. That function and its transform are

$$V_e(t) = (1/\tau_1)e^{-t/\tau_1}; \qquad t \geqq 0 \tag{26}$$

$$H_e(f) = \frac{1}{1 + j2\pi f\tau_1}. \tag{27}$$

A function in which there is theoretical interest[12,14] is one that occurs in diffusion processes. Kelly[12] found that the high-frequency asymptote for large-area flicker responses could be fitted well with a frequency function which one would find in diffusion that had no losses in the diffusing substance, namely with

$$H_e(f) = C_1 e^{(-|2\pi f\tau|^{\frac{1}{2}})}. \tag{28}$$

If the Laplace transform is taken as

$$H_e(s) = C_1 e^{-(2s\tau)^{\frac{1}{2}}}, \tag{29}$$

then the time function is [12,15]

$$V_e(t) = \frac{\tau^{\frac{1}{2}} e^{-\tau/2t}}{(2\pi)^{\frac{1}{2}} t^{\frac{3}{2}}}; \qquad t \geqq 0. \tag{30}$$

The six models which were compared with the experimental data are:

(*i*) Gaussian temporal/Cauchy spatial (*G/C*)

$$|S(\nu, f)| = Ae^{-2\pi^2 f^2\tau_1^2}(e^{-\nu\sigma_e} - ke^{-2\pi^2 f^2\tau_2^2 e^{-\nu\sigma_i}}), \tag{31}$$

(*ii*) Poissonian temporal/Cauchy spatial (*P/C*)

$$|S(\nu, f)| = \frac{A\{[e^{-\nu\sigma_e}(1 + 4\pi^2 f^2\tau_2^2) - ke^{-\nu\sigma_i}]^2 + (2\pi f\tau_2 ke^{-\nu\sigma_i})^2\}^{\frac{1}{2}}}{(1 + 4\pi^2 f^2\tau_2^2)(1 + 4\pi^2 f^2\tau_2^2)^{\frac{1}{2}}}, \tag{32}$$

(*iii*) Diffusion-Gaussian temporal/Cauchy spatial (*D-G/C*)

$$|S(\nu, f)| = A(e^{-(f\tau_1)^{\frac{1}{2}}}e^{-\nu\sigma_e} - ke^{-2\pi^2 f^2\tau_2^2}e^{-\nu\sigma_i}), \tag{33}$$

and three further models in which the Gaussian is substituted for the Cauchy response giving G/G, P/G, and D-G/G.

Note that in four of the models, G/C, P/C, G/G, and P/G, one time-lag stage is common to excitation and inhibition (Fig. 7a). The remaining two models, involving diffusion, have distinct paths for the two effects (Fig. 7b).

Each of the models differs from the others in its exact functional shapes but they are all similar in their form. Figure 8 illustrates the evolution of the point spread as given by the P/G model. The point spread function is shown at the instant of occurrence of the point impulse and at two subsequent time instants thereafter. In this, as in all the other models, the excitatory effect is confined to a smaller region and has a faster time course than the inhibitory effect.

III. SELECTION OF PARAMETERS

Each of the models chosen for comparison with the experimental data has six undetermined parameters: the gain A, time constants τ_1 and τ_2, space constants σ_e and σ_i, and the per unit inhibition k. The

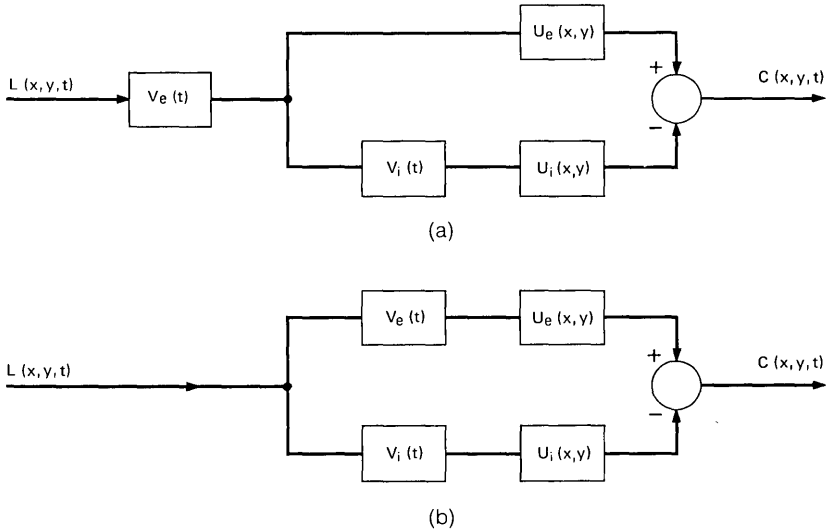


Fig. 7—System block diagram: (a) for P/G, G/G, P/C, and G/C models; (b) for D-G/G and D-G/C models.

parameters have to be given values to produce as close a fit as possible between model and data.

The following performance index may be used as an appropriate measure for the closeness of fit:

$$P = \sum_{i=1}^{N} \{[m_i - 1/|S(\nu_i, f_i)|]/\epsilon_i\}^2, \tag{34}$$

where m_i is the measured threshold modulation at the spatial frequency ν_i and temporal frequency f_i and ϵ_i is the estimated (standard) error of that measurement. The summation is over all N points measured at a given luminance L_o.

This will be called the aggregate-square fractional error, or ASFE, index. The ASFE index is perhaps the most defendable in light of the experimental procedure. However, if the aim is to obtain the best representation of data plotted as $(1/m)$ along a logarithmic scale (Figs. 1–5), then a better index is

$$P = \sum_{i} \{\log [|S(\nu_i, f_i)|/m_i]\}^2, \tag{35}$$

which can be called the aggregate-square log error, or ASLE, index.

Irrespective of index, the array of six parameter values can be looked upon as a vector $\mathbf{T}$ and the performance index as a real-valued function of it. Our object is then to find that location $\mathbf{T}_m$ in six-space at which $P(\mathbf{T})$ assumes its smallest value. However, there is no way of recognizing a global minimum and it is therefore impractical to insist on finding it. The object is rather to find as good a value for $\mathbf{T}$ as possible, while keeping computer expenditures within reasonable bounds.

Of the many possible parameter search routines we tried a gradient-dependent algorithm, random search, and a combination of the two. Random search proved the more successful, almost as good on its own as in combination with gradient techniques.

The gradient in question is

$$\nabla P = \sum_{n=1}^{6} \frac{\partial P}{\partial T_n} \mathbf{a}_n, \tag{36}$$

where $\mathbf{a}_n$ is the unit vector along the nth coordinate axis and T_n is the scalar $(\mathbf{T} \cdot \mathbf{a}_n)$. The components of the gradient were evaluated in one of two ways:

(*i*) approximate differentiation:

$$\frac{\partial P}{\partial T_n} \doteq \frac{P(\mathbf{T} + \Delta T_n \mathbf{a}_n) - P(\mathbf{T})}{\Delta T_n}; \tag{37}$$

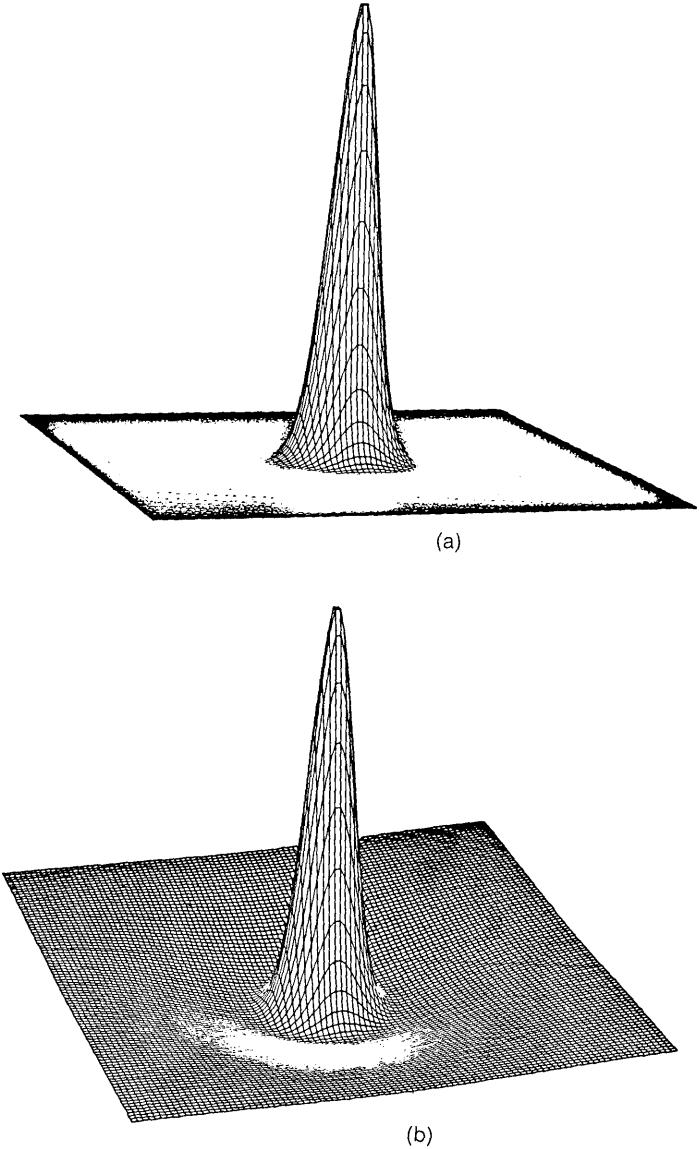
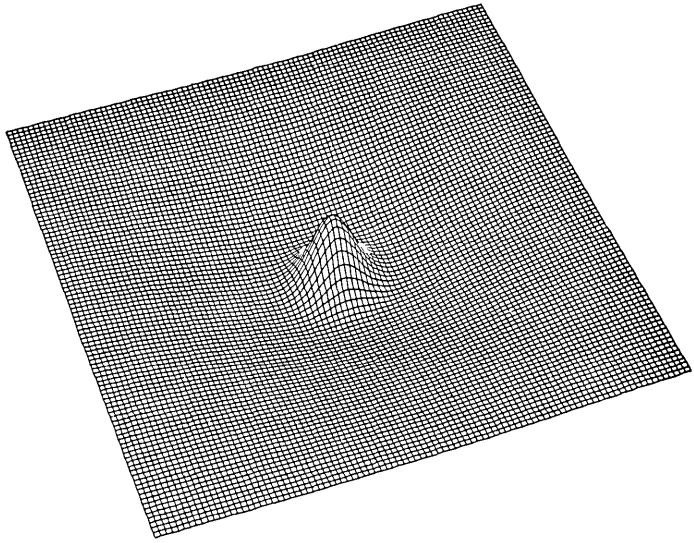


Fig. 8—Evolution of the point spread function in the Poissonian/Gaussian model. Inhibitory effect has been exaggerated. (a) at $t = 0$, (b) at $t = 45$ ms, (c) at $t = 150$ ms.



(c)

Fig. 8 (continued).

(*ii*) evaluation of exact expressions which, given (34), are

$$\frac{\partial P}{\partial T_n} = \sum_i 2\{[m_i - 1/|S(\nu_i, f_i)|]/\epsilon_i\} \times [1/\epsilon_i |S(\nu_i, f_i)|^2] \frac{\partial |S(\nu_i, f_i)|}{\partial T_n}. \quad (38)$$

Using (37), care had to be taken in choosing the size of ΔT_n.

Given the gradient at the vector location $\mathbf{T}_j$, the next location with a lower value of P should be at

$$\mathbf{T}_{j+1} = \mathbf{T}_j - K \nabla P|_{\mathbf{T}=\mathbf{T}_j}. \quad (39)$$

This will prove to be so, provided K is small enough. Improved convergence rates are possible by making K variable,[16] increasing its value with repeated improvements in P, and decreasing it with failures. The next location to be tested is then not given by (38) but by

$$\mathbf{T}_{j+1} = \mathbf{T}_b - K_j \nabla P|_{\mathbf{T}=\mathbf{T}_b}, \quad (40)$$

where $\mathbf{T}_b$ is the location at which the last lowest value of P was calculated and K_j has been determined from a starting value K_o by multiplication with either $\alpha (0 < \alpha < 1)$ or $\gamma (1 < \gamma)$, depending on outcomes of the j iterations thus far.

A difficulty with gradient-dependent search is that it may end at a local minimum which is far above the global, and that often proved to be so. A way around this is to alternate between the gradient-dependent search mode and random search. In random search the next location to be tested would be

$$\mathbf{T}_{j+1} = \mathbf{T}_b + K \sum_{n=1}^{6} G_n R_j(n) \mathbf{a}_n, \tag{41}$$

where $\mathbf{T}_b$ is again the last best location, K is a constant that scales the size of the search volume, G_n's are further scaling factors designed to make the search about equally sensitive along the different coordinates, and $R_j(n)$ is a Gaussian variate obtained from a (pseudo)-random number routine taking a fresh value for each component and each iteration.

Typically, a computational cycle would consist of gradient-dependent search to within a convergence test specification, taking some 20 to 100 iterations, followed by 100 iterations of random search. The number of cycles depended on progress and could be as many as 50.

Most of the performance improvements were found to come from the random search phases of the computational cycles. For that reason the gradient-dependent phase was dispensed with in many calculations, and then K of (40) became a variable similar to K_j of eq. (39). The calculation was still done in cycles, starting each cycle with a large value of K.

IV. RESULTS

Although there is no guarantee that the performance indexes finally arrived at are the lowest possible, in each case the chances are small that there would be anything substantially lower. Hence, Table I can be taken as a good guide for comparing the effectiveness of the different models in fitting the data. The table gives rms deviations D, which are calculated from P in accordance with

$$D = [P/(N - 6)]^{\frac{1}{2}}. \tag{42}$$

Division is by $(N - 6)$, because the parameters provide six degrees of freedom. For Table I, P was as defined by eq. (34), i.e., the ASFE criterion.

From the last column of Table I it can be seen that the best of the six models is the Diffusion-Gaussian/Cauchy and the worst the Gaussian/Cauchy. The Poissonian/Gaussian is somewhat worse than the average over the group. A comparison of the models by order of

TABLE I—RMS DEVIATIONS

Summary of rms deviations, D, derived from ASFE performance index [eq. (34)]. In computation, experimenter's estimates of experimental errors were used with Kelly's data and assumed errors with Robson's data.

Model \ Luminance (mL)	Kelly's Data				Robson's Data	Mean D for Model
	62.8	15.2	3.7	0.91	6.3	
Poissonian/Gaussian	7.0	4.3	3.9	7.6	3.7	5.3
Poissonian/Cauchy	7.2	4.4	4.0	7.9	3.4	5.4
Gaussian/Gaussian	4.0	4.6	4.7	8.8	2.9	5.0
Gaussian/Cauchy	6.1	6.2	5.6	9.8	4.2	6.4
Diff-Gauss/Gaussian	4.0	3.6	4.3	4.4	3.9	4.0
Diff-Gauss/Cauchy	3.4	3.7	4.5	4.5	2.8	3.8

rank within each set, and then over the sets, shows the two Diffusion-Gaussian models fit best, closely followed by the Poissonian/Gaussian model.

Except with Robson's data, where assumed error values were used, the actual magnitude of D in Table I has significance. With P by eq. (34) being measured relative to experimental errors one would expect with a perfect model fit a D value of unity. $(D - 1)$ is then the increase in relative error due to the model, and D can be thought of as error gain. In this sense all the models, including the best, give only poor fits.

The D-G/C model is shown fitted to Robson's data in Figs. 9a and b. According to Table I this ought to be about the best fit, but obviously is only fair. The same data is fitted by the P/G model in Figs. 10a and b. The P/G model is shown fitted to Kelly's data at 62.8 mL in Figs. 11a and b. According to Table I the P/G model represents nearly the worst fit.

Parameter values for the P/G model are given in Table IIA. These were determined using the relative error criterion. Table IIB gives parameter values for the same model but determined by the log departure criterion. The final mean log departures are shown in the bottom row. There are noticeable differences between the parameter values in Table IIA and Table IIB but, given the rather poor fit between model and data, agreement is good. Consistent trends are apparent in both sets: with decreasing luminance the gain (A) of the system decreases accompanied by a decrease in fractional inhibition

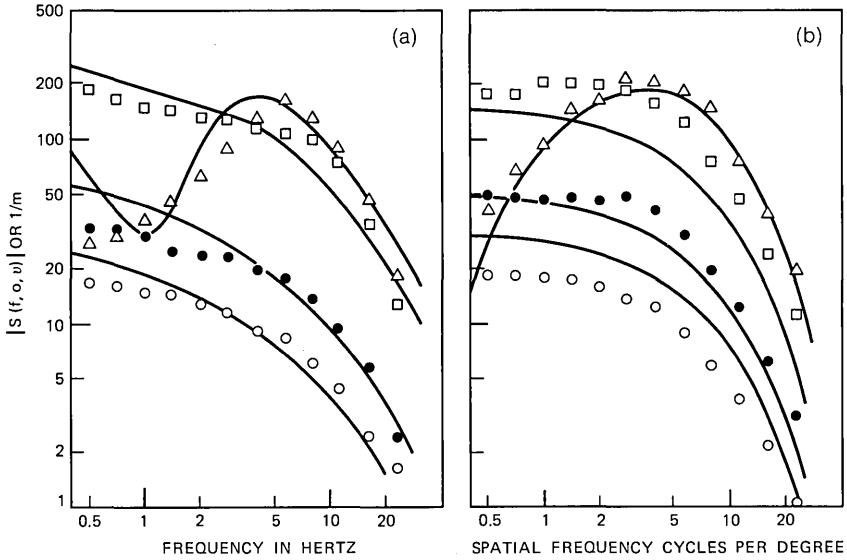


Fig. 9—Diffusion-Gaussian/Cauchy model applied to Robson's data. (a) Temporal frequency response, parameters as in Fig. 5a. (b) Spatial frequency response, parameters as in Fig. 5b.

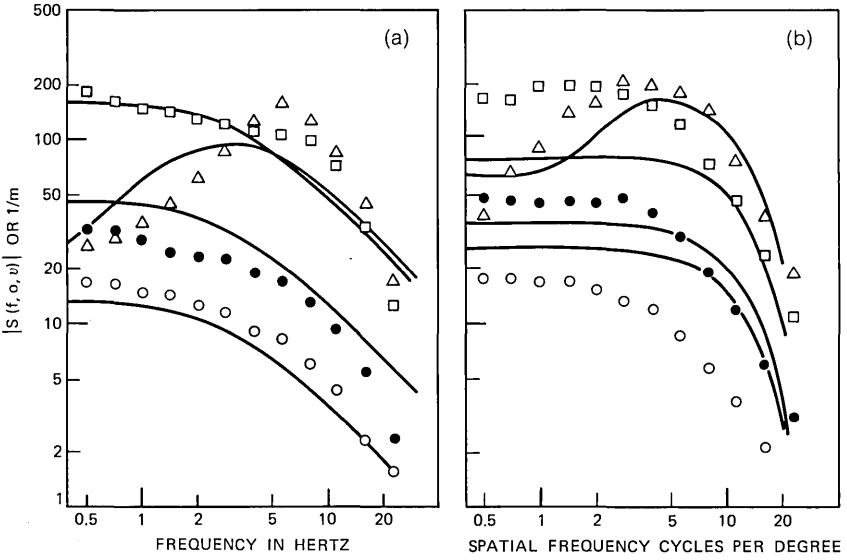


Fig. 10—Poissonian/Gaussian model applied to Robson's data. (a) Temporal frequency response, parameters as in Fig. 5a. (b) Spatial frequency response, parameters as in Fig. 5b.

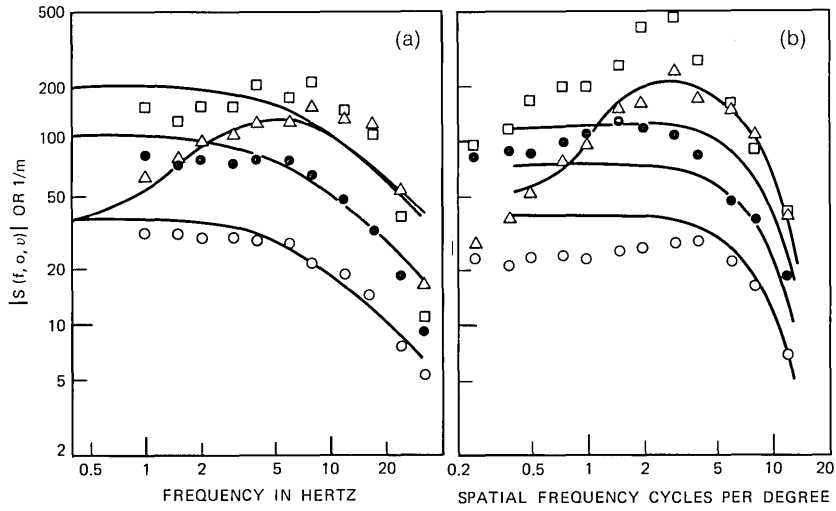


Fig. 11—Poissonian/Gaussian model applied to Kelly's data at 62.8 mL. (a) Temporal frequency response, parameters as in Fig. 1a. (b) Spatial frequency response, parameters as in Fig. 1b.

(k). The time constants tend to increase with lower luminance while the space constants remain unchanged. The parameter values for the D-G/C model obtained using the log of departure criterion are given in Table III. With this different model, parameter values are naturally very different, but the variations with luminance are similar to those with the P/G model and, indeed, with the remaining models.

TABLE IIA—PARAMETER VALUES IN POISSONIAN/GAUSSIAN MODEL DETERMINED WITH ASFE PERFORMANCE INDEX

	Parameter \ Luminance (mL)	Kelly's Data				Robson's Data
		62.8	15.2	3.7	0.91	6.3
1	A	298	236	145	116	219
2	τ_1 (ms)	39	32	32	61	45
3	τ_2 (ms)	63	43	70	102	52
4	σ_e (min arc)	1.48	1.55	1.49	1.49	1.01
5	σ_i (min arc)	9.82	4.72	10.1	6.19	5.62
6	k	0.9976	0.9831	0.9579	0.8150	0.9554
7	D	7.0	4.3	3.9	7.6	3.7
8	A/L_o	4.82	15.5	40.3	127.2	34.8
9	$(1 - k)$	0.0024	0.0169	0.0421	0.1850	0.0446

TABLE IIB—PARAMETER VALUES IN POISSONIAN/GAUSSIAN MODEL DETERMINED WITH ASLE CRITERION [EQ. (35)]

	Parameter \ Luminance (mL)	Kelly's Data				Robson's Data
		62.8	15.2	3.7	0.91	6.3
1	A	234	168	134	93	198
2	τ_1 (ms)	29	37	53	79	55
3	τ_2 (ms)	34	58	80	101	55
4	σ_e (min arc)	1.52	1.40	1.37	1.34	1.01
5	σ_i (min arc)	9.68	8.30	10.51	10.28	5.58
6	k	0.990	0.971	0.911	0.740	0.996
7	D (log units)	0.48	0.56	0.63	0.73	0.49
8	A/L_o	3.72	11.05	36.2	102	31.4
9	$(1 - k)$	0.010	0.029	0.089	0.260	0.004

V. DISCUSSION

Both the D-G/C and the P/G models will be useful in practice, particularly the latter when simplicity of computation is a major consideration. However, the fact that none of the models fits the data well enough to satisfy any fundamental inquiry prompts us to look again at the assumptions of Section 2.1.

One can scarcely doubt the interplay of excitation and inhibition in the visual mechanism, and that inhibition spreads over a wider

TABLE III—PARAMETER VALUES IN DIFFUSION-GAUSSIAN/CAUCHY MODEL DETERMINED WITH ASLE PERFORMANCE INDEX [EQ. (35)]

	Parameter \ Luminance (mL)	Kelly's Data				Robson's Data
		62.8	15.2	3.7	0.91	6.3
1	A	1596	943	810	372	853
2	τ_1 (ms)	472	489	649	656	496
3	τ_2 (ms)	74	75	74	111	98
4	σ_e (min arc)	9.33	8.01	6.47	7.43	8.59
5	σ_i (min arc)	12.38	11.45	6.50	8.27	32.4
6	k	0.517	0.479	0.351	0.236	0.677
7	D (log units)	0.45	0.48	0.50	0.61	0.33
8	A/L_o	25.4	62	219	409	135
9	$(1 - k)$	0.483	0.521	0.649	0.764	0.323

area and persists longer than the excitation, i.e., is confined to lower spatial and temporal frequencies. However, it is probably untrue that inhibition simply subtracts from the excitation. It is more likely[17] that it acts as a shunt, or a reduction in through-put gain, for which simple subtraction is only a first approximation. One could also expect a more precise characterization of inhibitory action to explain part of the adaptive changes. However, the model would be nonlinear and more complicated.

Apart from linearity, it is very probable that more separability of functions has been assumed than is warranted. The statement that excitation (or inhibition) is separable into space and time functions purports that, given a point flash, the form of the spatial response is independent of time, or that the shape of the time function is independent of distance from the stimulus point. This is probably true of the spread which is due to optical smearing of the retinal image. But it is probably untrue of the lateral spread of neural interactions. Since neural interactions predominate in the wider inhibitory spread, separability should be expected to be a poorer assumption for inhibition than for excitation. This seems to be borne out by the data.

The assumption of uniformity raises another question. To speak of isoplanatic patches is, of course, no more than a simplification. Even the central fovea varies substantially in receptor packing density within the space of less than a degree. It is therefore difficult to maintain the assumption of uniformity with data obtained for spatial frequencies of one cycle/degree or lower. To justify convolution in the presence of nonuniformity we only need to be sure that the spatial spread is small compared to the size of the "uniform" patch. However, we need uniformity over much more than $(1/f_c)$ in order to justify a Fourier transform to within f_c of the frequency origin. If this condition is not met, then with a sinusoidal input the output may, in the extreme, be nonsinusoidal even over only a part of a cycle. But our assumption of threshold is that a criterion value be exceeded by the peak-to-peak output and this then will not be related to the calculated transfer function.

The concept of detection needs to be examined, not only where lack of retinal uniformity is critical. It is unlikely that detection is based on a comparison of just two values, a maximum and a minimum in the output, and that this comparison is independent of how far apart in space and time these two values actually are. It is more likely that there should be a pooling of evidence and that there should be a decline in detectability, the further apart the relevant events.

However, it need not follow that, given a more complicated detection mechanism, the modeling done here would be invalidated. The detector with variable weighting of evidence could, in fact, be equivalent to a spatial/temporal filter in its own right, followed by the kind of decision stage assumed here. If this were so, then it would only mean that not all the filtering evident from threshold data can be attributed to peripheral processes, but that some of it is due to central neural activity. This is an important distinction where comparisons are made between the filtering evident in stimulus detection and in, say, perception of brightness. Inconsistencies of this nature have already been noted in the literature,[18] but have not been satisfactorily explained.

Higher-level filtering might also be responsible for the frequency-selective fatiguing discovered by Blakemore and Campbell.[19] It seems improbable that spatial filtering by optical and retinal spread constitutes spatial frequency channels which may be independently adapted, but higher-level filtering could, in fact, occur after a Fourier-like signal transformation. But again, the presence of any transformations like these would not affect the present modeling. They might however, affect adaptation effects.

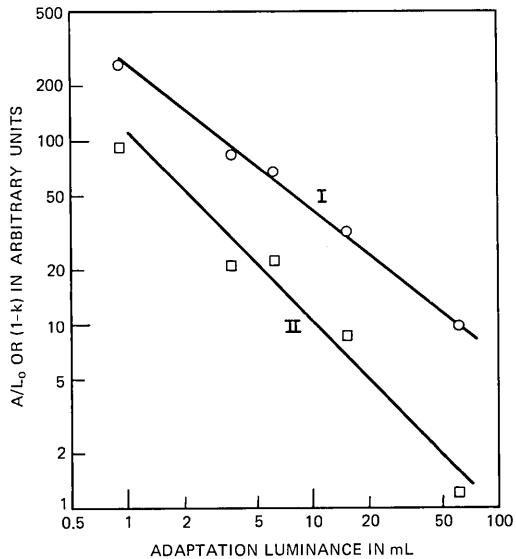


Fig. 12—Adaptation of gain parameters A/L_o (line I) and $(1 - k)$ (line II) against luminance as obtained in fitting Poissonian/Gaussian model to Kelly's and Robson's data.

Of the adaptive changes which are evident from the present modeling, the variation in gain requires comment. The fact that the gain constant A was seen to decrease with decreasing adaptive luminance might be taken to mean that the system becomes less sensitive in the dark. As is well known, visual sensitivity goes up markedly with darkness and the present results do not, in fact, contradict this. By eq. (8), $(1/m)$ equals $|S(u, 0, f)|$ only to within the multiplicative constant $T(p)/L_o$. Assuming that the threshold does not change, then to make the gain values at different luminances L_o comparable to each other they have to be divided by L_o. A/L_o does in fact go up with decreasing luminance as can be seen in row 8 of Table IIA and elsewhere. The actual A/L_o values are different across the models but the trend is always the same.

As the adaptation luminance decreases there is an additional increase in sensitivity restricted to low frequencies. This occurs because of the decline in fractional inhibition k. The zero-point value of $|S|$ is with all models $A(1 - k)/L_o$. The net excitation $(1 - k)$ is given in row 9 of Tables IIA, IIB, and III. A/L_o and $(1 - k)$ have also been plotted for the P/G model in Fig. 12. From the plot one can infer that for the P/G model and in the range $1.0 \leqq L_o \leqq 100$ mL

$$A/L_o = \text{const}_1 \times L_o^{-0.81}, \tag{43}$$

$$(1 - k) = \text{const}_2 \times L_o^{-1.03}, \tag{44}$$

so that

$$|S(0, 0, 0)| = \text{const}_3 \times L_o^{-1.84}. \tag{45}$$

The increase in low-frequency sensitivity with decreasing luminance is at the expense of bandwidth.

VI. CONCLUSION

Six spatio-temporal models of human visual filtering were tested against published experimental data on visual spatio-temporal sine-wave thresholds. These models arose as specific examples from a definite theoretical framework. It was assumed that thresholds could be related to a fixed peak-to-peak difference in a visually filtered version of the input stimulus, and that the filtering could be taken as time-invariant and spatially uniform and isotropic. Particular attention was directed to the question of whether the response was separable into functions of time and space. We showed that the total response is not so separable in this way. However, it was assumed that if the

response is expressed as an algebraic difference of two terms, excitation and inhibition, the individual terms would be separable.

Component functions which were tried were exponential, Gaussian, and diffusion-like functions of time, and Gaussian and Cauchy functions of space. The best fit was obtained with a model which has a diffusion-like time function for excitation, a Gaussian time function for inhibition, and Cauchy space functions for both. The diffusion function, as a model of the time course of excitation, has previously been advocated by Ives,[14] Kelly,[12] and others. The degree of fit obtained in the present study, involving both time and space, was however only moderate and no strong argument can be brought forward in favor of any of the functions, not even the best-fitting. In the best case the average departure from the model was three times larger than the average estimated experimental error. The present results do not exclude any of the functions either, for the fit was probably affected more by the restrictions of the framework than the choice of function.

In each of the models six parameter values had to be determined. These were gain, fractional inhibition, two time constants, and two space constants. Parameter searches consisted of up to 50 passes of gradient-dependent convergence and evolutionary random search. Random search was invariably found to be the more productive phase in all the computational passes.

With adaptation luminance between 1 and 60 mL, the time constants were found to be slightly larger at the low luminances than at the high, the space constants were almost nonvarying, and the gain and fractional inhibition decreased with decreasing luminance. As expected, the sensitivity, measured as gain divided by luminance, was found to go up with decreasing luminance. The reduction in fractional inhibition was shown to give a further increase in sensitivity with decreasing luminance, but only at low frequencies. With one model (P/G) the sensitivity at zero frequency was found to vary inversely as the 1.84 power of luminance, 0.81 of this being due to variation in overall sensitivity and the remainder due to changes in inhibition.

The major purpose of the present model fitting was to find a filter function for use in a program for predicting the subjective quality of visual signal coding schemes. Of the six models the most economical computational procedures are provided by the Poissonian/Gaussian model. The Poissonian, or negative exponential, time functions can be implemented recursively, using a delay of only one or two picture frames, and the Gaussian space functions, being themselves separable

into products of functions of x and y, can be implemented by two successive, modest transverse filter operations, instead of requiring one very large operation. This model was found to fit the data nearly as well as the best. Considering its computational advantages, it will no doubt be the one to find most use.

VII. ACKNOWLEDGMENTS

Both J. J. Robson of the University of Cambridge and D. H. Kelly of the Stanford Research Institute kindly explained their experiments and discussed their data. As noted, Dr. Kelly was also able to send copies of data and estimates of experimental errors.

REFERENCES

1. Ives, H. E., "Critical Frequency Relations in Vision," J. Opt. Soc. Am. and Rev. Sci. Instr., *6*, 1922, pp. 254–268.
2. Schade, O. H., "Optical and Photoelectric Analog of the Eye," J. Opt. Soc. Am., *46*, 1956, pp. 721–739.
3. Kelly, D. H., "J_0 Stimulus Pattern for Visual Research," J. Opt. Soc. Am., *50*, 1960, p. 1115.
4. Robson, J. J., "Spatial and Temporal Contrast-Sensitivity Functions of the Visual System," J. Opt. Soc. Am., *56*, 1966, pp. 1141–1142.
5. Kelly, D. H., "Frequency Doubling in Visual Responses," J. Opt. Soc. Am., *56*, 1966, pp. 1628–1633.
6. Kelly, D. H., "Adaptation Effects on Spatio-Temporal Sine-Wave Thresholds," Vision Res., *12*, 1972, pp. 89–101.
7. Budrikis, Z. L., "Visual Fidelity Criterion and Modeling," Proc. IEEE, *60*, 1972, pp. 771–779.
8. LeGrand, Y., *Light, Colour and Vision*, London: Chapman and Hall, 1968 (2nd Edition), p. 106.
9. Linfoot, E. H., *Fourier Methods in Optical Image Evaluation*, London and New York: The Focal Press, 1964, p. 15.
10. Westheimer, G., and Campbell, F. W., "Light Distribution in the Image Formed by the Living Human Eye," J. Opt. Soc. Am., *52*, 1962, pp. 1040–1045.
11. Ratliff, F., Hartline, H. K., and Miller, W. H., "Spatial and Temporal Aspects of Retinal Inhibitory Interaction," J. Opt. Soc. Am., *53*, 1963, pp. 110–120.
12. Kelly, D. H., "Theory of Flicker and Transient Responses, I. Uniform Fields," J. Opt. Soc. Am., *61*, 1971, pp. 537–546. Reference is given to earlier papers where high-frequency asymptote was first pointed out.
13. Graham, N. Y., "Perspective Drawing of Surfaces With Hidden Line Elimination." B.S.T.J., *51*, No. 4 (April 1972), pp. 843–861.
14. Ives, H. E., "A Theory of Intermittent Vision," J. Opt. Soc. Am. and Rev. Sci. Instr., *6*, 1922, pp. 343–361.
15. Roberts, G. E., and Kaufman, H., *Table of Laplace Transforms*, Philadelphia and London: Saunders, 1966, p. 246.
16. Cantoni, A., "Optimal Approximations with Piecewise Linear Functions," Ph.D. Thesis, The University of Western Australia, 1972.
17. Sperling, G., "Model of Visual Adaptation and Contrast Detection," Perception Psychophys., *8*, 1970, pp. 143–157.
18. Hay, G. A., and Chester, M. S., "Signal-Transfer Functions in Threshold and Super Threshold Vision," J. Opt. Soc. Am., *62*, 1972, pp. 990–998.
19. Blakemore, C., and Campbell, C. W., "Adaptation to Spatial Stimuli," Proc. Physiol. Soc., September 1968, pp. 11p–13p.

Contributors to This Issue

DAN L. BISBEE, B.S., 1965, Monmouth College; Bell Laboratories, 1955—. Mr. Bisbee has been involved in the measurement of optical transmission losses in bulk glass and optical fibers. He is presently engaged in developing techniques for splicing cables made up of optical fiber waveguides.

ZIGMANTAS L. BUDRIKIS, B.Sc., 1955, and B.E., 1957, University of Sydney; Ph.D., 1970, University of Western Australia; Research Laboratories of the Australian Post Office, 1958–1960; Aeronautical Research Laboratories of the Australian Department of Supply, 1961; University of Western Australia (currently Associate Professor of Electrical Engineering), 1962—. In 1968 Mr. Budrikis was Visiting Lecturer at the University of California at Berkeley and in 1972 he worked at Bell Laboratories. His chief interests are in the subjective effects of coding of visual messages and in the foundations of electromagnetic theory. Member, IEE (London), IREE (Australia), Optical Society of America.

EDWIN L. CHINNOCK, Stevens Institute of Technology; Bell Laboratories, 1939—. Mr. Chinnock has worked on microwave components, microwave radio relay, and helix waveguide fabrication. He is presently working on optical waveguide components.

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, 1966–67; Bell Laboratories, 1967—. Mr. Falconer has worked on problems in coding theory, communication theory, channel characterization, and high-speed data communication. Member, Tau Beta Pi, Sigma Xi, IEEE.

RICHARD L. FRANKS, B.S.E.E., 1963, University of Washington; M.S. (E.E.), 1969, and Ph.D. (E.E.), 1970, University of California, Berkeley; Bell Laboratories, 1970—. Mr. Franks has done work in control theory and algorithms. His current interest is in the modeling and analysis of telephone traffic systems. Member, IEEE, Tau Beta Pi, Sigma Xi.

D. Gloge, Dipl. Ing., 1961, Dr. Ing., 1964, Technical University of Braunschweig, Germany; Bell Laboratories, 1965—. Mr. Gloge's work has included the design and field testing of various optical transmission media and the application of ultra-fast measuring techniques to optical component studies. He is presently engaged in transmission research related to optical fiber communication systems.

Harry Heffes, B.E.E., 1962, City College of New York; M.E.E., 1964, and Ph.D., 1968, New York University; Bell Laboratories, 1962—. Mr. Heffes' work was previously in the areas of control theory and filtering theory. More recently, he has been concerned with modeling and analysis of telephone traffic systems. He has also been an Adjunct Associate Professor of Electrical Engineering at New York University. Member, Tau Beta Pi, Eta Kappa Nu.

Jack M. Holtzman, B.E.E., 1958, City College of New York; M.S., 1960, University of California (Los Angeles); Ph.D., 1967, Polytechnic Institute of Brooklyn; Hughes Aircraft Company, 1958–1963; Bell Laboratories, 1963—. Mr. Holtzman has worked in systems and control theory and is the author of *Nonlinear System Theory—A Functional Analysis Approach* (Prentice-Hall, 1970). More recently, he has been working on traffic theory problems. Member, ORSA.

Francis R. Magee, Jr., B.S., 1968, M.S., 1969, and Ph.D. (EE), 1972, Northeastern University; Bell Laboratories, 1972—. Mr. Magee is currently doing research on data communications problems. Member, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi, Sigma Xi, IEEE.

Enrique A. J. Marcatili, Aeronautical Engineer, 1947, and E.E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–54; Bell Laboratories, 1954—. Mr. Marcatili has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently, he has concentrated on optical transmission media. Fellow, IEEE.

David G. Messerschmitt, B.S. (Electrical Engineering), 1967, University of Colorado; M.S. (Electrical Engineering), 1968, and Ph.D. (Computer, Information, and Control Engineering), 1971, University of Michigan; Bell Laboratories, 1968—. Mr. Messerschmitt has been engaged in studies of new techniques for digital transmission systems. Member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

RAYMOND W. RISHEL, B.S., 1952, M.S., 1953, and Ph.D., 1959, University of Wisconsin; Department of Mathematics, Brown University, 1959–1960; Boeing Company, Seattle, Washington, 1960–1968; Department of Mathematics, Washington State University, 1968–1969; Bell Laboratories, 1969–1972; Department of Mathematics, University of Kentucky, 1972—. At Bell Laboratories, Mr. Rishel did research on stochastic control, and on the application of stochastic control to network management and to control of queuing systems involved in switching machines. Member, SIAM, American Mathematical Society.

PETER W. SMITH, B.Sc., Mathematics and Physics, 1958, and M.Sc. and Ph.D., Physics, 1961 and 1964, McGill University; Visiting Mackay Lecturer in Electrical Engineering, 1970, University of California, Berkeley; Bell Laboratories, 1964—. Mr. Smith has investigated a number of systems for obtaining single-frequency laser operation and is currently investigating the use of waveguide techniques for producing miniature gas lasers. Member, American Physical Society, Optical Society of America, IEEE.

K. K. THORNBER, B.S., 1963, M.S. (E.E.), 1964, Ph.D. (E.E.), 1966, California Institute of Technology; Research Associate, Stanford Electronics Laboratories, 1966–68; Research Assistant, Physics Department, University of Bristol, 1968–69; Bell Laboratories, 1969—. Mr. Thornber is a member of the Unipolar Integrated Circuit Laboratory. Member, Sigma Xi, Tau Beta Pi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 9, November, 1973
Printed in U.S.A.

B. S. T. J. BRIEF

The Accuracy of the Equivalent Random Method With Renewal Inputs*

J. M. HOLTZMAN

(Manuscript received July 19, 1973)

I. INTRODUCTION

The equivalent random method[1] (aso see Ref. 2) is widely used to approximate the blocking probabilities for non-Poisson traffic streams. Although much numerical experience and some analysis (e.g., Ref. 3) suggests that the method is usually reliable for superpositions of overflows, the reason for its accuracy (or errors) deserves further attention.

The equivalent random method first determines the mean M and variance V of the number of the trunks that would be occupied if the traffic were offered to an infinite trunk group. Then an overflow process with the same M and V is offered to the finite trunk group and its blocking calculated.† This blocking is taken as the approximation for the blocking seen by the original traffic.

In this Brief, we derive the range of the blocking probabilities which may be experienced by renewal streams characterized by the same M and V. Since this range may be rather wide, it follows that the success of equivalent random method cannot be explained solely by the constraints put on blockings by fixing M and V. Rather, one should factor in the special structure of the processes. Furthermore, it is seen that one cannot use an arbitrary renewal process to represent another process with the same mean and variance.

---

* A version of this Brief was presented at the Seventh International Teletraffic Congress, Stockholm, June 1973.

† That is, the blocking is calculated for the specific renewal process which is the overflow process from a Poisson input. Conceivably, other types of renewal processes could be used.

II. IMPLICATIONS OF THE EQUIVALENT RANDOM METHOD

Consider a nonlattice renewal process, with distribution function $F(t)$ for the interarrival times, offered to a group of N trunks. The holding times are mutually independent exponential random variables with unity mean (or the mean is the time unit). Blocked calls are cleared and the system is in equilibrium. Define

$$m = \int_0^\infty t dF(t), \tag{1}$$

$$\phi(x) = \int_0^\infty e^{-xt} dF(t). \tag{2}$$

Then it is known that the blocking probability is

$$B = \left\{1 + \binom{N}{1} \frac{1 - \phi(1)}{\phi(1)} + \cdots + \binom{N}{N} \frac{[1 - \phi(1)] \cdots [1 - \phi(N)]}{\phi(1)\phi(2) \cdots \phi(N)}\right\}^{-1} \tag{3}$$

(see, e.g., Ref. 4, Chap. 4). Observe that B depends on N values of $\phi(i)$, $i = 1, \cdots, N$, and that it is an increasing function of these $\phi(i)$. We shall show how the equivalent random method constrains these $\phi(i)$ by obtaining upper and lower bounds on them which, in turn, give upper and lower bounds on B.

The description of the equivalent random method in the Introduction leads to the question of how well M and V characterize a traffic. It turns out that they imply much more than is apparent at first glance. For our renewal input, we have the following relationships:

$$M = m^{-1}, \tag{4}$$

$$V = M\left[\frac{1}{1 - \phi(1)} - M\right] \tag{5}$$

(see, e.g., Ref. 4, Chap. 3*). Thus, (M, V) uniquely determines $(m, \phi(1))$ and vice versa. Specifically,

$$\phi(1) = \frac{V/M - 1 + M}{V/M + M}. \tag{6}$$

Hence, the equivalent random method fixes $\phi(1)$ which is particularly important in (3). Moreover, fixing V and M puts important constraints on the other $\phi(i)$, $i = 2, \cdots, N$, which, in turn, further constrains B.

* Also, see Ref. 5, p. 331/5, for an interesting characterization of peakedness.

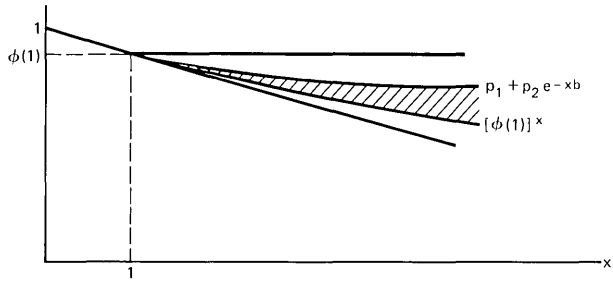


Fig. 1—Constraints on $\phi(x)$ for $x > 1$.

Figure 1 shows how $(m, \phi(1))$ constrains $\phi(x)$ for $x > 1$. The parameters p_1, p_2, b will be given in (16)–(18). All such $\phi(x)$ must lie within the shaded area. The upper bound is a least upper bound and the lower bound, a greatest lower bound. Also shown in Fig. 1 is a wedge for $x > 1$ which respresents simpler, cruder bounds for $\phi(x)$ which follow immediately from the decreasing convex nature of $\phi(x)$.

To derive the lower bound, let $y = e^{-t}$ in

$$E|y| \leqq E^{1/x}|y|^x, \qquad x > 1, \tag{7}$$

with ξ the renewal interarrival time.
We obtain

$$E^x(e^{-t}) \leqq E(e^{-x\xi}) \tag{8}$$

or, in other words,

$$[\phi(1)]^x \leqq \phi(x), \tag{9}$$

so that $\phi(x)$ must lie above the indicated curve for $x > 1$ in Fig. 1. To show that this is a sharp lower bound, let

$$dF(t) = [p_1\delta(t - a) + p_2\delta(t - b)]dt. \tag{10}$$

(p_1, a, p_2, b) must satisfy

$$p_1 + p_2 = 1, \tag{11}$$

$$p_1a + p_2b = m, \tag{12}$$

$$p_1e^{-a} + p_2e^{-b} = \phi(1). \tag{13}$$

By letting b get large and $p_1 \to 1$, we can show that

$$\phi(x) = p_1e^{-xa} + p_2e^{-xb} \to p_1e^{-xa} \to [\phi(1)]^x. \tag{14}$$

The sharp lower bound for $\phi(x)$ may also be derived using Theorem 2.1 on p. 472 of Ref. 6 (see Remark 2.3, p. 474). Use of this theorem*

* The problem to which we applied this theorem is to find sharp upper and lower bounds for $\int_0^\infty e^{-xt}dF(t)$ subject to $\int_0^\infty dF(t) = 1$, $\int_0^\infty t dF(t) = m$, and $\int_0^\infty e^{-t}dF(t) = \phi(1)$, a number fixed by (6).

also leads to the following sharp upper bound for $\phi(x)$:

$$\phi(x) \leqq \phi_m(x) = p_1 + p_2 e^{-xb} \tag{15}$$

with (p_1, p_2, b) satisfying

$$b = \frac{m(1 - e^{-b})}{1 - \phi(1)}, \tag{16}$$

$$p_2 = \frac{m}{b}, \tag{17}$$

$$p_1 = 1 - p_2. \tag{18}$$

We thus obtain that the true B satisfies

$$B_L \leqq B \leqq B_u, \tag{19}$$

where

$$B_L = \left\{1 + \binom{N}{1} \frac{1 - \phi(1)}{\phi(1)} + \cdots + \binom{N}{N} \frac{[1 - \phi(1)] \cdots [1 - \phi^N(1)]}{[\phi(1)]^{[N(N+1)/2]}}\right\}^{-1}, \tag{20}$$

$$B_u = \left\{1 + \binom{N}{1} \frac{1 - \phi_m(1)}{\phi_m(1)} + \cdots + \binom{N}{N} \frac{[1 - \phi_m(1)] \cdots [1 - \phi_m(N)]}{\phi_m(1) \cdots \phi_m(N)}\right\}^{-1}. \tag{21}$$

The blocking probability obtained by the equivalent random method, B_{er}, also satisfies (19) so that a bound on the error for the method is

$$\max \{B_u - B_{\text{er}}, B_{\text{er}} - B_L\}.$$

III. INTERPRETATION OF EXTREMAL SOLUTIONS

Some feeling for these bounds can be obtained by considering the maximum and minimum blocking probabilities attainable when only the mean interarrival time m is constrained (the equivalent random V is unspecified). It is shown in Ref. 7 that the minimum blocking is achieved when arrivals are regular with a separation of m. Our inf may be viewed as approaching that of regular arrivals but with a different mean [the impulse at b in (10) keeps the equivalent random M and V satisfied]. Observe that B_L is the blocking probability seen by a renewal input with constant interarrival times with mean m_1 determined from

$$e^{-m_1} = \phi(1). \tag{22}$$

It is shown in Ref. 7 that with a given m, blocking probabilities

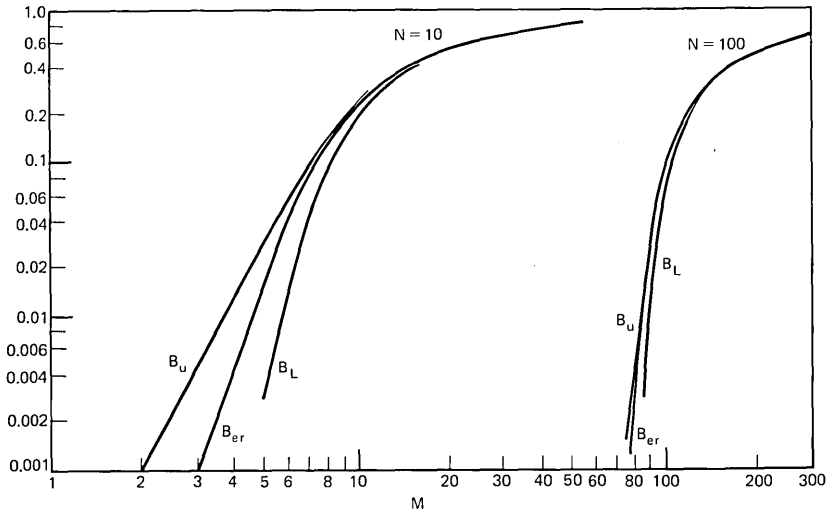


Fig. 2—B_u, B_{er}, B_L for $V/M = 1$.

arbitrarily close to unity may be obtained by having $F(t)$ consist of a step at t sufficiently small and another small step at t sufficiently large. This causes most of the arrivals to come tripping on each other's heels. Our maximum blocking may be viewed as trying to approach this but constrained by V to keep the second step at a finite t.

IV. EXAMPLES AND DISCUSSION

Some bounds are shown in Figs. 2 through 4. These results do not necessarily imply that the equivalent random method is commonly subject to errors of such magnitude. In practice, the method is usually applied to superpositions of overflows and these are a special class of processes, generally not renewal.* Nevertheless, the relatively large differences between the inf and sup blockings suggest that the apparent success of the equivalent random method for superpositions of overflows cannot be explained solely by the constraints put on blockings by fixing M and V. Rather, explanation of this accuracy should factor in the special structure of such processes. (It may be of interest to extend the results of this Brief to take special structures into account.) Furthermore, it is seen that one cannot use an arbitrary re-

* Teletraffic interest need not be confined to simple superposition of overflows from trunk groups; e.g., switching center congestion can alter a traffic.

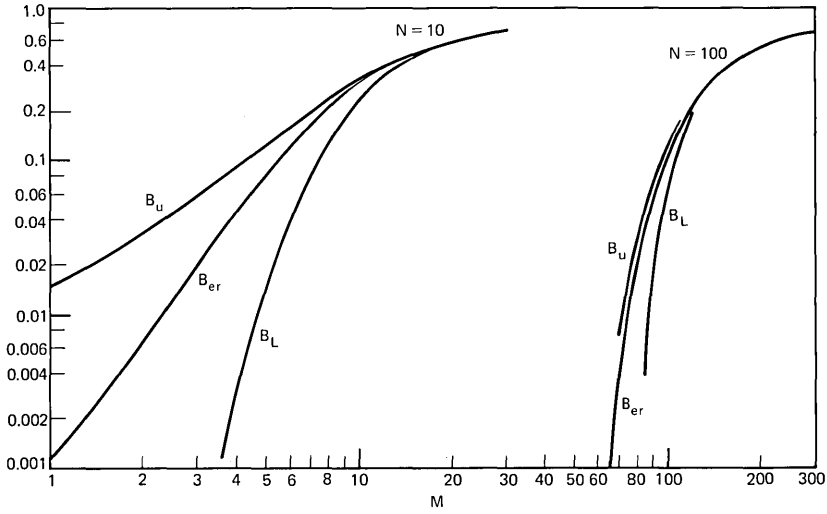


Fig. 3—B_u, B_{er}, B_L for $V/M = 2$.

newal process to represent another process with the same mean and variance.

As an aside, observe that if $V/M > 1$, the blocking B is bounded away from zero no matter how small M is. That is, (6) implies that

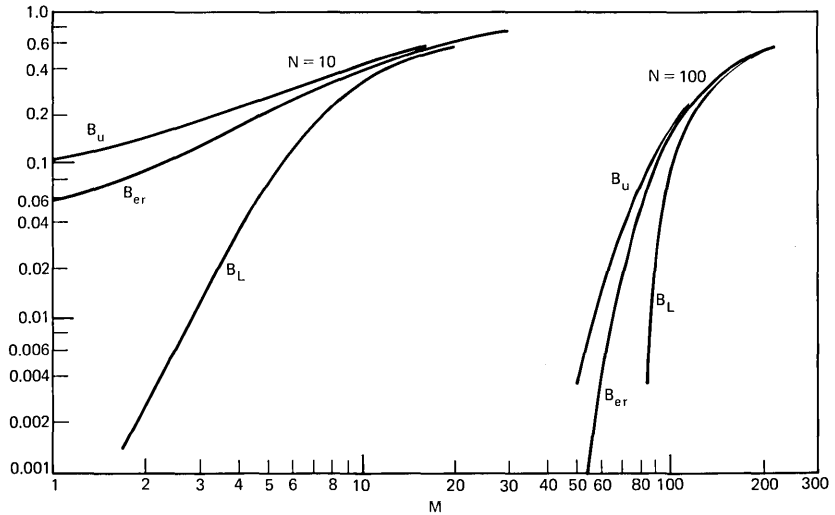


Fig. 4—B_u, B_{er}, B_L for $V/M = 4$.

B_L of (20) for fixed V/M cannot get below B_L evaluated with $\phi(1) = 1 - (M/V)$.

V. ACKNOWLEDGMENTS

Discussions with L. J. Forys and D. L. Jagerman are gratefully acknowledged.

REFERENCES

1. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U. S. A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
2. Bretschneider, G., "Die Berechnung von Leitungsgruppen für überfliebenden Verkehr in Fernsprechanlagen," Nachr.-techn. Z., *9*, 1956, pp. 533–540.
3. LeGall, P., "Le Trafic de Débordement," Anales des Telecommunications, *16*, 1961, pp. 226–238.
4. Tacács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
5. Descloux, A., "On Markovian Servers with Recurrent Input," Proc. Sixth Intl. Teletraffic Congress, 1970.
6. Karlin, S., and Studden, W. J., *Tchebycheff Systems with Applications in Analysis and Statistics*, New York: Interscience Publishers, 1966.
7. Beneš, V. E., "On Trunks with Negative Exponential Holding Times Serving a Renewal Process,"B.S.T.J., *38*, No. 1 (January 1959), pp. 211–258.

Bell System