THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 52 October 1973 Number 8

Volterra Systems With More Than One Input Port—Distortion in a Frequency Converter

By S. O. RICE

(Manuscript received April 19, 1973)

Consider a nonlinear system, with memory, which has two input ports and one output port. It is assumed that the system can be represented by a double Volterra series. Two results for such a system are stated in Part I. The first is a general expression for the sinusoidal components of the output $y(t)$ when the two inputs $x_u(t)$ and $x_v(t)$ are sums of sinusoidal terms. The second result is an expression for the power spectrum of $y(t)$ when $x_u(t)$ is a stationary Gaussian process and $x_v(t) = P \cos pt$. Part II is concerned with using results from the theory of Volterra series for multi-input systems to calculate the third-order distortion in an idealized frequency converter.

I. INTRODUCTION

This paper deals with nonlinear, time-invariant systems with memory which (*i*) have more than one input port, and (*ii*) are driven by inputs which are essentially sums of sine waves.

The paper consists of two parts. Part I is concerned with a system which has two inputs, $x_u(t)$ and $x_v(t)$, and one output $y(t)$. Two results for single-input systems are generalized: (*i*) an expression is given for an arbitrary frequency component of $y(t)$ when $x_u(t)$ and $x_v(t)$ are finite sums of sine waves, and *(ii)* an expression is given for the power spectrum of $y(t)$ when $x_u(t)$ is a stationary, zero-mean, Gaussian noise, and $x_v(t)$ is a single sine wave.

Although most of the discussion in Part I deals with systems having two input ports, many of the results can be formally generalized to systems with more than two inputs.

Part II is devoted to an example which shows how results given in Ref. 1 for a one-input Volterra system can be used to examine systems consisting of a single two-terminal nonlinear element imbedded in a linear network containing sources. The transformation from a multi-input to a single-input system is based upon Thévenin's theorem (see, for example, Anderson and Leon[2]). The example treated here is a frequency converter using a nonlinear capacitor. Particular attention is paid to computing the limiting form of the expression for the third-order distortion when the signal and pump amplitudes become small.

The procedure we use in Part II is essentially a systemization of a procedure used by Gardiner and Ghobrial[3] to study the distortion performance of a varactor frequency converter. As they point out, their treatment differs from the linear time-varying analysis usually employed to study frequency converters. It is appropriate to mention here that a promising new general method of computing distortion in frequency converters has been developed by R. B. Swerdlow.[4] His method is based upon the use of Volterra series with time-varying kernels.

Part I. Two Input Ports

When analysis of the type used to study Volterra systems is applied to nonlinear circuits having two input ports and one output port, some of the simpler results for one-input circuits can be generalized in a straightforward way. Here we state two such generalizations. In the first, the two inputs are sums of sine waves. In the second, one input is stationary, zero-mean Gaussian noise, and the other input is a single sine wave.

The derivations of the generalizations are not given here because they consist of rather straightforward, although lengthy, applications of the procedures used in Ref. 1 to deal with the one-input case.

II. DOUBLE VOLTERRA SERIES

Let $x_u(t)$ and $x_v(t)$ be the two inputs, and let the output $y(t)$ be given by the double Volterra series

$$y(t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty}{}' \frac{1}{m!n!} \int_{-\infty}^{\infty} du_1 \cdots \int_{-\infty}^{\infty} du_m \int_{-\infty}^{\infty} dv_1 \cdots \int_{-\infty}^{\infty} dv_n$$

$$\cdot g_{m;n}(u_1, \cdots, u_m; v_1, \cdots, v_n) \prod_{r=1}^{m} x_u(t - u_r) \prod_{s=1}^{n} x_v(t - v_s), \quad (1)$$

where the prime on $\sum'$ means that the term $m = n = 0$ is omitted. The product $\prod$ is understood to have the value 1 when the number of factors (m or n) is 0, and if n, say, is zero there are no v-integrations. The kernel $g_{m;n}$ is a symmetric function of $u_1, \cdots, u_m$ and of $v_1, \cdots, v_n$.

For the inputs that we shall consider, the $(m + n)$-fold Fourier transform

$$G_{m;n}(f_{u1}, \cdots, f_{um}; f_{v1}, \cdots, f_{vn}) = \int_{-\infty}^{\infty} du_1 \cdots \int_{-\infty}^{\infty} dv_n$$

$$\cdot g_{m;n}(u_1, \cdots; \cdots v_n) \exp\left[-j(u_1\omega_{u1} + \cdots + v_n\omega_{vn})\right] \quad (2)$$

plays an important role. Here $G_{0;0} \equiv 0$, $\omega = 2\pi f$, and $G_{m;n}$ is a symmetric function of $f_{u1}, \cdots, f_{um}$ and of $f_{v1}, \cdots, f_{vn}$.

Much as in Ref. 1, the "harmonic input" method can be used to determine $G_{m;n}$ from the system equations by setting

$$\begin{aligned} x_u(t) &= \sum_{r=1}^{m} \exp(j\omega_{ur}t), \\ x_v(t) &= \sum_{s=1}^{n} \exp(j\omega_{vs}t), \end{aligned} \quad (3)$$

where the ω's are incommensurable, and solving for the coefficient of $\exp[j(\omega_{u1} + \cdots + \omega_{um} + \omega_{v1} + \cdots + \omega_{vn})t]$ in the expansion of $y(t)$. This coefficient is equal to $G_{m;n}(f_{u1}, \cdots, f_{um}; f_{v1}, \cdots, f_{vn})$. Note that if the system output $y(t)$ is applied to the input of a linear transducer, the transducer output can also be expressed as a double Volterra series. The transducer output function corresponding to the transducer input function $G_{m;n}$ is

$$F[j(\omega_{u1} + \cdots + \omega_{vn})]G_{m;n}(f_{u1}, \cdots; \cdots f_{vn}),$$

where $F(j\omega)$ is the transfer function of the transducer.

If, say, n is zero and $m > 0$, $G_{m;0}(f_{u1}, f_{u2}, \cdots, f_{um};)$ is equal to the coefficient of $\exp[j(\omega_{u1} + \omega_{u2} + \cdots + \omega_{um})t]$ in the expansion of $y(t)$ when $x_v(t) \equiv 0$ and $x_u(t)$ is given by (3).

There is a resemblance between the double Volterra series (1) for the two-port inputs $x_u(t)$, $x_v(t)$ (Case A) and the single Volterra series for the special input $x(t) = x_u(t) + x_v(t)$ (Case B). For Case B,

$$y(t) = \sum_{k=1}^{\infty} \frac{1}{k!} \int_{-\infty}^{\infty} du_1 \cdots \int_{-\infty}^{\infty} du_k \hat{g}_k(u_1, \cdots, u_k) \cdot \prod_{r=1}^{k} [x_u(t - u_r) + x_v(t - u_r)]. \quad (4)$$

When the product is expanded and the symmetry of $\hat{g}_k$ is used, the product can be written as

$$\sum_{m=0}^{k} \binom{k}{m} \prod_{r=1}^{m} x_u(t - u_r) \prod_{s=1}^{k-m} x_v(t - u_{m+s}).$$

Setting $n = k - m$ and $u_{m+s} = v_s$ for $s = 1, 2, \cdots, n$ carries (4) into a form which goes into (1) when $\hat{g}_{m+n}(u_1, \cdots, u_m, v_1, \cdots, v_n)$ is replaced by $g_{m;n}(u_1, \cdots, u_m; v_1, \cdots, v_n)$.

The results stated below for Case A show a similar resemblance to the corresponding Case B stated in Ref. 1. For example, when $x_u(t)$ and $x_v(t)$ are the sums of sinusoidal terms, the expression for a particular component in $y(t)$ for Case A can be obtained from the corresponding expression for Case B by replacing $\hat{G}_{m+n}$ in Case B by $G_{m;n}$ and inserting a semicolon at the appropriate place in the string of arguments [as in eq. (9) below].

III. SINUSOIDAL INPUTS

When the inputs $x_u(t)$ and $x_v(t)$ are sums of sinusoidal waves, an expression for any particular component in $y(t)$ can be obtained by extending the analysis given in Section VI-B of Ref. 1. There, eqs. [1, (138), (139), (140)] [meaning eqs. (138), (139), and (140) of Ref. 1] show that if the input $x(t)$ to a one-input system is given by

$$x(t) = \sum_{r=1}^{\mu} P_r \cos \omega_r t, \quad (5)$$

where the ω_r's are incommensurable, then the $\exp [j(N_1\omega_1 + \cdots + N_\mu\omega_\mu)t]$, $N_r \geqq 0$, component of $y(t)$ is

$$\exp [j(N_1\omega_1 + \cdots + N_\mu\omega_\mu)t] \sum_{l_1=0}^{\infty} \cdots \sum_{l_\mu=0}^{\infty} \prod_{r=1}^{\mu} \left[\frac{(P_r/2)^{N_r+2l_r}}{(N_r + l_r)!\, l_r!} \right] \cdot G_n[(f_1)_{N_1+l_1}, (-f_1)_{l_1}, \cdots, (f_\mu)_{N_\mu+l_\mu}, (-f_\mu)_{l_\mu}], \quad (6)$$

where $G_0 \equiv 0$, $(f_\sigma)_k$ denotes the string of k arguments $f_\sigma, f_\sigma, \cdots, f_\sigma,$

and the subscript n on G has the value

$$n = \sum_{r=1}^{\mu} (N_r + 2l_r). \tag{7}$$

Here the notation of Ref. 1 has been changed to bring the statement of (6) in line with the notation used in the present paper.

Methods of computing (6) when the G_n's are constants, i.e., are independent of frequency, have been considered by several writers (see Kroupa[5] and Sea and Vacroux[6]).

To state the generalization of (6) let

$$x_u(t) = \sum_{r=1}^{\mu} P_r \cos \omega_r t, \qquad x_v(t) = \sum_{r=\mu+1}^{\lambda} P_r \cos \omega_r t, \tag{8}$$

where the ω_r's are incommensurable, $\omega_r = 2\pi f_r$, $\lambda = \mu + \nu$, and μ and ν are positive integers. Then the $\exp [j(N_1\omega_1 + \cdots + N_\lambda \omega_\lambda)t]$, $N_r \geqq 0$, component in $y(t)$ is

$$\begin{aligned} \exp [j(N_1\omega_1 + \cdots + N_\lambda\omega_\lambda)t] \sum_{l_1=0}^{\infty} \cdots \sum_{l_\lambda=0}^{\infty} \prod_{r=1}^{\lambda} \left[\frac{(P_r/2)^{N_r+2l_r}}{(N_r + l_r)!\, l_r!} \right] \\ G_{m;n}[(f_1)_{N_1+l_1}, (-f_1)_{l_1}, (f_2)_{N_2+l_2}, \cdots, (f_\mu)_{N_\mu+l_\mu}, (-f_\mu)_{l_\mu}; \\ \cdot (f_{\mu+1})_{N_{\mu+1}+l_{\mu+1}}, \cdots, (f_\lambda)_{N_\lambda+l_\lambda}, (-f_\lambda)_{l_\lambda}]. \end{aligned} \tag{9}$$

Here $G_{0;0} \equiv 0$, and if l or $N + l$ are 0 the corresponding arguments do not appear in $G_{m;n}$. The values of m and n are

$$m = \sum_{r=1}^{\mu} (N_r + 2l_r), \qquad n = \sum_{r=\mu+1}^{\lambda} (N_r + 2l_r). \tag{10}$$

The semicolon in the subscript of $G_{m;n}$ differs in meaning from the semicolon used in [1, (139), (140)]. The notation $(f_\sigma)_k$ is the same as that in (6) and in [1, (169)]. The series (9) may either converge or diverge, depending on the P's and G's.

Changing the signs of ω_1 and f_1 in (9) carries (9) into the expression for the $\exp [j(-N_1\omega_1 + N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t]$ component in $y(t)$, etc. [see the discussion below eq. (5) in Ref. 1]. When some of the ω_r's are commensurable, some of the components in $y(t)$ coalesce and can be treated by the method used in [1, (6), (7)].

To examine the case in which $x_u(t)$ contains a dc component, let f_1 and ω_1 tend to 0 in (8) and (9). Then P_1 is the dc component of $x_u(t)$ and the $\exp [j(N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t]$ component of $y(t)$ is the result of the coalescence (as $f_1 \rightarrow 0$) of (i) the components $\exp [j(N_1\omega_1$

$+ N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t]$ for $N_1 = 0, 1, 2, \cdots, \infty$ and (*ii*) $\exp [j(- N_1\omega_1 + N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t]$ for $N_1 = 1, 2, \cdots, \infty$. When (9) and (9) with $-f_1$ in place of f_1 are summed over the values of N_1, the double sum with respect to l_1 and N_1 can be reduced to a single sum by setting $k = N_1 + 2l_1$ and using the binomial theorem. The desired component, namely $\exp [j(N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t]$, in $y(t)$ when $x_u(t) = P_1 + P_2 \cos \omega_2 t + \cdots + P_\mu \cos \omega_\mu t$ and $x_v(t)$ is given by (8) is found to be

$$\exp [j(N_2\omega_2 + \cdots + N_\lambda\omega_\lambda)t] \sum_{k=0}^{\infty} \sum_{l_2=0}^{\infty} \cdots \sum_{l_\lambda=0}^{\infty} \frac{P_1^k}{k!} \prod_{r=2}^{\lambda} \left[\frac{(P_r/2)^{N_r+2l_r}}{(N_r + l_r)!\, l_r!} \right]$$
$$\cdot G_{m;n}[(0)_k, (f_2)_{N_2+l_2}, (-f_2)_{l_2}, \cdots; \cdots, (f_\lambda)_{N_\lambda+l_\lambda}, (-f_\lambda)_{l_\lambda}]. \quad (11)$$

Here n is given by (10), and $m = k + (N_2 + 2l_2) + \cdots + (N_\mu + 2l_\mu)$ when $\mu \geqq 2$ and $m = k$ when $\mu = 1$.

Equation (9) can be generalized to the case of three or more input ports in a straightforward way.

IV. $x_u(t)$ GAUSSIAN AND $x_v(t) = P \cos pt$

The case $x_v(t) = P \cos pt$ and $x_u(t) = I(t)$, where $I(t)$ is a stationary, zero-mean, Gaussian noise having the two-sided power spectrum $W_I(f)$, can be handled in much the same way as was the case $x(t) = I(t) + P \cos pt$ discussed in Section VII-C of Ref. 1.

The discrete sinusoidal components in $y(t)$ are given by the ensemble average

$$\langle y(t) \rangle = \sum_{n=-\infty}^{\infty} c_n \exp (jnpt), \quad (12)$$

where

$$c_n = \sum_{\sigma=0}^{\infty} \frac{(P/2)^{2\sigma+|n|}}{(\sigma + |n|)!\, \sigma!} S_{n,\sigma,0}(; f_p),$$

$$S_{n,\sigma,k}(f_1, \cdots, f_k; f_p) = \sum_{\nu=0}^{\infty} \frac{Q_\nu[W_I(f')]}{\nu!\, 2^\nu} G_{2\nu+k;2\sigma+|n|}[f_1', -f_1', \cdots, f_\nu',$$
$$-f_\nu', f_1, f_2, \cdots, f_k; (s_nf_p)_{\sigma+|n|}, (-s_nf_p)_\sigma]. \quad (13)$$

Here $2\pi f_p = p$, $s_n = 1$ for $n \geqq 0$, $s_n = -1$ for $n < 0$, and as in (6), $(s_nf_p)_\sigma$ denotes a sequence of σ arguments, all equal to s_nf_p. As explained in connection with [1, (145)], $Q_\nu[W_I(f')]$ denotes a ν-fold integration with respect to $f_1', \cdots, f_\nu'$ with limits $\pm \infty$. The integrand is $W_I(f_1') \cdots W_I(f_\nu')$ times the function [in (13) the function is G] of $f_1', \cdots, f_\nu'$ represented by all of the terms lying to the right of $Q_\nu[W_I(f')]$. $Q_0[W_I(f')]$ denotes the identity operator.

The two-sided power spectrum of $y(t)$ is

$$\begin{aligned} W_y(f) = &\sum_{n=-\infty}^{\infty} |c_n|^2\delta(f - nf_p) \\ &+ \sum_{k=1}^{\infty} \frac{Q_k[W_I(f)]}{k!} \sum_{n=-\infty}^{\infty} \delta(f - f_1 - \cdots - f_k - nf_p) \\ &\cdot \left| \sum_{\sigma=0}^{\infty} \frac{(P/2)^{2\sigma+|n|}}{(\sigma + |n|)!\sigma!} S_{n,\sigma,k}(f_1, \cdots, f_k; f_p)\right|^2. \end{aligned} \tag{14}$$

Replacing $G_{m;n}$ by G_{m+n} in eq. (13) for S and substituting in (14) gives the expression [1, (175)] for $W_y(f)$ in the single input port case $x(t) = I(t) + P\cos pt$. There is a corresponding similarity between the one input port formula [1, (16)] when the two-port expression (14) for $W_y(f)$ is written out.

V. EXAMPLE—COMPUTATION OF $G_{1;1}$

Consider the circuit shown in Fig. 1. The admittance $H(f)$ is linear but the resistor R and capacitor C are nonlinear. The voltage across R is

$$\alpha I_u + \beta I_u^2 \tag{15}$$

and the capacitance of C depends upon the charge Q, the capacitance being $a + bQ$. The output of interest is the voltage $y(t)$ across C:

$$\begin{aligned} Q &= (a + bQ)y, \\ I_u + I_v &= I = dQ/dt. \end{aligned} \tag{16}$$

The current $I_v(t)$ is given by

$$I_v(t) = \int_{-\infty}^{\infty} h(u)[x_v(t - u) - y(t - u)]du, \tag{17}$$

where $h(u)$ is the Fourier transform of $H(f)$.

Elimination of I_v leads to the circuit equations

$$\begin{aligned} x_u &= \alpha I_u + \beta I_u^2 + y, \\ dQ/dt &= I_u + \int_{-\infty}^{\infty} h(u)[x_v(t - u) - y(t - u)]du, \\ Q &= (a + bQ)y. \end{aligned} \tag{18}$$

The G's corresponding to y can be obtained from (18) by the harmonic input method. In using this method it is convenient to work with the notation $z_k = \exp(j\omega_k t)$ where the ω's are incommensurable.

In order to get $G_{1;0}(f_1;)$ we set $x_u = z_1$, $x_v = 0$, $y = c_1z_1 +$ higher harmonics, $I_u = i_1z_1 + \cdots$, and $Q = q_1z_1 + \cdots$. The harmonic input

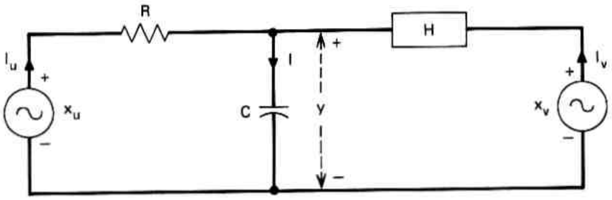

Fig. 1—Circuit with input voltages $x_u(t)$, $x_v(t)$, output voltage $y(t)$, and nonlinear R and C.

method states that $G_{1;0}\ (f_1;)$ is equal to c_1. Substituting in (18) and equating coefficients of z_1 gives

$$\begin{aligned} 1 &= \alpha i_1 + c_1, \\ j\omega_1 q_1 &= i_1 - H(f_1)c_1, \\ q_1 &= ac_1. \end{aligned} \tag{19}$$

Solving for c_1 gives $G_{1;0}\ (f_1;)$. Similarly, starting with $x_u = 0$ and $x_v = z_1$ gives $G_{0;1}(;f_1)$. The results are

$$\begin{aligned} G_{1;0}(f_1;) &= \left[\frac{1}{1 + \alpha H + j\omega a\alpha}\right]_{f_1}, \\ G_{0;1}(;f_1) &= \left[\frac{\alpha H}{1 + \alpha H + j\omega a\alpha}\right]_{f_1}, \end{aligned} \tag{20}$$

where the subscript f_1 means that $f = f_1$ is to be substituted in $\omega = 2\pi f$ and in $H(f)$.

To get $G_{1;1}(f_1;f_2)$ we set $x_u = z_1$, $x_v = z_2$, and assume

$$\begin{aligned} y &= c_1z_1 + c_2z_2 + c_{12}z_1z_2 + \cdots, \\ I_u &= i_1z_1 + i_2z_2 + i_{12}z_1z_2 + \cdots, \\ Q &= q_1z_1 + q_2z_2 + q_{12}z_1z_2 + \cdots. \end{aligned} \tag{21}$$

When (21) is substituted in the circuit equations (18), the coefficients of z_1 give the equations (19) and therefore $c_1 = G_{1;0}(f_1;)$. Similarly, $c_2 = G_{0;1}(;f_2)$. The coefficients of z_1z_2 give a set of equations which, upon solving for c_{12} and using $q_1 = ac_1$, $q_2 = ac_2$, $i_2 = -c_2/\alpha$, $i_1 = \cdots$, give

$$c_{12} = 2c_1c_2\left[\frac{(\beta/\alpha)(H + j\omega a)_{f_1} - j(\omega_1 + \omega_2)a\alpha b}{(1 + \alpha H + j\omega a\alpha)_{f_1+f_2}}\right]. \tag{22}$$

Replacing c_1c_2 by $G_{1;0}(f_1;)G_{0;1}(;f_2)$ gives the required expression for $G_{1;1}(f_1;f_2) = c_{12}$.

To get $G_{2;0}(f_1,f_2;)$ we start with $x_u = z_1 + z_2$, $x_v = 0$ and again make the substitutions (21) in the circuit equations (18). The coefficients of

z_1z_2 give the same set of equations as before because x_u and x_v appear only linearly in the circuit equations. We have $c_1 = G_{1;0}(f_1;)$, $i_1 = c_1[H(f_1) + j\omega_1 a]$, $q_1 = ac_1$, and $q_2 = ac_2$ as before, but now $c_2 = G_{1;0}(f_2;)$, $i_2 = c_2[H(f_2) + j\omega_2 a]$, and $c_{12} = G_{2;0}(f_1, f_2;)$.

To sum up, we have

$$G_{1;1}(f_1; f_2) = 2G_{1;0}(f_1;)G_{0;1}(;f_2) \times [\text{expression in brackets (22)}], \qquad (23)$$

where $G_{1;0}$ and $G_{0;1}$ are given by (20). Expressions for $G_{2;0}(f_1, f_2;)$ and $G_{0;2}(;f_1, f_2)$ are obtained by replacing the product $G_{1;0}G_{0;1}$ in (23) by $G_{1;0}G_{1;0}$ and $G_{0;1}G_{0;1}$, respectively, and changing the bracket slightly.

To get $G_{1;2}(f_1; f_2, f_3)$ we set $x_u = z_1$, $x_v = z_2 + z_3$ and proceed along the lines used to get $G_{1;1}$, and so on.

As an example of the use of (23), suppose that $x_u = P_1 \cos \omega_1 t$ and $x_v = P_2 \cos \omega_2 t$. Then the $\exp [j(\omega_1 \pm \omega_2)t]$ component in y is, from (9),

$$\exp [j(\omega_1 \pm \omega_2)t]\left[\frac{P_1P_2}{4} G_{1;1}(f_1; \pm f_2) + \cdots\right]. \qquad (24)$$

Similarly, the leading terms in the series for the components of frequency $2f_1$ and $2f_2$ are given by $G_{2;0}(f_1, f_1;)$ and $G_{0;2}(;f_2, f_2)$, respectively.

Part II. Analysis for a Simple Frequency Converter

Here the circuit shown in Fig. 2 is used as an example to show how a multi-input system can sometimes be analyzed by the single-input formulas of Ref. 1. The output of interest is the voltage $y(t)$ across the nonlinear capacitor C. Thévenin's theorem is used to replace the circuit of Fig. 2 by that of Fig. 3, and a recurrence relation is derived for the corresponding G_n's. The results are used to get an expression for the third-order distortion when Fig. 2 is regarded as the circuit for an up-converter.

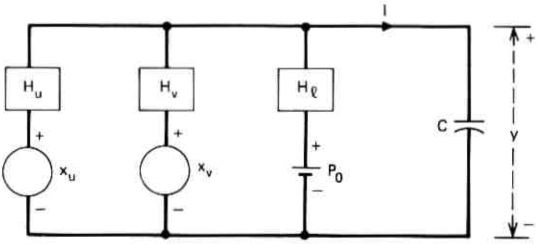

Fig. 2—Frequency converter with nonlinear capacitor C.

VI. REDUCTION OF A MULTIPLE-INPUT SYSTEM TO A SINGLE-INPUT SYSTEM

The system considered here and in the following sections is shown in Fig. 2. The admittances $H_u(f)$, $H_v(f)$, $H_l(f)$ are linear and C is the nonlinear capacitor used in Section V. The charge on C is $Q(t)$, the current $I = dQ/dt$ flows into C, the capacitance of C is $a + bQ$, and the voltage $y(t)$ across C is related to $Q(t)$ by

$$Q = (a + bQ)y. \tag{25}$$

P_0 is a biasing dc voltage.

The problem is to determine the components of $y(t)$ when

$$\begin{aligned} x_u(t) &= P_1 \cos \omega_1 t + P_2 \cos \omega_2 t, \\ x_v(t) &= P_p \cos \omega_p t, \end{aligned} \tag{26}$$

and ω_1, ω_2, and ω_p are incommensurable.

As far as $y(t)$ is concerned, the analysis of Fig. 2 can be reduced to that of the simpler circuit shown in Fig. 3. To accomplish this we apply Thévenin's theorem to the portion of Fig. 2 lying to the left of the terminals of C. As far as the exp $(j\omega_1 t)$ components of $y(t)$ and $I(t)$ are concerned, this portion of the system can be replaced by an admittance $H(f_1) = H_u(f_1) + H_v(f_1) + H_l(f_1)$ in series with the (open-circuit) voltage

$$\frac{P_1}{2} e^{j\omega_1 t} \left(\frac{H_u}{H_u + H_v + H_l} \right)_{f_1}.$$

Similar consideration of the remaining components shows that $I(t)$ and $y(t)$ can be computed from the circuit of Fig. 3 in which

$$\begin{aligned} H(f) &= H_u(f) + H_v(f) + H_l(f), \\ x(t) &= \rho_0 P_0 + \rho_1 P_1 \cos(\omega_1 t + \varphi_1) + \rho_2 P_2 \cos(\omega_2 t + \varphi_2) \\ &\qquad + \rho_p P_p \cos(\omega_p t + \varphi_p), \\ \rho_0 &= \frac{H_l(0)}{H(0)}, \qquad \rho_1 e^{j\varphi_1} = \frac{H_u(f_1)}{H(f_1)}, \\ \rho_2 e^{j\varphi_2} &= \frac{H_u(f_2)}{H(f_2)}, \qquad \rho_p e^{j\varphi_p} = \frac{H_v(f_p)}{H(f_p)}. \end{aligned} \tag{27}$$

The equation for $y(t)$, namely

$$\frac{d}{dt} \frac{ay}{1 - by} = \int_{-\infty}^{\infty} h(u)[x(t-u) - y(t-u)]du, \tag{28}$$

where $h(u)$ is the Fourier transform of $H(f)$, can be obtained by

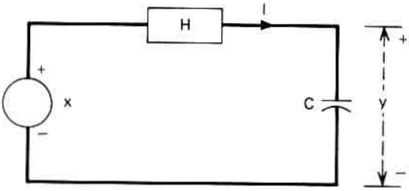

Fig. 3—Thévenin equivalent of Fig. 2.

equating two expressions for I, the one on the left being $I = dQ/dt$ in which $Q = (a + bQ)y = ay/(1 - by)$.

It is convenient to subtract out the dc components of $x(t)$ and $y(t)$ and apply the formulas of Ref. 1 to the portion $\hat{y}(t)$ of $y(t)$ which tends to 0 when $\hat{x}(t) \to 0$, $\hat{x}(t)$ being the time-varying component of $x(t)$. Therefore, in the system equation (28), we make the substitutions

$$\begin{aligned} x(t) &= x_0 + \hat{x}(t), \\ y(t) &= y_0 + \hat{y}(t), \end{aligned} \tag{29}$$

where x_0 is the dc value of $x(t)$ and y_0 is the (dc) value of $y(t)$ when $x(t) \equiv x_0$. From (27) $x_0 = \rho_0 P_0$, and substitution in (28) gives $0 = H(0)\ (x_0 - y_0)$. Assuming $H(0) \neq 0$ gives $y_0 = x_0 = \rho_0 P_0$. It should be noticed that the substitutions (29) are not strictly necessary because the dc component in $x(t)$ could be handled (at the cost of more work) by the analogue of (11).

Subtracting the result of substituting x_0 and y_0 in (28) from the result of substituting (29) in (28) and using

$$\frac{a(y_0 + \hat{y})}{1 - b(y_0 + \hat{y})} - \frac{ay_0}{1 - by_0} = \frac{\hat{a}\hat{y}}{1 - \hat{b}\hat{y}} \tag{30}$$

shows that the system to be analyzed by the single-input formulas of Ref. 1 is described by the equations

$$\frac{d}{dt}\frac{\hat{a}\hat{y}}{1 - \hat{b}\hat{y}} = \int_{-\infty}^{\infty} h(u)[\hat{x}(t - u) - \hat{y}(t - u)]du, \tag{31}$$

$$\begin{aligned} \hat{x}(t) = \rho_1 P_1 \cos(\omega_1 t + \varphi_1) + \rho_2 P_2 \cos(\omega_2 t + \varphi_2) \\ + \rho_p P_p \cos(\omega_p t + \varphi_p), \end{aligned} \tag{32}$$

where

$$\hat{a} = a/(1 - by_0)^2, \qquad \hat{b} = b/(1 - by_0). \tag{33}$$

Here $\hat{x}(t)$ and $\hat{y}(t)$ play the roles that $x(t)$ and $y(t)$ play in Ref. 1; and in the remainder of this paper the G_n's will refer to $\hat{x}(t)$ and $\hat{y}(t)$.

VII. CALCULATION OF THE G_n'S

The functions $G_1(f_1)$, $G_2(f_1,f_2),\cdots$ can be computed from (31) by the harmonic input method. A guide to the work is furnished by the resemblance of our problem to the one described by Fig. 3 of Ref. 1 and eqs. [1, (42), (43), (106)]. Expanding the left side of (31) as

$$\frac{d}{dt}\sum_{l=1}^{\infty} \hat{a}\hat{b}^{l-1}[\hat{y}(t)]^l \tag{34}$$

carries (31) into the form of [1, (106)] except for the operator d/dt. A procedure similar to the one used to deal with [1, (106)] gives

$$\begin{aligned} G_1(f_1) &= \left(\frac{H}{H+j\omega\hat{a}}\right)_{f_1}, \qquad K(f) = \frac{-2j\omega\hat{a}}{H(f)+j\omega\hat{a}}, \\ G_2(f_1, f_2) &= \hat{b}G_1(f_1)G_1(f_2)K(f_1+f_2), \\ G_3(f_1, f_2, f_3) &= \hat{b}^2G_1(f_1)G_1(f_2)G_1(f_3)K(f_1+f_2+f_3) \\ &\quad \cdot[K(f_1+f_2)+K(f_1+f_3)+K(f_2+f_3)+3], \end{aligned} \tag{35}$$

and the recurrence relation

$$G_n(f_1,\cdots, f_n) = \tfrac{1}{2}K(f_1+\cdots+f_n)\sum_{l=2}^{n} \hat{b}^{l-1}G_n^{(l)}(f_1,\cdots, f_n). \tag{36}$$

The $G_n^{(l)}$'s are the G_n's for the Volterra series for $[y(t)]^l$, and formulas for computing them are given in [1, (24) to (29)]. Equation (36) is a recurrence relation because $G_n^{(l)}$ can be expressed as the sum of products of G_1, G_2, $\cdots$, G_{n-1}. By starting with $G_1(f_1)$, the relation (36) can be used to compute G_2, G_3, $\cdots$ in succession. In the next section, (36) will be used to compute G_4.

VIII. COMPONENTS OF $y(t)$ OF FREQUENCY $f_1 + f_p$ AND $2f_1 - f_2 + f_p$

In this section we use (6) and the recurrence relation (36) to derive expressions for the $\exp[j(\omega_1+\omega_p)t]$ and $\exp[j(2\omega_1-\omega_2+\omega_p)t]$ components of $y(t)$ in Fig. 2 when (*i*) P_1, P_2, P_p are small, (*ii*) f_1 and f_2 are nearly equal, and (*iii*) H(f) is zero except for frequencies lying in narrow bands about the values

$$0,\ f_1,\ f_p,\ f_u, \tag{37}$$

where f_u denotes the upper sideband frequency $f_1 + f_p$.

The component of frequency $2f_1 - f_2 + f_p$ represents a typical third-order distortion product in an up-converter when f_1 and f_2 are signal frequencies, f_p the pump frequency, and $f_u = f_1 + f_p$, $f_2 + f_p$ the output frequencies.

Table I—Notation for Various Values of $H(f)$ and $K(f)$

Frequency, f	$H(f)$	$K(f)$
$0,\ f_1 - f_2$	H_0	0
$f_1,\ f_2,\ 2f_1 - f_2$	H_1	K_1
$f_p,\ f_1 - f_2 + f_p$	H_p	K_p
$f_1 + f_p(=f_u),\ 2f_1 - f_2 + f_p$	H_u	K_u
outside bands	0	-2

As mentioned in Section I, the procedure we use here can be regarded as a systemization of a method used by Gardiner and Ghobrial[3] to study the distortion performance of a varactor frequency converter. In our notation, the problem they solve is that of determining the $\exp[j(2\omega_1 - \omega_2 + \omega_p)t]$ component of the charge $Q(t)$ from the system equation

$$V(t) = aQ + bQ^2 + \int_{-\infty}^{\infty} k(u)I(t-u)du. \tag{38}$$

Here $V(t)$ is the sum of three sine waves [just as $\hat{x}$ is in (32)], $I = dQ/dt$, and $k(u)$ is the Fourier transform of the linear impedance $Z(f)$ in the Thévenin equivalent of the converter circuit. It can be shown from (38) that the G_n's corresponding to $Q(t)$ can be determined by recurrence from

$$G_1(f_1) = 1/(a + j\omega Z)_{f_1},$$
$$G_n(f_1,\cdots,f_n) = \left[\frac{-b}{a + j\omega Z}\right]_{f_1+\cdots+f_n} G_n^{(2)}(f_1,\cdots,f_n). \tag{39}$$

Now we return to our own problem. From the expression (32) for the input $\hat{x}(t)$ and the leading terms in the series (6) it follows that the $\exp[j(2\omega_1 - \omega_2 + \omega_p)t]$ and $\exp[j(\omega_1 + \omega_p)t]$ components of $y(t)$ are, respectively,

$$\exp\{j[(2\omega_1 - \omega_2 + \omega_p)t + 2\varphi_1 - \varphi_2 + \varphi_p]\} \cdot(\rho_1^2\rho_2\rho_pP_1^2P_2P_p/32)[G_4(f_1, f_1, -f_2, f_p) + \cdots], \tag{40}$$

$$\exp\{j[(\omega_1 + \omega_p)t + \varphi_1 + \varphi_p]\}(\rho_1\rho_pP_1P_p/4) \cdot[G_2(f_1, f_p) + (\rho_p^2P_p^2/8)G_4(f_1, f_p, f_p, -f_p) + \cdots]. \tag{41}$$

Only one G_4 term appears in (41) because we shall assume that ρ_1P_1/ρ_pP_p and ρ_2P_2/ρ_pP_p are small compared to one.

The function G_2 is given by (35), and the remaining problem is to compute G_4 from the formula obtained by setting $n = 4$ in the recurrence relation (36):

$$G_4(f_1, f_2, f_3, f_4) = \tfrac{1}{2}K(f_1 + f_2 + f_3 + f_4) \cdot \sum_{l=2}^{4} \hat{b}^{l-1}G_4^{(l)}(f_1, f_2, f_3, f_4). \quad (42)$$

As explained in Ref. 1 in connection with eqs. [1, (24) to (29)], we have

$$\frac{1}{4!}G_4^{(4)}(f_1, f_2, f_3, f_4) = (1)(2)(3)(4), \quad (43)$$

$$\frac{1}{3!}G_4^{(3)}(f_1, f_2, f_3, f_4) = (1)(2)(34) + (1)(3)(24) + (1)(4)(23) + (2)(3)(14) + (2)(4)(13) + (3)(4)(12), \quad (44)$$

$$\frac{1}{2!}G_4^{(2)}(f_1, f_2, f_3, f_4) = (1)(234) + (2)(134) + (3)(124) + (4)(123) + (12)(34) + (13)(24) + (14)(23), \quad (45)$$

where we have written "(2)," for example, for $G_1(f_2)$, "(34)" for $G_2(f_3, f_4)$, "(234)" for $G_3(f_2, f_3, f_4)$, and so on.

The next step is to compute the right-hand sides of (43), (44), (45) from the expressions (35) for G_1, G_2, G_3 when only frequencies in the bands indicated by (37) are allowed to flow. To aid in this, we introduce the notation shown in Table I for the values of the function $K(f) = -2j\omega\hat{a}/(H(f) + j\omega\hat{a})$ needed for the various G_2's and G_3's. We make the usual assumption that the admittance $H(f)$ remains constant in each band.

We consider first the $G_4(f_1, f_1, -f_2, f_p)$ in expression (40) for the $\exp[j(2\omega_1 - \omega_2 + \omega_p)t]$ component of $y(t)$. Equation (43) gives

$$G_4^{(4)}(f_1, f_1, -f_2, f_p) = 24G_1^2(f_1)G_1(-f_2)G_1(f_p), \quad (46)$$

where, from (35), $G_1(f) = [H/(H + j\omega\hat{a})]_f$. In eq. (44) for $G_4^{(3)}$ "(34)" now means $G_2(-f_2, f_p)$ and from the expression (35) for G_2 and Table I we get

$$(34) = \hat{b}G_1(-f_2)G_1(f_p)K(-f_2 + f_p) = \hat{b}G_1(-f_2)G_1(f_p)(-2).$$

Similarly, "(24)" means $G_2(f_1,f_p)$ and using the notation $K(f_1+f_p) = K_u$ gives

$$(24) = \hat{b}G_1(f_1)G_1(f_p)K_u,$$

and so on. Going through all six terms for $G_4^{(3)}$ in this way carries

(44) into

$$\frac{1}{3!} G_4^{(3)}(f_1, f_1, -f_2, f_p) = \hat{b} G_1^2(f_1) G_1(-f_2) G_1(f_p) \cdot [-2 + K_u + 0 + K_u + 0 - 2].$$

Going through all seven terms in $G_4^{(2)}$ carries (45) into

$$\frac{1}{2!} G_4^{(2)}(f_1, f_1, -f_2, f_p) = \hat{b}^3 G_1^2(f_1) G_1(-f_2) G_1(f_p) [K_p(K_u + 1) + K_p(K_u + 1) + (-2)(2K_u + 1) + K_1(1) + (-2)(-2) + (0)(K_u) + (K_u)(0)].$$

Substitution of the values of $G_4^{(l)}(f_1, f_2, -f_2, f_p)$, $l = 2, 3, 4$, in the expression (42) for G_4 and combining terms leads to

$$G_4(f_1, f_1, -f_2, f_p) = \hat{b}^3 G_1^2(f_1) G_1(-f_2) G_1(f_p) \cdot K_u[2(K_p + 1)(K_u + 1) + K_1]. \quad (47)$$

When this is put in (40) we get the approximation we have been seeking for the third-order distortion ($\exp [j(2\omega_1 - \omega_2 + \omega_p)t]$) component of $y(t)$.

The procedure used to obtain (47) can also be used to show that the G_4 in the expression (41) for the $\exp [j(\omega_1 + \omega_p)t]$ component of $y(t)$ has the value

$$G_4(f_1, f_p, f_p, -f_p) = \hat{b}^3 G_1(f_1) G_1^2(f_p) G_1(-f_p) \cdot K_u[2(K_1 + 1)(K_u + 1) + K_p]. \quad (48)$$

If we assume that the two series (40) and (41) converge at about the same rate, we can use (48) to get an idea of how large P_p can be before the leading term in (40) ceases to be a good approximation to the typical third-order distortion term. Thus, we expect the leading term in (40) to be a good approximation as long as the ratio

$$\tfrac{1}{8} | \rho_p P_p \hat{b} G_1(f_p) |^2 [2(K_1 + 1)(K_u + 1) + K_p] \quad (49)$$

of the first two terms in (41) is small compared to unity.

Note that setting $f_2 = f_1$ and then interchanging f_1 and f_p in the expression (47) for $G_4(f_1, f_1, -f_2, f_p)$ carries it into the expression (48) for $G_4(f_1, f_p, f_p, -f_p)$. This is to be expected since $G_4(f_1, f_2, f_3, f_4)$ is a symmetric function of f_1, f_2, f_3, f_4.

REFERENCES

1. Bedrosian, E., and Rice, S. O., "The Output Properties of Volterra Systems (Nonlinear Systems with Memory) Driven by Harmonic and Gaussian Inputs," Proc. IEEE, *59*, (December 1971), pp. 1688–1707.

2. Anderson, D. R., and Leon, B. J., "Nonlinear Distortion and Truncation Errors in Frequency Converters and Parametric Amplifiers," IEEE Trans. Circuit Theory, *CT-12*, (September 1965), pp. 314–321.
3. Gardiner, J. G., and Ghobrial, S. I., "Distortion Performance of the Abrupt-Junction Current-Pumped Varactor Frequency Converter," IEEE Trans. Microwave Theory Tech., *MTT-19*, (September 1971), pp. 741–749.
4. Swerdlow, R. B., unpublished work.
5. Kroupa, V., "Amplitude of the General Intermodulation Product," Proc. IEEE, *58*, (May 1970), pp. 851–852.
6. Sea, R. G., and Vacroux, A. G., "On the Computation of Intermodulation Products for a Power Series Nonlinearity," Proc. IEEE, *57*, (March 1969), pp. 337–338.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Picture Coding: The Use of a Viewer Model in Source Encoding*

By J. O. LIMB

(Manuscript received March 22, 1973)

A method is suggested for inserting viewer criteria directly into coding algorithms; any complex visual model may be used. The technique is applied to a DPCM-type coder, and a number of variations are compared on the basis of entropy, quality, and complexity. It is found that, using a simple one-dimensional filter model, the first-order entropy of the DPCM signal can be reduced by 30 percent for a high-detail picture with only a small reduction in picture quality. Furthermore, by means of a single threshold control, one can efficiently trade off bit-rate and picture quality over a large range for use in adaptive strategies.

I. INTRODUCTION

In early work in picture coding, Graham stressed the role of the viewer and Powers and Staras concluded that if large reductions in bit-rate are to be achieved they must come from "nonstatistical" (perceptual) redundancies.[1,2] However, there have been few attempts to explicitly incorporate the viewer in the encoder design. Unfortunately, there is no general method for handling complex viewer fidelity criteria, especially when one is concerned with how pleasing a picture appears.[†] Nevertheless *ad hoc* techniques have been proposed and evaluated and have achieved a certain measure of success.[3–8]

Source encoding, in its most general form, can be diagrammed as shown in Fig. 1. The first stage is an irreversible operation which generates a discrete signal as a result of a quite general multidimensional quantization process. The resulting discrete signal may still be redundant due to the presence of statistical dependencies; these are removed in the second stage of reversible processing in which a digital

[*] Presented, in part, at the 1972 IEEE International Symposium on Information Theory.

[†] See Ref. 9 for a discussion on viewer fidelity criteria.

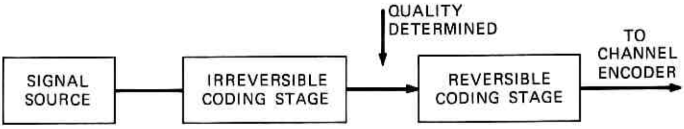

Fig. 1—General source-encoding model.

sequence is assigned to the output of the first stage. Thus, in the first stage the properties of the receiver together with the signal statistics are incorporated into the quantizing process so that the resulting signal just meets the required quality.

At the output of the first stage, picture quality is established and a discrete entropy can be measured. The actual transmission rate will then approach the entropy depending on how well the second encoding stage is designed to fit the statistics of the source.

1.1 *Receiver-Model Coding*

Algorithm: Components of a picture signal are estimated by some method. A test is made to see whether the estimate is adequate by testing the estimate on a model of the receiver. If so, the receiver is told (implicitly or explicitly) that the estimate is adequate. If not, a component is transmitted so as to meet the required criterion.

This type of algorithm will be referred to as "receiver-model" coding. Obviously, it is a rather general approach which can be appended to a larger number of existing algorithms; for example, the interpolators and predictors summarized by Kortman.[10] In this study we are interested in applying it to the differential quantizer (DPCM coder) although even here it can be applied in many ways.

In designing a coder to incorporate properties of the human observer the most important subjective effect is probably the large decrease in visual sensitivity that occurs adjacent to a change in luminance.[11–13] An attempt to design a coder based on this effect leads to some form of the familiar differential quantizer (DPCM coder).[3,7,8]

Probably the second most important subjective effect is the change in visual sensitivity with average luminance (Weber effect).[14] However, in the television situation nonlinearity between applied voltage and output luminance in most displays partially offsets this change in sensitivity, so that, roughly speaking, noise on an electrical television signal is nearly equally visible throughout the luminance range.[15,16]

Probably the third most important subjective effect is the spatial filtering of small-amplitude, luminance perturbations. It is this third

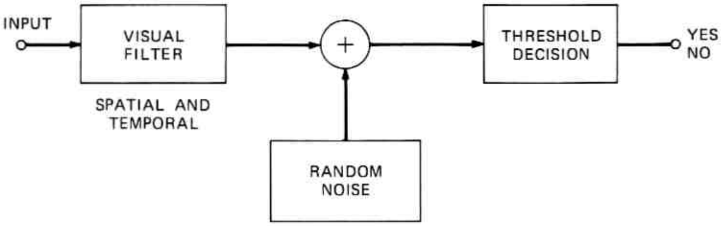

Fig. 2—Simple model of visual threshold filtering.

receiver property which we will attempt to capitalize on through the use of receiver-model coding in this paper.

A simple model that is reasonably successful in explaining the visibility of liminal stimuli is shown in Fig. 2.[13,17,18] Because we are dealing with very small perturbations (at least at the neural level) we will ignore nonlinearities. The input stimulus on which the model is developed is here a small luminance perturbation on a uniform background. The stimulus undergoes temporal and spatial filtering and in the process is corrupted by noise, represented as an additive random component. The filtered signal with the perturbation is compared with the filtered background signal. If the difference exceeds a certain threshold then the perturbation will be visible.* The model is quite accurate for variously shaped stimuli presented on a uniform background with the exception that if the stimulus is long (subtended angle >1 degree) and thin (subtended angle <5 minutes), it will be significantly more visible than the model predicts.[19,13]

The situation is more complex in the case of normal picture evaluation. First the perturbation is not presented against a uniform background and second the perturbation is not directly presented to the viewer; instead it is the difference between the coded picture and the viewer's memory of the original. Thus, although we will use this particular filter model it should be upgraded as we understand more about the visibility of perturbations in a complex scene.

In this study we will only be concerned with the spatial effect of the visual filter; different shapes have been postulated for the spatial impulse response and in one study the Gaussian function was found to fit as well as any.[13]† However, as we shall see, the performance of the algorithm is not sensitive to the exact shape that is used. The degree of spread, compared with the size of a picture element, is shown

* Because of the linearity assumption it does not matter if we filter the difference (error) signal or filter the two signals separately and then subtract.

† See also recent work of Ref. 20.

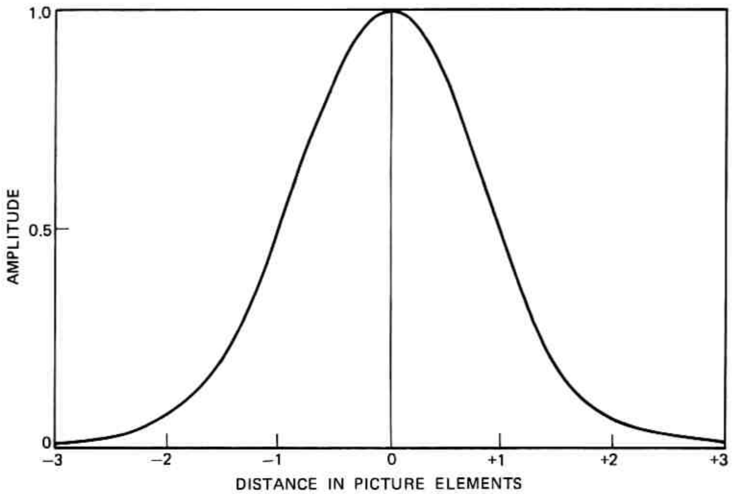

Fig. 3—Spatial impulse response of vision: visual point-spread function for a *Picturephone*®-type display reviewed at 36 inches.

in Fig. 3 for a *Picturephone*®-type display at a standard viewing distance of 36 inches. One should note that this filter is only appropriate to threshold vision; once a perturbation is much above threshold it may no longer be applicable.

Note that the efficacy of the filtering operation depends very much on viewing distance. Thus, one would expect that at smaller viewing distances the eliminated components would no longer be subliminal while at larger viewing distances the threshold filtering process could be taken further.

1.2 *Coding Algorithms*

Receiver-model coding will be applied to the differential quantizer by means of an interpolative algorithm.[10] Consider that sample i (Fig. 4) is the last nonzero sample that has been quantized and that sample $i + j$ is now being processed. The difference $X_{i+j} - \hat{X}_i$ is formed (where $\hat{X}_i$ is the differentially quantized value of X_i), it is quantized, and the discrete value of X_{i+j}, $\hat{X}_{i+j}$ is evaluated (i.e., normal differential quantizer operation). Interpolated values of the intermediate samples $\bar{X}_{i+1}, \cdots, \bar{X}_{i+j-1}$ are then formed from $\hat{X}_i$ and $\hat{X}_{i+j}$ and the error sequence associated with the interpolated values is calculated. This error sequence is than processed by the filter-threshold circuit to determine whether the errors are visually acceptable or not. If the error sequence associated with sample $i + j$

passes the test, the algorithm steps to sample $i + j + 1$ and no new value is transmitted. If the test fails, the run is terminated, that is, the quantized difference associated with sample $i + j - 1$ is transmitted.

There are two distinct forms which the coding algorithm may take; free-running or grid. In the free-running algorithm a maximum length-of-run is specified in advance for practical reasons. If the interpolation attempts to continue beyond the maximum length, a new sample is taken and a new run commenced. In most studies the maximum length-of-run is 10 pels. In the grid algorithm a fixed set of pels (grid elements) is always transmitted and interpolation or extrapolation is only applied to the intervening elements. Fixed patterns corresponding to every second or every fourth element along a line have been studied and the pattern is offset (staggered) from line to line. Grid algorithms are studied because in some forms they are very much simpler to implement.

Section II gives the experimental details and describes the basis for comparing different algorithms while Section III describes the operation and performance of a free-running interpolative algorithm and explores the effects of error filtering. In Section IV we describe and compare the operation of a number of different grid algorithms including one which involves but a minor modification to the normal differential quantizer.

II. EXPERIMENTAL ARRANGEMENT

The different algorithms were evaluated using a computer facility. The 8-bit digitized pictures are read from a digital disk, line at a time,

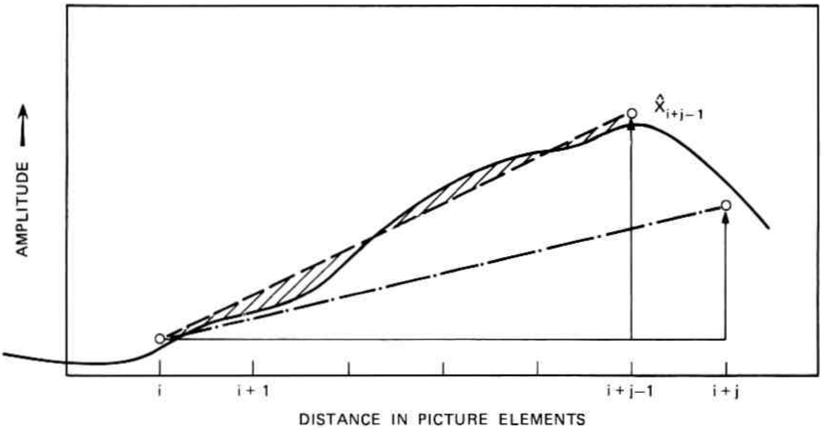

Fig. 4—In description of an extrapolative threshold coder.

processed, and then stored in a digital frame store for direct viewing on a television monitor. The picture consists of 250 lines with 210 elements in each line. The picture is generated and displayed as a 2:1 interlaced picture at 30 frames (60 fields) per second; hence adjacent lines in the picture originate in different fields. This format is similar to the *Picturephone* format.

In evaluating the picture we look at a single frame, repeated at 30 frames per second; thus temporal effects are not considered. The picture quality is slightly better when viewing a "frozen" frame of a differentially quantized picture since "edge busyness" and certain random noise components are noticeably less objectionable in the frozen situation contrary to the findings for white noise.[21]

2.1 *Differential Quantizer*

The normal differential quantizer is the vehicle with which the various algorithms will be tested. The 13-level companded quantizing characteristic is given in Table I. The differential quantizer has no integrator "leak" but the integrator is reset at the beginning of each line.

The results will be given mainly in terms of two different pictures. The first picture is the familiar "Karen" which by most measures would be regarded as active and is fairly difficult to code if both the soft hair and the sharp stripes are to be preserved. The second picture is much simpler having a large flat background and is referred to as

TABLE I—QUANTIZER CHARACTERISTIC OF 13-LEVEL DIFFERENTIAL QUANTIZER
(expressed in 1/128ths of the $p - p$ amplitude)

Level Number	Decision Level	Representative Level
0		0
	1	
±1		2
	3	
±2		4
	6	
±3		8
	11	
±4		14
	18	
±5		22
	27	
±6		32

"Lamp." A third picture ("Birdcage") is occasionally used; it is intermediate in complexity between the previous two.

The picture quality of the differentially quantized signal is only distinguishable from the 8-bit digital signal by careful comparison; there is a slight increase in background noise and very small amounts of slope overload and edge-busyness. The discrete, first-order entropies of the three pictures after coding by the differential quantizer are 3.10, 2.79, and 2.37 bits/pel for Karen, Birdcage, and Lamp, respectively. The second-order entropies are 2.92, 2.61, and 2.20 bits/pel, respectively. Thus, little would be gained in the second stage of coding, the reversible stage, by any attempt to remove higher-order redundancy.

2.2 *Quality*

One difficulty in documenting the performance of coders lies in specifying the quality of the processed pictures.

One can divide picture quality into different ranges by using a set of criteria. Consider the following three:

1. Difference just detectable by a skilled observer between the processed and unprocessed pictures in an A–B comparison with no restriction on viewing distance.
2. Defects just noticeable to a skilled observer at standard viewing distance (36 inches approximately $7H$) for a picture with which the observer is *familiar*.
3.[*] Defects just noticeable to a skilled observer, at standard viewing distance when the observer has *no knowledge* of the original picture.

The picture quality of criterion 1 is probably the most frequently used *ad hoc* criterion but it is unnecessarily severe for visual communication purposes and, if employed, would result in a significant increase in bit-rate over that required by criteria 2 and 3. In this study the author has attempted to specify the qualities of coded pictures using criteria 2 and 3. This is inevitably an approximate process and as a consequence a range is given rather then a specific value. Approximate as this process is, if it enables a rank ordering of coding strategies it will have served its purpose.

[*] Where the viewer was familiar with the test picture a conscious effort was made to disregard defects that depend on knowledge of the original picture. For example, noticing a loss of fine detail in the hair region of "Karen" depends on memory of the original; noticeable slope overload on the other hand generally appears as an unnatural distortion.

2.3 *Bit-Rate Calculation*

Picture quality and bit-rate are the two vital measures of coder performance. In this study we are concerned primarily with the first coding stage of Fig. 1, the multidimensional quantizing stage. But the final bit-rate will also depend on how thoroughly the second stage is implemented. However, what we will do is to calculate entropies of the signal after the first stage of coding, the rationale being that the figure represents a bound on what is obtainable in practice. In some instances variable wordlength coding, with buffering, will yield a data rate that is within a few percent of the entropy figure.[22,23] In other instances more complex coding will be required to approach the entropy figures, particularly for source alphabets which contain a highly probable event where something akin to runlength coding would be required.

The performance of the algorithms has been assessed by calculating the entropy under the assumption of two different types of reversible encoding. They are:[*]

Code I. All pels in the run are processed in the same way (with the same code). This is the simplest but most inefficient method. The bit-rate bound is obtained by calculating the first-order entropy of the signal;

$$H_1 = -\sum_{i=1}^{N} p_i \log p_i,$$

where p_i is the probability of occurrence of each event (a quantizer level or an interpolate command) and N is the total number of different types of event.

Code II. A separate code is used for each run position. That is, the first element in a run uses code 1, the second element in the run uses code 2, etc. The run is terminated by the sampled pel. The entropy is then given by

$$H_2 = \sum_{j=1}^{M} h_j q_j,$$

where h_j is the entropy of events in the jth position of the run and q_j is the probability that an event will be in the jth position of a run.

[*] In some of the algorithms to be discussed the information indicating that an element has been successfully estimated at the transmitter is obtained indirectly by the receiver from the coded bit stream. In other algorithms an additional code word is appended for this purpose.

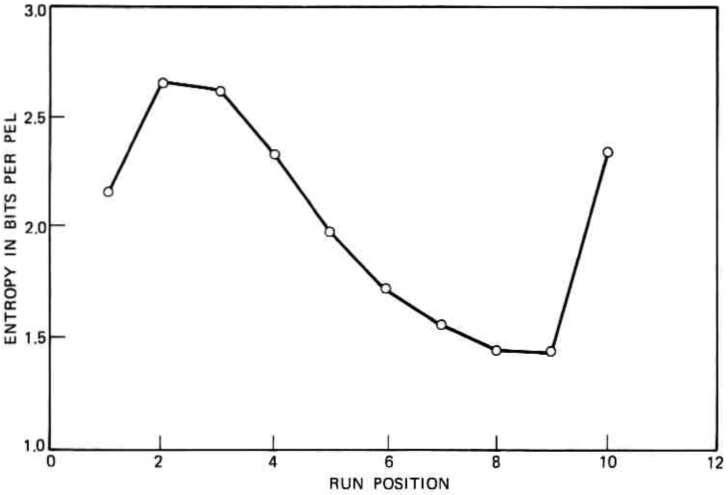

Fig. 5—Variation of entropy with the position of the element in the run; free-running interpolative algorithm with a maximum runlength of 10. Subject—Karen.

The entropy of the signal changes significantly depending on the position in the run. (This is shown in Fig. 5 where the first-order entropy of the differentially quantized signal is plotted as a function of the position in the run for a free-running interpolative algorithm having a maximum runlength of 10.) This change in entropy is exploited in code II (but not code I). Where the average length of a run is large, a practical realization of a code II coder could well result in a type of runlength encoding.

In summary, H_1 can be regarded as the lower bound on data rate when each element is coded in the same way while H_2 is a lower bound when run contiguity is exploited.

There are a great number of different techniques for reversibly coding the discrete output of the first coding stage (Fig. 1); by specifying the abovementioned two entropies we can concentrate more on the irreversible stage without getting overly involved in exactly how the second-stage coding will be achieved. The entropies are always given as bits/active (or unblanked) picture element.

III. RESULTS: FREE–RUNNING ALGORITHM

The details of the interpolative algorithm are summarized in the flow diagram of Fig. 6. Bookkeeping operations like entering a new line, testing for the end of a line, and gathering statistics are not

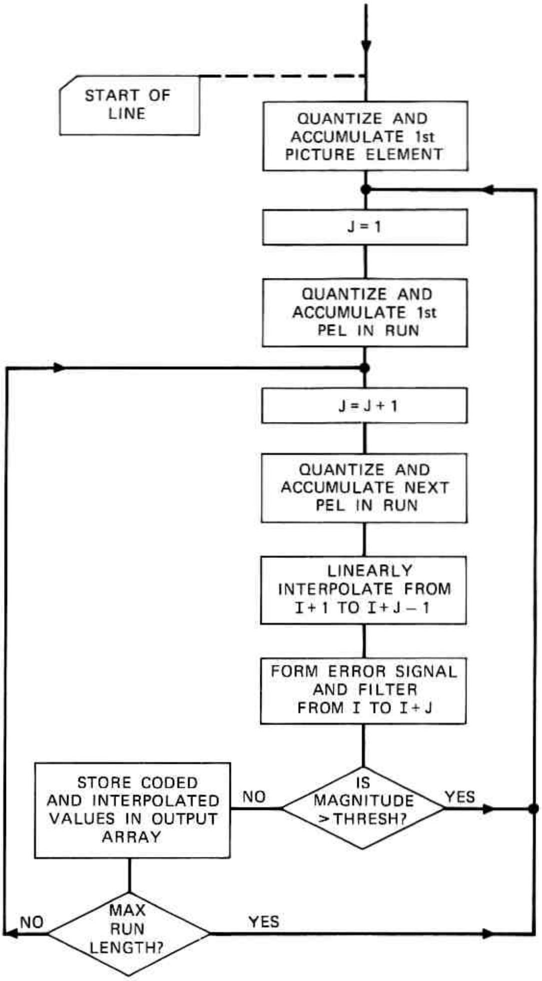
START OF LINE
QUANTIZE AND ACCUMULATE 1st PICTURE ELEMENT
J = 1
QUANTIZE AND ACCUMULATE 1st PEL IN RUN
J = J + 1
QUANTIZE AND ACCUMULATE NEXT PEL IN RUN
LINEARLY INTERPOLATE FROM I + 1 TO I + J − 1
FORM ERROR SIGNAL AND FILTER FROM I TO I + J
IS MAGNITUDE > THRESH?
YES
NO
STORE CODED AND INTERPOLATED VALUES IN OUTPUT ARRAY
MAX RUN LENGTH?
NO
YES

Fig. 6—Flow diagram for the element processing of the free-running interpolative algorithm. I denotes the last element in the previous run, J denotes the current length of the run being processed, and $I + J$ denotes the element being currently processed.

shown. We will first discuss (Section 3.1) the efficiencies obtained with the two methods of reversibly coding the discrete output. Neither the shape of the filter function nor the maximum length which the algorithms can run before a new run is forcibly commenced is varied in the above comparisons. The effect of varying these two parameters is described in Sections 3.2 and 3.3. Some observations are made on free-running algorithms in Section 3.4.

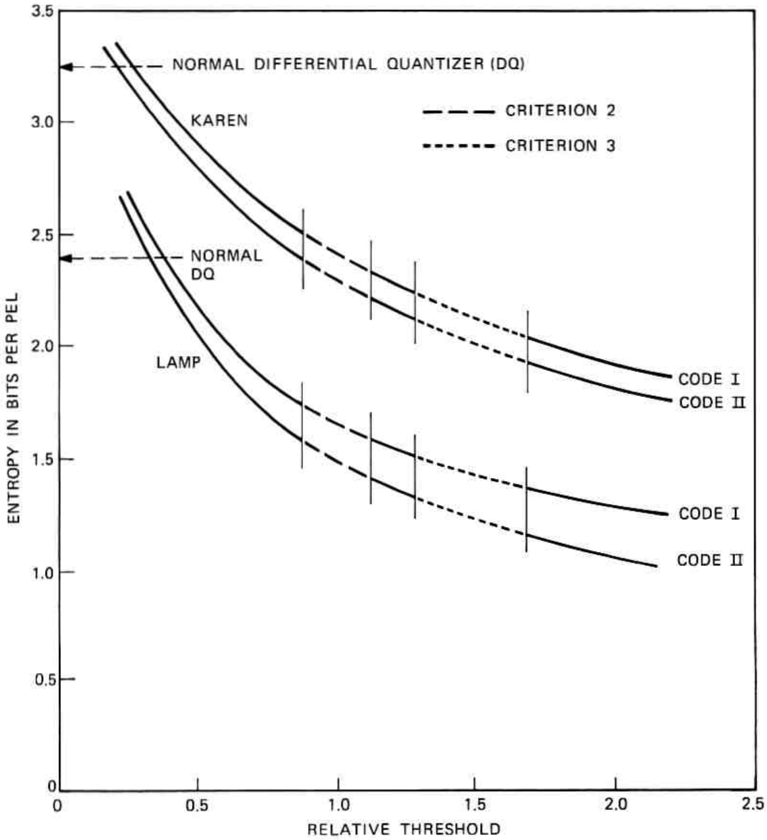

Fig. 7—Entropy of the free-running algorithm as a function of the threshold The measurements of entropy are made under the assumption of two different types of reversible code. The performance for the codes is very similar.

3.1 *Comparison of Reversible Coding Methods*

Figure 7 summarizes the results obtained by applying receiver-model coding interpolatively to the 13-level differential quantizer.* For computational simplicity, the filter used in this case has a rectangular impulse response three elements wide (i.e., corresponding to an average over three elements).

As the threshold is raised on the filtered error sequence, more and more elements are interpolated. Consequently probability distributions become more peaked and the entropy drops. At the same time

* The "relative" threshold is, in fact, one-fifth the threshold value, in 128ths, applied to the filtered error signal.

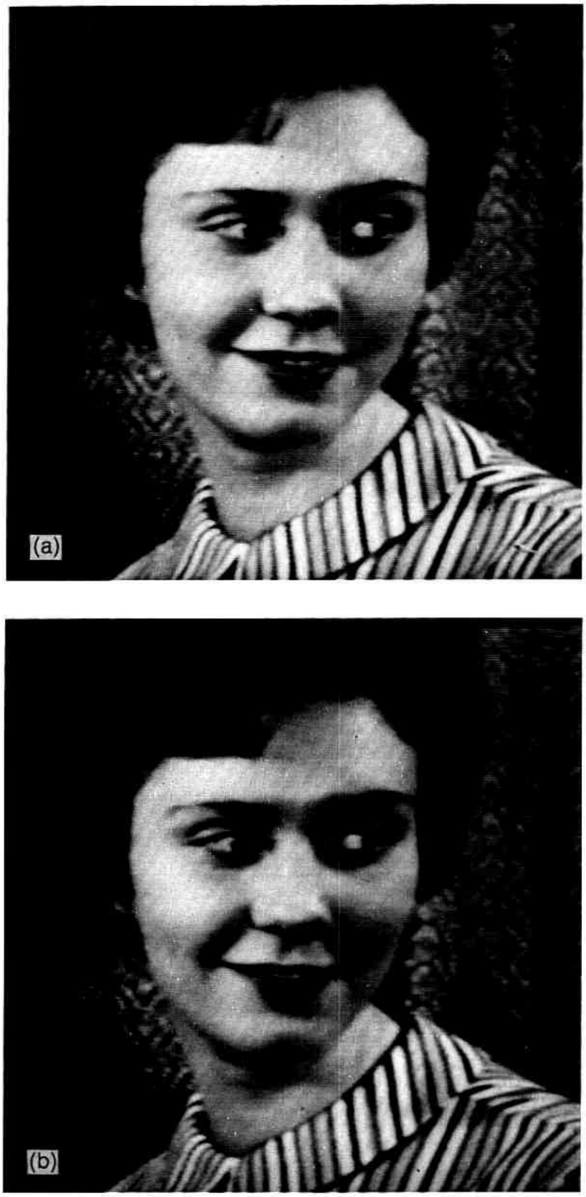

Fig. 8—(a) Karen—processed by normal 13-level differential quantizer, 1st order entropy 3.10 bits/pel. (b) Picture processed by free-running algorithm, 2.0 bits/pel; picture quality is criterion 3 or worse. (c) Unprocessed picture of "Lamp."

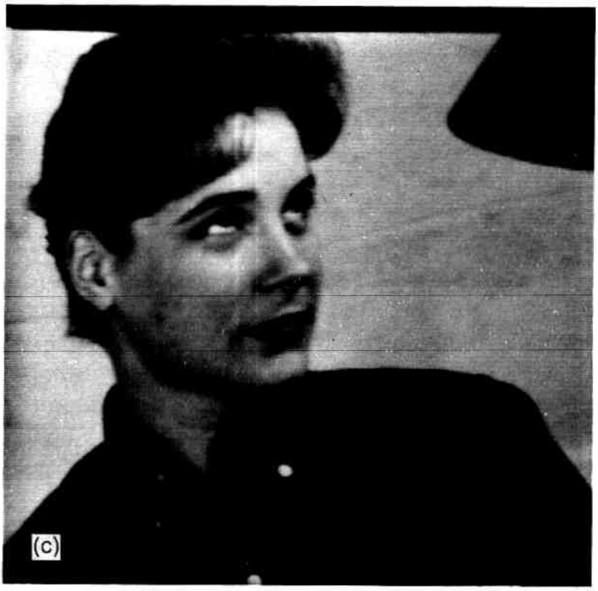

Fig. 8 (continued).

the picture quality is reduced in low-detail areas of the picture as soft detail and texture become blurred. Edges and high-detail areas, however, remain unaffected until very large thresholds are reached.

Consider the results for Karen. As the threshold is increased, the entropy drops from 3.1 bits/pel with a normal differential quantizer* to about 2 bits/pel at which point there is quite noticeable smearing in low-detail areas. Also shown on the curves are the criterion 2 and criterion 3 ranges (Section 2.2). Not until the threshold is raised to a value of 0.9 and the entropy has fallen to 2.4 bits/pel does the change in picture quality become visible when compared with a normal differential quantizer, other than by close A–B comparison. The normal differentially quantized picture is shown in Fig. 8a while the picture coded with a threshold of 1.5 (2.0 bits/pel) is shown in Fig. 8b.

The results obtained with the simpler picture "Lamp" (Fig. 8c) are similar to those obtained for Karen except that the advantage is somewhat greater; the rate is halved in going from the normal dif-

* It is necessary to send additional information to explicitly inform the receiver when to interpolate and when not to. It is this additional information which prevents the entropy of the coded signal from converging to the value of the normal differential quantizer.

ferential quantizer to the end of the criterion 3 range. It is to be expected (see Section V) that low-detail pictures will be more amenable to receiver-model coding given the present model.

There is surprisingly little difference in efficiency between the two reversible codes, particularly for Karen where the statistics for the highly detailed parts swamp the peaked distributions obtained in the low-detailed parts. In such instances an adaptive strategy would be of some help.[4] The complexity associated with implementing the simple code (code I) does not change with the maximum permitted length of run; for the variable code (code II) there is a proportional relationship since a code dictionary would need to be stored for each run position. Consequently, it is important to know how the entropy changes with the maximum length of run that is permitted. For the moment we may conclude that unless the more complex codes can be implemented simply or that channel capacity is at a premium then the simple code is probably adequate.

3.2 *Visual Filter Function*

The psychological literature is replete with different estimates of what the shape of the visual point-spread function should be. It was hoped that we could add something to the debate by investigating different functions in the coding model to see which shape gives the best results. In one experiment the shape of the function was varied keeping the spread of the function constant; the spread was measured by the first moment of the absolute value of the spread function. In a second experiment the spread of the filter was varied keeping the shape constant. Bear in mind that because our algorithm works only along the scan-line we cannot take full advantage of the two-dimensional, spatial, point-spread function. Consequently, we should really think of a line-spread function, the rationale being that in the worst-case situation the stimulus being filtered would have large vertical extent and hence the line-spread function would be appropriate.

3.2.1 *Effect of Shape*

Varying the shape of the filter function has little effect on coding efficiency (Fig. 9). The filter shape was varied from rectangular to quite peaked keeping both the area under the function and the first moment of the absolute value of the function constant. The threshold is also constant at 1.0. The square, crosses, and dots of Fig. 9 denote functions with widths of 3, 5, and 7 pels respectively. The values of the functions are given in Table II. For the interpolative algorithm

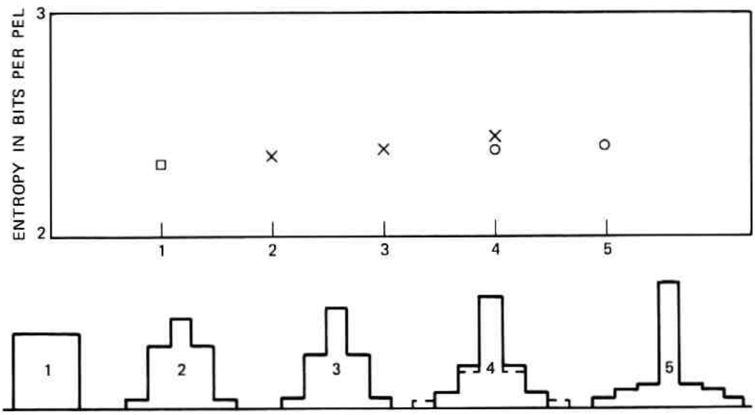

Fig. 9—Effect on entropy (code II) of varying the shape of the filter function. The width of the impulse response is: □—3 elements, ×—5 elements, ○—7 elements. Threshold decisions are very insensitive to the shape of the filter function.

with a maximum runlength of 10 pels there is an increase in bit-rate from 2.32 bits/pel for the rectangular function to 2.41 for the most peaked function; any accompanying change in picture quality was too small to notice.

3.2.2 *Effect of Spread*

The spread of the filter function, on the other hand, has far more effect on the picture quality and entropy than does the shape, as can be seen from Fig. 10a. A rectangular function was used and the spread was varied keeping the area under the impulse response constant and the threshold fixed at 1.0. The picture quality changed from almost

Table II—Weighting Coefficients of Transversal Filter (The filter shape is symmetrical with A being the central element)

Filter Number	A	B	C	D
1	0.333	0.333	0	0
2	0.4	0.275	0.025	0
3	0.45	0.231	0.044	0
4a	0.5	0.188	0.062	0
4b	0.5	0.156	0.062	0.031
5	0.55	0.103	0.081	0.041

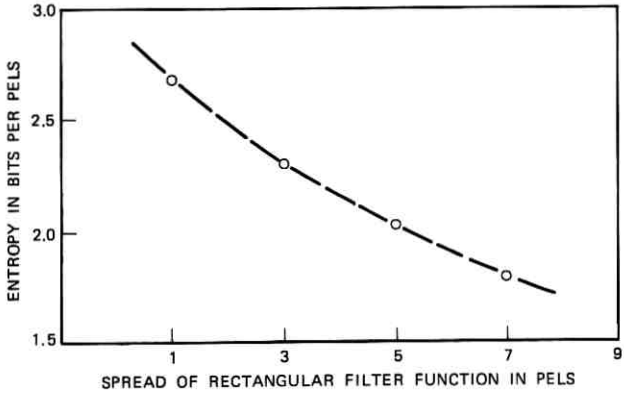

Fig. 10a—The effect on entropy (code II) of varying the width of the filter function. The overall spread of the function has a strong effect on entropy. Subject—Karen.

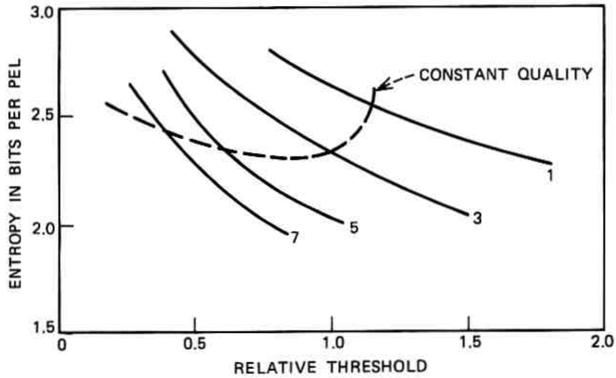

Fig. 10b—Curves of entropy versus threshold for filter functions having spreads of 1, 3, 5, and 7 elements for the free-running interpolative algorithm (Karen). The dashed curve passes through each of the full curves at points of approximately constant picture quality. A spread of between 3 and 5 elements gives the lowest bit-rate for standard viewing distance.

criterion 1 quality with a spread of 1 pel to worse than criterion 3 quality when the spread was 7 pels.

An attempt was made to determine the most suitable filter spread for a picture having criterion 2 quality (standard viewing distance). Figure 10b gives curves of entropy versus threshold for rectangular filter functions of different spread. The dashed curve is a line of approximately constant picture quality. It was determined by making pair-wise comparisons between a reference picture obtained using a threshold of 1.0 and a filter spread of three and pictures from the other spread curves. With the filter fixed at a particular value of spread the threshold was varied until the picture quality matched that of the reference picture. From the figure it can be seen that a spread of between 3 and 5 pels gives the lowest bit-rate for the standard viewing distance.

3.3 *Effect of Maximum Runlength*

The effect of changing the maximum permitted runlength is shown in Fig. 11. Interestingly, there is very little increase in bit-rate as the maximum runlength is reduced to as little as 4 pels, particularly for code II. Even for the low-detail picture (Lamp) where the average length of a run is much longer, the increase in entropy is still small. Bearing in mind that code II becomes much simpler to implement for short maximum runlengths there appears to be little reason to use long runlengths.

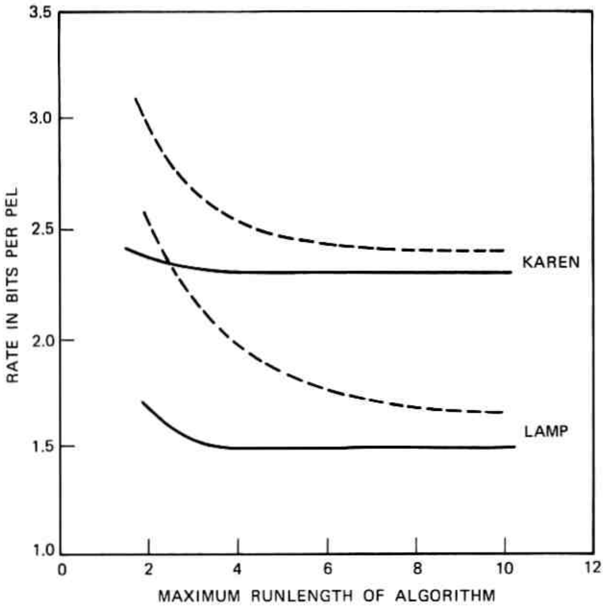

Fig. 11—Entropy as a function of the maximum runlength for Code I (dashed) and Code II (full). Note there is little increase in entropy for Code II as the maximum runlength is reduced from 10 to 4.

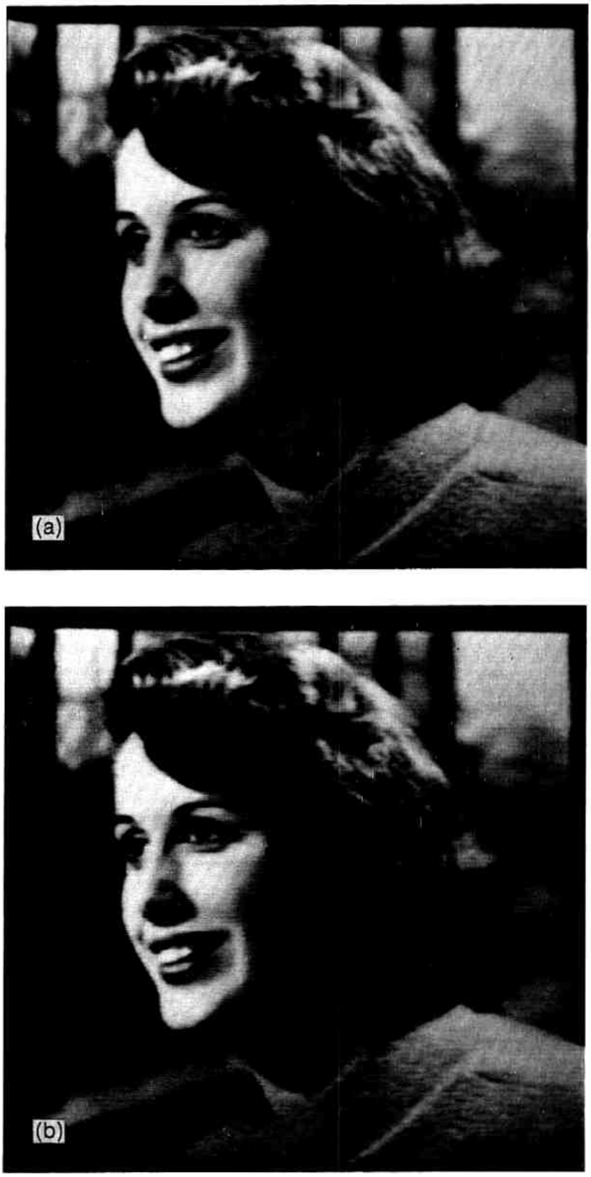

Fig. 12—Pictures showing the effect of changing the size of the picture element with the filter function, as measured at the eye, maintained constant: (a) original 8-bit signal, (b) processed, with threshold = 1.5 and H = 1.38 bits/pel, (c) original 8-bit signal, ½ lineal size, (d) processed with same threshold as in (b), H = 1.17 bits/pel. It is the quality difference between pictures of the same size that should be compared, not the relative quality of the two processed pictures.

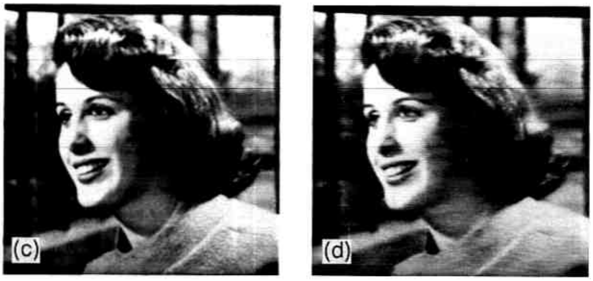

Fig. 12 (continued).

3.4 *Discussion*

The preceding experiments suggest two ways for decreasing bit-rate at the cost of decreased picture quality. First, it can be decreased by increasing the threshold as shown by Fig. 7. Second, it can be decreased by increasing the spread of the filter function as shown by Fig. 10a. The picture, Karen, was coded to have an entropy (Code II) of 1.81 bits/pel by reducing the quality (lower quality than criterion 3) in the two ways described above. For the first method the filter was rectangular with a spread of three elements while for the second method the filter was again rectangular but with a spread of seven elements. Both methods gave similar picture quality with the narrow-filter/high-threshold combination of the first method being, perhaps, slightly better. The improvement in sharpness of the first method was partly offset by the reduction in granularity and blotchyness of the second method.

If a particular filter, at normal viewing distance, produces a picture that is just distinguishable from a high-quality original then doubling the spread of the filter function should produce a picture at twice the viewing distance which is again just distinguishable from the original.

I have tried to demonstrate this prediction with Fig. 12 by reproducing a comparison pair of pictures at half-size to correspond to the situation where the viewing distance is doubled. It is the *difference* in quality between pairs of pictures at the same viewing distance that should be compared, not the comparative quality of the processed pictures.

One factor that could upset such a comparison is that the smaller picture has a greater scanning line density. The filter function operates in one dimension only and to the extent that deleted picture components are uncorrelated from line to line, vertical filtering taking place in the eye will tend to favor the smaller picture. An intuitive feel

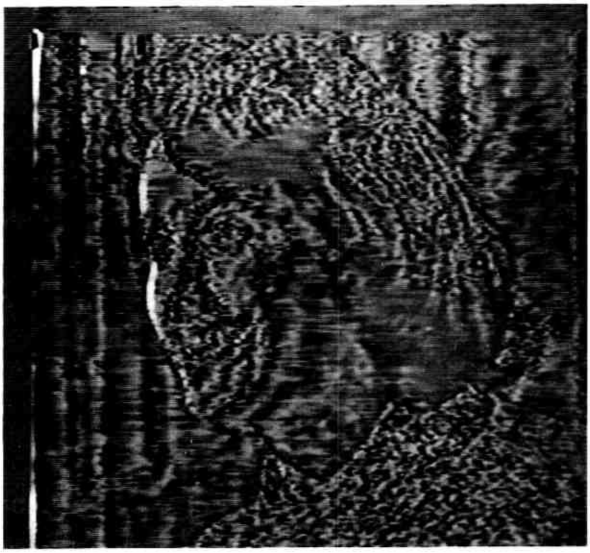

Fig. 13—Picture of the filtered error signal for the processed picture of Fig. 12b.

for the correlated nature of the error signal is obtained from Fig. 13, in which a certain amount of picture structure is evident.

IV. RECEIVER-MODEL CODING WITH GRID ALGORITHMS

4.1 *Introduction*

One can take advantage of the filtering action of vision without explicitly filtering the error signal. To appreciate this, let us consider the following grid algorithm. Every grid element (sampled point) is reproduced with full accuracy (e.g., 7 or 8 bits). The intermediate elements (referred to as "conditional points") are reproduced as the average of the adjacent pels, $\bar{X}_{i+1} = (X_i + X_{i+2})/2$, if the error $(\bar{X}_{i+1} - X_{i+1})$ is small (see Fig. 14). Otherwise, the error quantity is quantized and transmitted. In determining whether $\bar{X}_{i+1}$ is an adequate representation of X_{i+1}, the error signal adjacent to pel $(i + 1)$ must be filtered. However, the error at pels i and $(i + 2)$ is virtually zero so that for a filter that consists of a three-point average it is only necessary to examine the error introduced at pel $(i + 1)$.

Kretzmer[24] proposed a coding scheme similar to the above in which every fourth pel is always coded with 7-bit accuracy (i.e., 4:1 grid algorithm). The intermediate points are estimated by linear interpolation and the difference between the input and the estimate is quantized and transmitted. The midpoint in each quad is quantized more accurately than the quarter and three-quarter points. Fukushima and Ando[25] experimented with a very similar scheme in which every fourth point was transmitted with 6-bit accuracy and the intermediate points were transmitted using three levels. A final bit-rate of 2.7 bits/pel was achieved. They also investigated two-dimensional 4:1 algorithms. Connor has investigated a 2:1 grid algorithm (column coder) which uses two-dimensional prediction for differentially coding the grid points.[26] Pease[27] has applied what amounts to a 2:1 grid algorithm between fields of a television picture. All points in one field are estimated as the average of the four surrounding points coming from the previous and next fields. Only when this prediction breaks down is additional information sent about the interpolated field. In the presence of movement the four-way interpolation is less accurate and the number of pels that require correction increases somewhat. Notice that all the above schemes transmit two or more different types of amplitude information; the grid points are transmitted absolutely (or differentially, relative to one another) while the conditional points are transmitted as a *correction* to the estimation. These schemes will therefore be referred to as error transmission schemes.

In this section we will examine a number of grid coding schemes. For the most part they differ from the above schemes in that only one type of amplitude signal is transmitted so that all amplitude information is decoded in the same way (direct transmission). The distinction is best appreciated by considering a specific example. Take the inter-

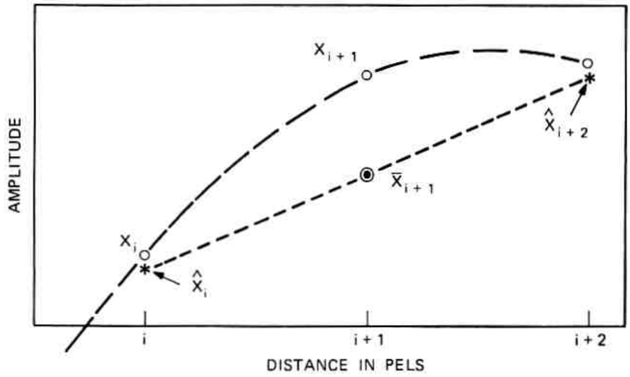

Fig. 14—Definition of locations and values of elements used in discussion of grid algorithms.

polative algorithm: if pel i has already been encoded (Fig. 14), pel $(i + 2)$ is then encoded differentially from pel i. From the encoded values of pel i and pel $(i + 2)$, $(\hat{X}_i, \hat{X}_{i+2})$, $\bar{X}_{i+1}$ is formed. The error signal $(X_{i+1} - \bar{X}_{i+1})$ is tested against the threshold; if it exceeds threshold, pel $(i + 1)$ is differentially coded from pel i and pel $(i + 2)$ is differentially recoded from pel $(i + 1)$. Thus it can be seen that the interpolated value $\bar{X}_{i+1}$ is only retained when the interpolation is adequate; otherwise it is discarded. Furthermore, the quantizing scales for pels i and $(i + 1)$ can be the same as for normal differential quantization since in high-detail areas the interpolation generally fails and each element is predicted from the previous element. In practice, a check is made to determine whether slope overload will occur in coding pel $(i + 2)$; if this can happen pel $(i + 1)$ is then coded and pel $(i + 2)$ is recoded, differentially, from pel $(i + 1)$. Thus, in high-detail parts of the picture, pel $(i + 1)$ is rarely interpolated and the coding operation differs little from normal differential quantization. In low-detail parts of the picture, where the interpolation process is usually adequate, again the coding process is normal differential quantization, but with twice the normal sample spacing.[4]

Errors will occur at pels i and $(i + 2)$ because differential quantization has been used and these errors will, because of the visual filtering action, affect the visibility or the error occurring at pel $(i + 1)$. Hence the encoding will be more efficient if filtering is used. But, as we will see, a three-point filter does not differ much from a single-point filter because the errors made at pels i and $(i + 2)$ are limited by the number and spacing of the quantizer levels and cannot be subjectively large if adequate quality is to be obtained.

In comparing the error transmission and direct transmission schemes, it can be seen that the decision on whether or not to transmit the conditional elements is the same in both cases. The error transmission scheme has the advantage that the estimate is a better prediction than the previous sample, and hence the correction signal, where it is necessary to transmit it, will be smaller. However, the disadvantage is that since the grid points are transmitted as differences from a point two pels away, the amplitude of the differences and hence the entropy associated with them will be larger. In practice this will increase complexity since the quantizer will need to have more levels to handle the larger changes. In Section 4.2 an error transmission scheme will be compared with a number of direct transmission algorithms and it will be seen that there is very little difference in performance between the two types of schemes. One would expect the performance to converge for low-detail pictures since the number of points which are not successfully interpolated becomes very small and the encoding of the remaining points is then very similar.

In the free-running algorithms a special code word was used to inform the receiver when to interpolate. For the grid algorithms an interpolate command has been inserted in a special manner. On the conditional samples only, the zero differential quantizer level is used to denote the interpolate command: this means that when the signal is not being interpolated the zero level cannot be used; instead the signal is forced to take on the next closest level, either the positive or negative inner level. This affects picture quality very little since, firstly, a zero level is rarely used on the conditional samples and, secondly, since interpolation generally fails in the vicinity of large luminance changes, the small error introduced by deleting the zero level is largely masked by the consequent luminance change.

Implementation of the grid algorithm becomes even simpler when we consider two variations, a modified form of the interpolative (MI) algorithm and an extrapolative algorithm. The MI algorithm is quite similar to the interpolative algorithm; the next grid point is *not* quantized prior to interpolation. This means that it is only necessary to quantize each element sequentially just as one does in normal differential quantization (when a pel is adequately interpolated, the classifier output is simply forced to a zero prior to processing by the local [and distant] decoder and the next element [pel $i + 2$] is processed in the normal manner [see Fig. 14]). In the extrapolative algorithm the method used to estimate the conditional sample is the same as the method of extrapolation for the coding process (i.e., previous sample prediction) and hence the need for an extrapolate command is obviated. The algorithm is then only slightly different from normal quantization, especially if the error occurring at the conditional sample is taken as the filtered value (the scheme described in Ref. 4 under the name "Level Variable Sampling Scheme").

4.2 *Comparison of Free-Running and Grid Algorithms*

The performance of both a 2:1 and a 4:1 MI, grid algorithm are compared with the free-running extrapolative algorithm in Fig. 15.

The maximum reduction that can be obtained with the 2:1 algorithm is a halving of the bit rate. Long before this point is reached the curve starts to flatten out and unless very large thresholds are used the picture quality remains high. Within the obtainable range of picture quality the 2:1 algorithm performs almost as well as the free-running

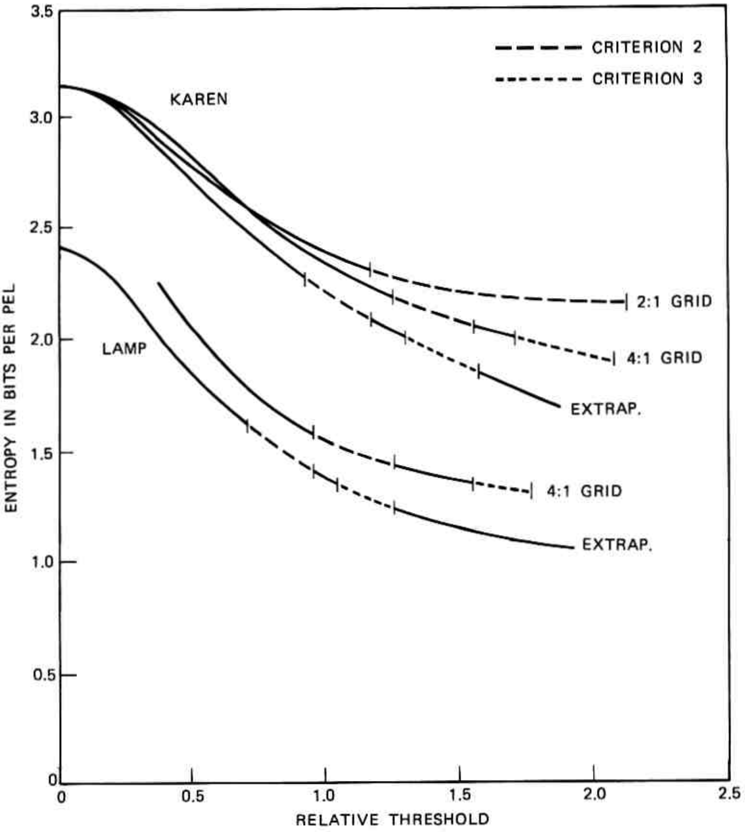

Fig. 15—Comparison of the performance of free-running and grid algorithms. The 4:1 grid algorithm performs equally as well as the free-running algorithm.

algorithm. By going to the 4:1 algorithm, a larger picture quality range can be accommodated without going to very large thresholds. In the criterion 2 range the grid algorithm seems slightly better than the extrapolative free-running algorithm while in the criterion 3 range the free-running algorithm is marginally better.

4.3 *Comparison of Three Grid Algorithms—Error-transmission, MI, and Extrapolative*

Since the MI and error-transmission algorithms are the most alike, we will compare them first. The error-transmission algorithm uses a 19-level differential quantizer. This is obtained from the 13-level quantizer by adding additional outer levels. The filtered error signal is obtained by summing the error at the estimation point and the two adjacent grid points. The MI algorithm uses the usual 13-level quantizer and the filtered error signal is the sum of only two error terms. The quantizing error occurring at the grid point to the right of the point being interpolated cannot be included since this point is not quantized until after a decision has been made on the conditional point.

The white markers in Fig. 16 indicate those conditional points in the two algorithms for which the filtered error signal is above threshold. Hence these points are not adequently represented by the estimate (the relative threshold is set at 1.5 for both algorithms). The distribution of markers is quite similar, especially when one bears in mind that the error summing procedure is different in the two cases. The picture quality and bit-rate is also very similar (see Fig. 17), which stands to reason since the signal is processed identically in those parts of the picture where there are no markers. The algorithms were evaluated on other pictures. In each case picture quality and bit-rate were very close.

The extrapolative (like the MI) algorithm uses a 13-level quantizer and sums the error over only two pels. The estimation procedure (zero-order-hold) is not as effective as linear interpolation and, as a result, the number of conditional points that need to be transmitted is very much larger for a specific threshold. A consequence is that the curve of entropy versus threshold lies above the other curves except at higher thresholds. Here, the curves converge since the only conditional points still being transmitted are edge points. The picture quality is not quite as high as that obtained with the other two algorithms with the defect appearing as a granularity in flat, dark regions of the picture. Although the granularity is also present for the other two algorithms it is significantly attenuated by the interpolative averaging.

4.4 *Effect of Filtering*

As indicated previously, the effect of filtering for the 2:1 grid algorithm will not be very strong since when the error is evaluated at each conditional point the error permitted at the adjacent points, which are quantized with full accuracy, will be quite small. Even so, there is a small increase in the number of conditional flags that are transmitted in going from the single-point filtering to the two-point

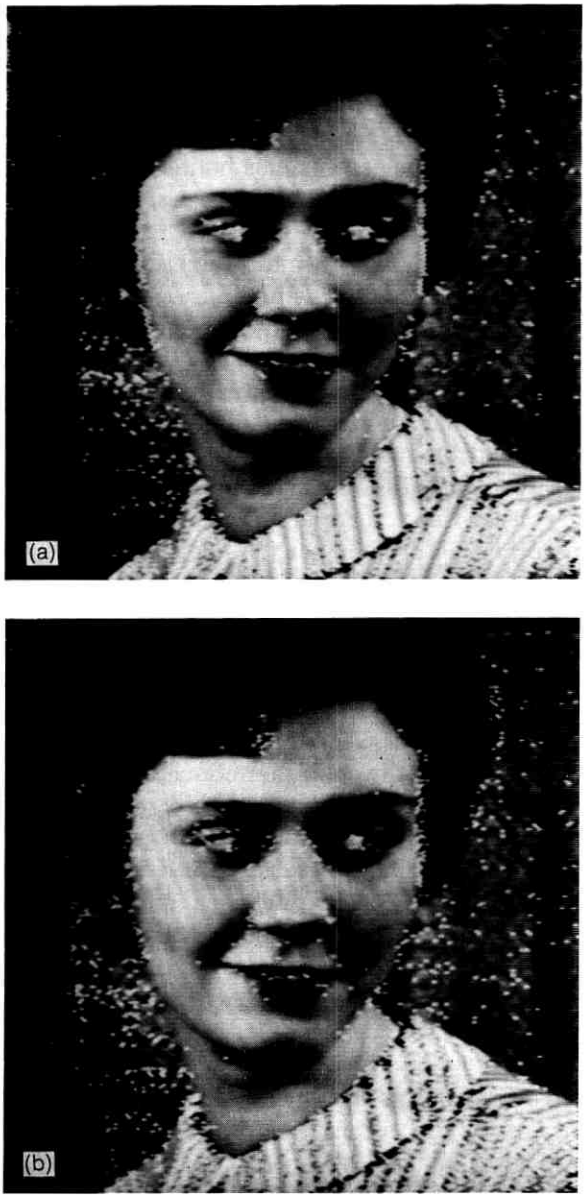

Fig. 16—Markers showing conditional points that were updated with a threshold of 1.5: (a) error transmission algorithm, (b) MI algorithm.

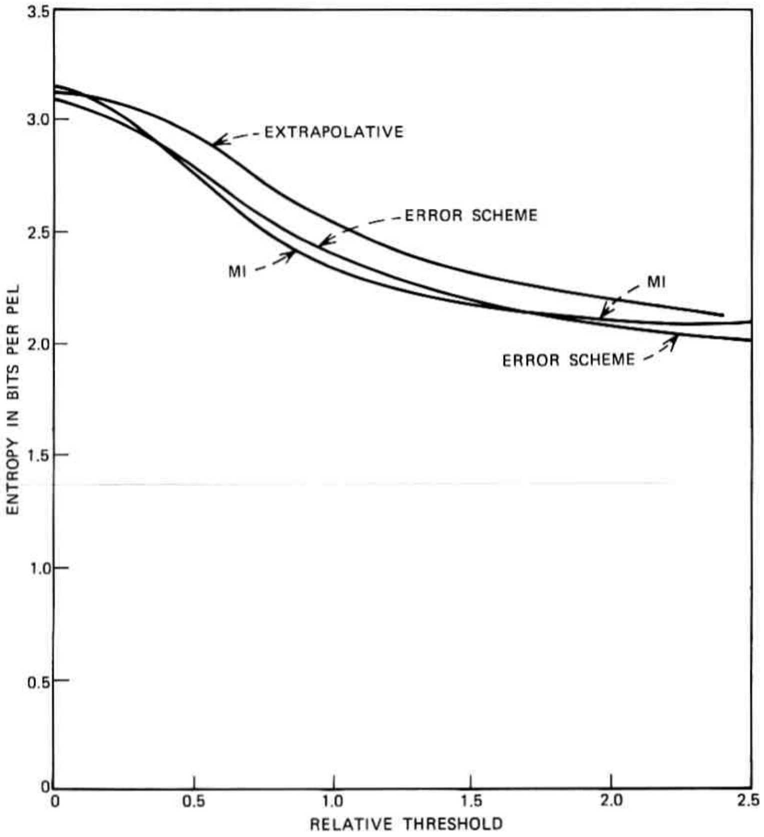

Fig. 17—Relative performance of three different 2:1 grid algorithms. The extrapolative algorithm is slightly inferior to the MI and dual-mode algorithms. Subject—Karen.

filtering (error at conditional point plus the error at previous point). This, in turn, results in a small increase in entropy (from 2.17 to 2.20 bits/pel).

For the 4:1 fixed-point algorithm the difference between single-point and three-point filtering is larger. The conditional points that are transmitted have been marked in Fig. 18 where, for single-point filtering, the threshold is 0.9 and the entropy is 2.08 bits/pel and for three-point filtering the threshold is 1.5 and the entropy is 2.04 bits/pel. In this case, however, the effect on picture quality is more noticeable. With the broader filter low-detail areas are reproduced better while medium-detail areas appear more noisy. At normal viewing dis-

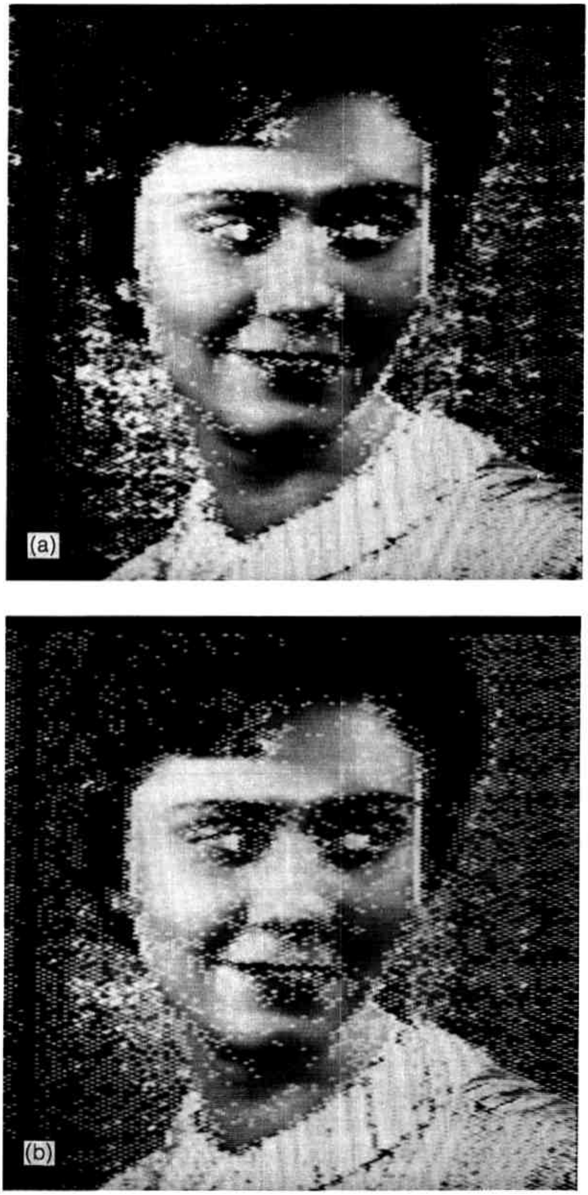

Fig. 18—Markers showing the conditional points that are updated for 4:1 grid algorithm: (a) single-point filtering, threshold = 0.9, $H = 2.08$, (b) three-point filtering, threshold = 1.5, $H = 2.04$.

tances the broad filter is preferable while for close scrutiny the single-point filter is better.

There is no reason why filtering could not take place in two or three dimensions in which case more elements would be involved and the accuracy with which the picture was encoded could more accurately match perceptual requirements for a given viewing situation.

V. DISCUSSION

As we have seen, the receiver-model coding algorithm with the simple threshold model of Fig. 2 tends to work best on low-detailed pictures. There are two reasons for this: (*i*) In detailed parts of the picture the estimation procedure is not as good as in low-detail areas; (*ii*) The threshold model, as described, is a simple, low-pass filter model and does not incorporate the effects of masking by adjacent signal components such as occurs when an element lies close to a large change (spatially or temporally) in luminance.*

The receiver-model coding concept, as stated, does not depend on any specific receiver model. As better models of the human viewer are obtained they can be incorporated directly into the encoding operation. In essence it is a three-step operation: estimation, testing, and, if necessary, more accurate recoding. There is an intrinsic separation between the source-property operation (estimation) and the receiver-property operation (testing) and as such the technique will be suboptimum. Performance could undoubtedly be improved by cycling through the estimate-test-recode sequence iteratively.[28] The interesting, practical question would be, is the improved performance worth the added complexity?

In all the coders described here the bit-rate—picture-quality operating point is determined by means of a single threshold control. This means that it is a relatively simple matter to dynamically alter the operating point in response to some system requirement. An example occurs in frame-to-frame coding where the moving area is transmitted as an element-differentially-quantized signal. As the buffer fills in response to increased movement the threshold is raised so as to keep the data-generation rate more uniform.[29]

VI. SUMMARY AND CONCLUSIONS

Receiver-model coding is a powerful, though not optimal, technique for incorporating properties of the human observer into the picture

* Some practical coding strategies have been developed that take advantage of spatial masking effects.[4,6]

encoding process. In essence, components of the signal are estimated according to some algorithm. The difference between the actual signal and the estimate is processed in a model of the receiver to determine if the estimate is adequate. If so, the receiver is informed of this; if not, additional information is transmitted to improve the estimate.

The receiver-model coding concept may be applied in many different ways and the visual model may range from very simple to very complex. In this paper I have used the differential quantizer (DPCM coder) as the basic vehicle with which to investigate receiver-model coding, and the visual model is a one-dimensional low-pass filter. Three types of estimation are investigated: extrapolation, interpolation, and a simplified form of interpolation referred to as "modified interpolation." It is important to bear in mind that the estimation is used to help determine *which* components need to be transmitted and does not indicate *how* the components are transmitted. In nearly all examples considered here the transmitted component is a simple difference signal which is decoded by adding the difference to the last decoded value.

Coders are divided into two separate classes, free-running algorithms and grid algorithms. In the free-running algorithms the estimation procedure may continue in a single run until the estimate fails with the proviso that the length of the run may not exceed a specified maximum. With the grid algorithm a fixed set of elements is always transmitted (e.g., every second or every fourth element). The interest in grid algorithms stems from the fact that they are more easily implemented.

The free-running interpolative algorithm gives a reduction in entropy of approximately 30 percent for high-detail pictures and 50 percent for low-detail pictures for a small loss in picture quality when the picture is evaluated by observing a single "frozen" frame on a high-quality CRT display.

Two reversible coding strategies were explored for converting the quantizer output to a binary code. Code II gives an advantage of between 0.1 and 0.15 bits/pel over Code I when using a maximum runlength of ten elements; the relative advantage of Code II over Code I about doubles when the maximum runlength is reduced to four elements.

The effect of the threshold filter function on the coding operation was explored by varying the shape of the filter function while keeping the spread of the function constant and then, in a second experiment, keeping the shape constant and varying the amount of spread. While the exact shape of the filter function affected performance very little, the spread of the function had a large effect; the most suitable spread appears to be about three elements for the normal viewing distance.

As the maximum permitted length of run is decreased from 10, it is found that there is very little increase in entropy for Code II for a maximum runlength even as short as 4, suggesting that a 4:1 grid algorithm may perform almost as well as free-running algorithms.

The 2:1 grid algorithm (modified interpolative) does not permit operation at lower picture qualities and bit rates; the 4:1 algorithm has a larger range. However, over their range of operation, the grid algorithms perform at least as well as the best free-running algorithm and in view of their simpler implementation appear to be the most promising.

Three different 2:1 grid algorithms were compared, an error-transmission technique in which the correction signal is sent as a difference between the estimate and the input, the modified interpolative algorithm, and the extrapolative algorithm. Extrapolation was slightly inferior to the other two methods and of these the modified interpolative method is more simply implemented.

The emphasis in this paper has been on obtaining an efficient discrete representation of a picture signal rather than presenting a complete coding system. Consequently, there are a number of considerations such as sensitivity to transmission errors which are not discussed in the paper but nevertheless bear importantly on the feasibility of any practical coder.

VII. ACKNOWLEDGMENTS

This study would have been extraordinarily difficult without the picture processing facility, both hardware and user-oriented system software, of R. C. Brainard and J. D. Beyer. I thank them for their gentle guidance.

REFERENCES

1. Graham, R. E., "Subjective Experiments in Visual Communication," IRE Conv. Rec., 1958, pp. 100–106.
2. Powers, K. H., and Staras, H., "Some Relations Between Television Picture Redundancy and Bandwidth Requirements," Trans. Amer. IEE, *76* (1957), p. 492.
3. Graham, R. E., "Predictive Quantizing of Television Signals," IRE Wescon Conv. Rec., Part 4, 1958, pp. 142–157.
4. Limb, J. O., "Adaptive Encoding of Picture Signals," in Huang, T. S., and Tretiak, O. J., eds., *Picture Bandwidth Compression*, Gordon and Breach, Science Publishers, 1972.
5. Wintz, P. A., and Tasto, M., "Picture Bandwidth Compression by Adaptive Block Quantization," Purdue University School of Electrical Engineering, TR-EE 70-14, July 1970.

6. Brown, E. F., and Kaminski, W., "An Edge-Adaptive Three-Bit Ten-Level Differential PCM Coder for Television," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (December 1971), pp. 944–947.
7. Limb, J. O., "Source-Receiver Encoding of Television Signals," Proc. IEEE, *55*, March 1967, pp. 364–379.
8. Candy, J. C., and Bosworth, R. H., "Methods for Designing Differential Quantizers Based on Subjective Evaluations of Edge Busyness," B.S.T.J., *51*, No. 7 (September 1972), pp. 1495–1516.
9. Budrikis, Z. L., "Visual Fidelity Criterion and Modeling," Proc. IEEE, *60*, No. 7 (July 1972), pp. 771–779.
10. Kortman, C. M., "Redundancy Reduction—A Practical Method of Data Compression," Proc. IEEE, *55*, No. 3 (March 1967), pp. 253–263.
11. Novak, S., and Sperling, G., "Visual Thresholds Near a Continuously Visible or Briefly Presented Light-Dark Boundary," Optical Acta, *10*, No. 2 (April 1963), pp. 87–91.
12. Fiorentini, A., "Mach Band Phenomena," in Jameson, D., and Hurvich, L. M., eds., *Handbook of Sensory Physiology*, Vol. VII/4, Springer-Verlag, 1972.
13. Limb, J. O., "Vision Oriented Coding of Visual Signals," Ph.D. Thesis, University of Western Australia, 1966.
14. Moon, P., and Spencer, D. E., "The Visual Effect of Non-Uniform Surrounds," J. Opt. Soc. Amer., *35*, March 1945, pp. 233–248.
15. Hacking, K., "The Relative Visibility of Random Noise Over the Grey-Scale," J. Brit. IRE, *23*, No. 4 (April 1962), p. 307.
16. Newell, G. F., and Geddes, W. K. E., "Visibility of Small Luminance Perturbations in Television Displays," BBC Research Department, Report T106, 1963.
17. Budrikis, Z. L., "Visual Threshold and the Visibility of Random Noise in TV," Proc. IRE Australia, *22*, December 1961, pp. 751–759.
18. Blackwell, H. R., "Neural Theories of Simple Visual Discriminations," J. Opt. Soc. Amer., *53*, January 1963, pp. 129–160.
19. Kristofferson, A. B., "Visual Detection as Influenced by Target Forms," in Wulfeck, J. W., and Taylor, J. H., eds., *Form Discrimination as Related to Military Problems*, National Academy of Sciences—National Research Council Publication, 561, 1957, pp. 109–126.
20. Budrikis, Z. L., "Model Approximations to Visual Spatio-Temporal Sine-Wave Threshold Data," to be published in November 1973 B.S.T.J.
21. Mounts, F. W., and Pearson, D. E., "Measurements of the Apparent Increase in Noise Level Resulting from Frame-Repetition of Low-Resolution TV Pictures," B.S.T.J., *48*, No. 3 (March 1969), pp. 527–539.
22. Limb, J. O., "Efficiency of Variable-length Binary Codes," Proc. Univ. Missouri, Rolla, M. J. Kelly Communications Conf., 1970, pp. 13-3-1, 13-3-9.
23. Rice, R. F., and Plaunt, J. R., "Adaptive Variable-length Coding for Efficient Compression of Spacecraft Television Data," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (December 1971), pp. 889–897.
24. Kretzmer, E. R., "Reduced Bandwidth Transmission System," U. S. Patent No. 2,949,505, August 16, 1960.
25. Fukushima, K., and Ando, H., "Television Band Compression by Multimode Interpolation," Technical Research Laboratories, Japan Broadcasting Corporation, Japan.
26. Connor, D. J., "Techniques for Reducing the Visibility of Tranmission Errors in Digitally Encoded Video Signals," IEEE Trans. Commun. Tech., *COM-21*, No. 3 (June 1973).
27. Pease, R. F. W., "Conditional Vertical Subsampling—A Technique to Assist in the Coding of Television Signals," B.S.T.J., *51*, No. 4 (April 1972), pp. 787–802.
28. Viterbi, A. J., "Convolutional Codes and Their Performance in Communication Systems," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (October 1971), pp. 751–772.
29. Limb, J. O., Pease, R. F. W., and Walsh, K. A., "Combining Intraframe and Frame-to-frame Coding for Television," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Statistical Properties of Gilbert's Burst Noise Model

By MIN-TE CHAO

(Manuscript received April 17, 1973)

Simple statistical procedures for analyzing error data, e.g., in digital data transmission systems, are usually based on the assumption of independence. This paper studies the performance and potential utility of such simple statistical procedures in the case of nonindependent error occurrences. The burst noise model is selected for this purpose because of its neatness, its mathematical tractability, its built-in structure of dependence, and its importance in communication theory. We show that statistical procedures designed under the assumption of independence tend to be conservative for the burst noise model. For example, the usual binomial test will reject, on the average, more channels with small error rates than it would if the errors were independent. The case that the sample size n and the error rate ρ converge in such a way that $n\rho \rightarrow \mu_0$ is also studied. It is shown that the error process can be approximated by a compound Poisson process in continuous time t. The statistical implications of this fact are also discussed.

I. INTRODUCTION

A dilemma long existing in the theory and applications of digital data transmission is the precise determination of the error structure. On the one hand, it is a well-recognized fact that errors do not occur independently; on the other hand, only the assumption of independence offers us a model sufficiently tractable that ordinary statistical procedures can be designed accordingly. A direct consequence is, of course, that we are using statistical methods designed for independent observations to make statistical inferences on dependent data.

The fact is, we do not have much knowledge of the error structure of data transmission channels. Mathematical models have been constructed for fitting observed data streams containing errors, noticeably the burst noise model of Gilbert,[1] the Markov error process and renewal error process of Elliott,[2,3] and the binary regenerative model of McCullough.[4]

One of the most pertinent models with a built-in dependence structure is Gilbert's burst noise model. It is this model that we shall study in this paper. One of the prime concerns of this study is the behavior of various statistical procedures under the burst noise model.

Gilbert[1] constructs a model for burst noise as follows. An input binary signal (0 or 1) is transmitted through a noisy channel with noise z (0 or 1) so that the output is given by

$$\text{output} \equiv \text{input} + z \pmod{2}.$$

The channel can be in either of the two states, good (G) or bad (B). If, at time n, the channel is in G, there is no noise so $z_n = 0$; if the channel is in B, a "coin" with $P[\text{head}] = h$ is tossed and $z_n = 1$ is identified with a tail outcome.

The channel can shift from a good state to a bad state and vice versa. Identify 1 as G and 2 as B and let X_n denote that state of the channel at time n. It is assumed that the process $\{X_n : n \geqq 1\}$ is a two-state Markov chain with stationary transition probabilities

$$T = \begin{bmatrix} 1 - P & P \\ p & 1 - p \end{bmatrix} \tag{1}$$

and initial distribution (π_1, π_2).

Let $Z_n = z_1 + \cdots + z_n$ denote the number of errors through the nth-bit output (0 or 1) digits of the channel where $z_i = 1$ if and only if an error occurs at the ith bit. The statistic Z_n is obviously the quantity that will be used in any statistical procedure concerning the bit error rate. The statistical behavior of Z_n will be studied extensively in this work.

In Section II, we derive most of the exact formulas concerning Z_n, including explicit expressions for its probability-generating function and its first and second moments. The exact form of the probability distribution of Z_n is quite involved in general. For the special case $p + P = 1$, Z_n reduces to the binomial variable. The quantity $\lambda = 1 - p - P$ can thus be used as a measure of dependence; most of the complications in this work are caused by the presence of a nonzero λ. The effect of dependence is discussed in some detail in Section III. Transmission in blocks of digits is considered; one of our major results is that it can be shown in this model that the block error and the bit error have essentially the same covariance structure. Thus, most results concerning bit error rate can be transferred easily to results about block error rate. As a corollary, the variance for Z_n is obtained as a sum of two components, one due to the sum of variances (as if the z's were independent) and the other due to the fact that $\lambda \neq 0$ (the effect of dependence).

Since Z_n is known to be asymptotically normally distributed, the variance formula of Z_n can be used to judge the effect of dependence on the robustness of statistical procedures (i.e., on how well procedures based on the independence assumption perform if this assumption is violated). A general conclusion of Section IV is that statistical procedures designed under the assumption of independence tend to be conservative for the burst noise model. For example, the usual binomial test will reject, on the average, more channels with small error rate then it is supposed to.

It is shown in Section V that if the bit error rate $\rho \rightarrow 0$ in such a way that $n\rho \rightarrow \mu_0 > 0$, then Z_n converges in distribution to a compound Poisson distribution. The statistical implications of this fact are also discussed. In particular, Z_n is a minimal sufficient statistic for $\mu_0(\rho)$ in some approximate sense. This justifies the use of Z_n in any statistical decision procedures concerning the error rate ρ.

Despite the model's simplicity, the insight we gained in studying this burst noise model enables us to investigate more deeply the structure of error processes. For example, it is possible to treat the underlying Markov chain $\{X_n\}$ as an s-state stationary Markov chain. Details of this and other extensions and their implications will be discussed in a forthcoming report.

II. STATISTICAL PROPERTIES OF Z_n

We shall assume, for simplicity and without loss of too much generality, in the sequel that the initial distribution (π_1, π_2) of the two-state Markov chain $\{X_n\}$ agrees with its absolute stationary distribution $(p/(p + P), P/(p + P))$. Under this assumption, $\{X_n\}$ is strictly stationary.

Let

$$g_n = P[Z_n = 0].$$

Note that the bit error rate ρ is given by

$$\begin{aligned} \rho_1 &= 1 - g_1 \\ &= P[Z_1 \neq 0] \\ &= P[z_1 = 1] \\ &= (1 - h)P/(p + P); \end{aligned} \tag{2}$$

and the block error rate ρ_k, the probability that a block of size k contains at least one error, is

$$\rho_k = 1 - g_k. \tag{3}$$

Thus, $\rho_1 = \rho$.

Since the event $[z_i = 1]$ implies $[X_i = 2]$ and thus signifies a return to a bad state (a recurrent event), it is possible to utilize the renewal equation to derive an exact expression for g_n. The following theorem is essentially due to Gilbert [Ref. 1, eq. (14)].

Theorem 1: For $n \geqq 1$,

$$g_n = \frac{A_1\alpha_1^{n+1}}{1-\alpha_1} + \frac{A_2\alpha_2^{n+1}}{1-\alpha_2}, \tag{4}$$

where

$$\alpha_1 = \tfrac{1}{2}[-(1-h)(1-p) - (p+P-2) + \sqrt{[(1-P) - h(1-p)]^2 + 4pPh}]$$
$$\alpha_2 = \tfrac{1}{2}[-(1-h)(1-p) - (p+P-2) - \sqrt{[(1-P) - h(1-p)]^2 + 4pPh}]$$
$$A_1 = \rho[\alpha_1 + (p+P-1)]/\alpha_1(\alpha_1 - \alpha_2)$$
$$A_2 = \rho[\alpha_2 + (p+P-1)]/\alpha_2(\alpha_2 - \alpha_1).$$

A proof of Theorem 1 different from that of Gilbert (and the proofs of all other theorems) will be presented in the appendix. We remark here that since a broader view and a more systematic approach is adopted in our new proof, it is possible to extend our method readily to a more general framework than a two-state Markov chain.

Relation (4) can be viewed as a relation between bit error rate and block error rate. If $\lambda = 1 - p - P > 0$, it can be shown that $0 < \alpha_2 < \alpha_1 < 1$ so that $g_n \to 0$ exponentially fast. One effect of dependence in this model is reflected in (4), namely that g_n is the sum of two exponential terms instead of one. In general, if the underlying Markov chain is s-state, g_n will be a sum of s exponential terms.

The right-hand side of (4) is a function of p, P, and h. We shall write $g_n = g_n(p, P, h)$ when we want to emphasize this point. An important connection between g_n and Eu^{Z_n}, the probability-generating function (PGF) of Z_n, is stated in Theorem 2.

Theorem 2: The probability-generating function of Z_n *is given by*

$$Eu^{Z_n} = g_n(p, P, H), \tag{5}$$

where

$$H = (1-h)u + h.$$

Thus, replacing each h by $(1 - h)u + h$ in (4), we obtain the PGF of Z_n. The exact expressions for $P[Z_n = i]$ are involved unless i is small. Using (4) and the fact that $0 < \alpha_2 < \alpha_1 < 1$, it is possible to express $P[Z_n = i]$ approximately in terms of its leading term as

$$P[Z_n = i] \approx \frac{A_1^{i+1}}{1 - \alpha_1} \binom{n}{i} \alpha_1^{n+1}. \tag{6}$$

Relation (6) can be used to establish the Poisson convergence of Z_n if $\rho = \mu_0/n \to 0$. However, an indirect proof will be presented later.

Moments of Z_n can be obtained by differentiating the right-hand side of (5) and setting $u = 1$. Specifically, we have

$$EZ_n = n\rho, \tag{7}$$

$$\text{Var } Z_n = n\rho(1 - p) + 2C\left[\frac{(n-1)\lambda}{1 - \lambda} - \frac{\lambda^2(1 - \lambda^{n-1})}{(1 - \lambda)^2}\right], \tag{8}$$

where

$$C = (1 - h)^2 \pi_1 \pi_2$$
$$\lambda = 1 - p - P.$$

Relation (8) also can be obtained by other methods which we shall discuss in Section III.

III. MEASURE OF DEPENDENCE AND ITS EFFECT

If the transition matrix of a Markov chain has identical rows, then this Markov chain is merely a sequence of independent and identically distributed (iid) random variables. For the two-state Markov chain $\{X_n\}$ underlying this burst noise model, the matrix T in (1) has identical rows if and only if $p + P = 1$. Letting $\lambda = 1 - p - P$, we see that $|\lambda| \leqq 1$ and that $\lambda = 0$ if and only if the channel is memoryless.

The eigenvalues of the transition matrix play important roles in the theory of Markov chains. The largest (in absolute value) eigenvalue is always 1; in general, it is the second largest eigenvalue that affects all the essential features of a Markov chain. The parameter λ defined earlier is the second largest eigenvalue of the matrix T in (1).

The significance of the parameter λ can be interpreted intuitively. If p and P are small, the underlying Markov chain $\{X_n\}$ tends to stay in a certain state (G or B) once it enters this state; hence, $\lambda > 0$ indicates the tendency of producing bursty errors. If both p and P are large, then $\{X_n\}$ tends to shift between the good state and bad state alternatively. Since the latter case is obviously not very interesting, we shall always assume $\lambda \geqq 0$ in the sequel.

Let Π denote the 2×2 matrix with identical rows

$$\Pi = \begin{bmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{bmatrix},$$

where (π_1, π_2) is the absolute stationary distribution of $\{X_n\}$. By the definition of the absolute stationary distribution and by some simple calculations, it can be seen that

$$\Pi T = T\Pi = \Pi^2 = \Pi. \tag{9}$$

It follows from (9) and simple induction that, for $n \geqq 1$,

$$\begin{aligned} T^n - \Pi &= (T - \Pi)^n \\ &= \lambda^n \begin{bmatrix} \pi_2 & \pi_2 \\ \pi_1 & \pi_1 \end{bmatrix}. \end{aligned} \tag{10}$$

Relation (10) allows us to calculate the ℓ-step transition probabilities of $\{X_n\}$ accurately. It can also be used to find the covariance of z_i and z_j. We restate eq. (17) of Ref. 1 as follows:

Theorem 3: The covariance of z_i, z_j $(i \neq j)$ is given by

$$\operatorname{Cov}(z_i, z_j) = C\lambda^{|i-j|}, \tag{11}$$

where $C = (1 - h)^2\pi_1\pi_2$.

Corollary:

$$\operatorname{Var}(Z_n) = n\rho(1 - \rho) + 2C\left[\frac{(n-1)\lambda}{1-\lambda} - \frac{\lambda^2(1-\lambda^{n-1})}{(1-\lambda)^2}\right]. \tag{12}$$

Define, for $i = 1, 2, \cdots$,

$$\begin{aligned} T_i &= 1 \quad \text{if} \quad z_{(i-1)k+1} + z_{(i-1)k+2} + \cdots + z_{ik} \geqq 1 \\ &= 0 \quad \text{otherwise;} \end{aligned} \tag{13}$$

namely, $T_i = 1$ if and only if the ith block of length k is not error-free. It is possible to extend eq. (11), and therefore (12), to the corresponding equations involving the T's.

Theorem 4: There exists $0 < C_1 < \infty$ such that

$$\operatorname{Cov}(T_i, T_j) = C_1\lambda^{|i-j|k}. \tag{14}$$

The value of C_1 can be found explicitly. However, we shall be satisfied with a crude estimate $C_1 = C_2\pi_1\pi_2\lambda^{1-k}$ where $|C_2| \leqq \frac{1}{4}$.

Note that $T_i = z_i$ if $k = 1$. In this case, eq. (14) reduces to (11). Theorem 4 not only states that the T's are "less dependent" than the z's but it also tells us, in some sense, how much less dependent the

T's are. Let

$$S_n = T_1 + T_2 + \cdots + T_n.$$

The statistic S_n is the obvious statistical quantity to analyze if digits are transmitted in blocks of size k. For example, in the 1969–70 Connection Survey[5,6] on the Bell System Switched Telecommunications Network conducted by Bell Laboratories, statistics of block errors are presented for both high-speed and low-speed data transmission. Hence, the more important implication of Theorem 4 is that eq. (14) exhibits the same general structure as eq. (11). For example, replacing C by C_1, ρ by ρ_k, and λ by $\lambda^* = \lambda^k$ in (12), we immediately obtain the formula for Var (S_n).

Corollary:

$$\text{Var}\,(S_n) = n\rho_k(1 - \rho_k) + 2C_1\left[\frac{(n-1)\lambda^*}{1-\lambda^*} - \frac{\lambda^{*2}(1-\lambda^{*n-1})}{(1-\lambda^*)^2}\right]. \qquad (15)$$

Consequently, statistical procedures using S_n and concerning the inferences on the block error rate ρ_k should have essentially the same behavior as those procedures using Z_n and concerning the bit error rate ρ. The above reasoning implies that, at least as far as the large sample properties are concerned, it is sufficient to consider inference on ρ only.

Both the law of large numbers and the central limit theorem hold true for the sum of Markovian random variables; see, for example, Ref. 7. Hence,

$$\frac{S_n}{n} \to \rho_k \qquad (16)$$

with probability 1; and

$$P\left[\frac{S_n - n\rho_k}{\sqrt{\text{Var}\,S_n}} \leqq v\right] \to \Phi(v) \qquad (17)$$

for each $-\infty < v < \infty$, where $\Phi(v)$ denotes the cumulative distribution function of an $N(0, 1)$ random variable. Relations (16) and (17) will be used in Section IV to discuss the robustness of some statistical procedures concerning inferences on ρ_k.

IV. STATISTICAL INFERENCES ON ρ_k

For simplicity, we shall consider the special case $k = 1$ and concentrate our discussion on problems of statistical estimation and hypothesis testing of $\rho = \rho_1$. As remarked earlier in Section III, the restriction $k = 1$ can easily be extended to the general case.

Since $\{X_n\}$ is assumed to be stationary, so is $\{z_n\}$; we have seen that

$$E[Z_n] = n\rho, \tag{18}$$

so that the obvious estimator $\hat{\rho}_n = Z_n/n$ of ρ is unbiased. Relation (16), specialized for the case $k = 1$, states that $\hat{\rho}_n$ is a strongly consistent estimator of ρ.

Very few (optimal) small sample properties of $\hat{\rho}_n$ can be stated, however. For $n \geqq 3$, it can be shown that no uniformly minimum variance estimator of ρ exists. Nevertheless, it is intuitively obvious that $\hat{\rho}_n$ is about the best we can do if the z's are the only observables. From (12),

$$\begin{aligned} n \operatorname{Var}(\hat{\rho}_n) &= \rho(1-\rho) + 2C\frac{\lambda}{1-\lambda} + o(1) \\ &\triangleq \sigma^2 + A, \end{aligned} \tag{19}$$

where

$$\begin{aligned} \sigma^2 &= \rho(1-\rho) \\ A &= 2(1-h)^2\pi_1\pi_2\lambda/(1-\lambda). \end{aligned}$$

Note that the term $\rho(1-\rho)$ in eq. (19) corresponds to $n \operatorname{Var}(\hat{\rho}_n)$ if the z's were independent. Since we have assumed that $\lambda \geqq 0$, it follows that $A \geqq 0$ and $n \operatorname{Var}(\hat{\rho}_n) \geqq \sigma^2$. Thus, the presence of a positive λ actually causes loss of efficiency in estimating ρ. Writing $A = \tau^2$, we see that if the parameters h, p, and P (hence σ^2 and τ^2) can be estimated from the data, the loss of efficiency due to dependence can be estimated as the ratio $\hat{\tau}/\hat{\sigma}$, where $\hat{\tau}$ and $\hat{\sigma}$ denote the estimates of τ and σ from the sample. Hence, if control or confidence limits are used to evaluate the channel performance, the actual 3 standard deviation or 2 standard deviation limits should be wider by $100(\tau/\sigma)$ percent.

We may also consider the loss of power for statistical tests for H_0: $\rho \leqq \rho_0$ of the form

$$\text{reject } H_0 \quad \text{if} \quad Z_n \geqq C^*.$$

Based on the assumption of independence, the power function is approximately

$$\beta_I = 1 - \Phi\left(\frac{C^* - n\rho}{\sqrt{n\sigma^2}}\right); \tag{20}$$

whereas for our model, the power is approximately

$$\beta_D = 1 - \Phi\left(\frac{C^* - n\rho}{\sqrt{n(\sigma^2 + \tau^2)}}\right). \tag{21}$$

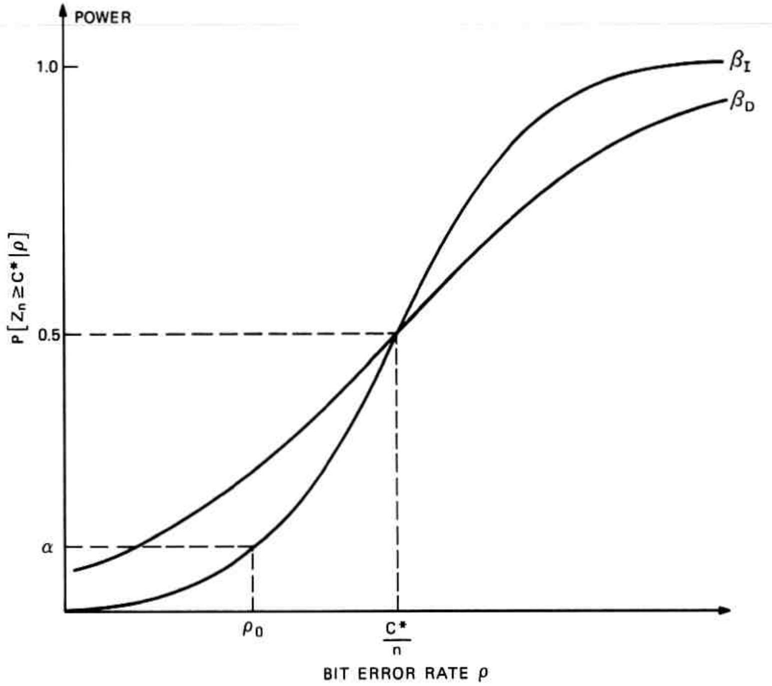

Fig. 1—Comparison of power functions.

If the first-type error $\alpha \leqq \frac{1}{2}$, we see from (20) that $C^* - n\rho_0 \geqq 0$. We see that $\beta_I \leqq \beta_D$ if $C^* - n\rho \geqq 0$ and $\beta_I \geqq \beta_D$ otherwise. This means that it might be possible to design more powerful tests for H_0 based on the knowledge that the dependent model obtains. On the other hand, the test is conservative in the sense that it may reject more channels than expected if the bit error rate ρ is close to the service objective ρ_0 and if the dependent model obtains. The rules of the game shift in the other direction if $C^* - n\rho < 0$. However, it is the smaller values of ρ that we are really concerned with and we may claim that the test based on the assumption of independence gives a pessimistic estimate of channel reliability (see Fig. 1).

V. POISSON APPROXIMATIONS

The bit error rates of high-speed digital channels are usually very small, say 10^{-5} to 10^{-8}; therefore, the normal approximation and the statistical theory discussed earlier may not be too helpful in practice unless n is large. In this section, we prove that Z_n converges in distribution to a Poisson distribution if $n\rho \rightarrow \mu_0$ in a suitable way. Using this result, we construct a Poisson process in continuous time t that approximates the process $\{Z_n(t):t > 0\}$ where n denotes the number of transmitted digits per unit time.

We have shown earlier in (2) that the error rate ρ is given by

$$\rho = (1 - h)P/(p + P). \tag{22}$$

If $\rho \rightarrow 0$ in such a way that $n\rho \rightarrow \mu_0 > 0$, what do we expect to be the asymptotic distribution for $Z_n = z_1 + \cdots + z_n$, the number of errors in the first n digits? Note that we have quite a few choices for the convergence $n\rho \rightarrow \mu_0$. For example, keeping p fixed and letting $P = (\mu/n)^{\epsilon_1}$, $1 - h = (\mu/n)^{\epsilon_2}$, $\epsilon_1 + \epsilon_2 = 1$, we have, by (8),

$$\begin{aligned} \text{Var}\,(Z_n) &\approx \mu_0(1 - \rho) + 2Cn\cdot\frac{\lambda}{1 - \lambda} \\ &\approx \mu_0 + \frac{2}{n^{\epsilon_2}}\cdot\mu^{1+\epsilon_2}\frac{1 - p}{p^2}, \end{aligned} \tag{23}$$

where $\mu_0 = \mu/p$. Also,

$$\begin{aligned} EZ_n &= n\rho \\ &= \mu_0. \end{aligned}$$

Hence, if $\epsilon_2 = 0$ is selected, we see that for large n, Var $Z_n \neq EZ_n$ so that the limiting distribution of Z_n cannot be Poisson.

In order that $\rho = (1 - h)P/(p + P) \approx \mu_0/n$, the most general choice of h and P would be

$$\begin{aligned} 1 - h &= a_1x + a_2x^2 + a_3x^3 + \cdots, \\ P &= b_1y + b_2y^2 + b_3y^3 + \cdots, \end{aligned} \tag{24}$$

where $x = n^{-\epsilon_2}$, $y = n^{-\epsilon_1}$, $\epsilon_1 + \epsilon_2 = 1$, $\epsilon_1 \geqq 0$, $\epsilon_2 > 0$, and $a_1b_1/p = \mu_0$ (the case $\epsilon_2 = 0$ is of particular interest and will be considered separately later). We state the main theorem of this section as follows:

Theorem 5: If $\rho \rightarrow 0$ in such a way that (24) *holds, then*

$$P[Z_n = i] \rightarrow \frac{1}{i!}\,\mu_0^i e^{-\mu_0}$$

as $n \rightarrow \infty$, where $\mu_0 = a_1b_1/p$. Furthermore, the convergence is uniform in $i = 0, 1, 2, \cdots$.

By using the result of Theorem 5, we may construct a Poisson process in continuous time t as an approximation to the process of partial sums $\{Z_n\colon n \geqq 1\}$. Suppose the underlying channel can transmit n digits per unit time. Let $Z_n(t)$ denote the number of errors in $(0, t)$. Theorem 5 states that, for $i = 0, 1, 2, \cdots$,

$$P[Z_n(t) = i] \to \frac{1}{i!} (\mu t)^i e^{-\mu t};$$

here μ denotes the limiting error rate per unit time. Let $Z(t)$ denote the number of errors in $(0, t)$ in the limiting case. The fact that $Z(t)$ is a process with independent increments, namely that $Z(t)$ is indeed a Poisson process, is easy to prove and we shall omit it.

Theorem 5 implies that Z_n is asymptotically a minimum sufficient statistic for the bit error rate ρ if (24) can be justified; this provides theoretical support for the use of Z_n in any statistical inferences concerning ρ. We remark here that, by replacing ρ by ρ_k and Z_n by S_n, the same comment applies for block error rates. Another consequence of Theorem 5 is that

$$\text{Var } (Z_n) \to \mu_0 = a_1 b_1 / p. \tag{25}$$

Note that if $\lambda = 1 - p - P = 0$ (the independent case), (24) implies $P \to 0$ and this in turn implies $p \to 1$. From (25), we see that Var (Z_n) is minimized in the independent case. The increase of variance due to dependence is therefore $100(1/p - 1)$ percent. Hence, in the dependent case, the confidence interval for ρ should be wider than we thought in the independent case.

The null hypothesis $H_0 : \rho \leqq \rho_0$ becomes $H_0' : \mu_0 \leqq \mu_0^*$ in the limiting case. The uniformly most powerful test for H_0' exists and is given by the rule:

$$\text{reject } H_0' \quad \text{if} \quad Z_n \geqq C^*.$$

Based on the approximation that Z_n is Poisson, we may compute the power functions as

$$\begin{aligned} \beta_D(\mu_0) &= P[Z_n \geqq C^* \mid \mu_0] \\ &= \sum_{i=C^*}^{\infty} \frac{1}{i!} \mu_0^i e^{-\mu_0} \\ &= \int_0^{\mu_0} \frac{1}{(C^* - 1)!} e^{-x} x^{C^*-1} dx \end{aligned}$$

and

$$\beta_I = \int_0^{a_1 b_1} \frac{1}{(C^* - 1)!} e^{-x} x^{C^*-1} dx.$$

It follows that $\beta_I \leqq \beta_D$ so that a test for H_0 based on the assumption of independence and used when dependence is present rejects more channels than it should. In other words, tests designed for independent observations protect customers in the sense that channels they are using may have better quality than inferred.

The effect of dependence reported for both the binomial and the Poisson cases has an intuitive explanation. By using Z_n or S_n, we are actually abandoning some of the information contained in the sequence $z_1, z_2, \cdots$, so that statistical inferences based on Z_n or S_n tend to be more conservative in the sense that channel reliability is estimated pessimistically.

We now return to (24) and consider the special case $\epsilon_2 = 0$. This case cannot be ignored because previous papers, for example Ref. 1, indicate that sometimes $h \approx 0.5$ (rather than 0.999) is a reasonable value. The fact that Poisson processes do not describe certain error processes well has also been reported in the literature.

If $\epsilon_2 = 0$, eq. (24) reduces to

$$P = [b_1 + o(1)]/n \qquad b_1 > 0. \tag{26}$$

We have

Theorem 6: If (26) *holds, then*

$$Eu^{Z_n} \to \exp\left[-\frac{b_1(1-H)}{1-H+pH}\right], \tag{27}$$

where $H = (1-h)u + h$.

We remark here that the limiting value in eq. (27) is the PGF of a compound Poisson process. More specifically, let N be a Poisson variable with mean $b_1/(1-p)$, and let $W_1, W_2, \cdots$ be iid random variables with the geometric distribution

$$P[W_1 = i] = \left[\frac{p}{1-(1-p)h}\right]\left[\frac{(1-p)(1-h)}{1-(1-p)h}\right]^i, \qquad i = 0, 1, 2, \cdots. \tag{28}$$

If the W's are independent of N, then the left-hand side of (27) is simply $Eu^{W_1+W_2+\cdots+W_N}$. It is of course possible to introduce a continuous time parameter t and consider the following random mechanism which describes the bursty nature of this error process vividly. The bursts are generated by a Poisson process; given that a burst occurs, the errors are generated by a geometric distribution.

From the right-hand side of (27), it is possible to compute the moments of the limiting distribution of Z_n. We have

$$E(W_1 + W_2 + \cdots + W_N) = \frac{b_1(1-h)}{p} \tag{29}$$

$$\operatorname{Var}(W_1 + W_2 + \cdots + W_N) = \frac{b_1(1-h)}{p} + \frac{2b_1(1-h)^2(1-p)}{p^2}. \tag{30}$$

Note that the variance is always larger than the mean in this case. Note also that, as h approaches 1, the second term on the right-hand side of eq. (30) is of higher order and vanishes in the limiting case. Another interesting thing is that it is possible to show that the right-hand side of (30) is minimized at $p = 1$, and as p approaches 1, the limiting distribution is Poisson.

Branching renewal processes have been suggested in the literature[8] as a model for series of events. The basic structure for branching renewal processes can be described in terms of our problem as follows: The series of primary events (bursts) are generated by a Poisson process. Each of these primary events generates a subsidiary series of events (bit errors), separated by the waiting time $Y_1, Y_2, \cdots, Y_S$, where S is random. If we assume that these subsidiary series of events take no time, then the branching renewal process reduces to the compound Poisson process.

VI. CONCLUSIONS AND FURTHER EXTENSIONS

(*i*) The burst noise model of Gilbert discussed in this paper provides a vehicle for studying the robustness of some fixed sample size statistical procedures. The general result is that the presence of dependence increases the variance of the random variable Z_n, for the case where the bit error rate ρ is fixed and the case in which $\rho = [\mu_0 + o(1)]/n$. Thus, use of statistical tests based on the assumption of independence increases the power at the cost of rejecting more satisfactory channels than would be rejected if dependence were absent. The use of blocks does reduce the covariances among errors compared with bits or smaller blocks. However, the covariance structure among the blocks is essentially the same as that among the bits.

(*ii*) Although the dependence structure of the Gilbert's burst noise model is a simple one, it is by no means a trivial one. In fact, from the insight gained through this study, many results obtained in this paper have generalizations in error processes defined over an s-state Markov chain as well. A unified treatment on channels with Markov type of memory will be reported elsewhere.

(*iii*) The second largest eigenvalue (in absolute value) of the $(s \times s)$ transition matrix of the underlying Markov chain is a parameter which should not be overlooked. It can be viewed as a measure of dependence of a Markovian model. The effect of this parameter $(=\lambda$ in this work) is visible in many important formulas, for example, in (14).

(*iv*) Another important question to ask is what kind of stochastic process can be used to approximate the error process of a binary channel with memory. If the bit error rate is small, we can extend the proof of Theorem 6 (in a nontrivial way) to find an important conclusion: the compound Poisson process can serve the purpose.

(*v*) The by-products of this work are also fruitful. For example, the variance formula of Z_n can be generalized to find the variance of $T_n = f(X_1) + \cdots + f(X_n)$ where $\{X_i\}$ is an s-state Markov chain, $s \leqq \infty$, and f is an arbitrary function. Since many continuous sampling plans, such as CSP1, CSP2, CSP3, can be described as random walks of the form T_n (see Refs. 9 and 10), the application of this formula to quality assurance is evident.

(*vi*) Mathematically speaking, there is an essential difference between Gilbert's original treatment and our generalizations to the s-state Markov chain. More specifically, Gilbert viewed his problem as one of the renewal type whereas the s-state Markov case should be handled by the semigroup property (of taboo probabilities). We remark here that many results of the theory of recurrent events (see, for example, Ref. 11) can be applied to Gilbert's model. We also remark that the renewal process is a one-state semi-Markov process. A general question can be raised at this point: What is the behavior of an s-state semi-Markov channel? Since it is known that distributions other than the exponential (for example, the Pareto distribution, see Ref. 12) describe the waiting time distribution well, the question raised is a realistic one and should not be merely considered as an attempt at mathematical generality.

VII. ACKNOWLEDGMENTS

The author wishes to thank J. A. Tischendorf for his critical reading. He also wants to thank L. B. Boza for many helpful discussions.

APPENDIX

A.1 *Proof of Theorem 1*

Consider $Y_n = (X_n, z_n)$ as a three-state Markov chain with transition matrix

$$\begin{array}{cc} & \begin{array}{ccc} (\mathrm{G}, 0) & (\mathrm{B}, 0) & (\mathrm{B}, 1) \end{array} \\ \begin{array}{c} (\mathrm{G}, 0) \\ (\mathrm{B}, 0) \\ (\mathrm{B}, 1) \end{array} & \left[\begin{array}{ccc} 1 - P & hP & (1-h)P \\ p & h(1-p) & (1-h)(1-p) \\ p & h(1-p) & (1-h)(1-p) \end{array}\right] \end{array} = Q = (q_{ij}), \quad (31)$$

say. We have

$$
\begin{aligned}
Eu^{Z_n} &= Eu^{z_1+z_2+\cdots+z_n}\\
&= \textstyle\sum_{y_0}\sum_{y_1}\sum_{y_2}\cdots\sum_{y_n} u^{z_1+\cdots+z_n} q_{y_{n-1}y_n}\\
&\qquad\qquad\qquad\qquad \cdot q_{y_{n-2}y_{n-1}}\cdots q_{y_0y_1}\cdot\lambda_{y_0}, \qquad (32)
\end{aligned}
$$

where $\lambda_{y_0} = P[Y_0 = y_0]$. Note that the value of z_ℓ is completely determined by the value $Y_\ell = y_\ell$. Let

$$r_{y_{\ell-1},y_\ell} = u^{z_\ell} q_{y_{\ell-1},y_\ell},$$

and let

$$R = (r_{ij}).$$

Relation (32) can then be written as

$$Eu^{Z_n} = \boldsymbol{\lambda}' R^n \mathbf{1}, \qquad (33)$$

where

$$
\begin{aligned}
\boldsymbol{\lambda}' &= (\lambda_{(G,0)}, \lambda_{(B,0)}, \lambda_{(B,1)}),\\
\mathbf{1}' &= (1, 1, 1),\\
R &= (r_{ij})\\
&= Q\begin{bmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & u \end{bmatrix}.
\end{aligned}
$$

We remark here that eq. (33) can be extended to the case of an s-state Markov chain easily. Letting $u = 0$ in eq. (33), the PGF of Z_n, we have

$$P[Z_n = 0] = \boldsymbol{\lambda}'\left[Q\begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0 \end{pmatrix}\right]^n \mathbf{1}\cdot \qquad (34)$$

The last column of the 3×3 matrix in eq. (34) is always a zero vector for every $n \geqq 1$. Hence, the right-hand side of (34) is essentially the nth power of a 2×2 matrix. The explicit formula for g_n in eq. (4) follows from (34) by straightforward calculations.

A.2 *Proof of Theorem 2*

The z's are conditionally independent if the values of the X's are given. Hence,

$$
\begin{aligned}
P[Z_n = 0] &= E\{P[z_1 = z_2 = \cdots = z_n = 0 \mid X_1, X_2, \cdots, X_n]\}\\
&= E\left\{\prod_{i=1}^{n} P[z_i = 0 \mid X_i]\right\}\\
&= E\prod_{i=1}^{n} h^{X_i - 1}\\
&= Eh^{X_1+X_2+\cdots+X_n-n}. \qquad (35)
\end{aligned}
$$

Similarly,

$$\begin{aligned} Eu^{Z_n} &= E\{E[u^{z_1+z_2+\cdots+z_n}|X_1, X_2, \cdots, X_n]\} \\ &= E\left\{\prod_{i=1}^{n} E[u^{z_i}|X_i]\right\} \\ &= E\left\{\prod_{i=1}^{n} [h + (1-h)u]^{X_i-1}\right\} \\ &= EH^{X_1+X_2+\cdots+X_n-n}, \end{aligned} \tag{36}$$

where $H = h + (1-h)u$. By comparing eqs. (35) and (36), Theorem 2 follows.

A.3 *Proof of Theorem 3*

By eq. (10),

$$P[X_n = 2|X_0 = 2] = \pi_2 + \lambda^n\pi_1.$$

Hence,

$$\begin{aligned} \text{Cov}\,(z_i, z_j) &= P[z_i = z_j = 1] - \rho^2 \\ &= (1-h)^2P[X_i = X_j = 2] - \rho^2 \\ &= (1-h)^2\pi_2(\pi_2 + \pi_1\lambda^{|i-j|}) - \rho^2 \\ &= \pi_1\pi_2(1-h)^2\lambda^{|i-j|}. \quad \text{QED.} \end{aligned}$$

A.4 *Proof of Theorem 4*

Let us compute a special case first. Consider $P[T_1 = 0, T_n = 0]$. A typical path of the underlying Markov chain $\{X_1, X_2, \cdots, X_{kn}\}$ may be of the following form:

$$\begin{aligned} &b(x_k, x_{k+1}, x_{(n-1)k}, x_{(n-1)k+1}) \\ &\qquad = (\underbrace{x_1x_2\cdots x_k}_{\text{first block of size } k}x_{k+1}\cdots x_{(n-1)k}\underbrace{x_{(n-1)k+1}x_{(n-1)k+2}\cdots x_{nk}}_{\text{last block of size } k}). \end{aligned} \tag{37}$$

The rest of the x's in b (and in W_1, W_2 later) are omitted for typographical reasons. Note that the values of $x_{k+2}, \cdots, x_{(n-1)k-1}$ are deliberately unspecified; also, $n > 2$ is assumed *pro tem.*

For fixed first block and last block, there are four different kinds of paths, according to the values of x_{k+1}, $x_{(n-1)k}$. Let m denote the number of 2's in the first and last blocks together. We have

$$P[T_1 = 0, T_n = 0|b(x_k, x_{k+1}, x_{(n-1)k}, x_{(n-1)k+1})] = h^m \tag{38}$$

and

$$\begin{aligned} &P[b(x_k, x_{k+1}, x_{(n-1)k}, x_{(n-1)k+1})] \\ &\qquad = W_1(x_k)p_{x_kx_{k+1}}p_{x_{k+1}x_{(n-1)k}}^{[(n-2)k-1]}p_{x_{(n-1)k}x_{(n-1)k+1}}W_2(x_{(n-1)k+1}), \end{aligned} \tag{39}$$

where

$$W_1(x_k) = P[X_1 = x_1, X_2 = x_2, \cdots, X_{k-1} = x_{k-1}, X_k = x_k]$$

and

$$W_2(x_{(n-1)k+1}) = P[X_{(n-1)k+2} = x_{(n-1)k+2}, \cdots, X_{nk} = x_{nk} \\ \mid X_{(n-1)k+1} = x_{(n-1)k+1}].$$

We may also find $P[T_1 = 0] \cdot P[T_n = 0]$ by considering their conditional probabilities over the first and the nth blocks. It is not difficult to see that

$$P[T_1 = 0]P[T_n = 0] = \sum h^m W_1(x_k) W_2(x_{(n-1)k+1}) \cdot \pi_{x_{(n-1)k+1}}, \tag{40}$$

where the summation ranges over all 2^{2k} possible blocks. The expression for $P[T_1 = 0, T_n = 0]$ can be obtained by taking the product of (38) and (39) and summing over all 2^{2k+2} possibilities. The 2^{2k+2} terms in this form of $P[T_1 = 0, T_n = 0]$ outnumbered the terms in (40) by a margin of 4 to 1, and there is an obvious 4:1 correspondence between these terms. Consider

$$\begin{aligned} \text{Cov}\,(T_1, T_n) &= \text{Cov}\,(1 - T_1, 1 - T_n) \\ &= P[T_1 = 0, T_n = 0] - P[T_1 = 0]P[T_n = 0]. \end{aligned} \tag{41}$$

For fixed first and last blocks, a typical difference between the (4:1) correspondent terms is

$$\begin{aligned} h^m[W_1(x_k) W_2(x_{(n-1)k+1})][p_{x_k 1} p_{11}^{[(n-2)k-1]} p_{1x_{(n-1)k+1}} \\ + p_{x_k 1} p_{12}^{[(n-2)k-1]} p_{2x_{(n-1)k+1}} + p_{x_k 2} p_{21}^{[(n-2)k-1]} p_{1x_{(n-1)k+1}} \\ + p_{x_k 2} p_{22}^{[(n-2)k-1]} p_{2x_{(n-1)k+1}} - \pi_{x_{(n-1)k-1}}]. \end{aligned} \tag{42}$$

By (10), it can be shown that the third factor of (42) becomes

$$(p_{x_k 1} p_{1x_{(n-1)k+1}} + p_{x_k 1} p_{2x_{(n-1)k+1}} + p_{x_k 2} p_{1x_{(n-1)k+1}} + p_{x_k 2} p_{2x_{(h-1)k+1}}) \lambda^{(n-2)k-2}$$

$$= \begin{cases} \dfrac{P}{p + P} \lambda^{(n-2)k+1} & \text{if} \quad (x_k, x_{(n-k)+1}) = (1, 1) \\ -\dfrac{P}{p + P} \lambda^{(n-2)k+1} & \qquad = (1, 2) \\ -\dfrac{p}{p + P} \lambda^{(n-2)k+1} & \qquad = (2, 1) \\ \dfrac{p}{p + P} \lambda^{(n-2)k+1} & \qquad = (2, 2). \end{cases} \tag{43}$$

Note that in all terms we have a common factor $\lambda^{(n-2)k+1}$. By factoring out this common factor, Theorem 4 follows immediately.

We may even push the computations further to find an exact expression for the constant C_1 in Theorem 4. Note that we have four types of combinations of blocks, according to the values of x_k and $x_{(n-k)+1}$. The quantity in (42) becomes

$$\begin{aligned}
&(1, 1) \Rightarrow h^m W_1(1) W_2(1) \frac{P}{p+P} \lambda^{(n-2)k+1}, \\
&(1, 2) \Rightarrow -\ h^m W_1(1) W_2(2) \frac{P}{p+P} \lambda^{(n-2)k+1}, \\
&(2, 1) \Rightarrow -\ h^m W_1(2) W_2(1) \frac{p}{p+P} \lambda^{(n-2)k+1}, \\
&(2, 2) \Rightarrow h^m W_1(2) W_2(2) \frac{p}{p+P} \lambda^{(n-2)k+1},
\end{aligned} \tag{44}$$

and Cov (T_1, T_n) is the sum of all 2^{2k} terms in (44).

Let

$$T_1' = z_1 + z_2 + \cdots + z_{k-1}, \qquad T_2' = z_{k+2} + \cdots + z_{2k}.$$

(We should use $\mathrm{T}_n' = z_{(n-1)k+2} + \cdots + z_{nk}$; however, the distribution of T_n' is independent of n so we may take $n = 2$.) Then

$$\begin{aligned}
P[T_1' &= 0 \mid X_k = 1] P[T_2' = 0 \mid X_{k+1} = 1] \\
&= \sum_{2^{2(k-1)}} h^m P[X_1 = x_1, \cdots, X_{k-1} = x_{k-1} \mid X_k = 1] \\
&\qquad\qquad \cdot P[X_{k+2} = x_{k+2}, \cdots, X_{2k} = x_{2k} \mid X_{k+1} = 1] \\
&= \frac{1}{\pi_1} \sum_{2^{2(k-1)}} h^m W_1(1) W_2(1),
\end{aligned}$$

where m denotes the number of 2's in the sequence $x_1, x_2, \cdots, x_{k-1}, x_{k+2}, \cdots, x_{2k}$, which is equal to the number of 2's in the sequence $x_1, x_2, \cdots, x_{2k}$ in the case $x_k = x_{k+1} = 1$. Thus, the sum of terms of the type (1, 1) in (44) is simply

$$\pi_1 P[T_1' = 0 \mid X_k = 1] P[T_2' = 0 \mid X_{k+1} = 1] \cdot \pi_2 \lambda^{(n-1)k+1}.$$

Similarly, we may find the sums of other types of terms in (44). We have

$$\begin{aligned}
&(1, 2) \Rightarrow -\ \pi_1 h P[T_1' = 0 \mid X_k = 1] P[T_2' = 0 \mid X_{k+1} = 2] \pi_2 \lambda^{(n-2)k+1} \\
&(2, 1) \Rightarrow -\ \pi_2 h P[T_1' = 0 \mid X_k = 2] P[T_2' = 0 \mid X_{k+1} = 1] \cdot \pi_1 \lambda^{(n-2)k+1} \\
&(2, 2) \Rightarrow \pi_2 h^2 P[T_1' = 0 \mid X_k = 2] P[T_2' = 0 \mid X_{k+1} = 2] \pi_1 \lambda^{(n-2)k+1}.
\end{aligned}$$

Thus, if $n > 2$, $i \geqq 1$,

$$\begin{aligned}\operatorname{Cov}(T_1, T_n) &= \operatorname{Cov}(T_i, T_{i+n-1}) \\ &= \pi_1\pi_2\lambda^{(n-2)k+1}\{P[T_1' = 0 \mid X_k = 1]P[T_2' = 0 \mid X_{k+1} = 1] \\ &\qquad - hP[T_1' = 0 \mid X_k = 1]P[T_2' = 0 \mid X_{k+1} = 2] \\ &\qquad - hP[T_1' = 0 \mid X_k = 2]P[T_2' = 0 \mid X_{k+1} = 1] \\ &\qquad\quad + h^2P[T_1' = 0 \mid X_k = 2]P[T_2' = 0 \mid X_{k+1} = 2]\}. \end{aligned} \tag{45}$$

The case $n = 2$ should be considered separately; this is because $(n - 2)k - 1 < 0$ if $n = 2$ so that (39) simplifies to

$$P[b(x_k, x_{k+1})] = W_1(x_k)p_{x_kx_{k+1}}W_2(x_{k+1}). \tag{46}$$

In this case, the number of terms in $P[T_1 = 0, T_2 = 0]$ equals the number of terms in the product $P[T_1 = 0]P[T_2 = 0]$ and there is an obvious one-to-one correspondence between the terms. Consider the difference $P[T_1 = 0, T_2 = 0] - P[T_1 = 0]\ P[T_2 = 0]$. For fixed first and last (second) blocks, the term-wise difference is

$$h^mW_1(x_k)W_2(x_{k+1})[p_{x_kx_{k+1}} - \pi_{x_{k+1}}]. \tag{47}$$

The last factor in (47) can be computed. We have

$$\begin{array}{rlll} p_{x_kx_{k+1}} - \pi_{x_{k+1}} = & \dfrac{P}{p+P}\lambda & \text{if} & (x_k, x_{k+1}) = (1, 1) \\ = & -\dfrac{P}{p+P}\lambda & & = (1, 2) \\ = & -\dfrac{p}{p+P}\lambda & & = (2, 1) \\ = & \dfrac{p}{p+P}\lambda & & = (2, 2). \end{array} \tag{48}$$

Note that (43) reduces to (48) if $n = 2$; hence, all arguments leading to (45) hold true even if $n = 2$.

Let C_2 be the quantity in the large square bracket of (45); we may write (45) as

$$\operatorname{Cov}(T_i, T_j) = C_2\pi_1\pi_2\lambda^{(|i-j|-1)k+1} \tag{49}$$

if $i \leqq j$.

It is possible to find the value of C_2 through an argument similar to that of finding g_n in Theorem 1. However, we shall be satisfied with a crude estimate

$$|\pi_1\pi_2C_2| \leqq 1,$$

which follows from $\pi_1\pi_2 = \pi_1(1 - \pi_1) \leqq \frac{1}{4}$ trivially.

A.5 *Proof of Theorem 5*

Let $H = (1 - h)u + h$ and let $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{A}_1$, $\hat{A}_2$, $\hat{\rho}$ be the quantities obtained from α_1, α_2, A_1, A_2, ρ by replacing each h with H respectively. As $n \to \infty$, $H \to 1$. Hence,

$$\begin{aligned} \hat{\alpha}_1 &\to 1 \\ \hat{\alpha}_2 &\to 1 - p < 1 \\ \hat{A}_1 &\to 0 \\ \hat{A}_2 &\to 0 \\ \hat{\rho} &\to 0. \end{aligned}$$

It follows from Theorem 2 that

$$\lim_{n \to \infty} Eu^{Z_n} = \lim_{n \to \infty} \frac{\hat{A}_1 \hat{\alpha}_1{}^{n+1}}{1 - \hat{\alpha}_1}. \tag{50}$$

Let $\Delta^2 = [1 - P - H(1 - p)]^2 + 4pPH$ in the expression of $\hat{\alpha}_1$. An important step in our argument is to find the value of Δ. By substituting (24) into the expression for Δ^2, we have

$$\Delta^2 = p^2 + \sum_{n=1}^{\infty} \gamma_n x^n + \sum_{n=1}^{\infty} \delta_n y^n + \epsilon xy + o(xy), \tag{51}$$

where

$$\begin{aligned} \gamma_{2n} &= (1 - p)^2(1 - u)^2(a_n^2 + 2a_1a_{2n-1} + 2a_2a_{2n-2} + \cdots \\ &\qquad + 2a_{n-1}a_{n+1}) + 2p(1 - p)(1 - u)a_{2n} \\ \gamma_{2n+1} &= (1 - p)^2(1 - u)^2(2a_1a_{2n} + 2a_2a_{2n-1} + \cdots + 2a_na_{n+1}) \\ \delta_{2n} &= b_n^2 + 2b_1b_{2n-1} + 2b_2b_{2n-2} + \cdots + 2b_{n-1}b_{n+1} \\ \delta_{2n+1} &= 2b_1b_{2n} + 2b_2b_{2n-1} + \cdots + 2b_nb_{n+1} \\ \epsilon &= -2(1 - p)(1 - u)a_1b_1 - 4p(1 - u)a_1b_1. \end{aligned}$$

Let

$$\Delta = p + \sum_{n=1}^{\infty} (d_nx^n + e_ny^n) + fxy + o(xy). \tag{52}$$

By comparing the Δ^2 in (52) with the same quantity in (51), it is not difficult to see that

$$\begin{aligned} d_k &= (1 - p)(1 - u)a_k \\ e_k &= b_k \end{aligned} \tag{53}$$

for $k = 1, 2, \cdots$. Also, it is easy to find that

$$f = -\frac{2}{p}(1 - u)a_1b_1. \tag{54}$$

Using (53), it can be seen that in the expression of $\hat{\alpha}_1$, the coefficients of x^k, y^k are zero for all k. Hence, recall that $xy = 1/n$,

$$\begin{aligned}\hat{\alpha}_1 &= 1 - \frac{(1-u)a_1b_1}{p}\,xy + o(xy)\\ &= 1 - \frac{(1-u)a_1b_1}{np} + o\left(\frac{1}{n}\right).\end{aligned} \tag{55}$$

By (55) and (24), it can be seen that

$$\hat{A}_1 = \frac{a_1b_1(1-u)}{pn} + o\left(\frac{1}{n}\right). \tag{56}$$

By (50), (55), and (56), we have

$$\lim_{n\to\infty} Eu^{Z_n} = \exp\left(\frac{a_1b_1(u-1)}{p}\right),$$

which is the PGF of the Poisson distribution with mean equal to a_1b_1/p.

A.6 *Proof of Theorem 6*

Using (26), we may express $\hat{\alpha}_1$, $\hat{\alpha}_1$ in terms of powers of $1/n$ as

$$\hat{\alpha}_1 = 1 - \frac{\alpha}{n} + o\left(\frac{1}{n}\right), \tag{57}$$

$$\hat{A}_1 = \frac{\alpha}{n} + o\left(\frac{1}{n}\right),$$

where

$$\begin{aligned}\alpha &= \frac{b_1(1-H)}{1-H+pH}\\ H &= (1-h)u + h.\end{aligned} \tag{58}$$

By eqs. (50), (57), and (58), we have

$$\begin{aligned}\lim_{n\to\infty} Eu^{Z_n} &= \lim_{n\to\infty}\frac{\hat{A}_1}{1-\hat{\alpha}_1}\lim_{n\to\infty}\hat{\alpha}_1{}^{n+1}\\ &= \exp\left[-\alpha\right]\\ &= \exp\left[-\frac{b_1(1-H)}{1-H+pH}\right].\end{aligned} \tag{59}$$

This proves Theorem 6.

REFERENCES

1. Gilbert, E. N., "Capacity of a Burst-Noise Channel," B.S.T.J., *39*, No. 5 (September 1960), pp. 1253–1266.

2. Elliott, E. O., "Estimates of Error Rates for Codes on Burst-Noise Channels," B.S.T.J., *42*, No. 5 (September 1963), pp. 1977–1998.
3. Elliott, E. O., "A Model of the Switched Telephone Network for Data Communications," B.S.T.J., *44*, No. 1 (January 1965), pp. 89–109.
4. McCullough, R. H., "The Binary Regenerative Channel," B.S.T.J., *47*, No. 8 (October 1968), pp. 1713–1735.
5. Balkovic, M. D., Klancer, H. W., Klare, S. W., and McGruther, W. G., "1969–70 Connection Survey: High-Speed Voiceband Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1349–1384.
6. Fleming, H. C., and Hutchinson, R. M., Jr., "1969–70 Connection Survey: Low-Speed Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1385–1405.
7. Chung, K. L., *Markov Chains with Stationary Transition Probabilities*, Part I, Sections 15–16, New York: Springer-Verlag, 1967.
8. Cox, D. R., and Lewis, P. A. W., *The Statistical Analysis of Series of Events*, Chapter 7, London: Methuen & Co. Ltd., 1966.
9. Roberts, S. W., "States of Markov Chains for Evaluating Continuous Sampling Plans," Trans. 17th Annual All-Day Conference on Quality Control, Metropolitan Section ASQC and Rutgers University, New Brunswick, New Jersey (1965), pp. 106–111.
10. Lieberman, G. J., and Solomon, H., "Multilevel Continuous Sampling Plans," Ann. Math. Stat., *26* (1955), pp. 686–704.
11. Feller, W., "Fluctuation Theory of Recurrent Events," Trans. Amer. Math. Soc., *67* (1949), pp. 98–119.
12. Sussman, S. M., "Analysis of the Pareto Model for Error Statistics on Telephone Circuits," IEEE Trans. Commun. Syst., *11*, No. 2 (June 1963), pp. 213–221.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

The Metallurgy of Remendur: Effects of Processing Variations

By M. R. PINNEL and J. E. BENNETT

(Manuscript received May 14, 1973)

A recent development effort in telecommunications switching apparatus has been directed toward the production of a remanent reed, dry, sealed contact (remreed). Remendur, a medium-hard magnetic alloy nominally composed of equal parts iron and cobalt and 2.7-wt. percent vanadium, was chosen as the reed material in this contact. However, the application required the alloy to possess rather specific magnetic and mechanical properties and considerable difficulty was experienced in consistently processing Remendur into wire with these specified properties. To ascertain the sensitivity of these properties to variations in processing times and temperatures, and vanadium content, two melts of Remendur (2.5-percent V and 3.0-percent V) were processed with selected alterations in annealing temperatures at several stages. Microstructures were characterized following each step by light microscopy and were correlated with the appropriate ternary equilibrium diagram. Results demonstrate that microstructures developed by anneals between 900°C and 950°C are extremely sensitive to the precise temperature of the anneal and composition of the alloy. The microstructure, which strongly influences magnetic and mechanical properties, can be varied over the limits of the two-phase $\alpha_1 + \gamma$ *region by variations in vanadium content of only 0.5 wt. percent and by the small 50°C temperature range.*

I. INTRODUCTION

Historically, considerable difficulty has been experienced in consistently processing the Fe-Co-2.7-wt. percent V alloy (Remendur) into wire with specified magnetic and mechanical properties. A major cause for these problems has been that the metallurgy of this ternary system was not sufficiently understood. A number of investigations have provided information on various aspects of phase equilibria[1–5] and kinetics and mechanisms of phase transformations[6–10] in this system.

The most notable of these are the work of Köster and Schmid[1] on the phase equilibrium of the system and the work of English[6] and Martin and Geisler[2] on the (FCC) ⇆ (BCC) transformation. However, discrepancies still exist on aspects of phase equilibrium,[1,2,4,7] mechanical ductility,[2,6,7,11] transformation kinetics,[2,6,7,10] and the influence of these factors on the development of magnetic properties.[2,6,7,10,12,13] These discrepancies made analysis of the sensitivity of the mechanical and magnetic properties to composition and to heat treating times and temperatures during processing unreliable with further investigation. This analysis is vital since 2.7-percent V Remendur is being used as the reed material in the remreed contact.[14] This paper describes the characterization of low-vanadium (2–3-wt. percent) Remendur, mainly by light microscopy, for the various processing steps from 6.35-mm (0.25-inch) rod through 0.53-mm (0.021-inch) wire to 0.18-mm (0.007-inch) flattened strip. Correlation of the microstructures with the equilibrium phase diagram of Köster and Schmid[1] is provided. Aspects of the microstructures in relation to cold ductility and magnetic properties are discussed.

II. MATERIALS AND EXPERIMENTAL PROCEDURES

Two melts of Remendur, each with a different vanadium content, were used in these experiments. The alloys were prepared by Battelle Memorial Institute with nominal compositions of 2.5-percent V-balance equal Fe/Co and 3-percent V-balance equal Fe/Co. The actual analyses as determined by atomic absorption spectroscopy are given in Table I. The typical Remendur processing sequence from melting through final reed fabrication is outlined in Fig. 1. The influence on cold ductility, yield and tensile strength, resistivity, and magnetic properties of several steps in this processing sequence was unclear. These steps included temperature of and rate of cooling from the 6.35-mm rod anneal, the need for intermediate strand anneals, and the temperature of the 0.53-mm wire strand anneal. Also, the influence of the stamping operation relative to the unworked shank on final reed properties was of interest.

TABLE I—REMENDUR COMPOSITIONAL ANALYSES
(wt. %)

	V	Co	Fe
3% V Battelle	2.97	48.70	47.34
2.5% V Battelle	2.46	48.36	48.27

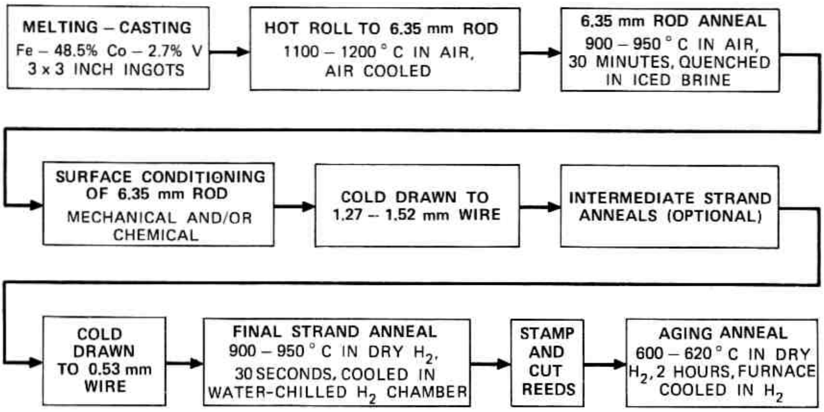
MELTING – CASTING
Fe – 48.5% Co – 2.7% V
3 x 3 INCH INGOTS
HOT ROLL TO 6.35 mm ROD
1100 – 1200 °C IN AIR, AIR COOLED
6.35 mm ROD ANNEAL
900 – 950 °C IN AIR, 30 MINUTES, QUENCHED IN ICED BRINE
SURFACE CONDITIONING OF 6.35 mm ROD
MECHANICAL AND/OR CHEMICAL
COLD DRAWN TO 1.27 – 1.52 mm WIRE
INTERMEDIATE STRAND ANNEALS (OPTIONAL)
COLD DRAWN TO 0.53 mm WIRE
FINAL STRAND ANNEAL
900 – 950 °C IN DRY H2, 30 SECONDS, COOLED IN WATER-CHILLED H2 CHAMBER
STAMP AND CUT REEDS
AGING ANNEAL
600 – 620 °C IN DRY H2, 2 HOURS, FURNACE COOLED IN H2

Fig. 1—Typical Remendur processing.

An experimental program was developed to provide clarification of the influences of these factors. A synopsis of this program is given in Fig. 2. It involved carrying two compositions of Remendur through the entire processing sequence with appropriate variations in parameters at each critical stage. For temperature variation, a matrix of temperatures between 900°C and 950°C was investigated. This range was based on previous statements by Gould and Wenny that cold ductility could be obtained by an iced brine quench from 925°C[12] and the current use of this temperature by K. Olsen (Bell Laboratories–Murray Hill) for wire processing.[15]

Microstructures were evaluated primarily by optical microscopy. Metallographic preparation was routine with a 5-percent Nital solution used to lightly etch (10–50 seconds) the polished surfaces. Observation and photography were carried out on a Ziess Ultraphot III metallograph using Nomarski Differential Interference Contrast (DIC). Indications of the ductility of both hot-rolled and annealed 6.35-mm rod were provided by the drawability of the material.

III. RESULTS AND DISCUSSION

3.1. *Hot-Rolled 6.35-mm (0.25-inch) Rod*

A vertical section through a portion of the Fe-Co-V ternary equilibrium phase diagram (Köster and Schmid)[1] near the nominal composition of Remendur is given in Fig. 3. At 2.7-percent vanadium and temperatures above 950°C the alloy is single-phase FCC (γ). This is the structure of Remendur during hot rolling. The typical micro-

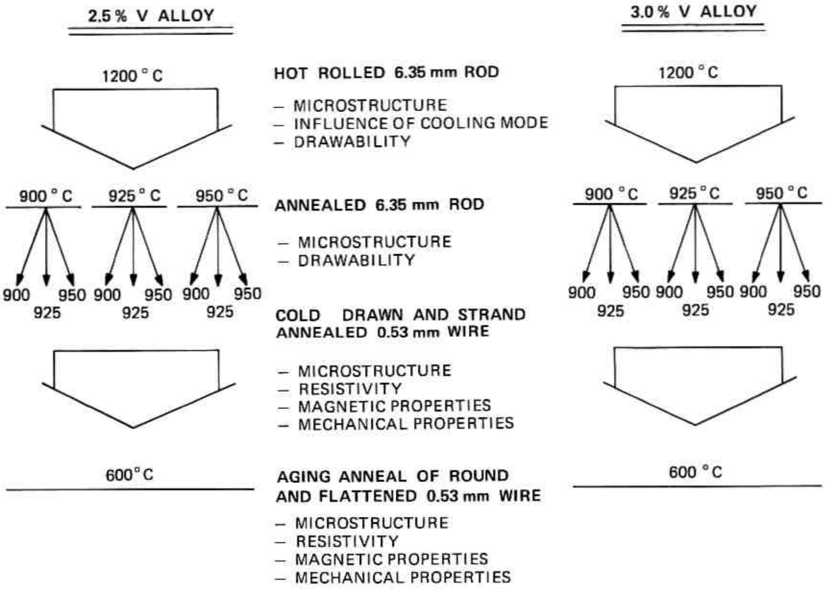

Fig. 2—Synopsis of Remendur processing experiment.

structure developed by cooling from the hot-rolling stage is shown in Fig. 5a. This microstructure is irregular and unconventional. English[6] and Chen[7,8] have amply reviewed the characteristics of this structure and only the most pertinent factors will be restated. Although this composition passes into the two-phase $\alpha_1 + \gamma$ field on cooling, the transformation is sluggish. As a result, with rapid cooling a nonequilibrium single-phase structure which is entirely BCC (α_1), supersaturated in vanadium, is developed (Fig. 5a). No retained γ was detected by X-ray analysis. This has been referred to as a "massive" transformation by English[6] and designated as the α_2 structure. This phase, being a nonequilibrium structure, does not appear on a conventional equilibrium phase diagram. Chen[7,8] proposed that the transformation is actually the more common "martensitic" type. However, the structure of the transformed material is similar for both types[6] and, hence, this will not be discussed here.

The ductility of Remendur with the all α_2 microstructure and as air cooled was tested in a qualitative manner. An attempt was made to draw 6.35-mm rod in this condition (i.e., as hot rolled) into 0.53-mm wire by successive reductions of approximately 20 percent in area. The rod could not survive the first draw pass due to extreme brittleness.

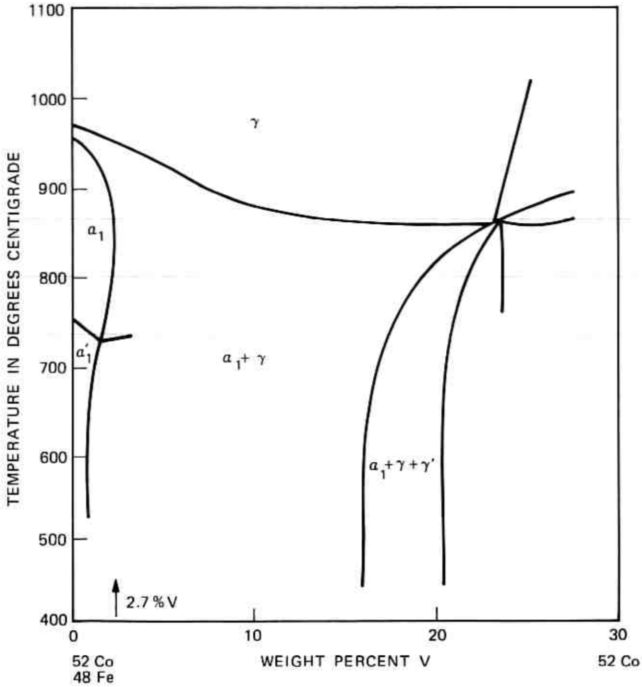

Fig. 3—Vertical section through Fe-Co-V ternary phase diagram, from the work of Köster and Schmid.[1]

3.2 *Annealed 6.35-mm (0.25-inch) Rod*

The brittleness of the air-cooled α_2 microstructure dictates the need for an additional heat treatment to create drawability. A treatment of one-half hour at 925°C (in the two-phase $\alpha_1 + \gamma$ field) followed by an iced brine quench had been found satisfactory in producing cold ductility.[12,15] However, the sensitivity of ductility to this precise time and temperature and the influence of the structure produced by this anneal on subsequent properties developed in the drawn wire had yet to be investigated.

Samples of both the 2.5-percent V and 3.0-percent V alloys in the hot-rolled and air-cooled condition were annealed at 950°C, 925°C, and 900°C for one-half hour and quenched in iced brine. The corresponding microstructures are shown in Figs. 4 and 5. It is apparent that a significant variation in structure occurs as a function of temperature over the relatively small range of 900 to 950°C. At the same annealing temperature, a significant variation as a function of the

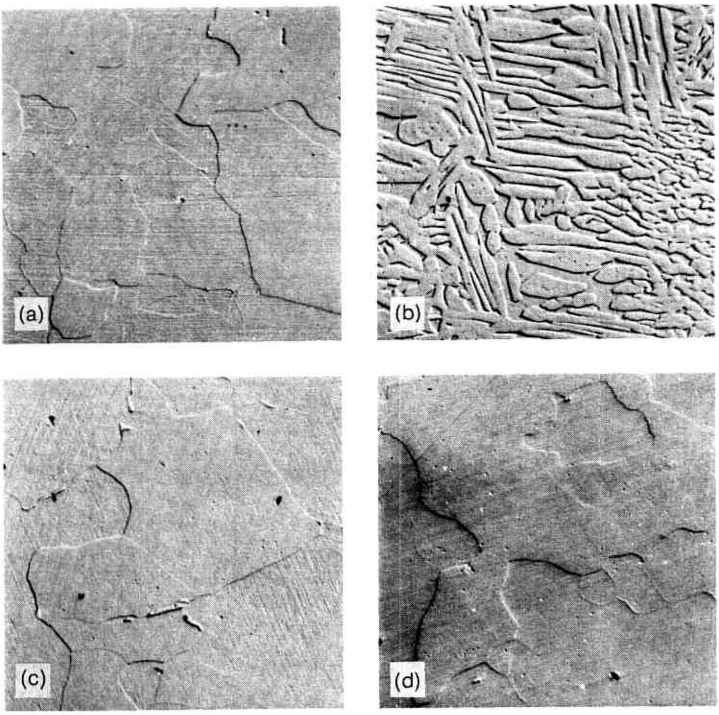

Fig. 4—Quenched microstructures of 2.5% V–6.35-mm Remendur rod in hot-rolled and various annealed conditions; etched; DIC; 600X. (a) Hot-rolled. (b) Annealed; ½ hour; 950°C. (c) Annealed; ½ hour; 925°C. (d) Annealed; ½ hour; 900°C.

small difference in vanadium composition (2.5 to 3.0 percent) is also obvious.

Analysis of the equilibrium phase diagram for this ternary system is necessary to understand the critical nature of temperature and composition for transformations in this range. Based on the vertical sections of the ternary space diagram provided by Köster and Schmid,[1] isotherms (horizontal sections) at temperatures of interest were constructed by the authors. The Fe-Co-rich portions of isotherms at 950, 925, and 900°C are drawn in Fig. 6. The nominal composition of Remendur (2.7-wt. percent V-balance equal Fe/Co) is represented by an X. At all three temperatures the equilibrium structure is in the two-phase ($\alpha_1 + \gamma$) region, but the amount of each phase can be seen to change radically with temperature. This is a consequence of the γ phase field shrinking rapidly toward the Co corner and the $\alpha_1 + \gamma$ phase field increasing significantly in area as the temperature decreases

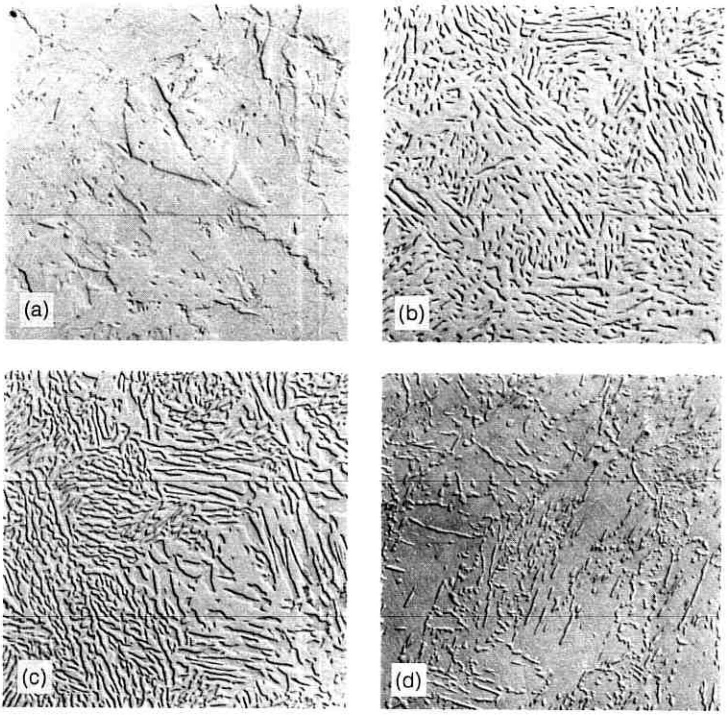

Fig. 5—Quenched microstructures of 3.0% V–6.35-mm Remendur rod in hot-rolled and various annealed conditions; etched; DIC; 600X. (a) Hot-rolled. (b) Annealed; ½ hour; 950°C. (c) Annealed; ½ hour; 925°C. (d) Annealed; ½ hour; 900°C.

from 950°C to 900°C. This α_1 phase (BCC) is an equilibrium solid solution of V in Fe-Co, in contrast to the supersaturated α_2 structure.

The microstructures of Figs. 4 and 5 can be correlated with these diagrams (Fig. 6). At 950°C, the compositions are near the γ phase boundary, hence the two-phase structure is primarily γ with some α_1 (Figs. 4b and 5b).*,† The 3.0-percent V material is practically all γ

* The γ phase formed by these anneals transforms to α_2 on quenching identical to the transformation from the single-phase γ region after the hot-rolling stage. Therefore, the two-phase microstructure is actually $\alpha_1 + \alpha_2$ at ambient temperature as photographed. However, the point of interest is the percentage of each phase formed at the elevated temperature which is identical to that existing at ambient for a quenched sample. Thus the designation $\gamma + \alpha_1$ will be maintained for clarity in relating the microstructures to the phase diagrams.

† The α_1 etches preferentially to the $\alpha_2(\gamma)$ using the Nital etch. And, under the interference contrast conditions used in all cases, the α_2 phase (raised) appears bright on the top edge as if the light source were shining from the 12-o'clock position. Therefore, for example, in viewing Figs. 4b and 5b the primary or major phase should appear raised, whereas in viewing Fig. 5d the minor phase or particles should appear raised.

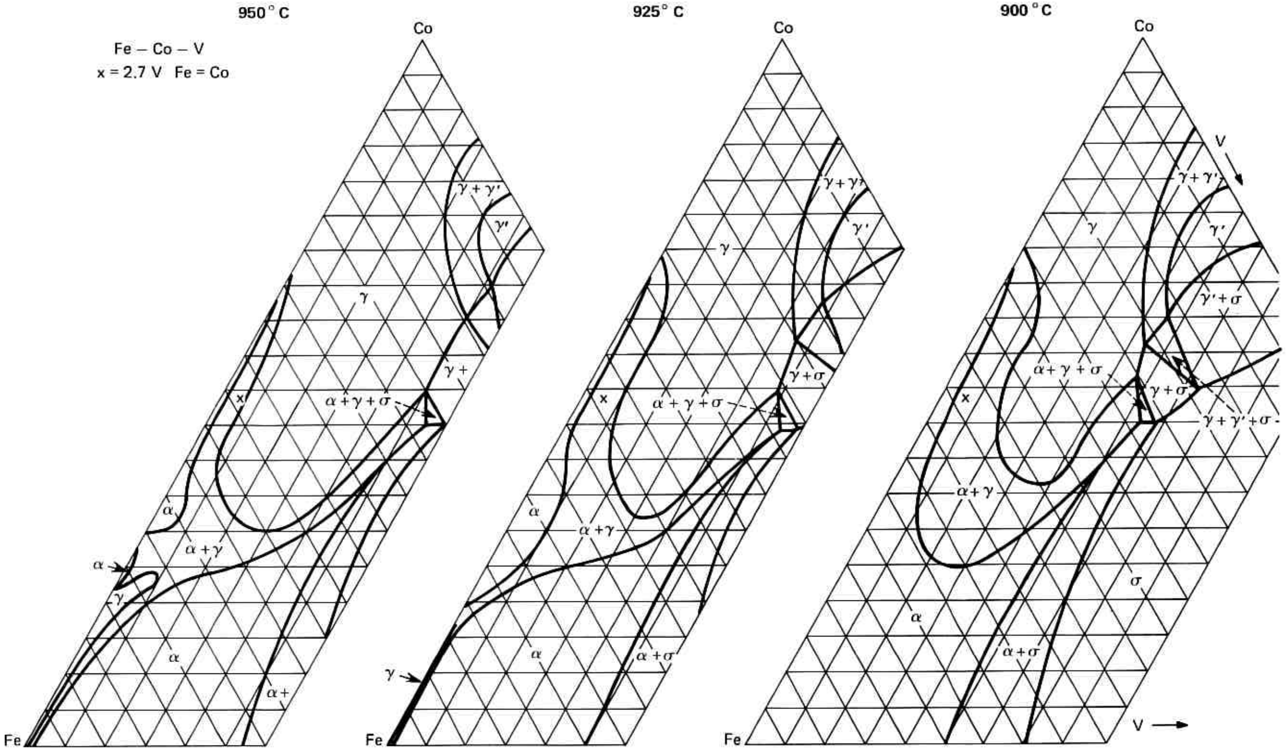

Fig. 6—Isotherms of the Fe-Co-rich portions of the Fe-Co-V ternary equilibrium diagram constructed from the vertical sections in the work of Köster and Schmid.[1]

(Fig. 5b) as this composition almost coincides with the phase boundary. At 900°C, the alloys are near the α_1 phase boundary, hence the two-phase structure is primarily α_1 with some γ (Figs. 4d and 5d). In this instance the lower-vanadium (2.5-percent) alloy is apparently all α_1 (Fig. 4d) as this composition almost coincides with the α_1 phase boundary. For anneals at 925°C, these alloys should have nearly equal proportions of α_1 and γ. This can be seen to be the case for the higher-vanadium alloy (Fig. 5c). However, very little γ is present in the low-vanadium material (Fig. 4c). This may be an indication of a non-equilibrium character in these anneals. Apparently, the more highly strained α_2 lattice of the 3.0-percent V alloy aids sufficiently in the kinetics of nucleation of the γ phase, such that this alloy approaches equilibrium more rapidly than the 2.5-percent V alloy.

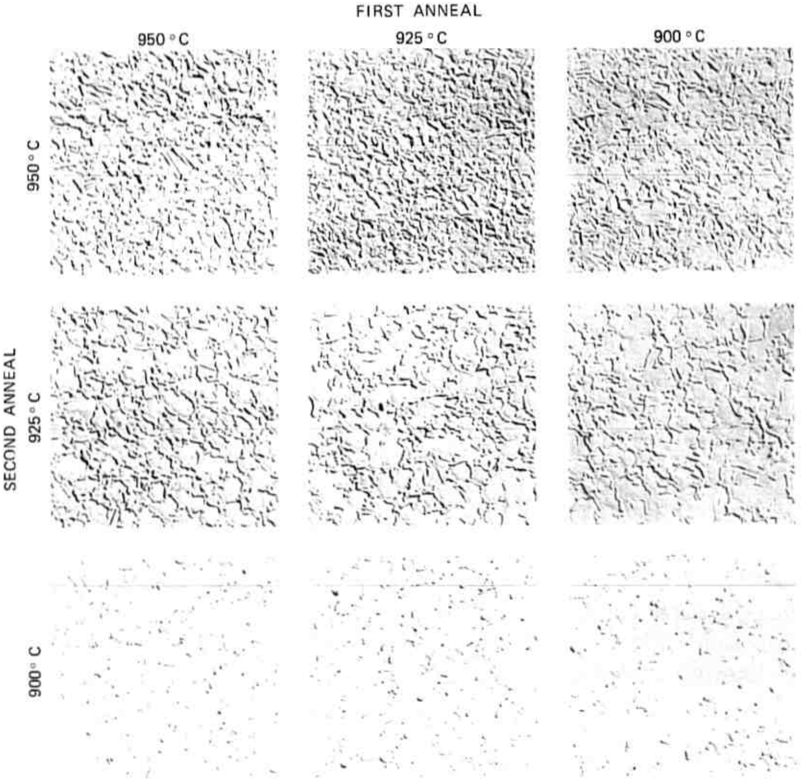

Fig. 7—Quenched microstructures of 2.5% V–0.53-mm Remendur wire; etched; DIC; 600X. First anneal indicates temperature of ½ hour anneal of 6.35-mm rod. Second anneal indicates temperature of ½ minute strand anneal of 0.53-mm wire.

A second influence of the higher vanadium content is apparent in comparing Figs. 4 and 5. A significant degree of grain size refinement occurs for the 3.0-percent V alloy.

All six of the annealed microstructures were found to be ductile. They were drawn from 6.35-mm rod to 0.53-mm wire by successive 20-percent reductions (23 passes) without any failures or evidence of cracking. This result clarifies several points with regard to ductility in Remendur. It has previously been shown that the air-cooled α_2 phase is brittle. English has shown that the α_1 phase orders (α_1') on other than very rapid rates of cooling[9] and alludes to the enhanced brittleness of this microconstituent.[6] Stoloff and Davies presented more complete evidence of the significant loss of ductility due to ordering.[11] Chen[8] and Davies[16] have shown that some degree of ductility also can be created in the α_2 constituent by the rapid cooling provided by a quench. However, Chen[8] demonstrated that the quenched single-phase α_2 structure does not provide the degree of ductility of the quenched two-phase structure. The influence of the quench on the α_2 constituent may also be related to the ordering phenomena although this has not yet been verified. Thus the two-phase structures developed by anneals between 900 and 950°C and followed by a quench are believed to provide the maximum degree of ductility attainable in this alloy. Anneals at temperatures below 900°C but above the ordering temperature ($\approx$ 720°C) followed by a quench may also produce a ductile structure. However, structures produced by lower-temperature anneals for drawability may be undesirable for developing proper magnetic properties in the final reed.

3.3 *Annealed 0.53-mm (0.021-inch) Wire*

Remendur rods of both compositions and all three annealing temperatures were drawn directly to 0.53-mm wire (6 samples). Sections of each sample were strand-annealed at furnace set temperatures of 900, 925, and 950°C for 30 seconds to produce a series of 18 samples with different combinations of composition, 6.35-mm anneal temperature, and 0.53-mm anneal temperature. The actual peak temperatures as determined by drawing a thermocouple through the furnace were 935°C, 913°C, and 890°C. The appropriate microstructures are shown in Figs. 7 and 8. The most significant points to be noted are as follows:

(*i*) For the *low*-vanadium alloy (Fig. 7) the second anneal uniquely determines the percentages of α_1 and γ. (Compare the identical microstructures of each horizontal row.) This is due to the fact that, con-

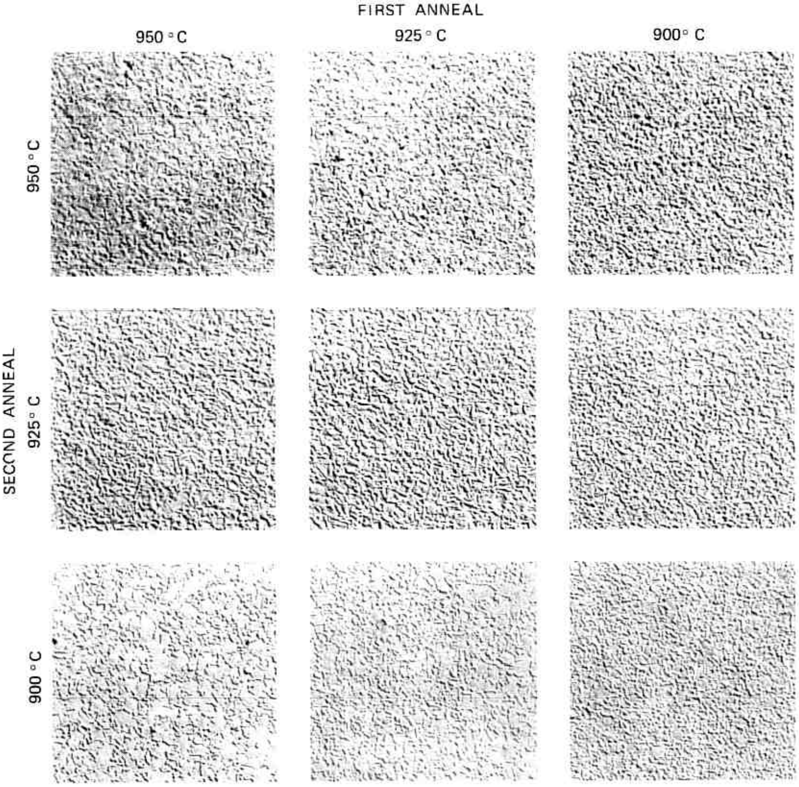

Fig. 8—Quenched microstructures of 3.0% V–0.53-mm Remendur wire; etched; DIC; 600X. First anneal indicates temperature of ½ hour anneal of 6.35-mm rod. Second anneal indicates temperature of ½ minute strand anneal of 0.53-mm wire.

trary to the case for the anneals of the 6.35-mm rod, the anneals in the 0.53-mm wire apparently produce a nearly equilibrium two-phase structure. This is likely a consequence of the aid to kinetics of nucleation provided by the deformation of wire drawing.

(*ii*) For the *high*-vanadium alloy (Fig. 8) the second anneal also primarily determines the percentages of $\alpha_1 + \gamma$ in qualitative agreement with the equilibrium diagrams (Fig. 6). The first anneals have little effect on phase percentages in the annealed wires indicating the 0.53-mm, 30-second strand anneals *may* produce equilibrium structures.

(*iii*) The same refinement in grain size of the high-V alloy compared to the low-V alloy previously noted is again present. An additional and substantial refinement occurs in the reduction of the wire from 6.35 to 0.53 mm (compare Figs. 5 and 8).

3.4. *600°C-Annealed 0.53-mm Wire and 0.18-mm Ribbon*

Each of the 18 samples was cut into two lengths. One length was rolled to a thickness of 0.18 mm (0.007 inch) and the other length remained as 0.53-mm wire. These conditions simulate the paddle and shank, respectively, of a typical reed. All 36 samples were given the standard 600°C, two-hour anneal in dry H_2 which is used to produce a square magnetic hysteresis loop with a specified range of coercive force.

The values of the coercive force developed for the various combinations of strand-annealing temperature and vanadium content in the round and flat sections are listed in Table II, and representative microstructures for both round and flattened sections are shown in Figs. 9 and 10. The details of the phase transformation which occurs during this 600°C anneal are presented in another paper[17] and will not be restated here. It is sufficient to note that the final microstructure developed in the round (undeformed) wire by this anneal is a strong

TABLE II—COERCIVE FORCE IN OERSTEDS OF 0.53-MM ROUND WIRE AND 0.18-MM FLAT RIBBON IN THE 600°C–ANNEALED CONDITION

(2.5 wt. % V)
First Anneal

		950°C		925°C		900°C	
		F	R	F	R	F	R
Second Anneal	950°C	22.9	11.9	23.7	11.9	24.1	11.5
	925°C	21.8	11.0	22.9	11.5	23.7	11.5
	900°C	22.6	11.9	22.9	8.3	23.3	5.1

(3.0 wt. % V)
First Anneal

		950°C		925°C		900°C	
		F	R	F	R	F	R
Second Anneal	950°C	29.7	18.2	30.8	20.2	30.8	20.6
	925°C	30.8	23.7	32.0	24.9	32.0	24.6
	900°C	30.5	17.0	30.8	17.8	31.3	21.0

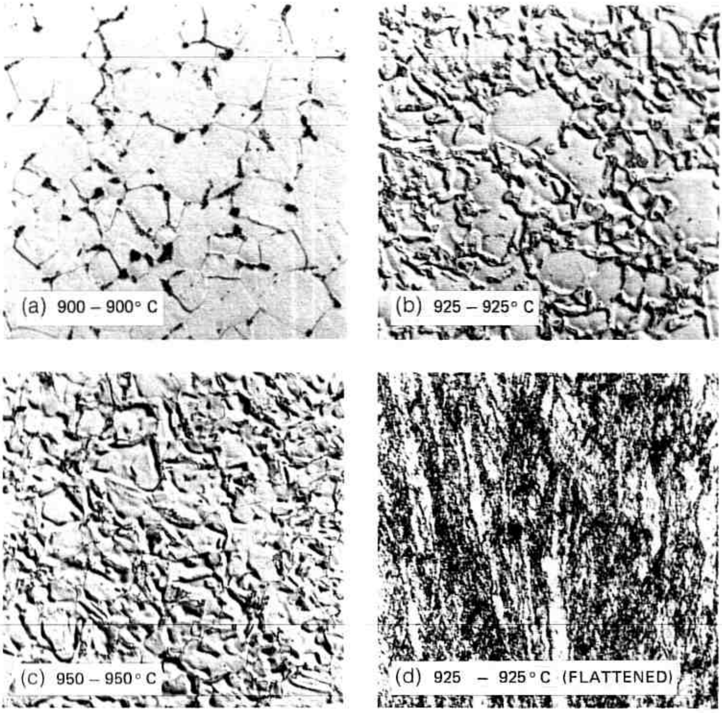

Fig. 9—Microstructures of 0.53-mm round wire and 0.18-mm flattened ribbon following 600°C, 2-hour anneal of 2.5% V alloy; etched; DIC; 1500X. Subcaptions specify temperatures of first and second anneals prior to 600°C anneal.

function of the microstructure from the strand-annealed state. This may be seen by comparing Fig. 7 to Fig. 9 and Fig. 8 to Fig. 10 for the low- and high-vanadium alloys, respectively. For the low-V alloy strand annealed at 900°C (Fig. 9a) the large grain α_1' matrix with particles of γ at the grain boundaries produces a coercive force of only 5.1 oersteds. The other structures show a coercive force of approximately 11.0 oersteds with a maximum value of only 11.9 oersteds. This demonstrates the significant influence of vanadium content on coercivity and the need to keep the minimum permissible vanadium content above the 2.5-wt. percent level to achieve the required minimum value of 18 oersteds on the round section of the reed for remreed applications.

The high-V alloy shows a clear correlation between magnetic properties after the 600°C anneal and the prior strand-annealed microstructure. As shown in Table II a peak in coercivity of the round wire is attained for the 925°C strand anneal of 0.53-mm wire. This is the

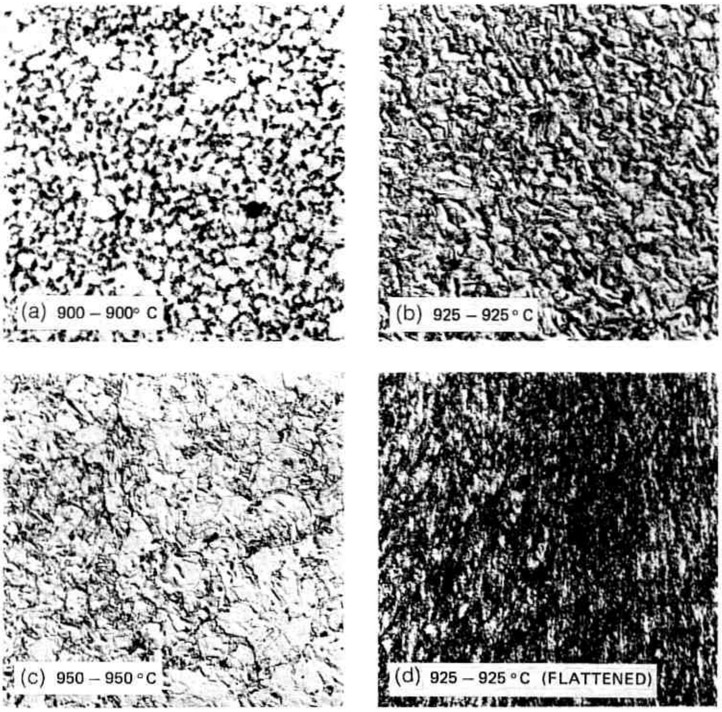

Fig. 10—Microstructures of 0.53-mm round wire and 0.18-mm flattened ribbon following 600°C, 2-hour anneal of 3.0% V alloy; etched; DIC; 1500X. Subcaptions specify temperatures of first and second anneals prior to 600°C anneal.

state of a nearly equal distribution of the two-phase $\alpha_1 + \alpha_2$ structure (Fig. 8) which produces a more uniform distribution of α_1' (ordered α_1) and γ (Fig. 10b) during the subsequent 600°C anneal. Coercivity decreases for strand anneals at higher or lower temperatures where the α_1 or α_2 phase predominates.

The flattening operation eliminates most of the microstructural influence on final magnetic properties as shown in Table II. For the high-V alloy, all values are in the range of 30 to 32 oersteds and for the low-V alloy in the range of 22 to 24 oersteds. All microstructures are similar to that shown in Figs. 9d and 10d which consists of the very fine distribution of γ in an α_1' matrix.

IV. CONCLUSIONS

The primary conclusion from this study is that the microstructure of 2–3-wt. percent vanadium Remendur is extremely sensitive to annealing temperature and vanadium content. Magnetic characteristics of Remendur wire and the remreed contact and the mechanical properties, particularly drawability, of the material are dependent on the precise microstructure developed during processing. More specific conclusions are listed as follows.

(*i*) Anneals in the range of 900 to 950°C produce a two-phase α_1 (BCC) $+ \gamma$ (FCC) structure which changes to a two-phase α_1 (BCC) $+\alpha_2$ (supersaturated BCC) structure on rapid cooling. The percentages of α_1 and γ formed are a strong function of temperature with increasing temperatures yielding greater percentages of γ (α_2).

(*ii*) The one-half-hour anneal of low-vanadium 6.35-mm rod does not produce the equilibrium two-phase structure defined by the phase diagram. Increased vanadium content assists in producing a nearly equilibrium structure as does the deformation of wire drawing for the anneal of 0.53-mm high- and low-vanadium wire.

(*iii*) The percentages of α_1 and γ (α_2) formed in the 0.53-mm wire are primarily a function of the temperature of the anneal. The temperature of the 6.35-mm rod anneal has little influence on determining these final percentages.

(*iv*) The all α_2 structure created by slow to moderate rates of cooling from the γ phase region is brittle and cannot be cold drawn. However, a two-phase mixture of $\alpha_1 + \alpha_2$, developed by quenching from anneals in the range of 900 to 950°C (the two-phase $\alpha + \gamma$ region), is sufficiently ductile to be drawn from 6.35-mm rod directly to 0.53-mm wire. If the rate of cooling is not sufficiently rapid from the 900 to 950°C range, the α_1 structure orders[9] to form α_1' which has also been shown to be brittle and, therefore, nondrawable to wire.[11]

(*v*) Increased vanadium content and the deformation of cold drawing both are influential in refining the grain size of the two-phase Remendur structure. The refined grain size and higher vanadium content influence the magnetic properties of 0.53-mm Remendur wire which has been annealed at 600°C for 2 hours by increasing the coercivity.

V. ACKNOWLEDGMENTS

The authors wish to express their appreciation to the following for assistance in various aspects of the study: G. V. McIlhargie and D. Heath for metallographic preparation; K. M. Olsen for processing of wire and magnetic data; W. E. Archer and J. A. Smiley for magnetic data; J. O. Rasmuson for compositional analysis; and A. T. English for helpful discussions on his prior studies of this topic.

REFERENCES

1. Köster, W., and Schmid, H., "Das Dreistoffsystem Eisen-Kobalt-Vanadin, Teil I and Teil II," Archiv f. das Eisenhüttenwesen, *26*, (1955), pp. 345–353, 421–425.
2. Martin, D. L., and Geisler, A. H., "Constitution and Properties of Cobalt-Iron-Vanadium Alloys," Trans. ASM, *44*, (1952), pp. 461–483.
3. Baer, G., and Thomas, H., "Eigenschaften und Aufbau von Eisen-Kobalt-Legierungen mit 50% Co and Kleinen Vanadinzusätzen," Z. Metallkde, *45*, (1954), pp. 651–655.
4. Fielder, H. C., and Davis, A. M., "The Formation of Gamma Phase in Vanadium Permendur," Met. Trans., *1*, (1970), pp. 1036–1037.
5. Koster, W., and Lang, K., "Die Kobaltecke des Systems Eisen-Kobalt-Vanadin," Z. Metallkde., *30*, (1938), pp. 350–352.
6. English, A. T., unpublished work, Bell Telephone Laboratories, Murray Hill, N. J., 1965.
7. Chen, C. W., "Metallurgy and Magnetic Properties of an Fe-Co-V Alloy," J. Appl. Phys., *32*, (1961), pp. 348s–355s.
8. Chen, C. W., "Soft Magnetic Cobalt-Iron Alloys," Cobalt, No. 22, (March 1964), pp. 3–21.
9. English, A. T., "Long-Range Ordering and Domain Coalescence Kinetics in Fe-Co-2V," Trans. AIME, *236*, (1966), pp. 14–20.
10. Josso, E., "Relations Between Structure and Properties in Magnetically Hard Fe-Co-V Alloys: Deficiencies of Equilibrium Diagrams," IEEE Trans. Magnetics, *MAG-6*, (1970), pp. 230–232.
11. Stoloff, N. S., and Davies, R. G., "The Plastic Deformation of Ordered FeCo and Fe_3Al Alloys," Acta Met., *12*, (1964), pp. 473–485.
12. Gould, H. L. B., and Wenny, D. H., "Supermendur: A New Rectangular-Loop Magnetic Material," Elect. Eng., *76*, (1957), pp. 208–211.
13. Ellis, W. C., and Greiner, E. S., "Equilibrium Relations in the Solid State of the Iron-Cobalt System," Trans. ASM, *29*, (1941), pp. 415–434.
14. Renaut, P. W., private communication, Bell Laboratories, Columbus, Ohio, March 12, 1971.
15. Olsen, K., private communication, Bell Laboratories, Murray Hill, N. J., October 21, 1971.
16. Davies, R., private communication, Western Electric Company, Allentown, Pennsylvania, April 26, 1972.
17. Bennett, J. E., and Pinnel, M. R., to be published in Met. Trans.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Optimum Mean-Square Decision Feedback Equalization

By J. SALZ

(Manuscript received April 20, 1973)

In this work we report new results relating to decision feedback equalization. The equalizer and the transmitting filter are optimized in a PAM data communication system operating over a linear noisy channel. We use a mean-square error criterion and impose an average power constraint at the transmitter. Assuming correct past decisions, an explicit formula for the minimum attainable mean-square error is given. The possible advantages of signaling faster than the Nyquist rate while decreasing the number of levels to maintain the same information rate are investigated. It is shown that, in all cases of practical interest, signaling faster than the Nyquist rate, while keeping fixed the information rate, increases the mean-square error. Finally, to illustrate the use of the results, application is made to a cable channel where the loss in dB varies as the square root of frequency. Various asymptotic formulas and curves are provided to exhibit the relationships between the quantities of interest.

I. INTRODUCTION

A great deal of research, particularly in the past decade, has been expended on the problem of linear equalization. This has yielded a considerable body of theory and technology making possible the design of apparatus for successfully combating intersymbol interference in PAM data transmission systems operating over noisy linear channels where delay distortion predominates. Since linear equalizers must compensate for the channel characteristics in the presence of noise, they cannot be expected to perform well over severely frequency-attenuating channels or channels possessing nulls in the amplitude characteristic.

Interest in the high data rates over voiceband and cable channels inevitably leads to the search for more effective equalization methods. Faster pulse rates place signal energy well within the badly attenuated portion of the transmission spectrum, resulting in severe intersymbol interference correctable by linear methods only at the expense of a significant enhancement of the noise.

A "bootstrap" technique, commonly referred to as "decision feedback," when combined with linear equalization can yield significant performance improvement.[1,2] In this method the samples of the pulse tails (postcursors) interfering with subsequent or future data symbols are subtracted without incurring a significant noise penalty. The effect of pulse tails (precursors) which occur prior to detection and interfere with past symbols is minimized by a conventional linear equalizer.

Much has been written about this subject. In a fundamental paper where an excellent bibliography can be found, Robert Price[3] demonstrated quantitatively the merits of decision feedback equalization in certain applications.

In this work we jointly optimize the receiving and transmitting filters in a PAM data transmission system employing decision feedback. The chief difference between our work and Price's is in the choice of performance criterion. We minimize mean-square error while Price maximized the signal-to-noise ratio under the constraint that the overall intersymbol interference be zero. Our criterion is not as stringent as Price's and allows trade-offs between added noise and intersymbol interference. Monsen[4] also investigated some aspects of our problem but did not arrive at a complete solution. Our chief contribution is an explicit formula for the minimum mean-square error (MSE). The simplicity of the formula makes possible detailed investigation of optimized system performance.

In Section II the model is stated and the problem is formulated. In Section III the receiving filter is optimized and in Section IV the transmitter filter is optimized. In Section V we examine the problem of signaling faster than the Nyquist rate and finally in Section VI we use our results to investigate in detail the performance of a data system operating over a cable channel.

II. THE MODEL AND PROBLEM FORMULATION

The system model under investigation is depicted in Fig. 1. The data signal denoted by $D(t)$ is passed through the transmitting filter having an impulse response $s(t)$ and giving rise to an average transmitted power P. The data symbols $\{a_n\}_{-\infty}^{\infty}$ are independently picked at the rate $1/T$ and take on values with equal probability from the set $\{\pm 1, \pm 3 \pm 5 \cdots \pm (L-1)\}$ where L is an even integer. The resulting signal is admitted to a linear channel characterized by an impulse response $h(t)$. The received signal plus noise is processed by the equalizer which is comprised of a linear filter having impulse

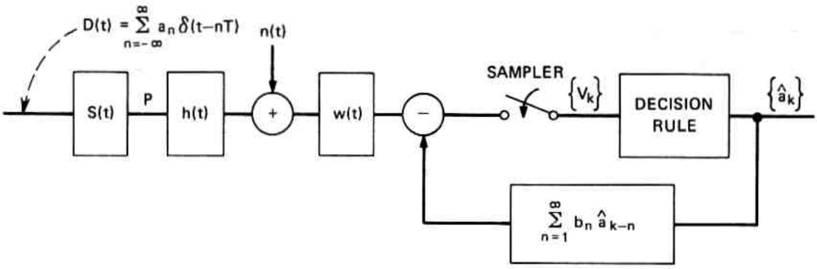
$D(t) = \sum_{n=-\infty}^{\infty} a_n \delta(t-nT)$
n(t)
S(t)
P
h(t)
+
w(t)
−
SAMPLER
{V_k}
DECISION RULE
{â_k}
$\sum_{n=1}^{\infty} b_n \hat{a}_{k-n}$

Fig. 1—Block diagram of the model.

response $w(t)$, a sampler, a decision rule, and a feedback digital filter characterized by the infinite set of real numbers $\{b_n\}_1^\infty$. The added noise $n(t)$ is a zero-mean white random process with double-sided spectral density $N_0/2$. The output data symbols are denoted by $\hat{a}_n$.

The general problem we would like to solve is the minimization of

$$\text{MSE} = E\{v_k - \hat{a}_k\}^2$$

with respect to the set of square integrable functions $\{s(t), w(t)\}$ and the infinite sequence of numbers $\{b_n\}$. This is to be carried out when the channel impulse response $h(t)$, transmitted power P, and a decision rule are given. The symbol $E(\cdot)$ denotes expectation with respect to all the random variables.

The nonlinear relation between the estimated symbols $\{\hat{a}_n\}$ and the input symbols $\{a_n\}$ makes this problem mathematically intractable. However, by assuming that past decisions have been correct, we can begin to approach the problem. The resulting MSE must then be interpreted as a lower bound on the true MSE. Alternatively, if no errors have occurred in the past, the MSE under this assumption provides an indication of the noise immunity of the system (including residual intersymbol interference).

Let $r(t) = s(t)*h(t)*w(t)$ denote the overall impulse response, where $*$ denotes convolution. Under the assumption of correct past decisions, the received sample taken at time $t = kT$ is

$$v_k = \sum_{n=-\infty}^{\infty} r_n a_{k-n} - \sum_{n=1}^{\infty} b_n a_{k-n} + n(t)*w(t)\,|_{t=kT}$$

and the mean-square error is then by definition

$$\text{MSE} = E\left\{ \sum_{n=-\infty}^{-1} r_n a_{k-n} + \sum_{n=1}^{\infty} (r_n - b_n) a_{k-n} \right. \\ \left. + n(t)*w(t)\,|_{t=kT} + (r_0 - 1)a_k \right\}^2.$$

A straightforward calculation gives

$$\text{MSE} = \sigma_a^2 \sum_{n=-\infty}^{-1} r_n^2 + \sigma_a^2(r_0 - 1)^2 + \sigma_a^2 \sum_{n=1}^{\infty} (r_n - b_n)^2 + \sigma^2,$$

where

$$\sigma_a^2 = E\{a_n\}^2 = \frac{L^2 - 1}{3}$$

and

$$\sigma^2 = \frac{N_0}{2} \int_{-\infty}^{\infty} w^2(t)dt.$$

It can be immediately concluded that the mean-square error is minimized by setting $b_n = r_n$, $n = 1, 2, \cdots$, which eliminates the feedback coefficients $\{b_n\}$ from further consideration.

The problem we now confront is the dual minimization of

$$\text{MSE}[s(t), w(t)] = \sigma_a^2 \left[\sum_{n=-\infty}^{0} r_n^2 - 2r_0 + 1 + \frac{\sigma^2}{\sigma_a^2} \right]$$

with respect to $s(t)$ and $w(t)$ when a constraint is imposed on the average transmitted power.

The above expression indicates that, under the assumption of perfect past decisions, the mean-square error is minimized by minimizing both the pulse precursors in the overall impulse response and the output noise power, while keeping r_0 close to unity.

We give a precise formulation and solution to this problem in Section III.

III. RECEIVER OPTIMIZATION

Writing the MSE in detail we obtain

$$\frac{\text{MSE}}{\sigma_a^2} = 1 + \sum_{n=-\infty}^{0} \left[\int_{-\infty}^{\infty} w(\tau)p(nT - \tau)d\tau \right]^2 - 2 \int_{-\infty}^{\infty} w(\tau)p(-\tau)d\tau + \frac{N_0}{2\sigma_a^2} \int_{-\infty}^{\infty} w^2(\tau)d\tau, \quad (1)$$

where

$$p(t) = s(t)*h(t).$$

Keeping $s(t)$ fixed, and using a standard calculus-of-variation approach, results in an integral equation for $w(t)$

$$p(-t) = N_0'w(t) + \int_{-\infty}^{\infty} w(\tau) \left(\sum_{n=-\infty}^{0} p(nT - \tau)p(nT - t) \right) d\tau, \quad (2)$$

where

$$N_0' = \frac{N_0}{2\sigma_a^2}.$$

If in eq. (2) we set

$$U_n = \int_{-\infty}^{\infty} w(\tau)p(nT - \tau)d\tau,$$

we see that the optimum solution must have a representation in the form

$$w(t) = \sum_{n=-\infty}^{0} g_n p(nT - t), \tag{3}$$

where

$$g_0 = \frac{1}{N_0'}(1 - U_0)$$

and

$$g_n = -\frac{U_n}{N_0'}, \qquad n = \leqq -1.$$

In (3) is revealed the structure of the optimum receiving filter. It is composed of a *matched filter having impulse response $p(-t)$ followed by a one-sided (anticausal) tapped delay line with weights equal to g_n.*

Linear equations involving the set $\{U_n\}$ can be obtained by first multiplying both sides of (2) by $p(kT - t)$, $k \leqq 0$, then integrating from minus to plus infinity. The resulting linear system of equations is

$$R_k = N_0'U_k + \sum_{n=-\infty}^{0} R_{n-k}U_n, \qquad k = 0, -1 \cdots, \tag{4}$$

where

$$R_k = R_{-k} = \int_{-\infty}^{\infty} p(-t)p(kT - t)dt.$$

The system of eqs. (4) can be solved by standard Wiener-Hopf techniques and the details are given in Appendix A. The solution in terms of the discrete Fourier transform of the sequence $\{U_n\}_{-\infty}^{0}$ is

$$\begin{aligned} U(\theta) &= \sum_{n=-\infty}^{0} U_n e^{in\theta} \\ &= 1 - \frac{N_0'}{M^-(\theta)\gamma_0}, \end{aligned} \tag{5}$$

where

$$M(\theta) = M^+(\theta)M^-(\theta) = R(\theta) + N_0' = \sum_{n=-\infty}^{n=\infty} M_n e^{in\theta},$$

$$M^+(\theta) = \sum_{n=0}^{\infty} \gamma_n e^{in\theta},$$

and

$$M^-(\theta) = M^+(-\theta).$$

This is standard procedure making use of the well-known factorization property of covariance functions. Methods for obtaining the sequence $\{\gamma_n\}_0^\infty$ from the given sequence $\{R_n\}_{-\infty}^{\infty}$ are well documented in the literature. One method is summarized in Appendix A.

Having specified the optimum receiving filter, we now obtain a formula for the minimized mean-square error. The availability of this simple formula will allow further optimization of the transmitting filter.

Let $w_0(t)$ be the impulse response of the optimum receiving filter. [This function solves the integral equation (2).] Substitute $w_0(t)$ into (2), multiply both sides by $w_0(t)$, and integrate from minus infinity to plus infinity to obtain

$$\int_{-\infty}^{\infty} p(-t)w_0(t)dt = N_0' \int_{-\infty}^{\infty} w_0^2(t)dt + \sum_{n=-\infty}^{0} \left(\int_{-\infty}^{\infty} p(nT - t)w_0(t)dt \right)^2. \tag{6}$$

Putting this into (1) with $w(t)$ replaced by $w_0(t)$ we get a formula for the optimized MSE

$$\text{MSE}[w_0(t)] = \sigma_a^2(1 - U_0) = N_0'\sigma_a^2 g_0. \tag{7}$$

This result was obtained by Monsen[4] but unfortunately he did not go any further. As it turns out, a much richer formula than (7) can be obtained since U_0 can be expressed directly in terms of the spectrum of the channel characteristics in cascade with the transmitting filter. To carry this further, observe from (5) that

$$\begin{aligned} U_0 &= \text{dc term in } U(\theta) \\ &= 1 - \frac{N_0'}{\gamma_0^2} \end{aligned} \tag{8}$$

and consequently

$$\text{MSE} = \sigma_a^2 \frac{N_0'}{\gamma_0^2}. \tag{9}$$

As it turns out, γ_0 is functionally related to $M(\theta)$ in a rather simple manner. This relationship can be found in the literature but the derivation is short and so we briefly outline the approach.

Under very mild conditions on $M(\theta)$ (see Doob,[5] pp. 159–161)

$$M(\theta) = \left| \sum_{n=0}^{\infty} \gamma_n e^{in\theta} \right|^2, \tag{10}$$

where γ_0 is real and positive. Since $M(\theta) > 0$, consider

$$\int_{-\pi}^{\pi} \ln M(\theta) dt = \int_{-\pi}^{\pi} \ln \left[\gamma_0 + \sum_{n=1}^{\infty} \gamma_n e^{in\theta} \right] d\theta + \int_{-\pi}^{\pi} \ln \left[\gamma_0 + \sum_{n=1}^{\infty} \gamma_n e^{-in\theta} \right] d\theta. \tag{11}$$

When the ln's on the r.h.s. of (11) are expanded in a power series and the integrations are carried out (recognizing that all integrals involving powers of exp $\{in\theta\}$, $n \neq 0$, vanish) we get

$$\gamma_0^2 = \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln[R(\theta) + N_0'] d\theta \right\}, \tag{12}$$

where

$$R(\theta) = \sum_{n=-\infty}^{\infty} R_n e^{in\theta},$$

$$R_n = \int_{-\infty}^{\infty} |P(\omega)|^2 e^{i\omega n T} \frac{d\omega}{2\pi},$$

and

$$P(\omega) = \int_{-\infty}^{\infty} s(t) * h(t) e^{i\omega t} dt.$$

After minor algebraic manipulations and changes of variables we obtain by substituting (12) into (9)

$$\text{MSE} = \sigma_a^2 \exp \left\{ -\frac{\text{T}}{2\pi} \int_{-\pi/T}^{\pi/T} \ln[Y(\omega) + 1] d\omega \right\}, \tag{13}$$

where

$$Y(\omega) = \frac{1}{N_0' T} \sum_{n=-\infty}^{\infty} \left| P\left(\omega - \frac{2\pi n}{T}\right) \right|^2. \tag{14}$$

This formula, as far as can be determined, is new and its simple form will enable us in Section IV to carry out an additional optimization with respect to the transmitting filter.

It is instructive at this point to compare this formula with the one obtained for a linear equalizer without decision feedback. Berger and

Tufts[6] have found such a formula and from their paper we have that

$$(\mathrm{MSE})_{\mathrm{linear}} = \sigma_a^2 \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} [Y(\omega) + 1]^{-1} d\omega$$
$$= \left\langle \frac{1}{Y(\omega) + 1} \right\rangle, \tag{15}$$

where

$$\langle \cdot \rangle = \sigma_a^2 \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} [\cdot] d\omega.$$

In terms of the same notation, (13) can be put into the form

$$\mathrm{MSE} = \exp\{-\langle \ln[Y(\omega) + 1] \rangle\} \tag{16}$$

from which we get immediately that

$$\mathrm{MSE} \leqq \left\langle e^{-\ln[Y(\omega)+1]} \right\rangle = \left\langle \frac{1}{Y(\omega) + 1} \right\rangle = (\mathrm{MSE})_{\mathrm{linear}}. \tag{17}$$

As expected, the mean-square error with decision feedback is always smaller than the MSE of a linear equalizer. Comparing (15) with (16)

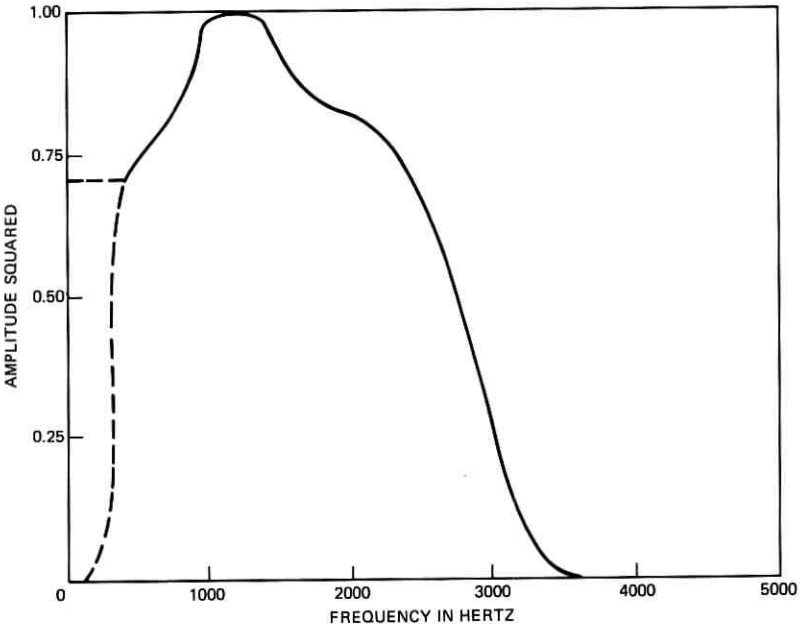

Fig. 2—Amplitude-squared characteristic of typical voiceband channel.

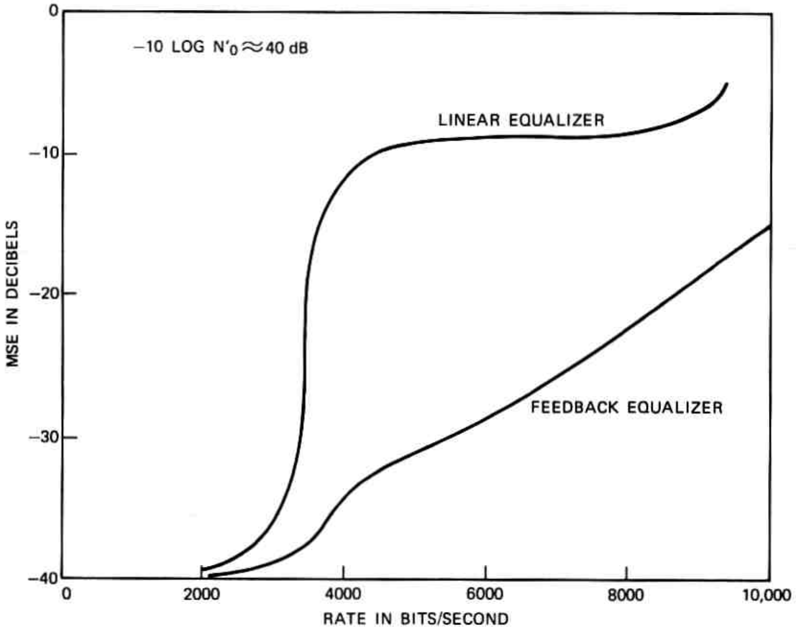

Fig. 3—MSE in dB vs binary data rate for channel shown in Fig. 2 without dc transmission.

shows that both equalization methods yield the same MSE if and only if $Y(\omega)$ is a constant, i.e., there is no intersymbol interference.

Prior to optimizing the transmitting filter we wish to illustrate the behavior of (15) and (16) for a typical voiceband channel as the signaling rate $1/T$ is allowed to increase. An amplitude-squared characteristic for a typical voiceband telephone channel is shown in Fig. 2 (the dashed line with zero transmission at zero frequency is typical). Figure 3 shows the resulting MSE vs pulse rate for both a linear equalizer and a decision feedback equalizer when $\sigma_a^2 = 1$ (binary data) and $-10 \log N_0' \approx 40$ dB. The calculations were done numerically by using (15) and (16). We note that the performance of the linear equalizer deteriorates rapidly when the rate is greater than ≈ 3000 bits/second while the decision feedback equalizer deteriorates gracefully. The reason the linear equalizer shows such a poor performance is that it must compensate for the missing energy around dc. In practice, of course, modulation is used to place the data energy at a more suitable location in the passband spectrum to avoid this severe null. To make the comparison fair, we artificially extended the channel

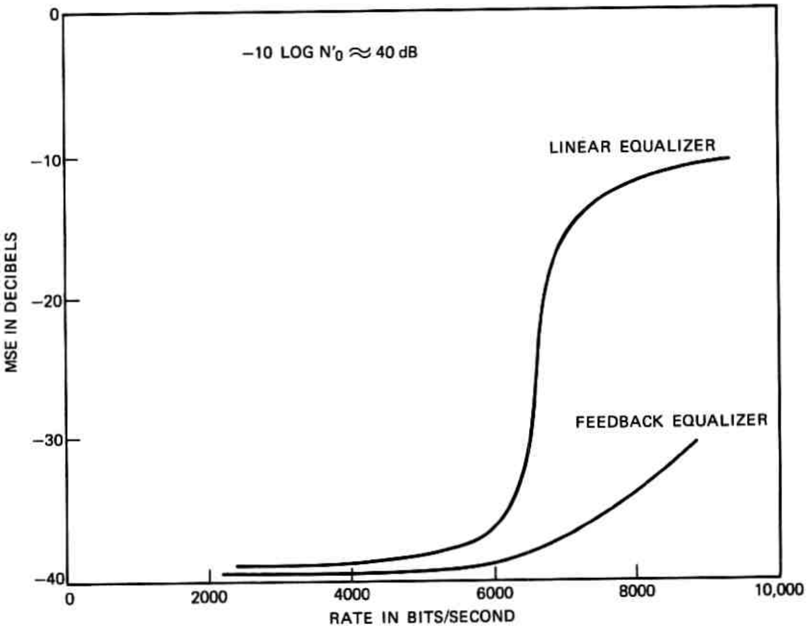

Fig. 4—MSE in dB vs binary data rate for channel shown in Fig. 2 with dc transmission.

characteristic from about 300 Hz to 0 Hz to a constant transmission. This is indicated in Fig. 2 by the dashed line parallel to the frequency axis. Figure 4 shows the comparisons for this atypical channel. As expected, the linear equalizer has a sharp threshold at approximately the Nyquist rate but, as before, the decision feedback equalizer deteriorates much more gracefully.

IV. TRANSMITTER OPTIMIZATION

The problem we address here is the optimization of (13) with respect to the transmitting filter characteristics subject to an average power constraint.

Let $S(\omega)$ and $H(\omega)$ be the Fourier transforms of $s(t)$ and $h(t)$ respectively. The average power at the output of $s(t)$ is

$$\begin{aligned} P &= \frac{\sigma_a^2}{T}\int_{-\infty}^{\infty} s^2(t)dt = \frac{\sigma_a^2}{T}\frac{1}{2\pi}\int_{-\infty}^{\infty} S^2(\omega)d\omega \\ &= \frac{\sigma_a^2}{T}\frac{1}{2\pi}\int_{-\pi/T}^{\pi/T}\left(\sum_{n=-\infty}^{\infty} S^2\left(\omega - \frac{2\pi n}{T}\right)\right)d\omega, \end{aligned} \tag{18}$$

where by $S^2(\omega)$ we mean $|S(\omega)|^2$.

The problem at hand is to maximize the functional

$$I = \int_{-\pi/T}^{\pi/T} \ln[K \sum_n S_n^2(\omega) H_n^2(\omega) + 1] d\omega + \lambda \int_{-\pi/T}^{\pi/T} [\sum_n S_n^2(\omega)] d\omega \quad (19)$$

with respect to the infinite set of functions $\{S_n^2(\omega),\ n = \text{all integers}\}$. Where

$$S_n^2(\omega) = \left| S\left(\omega - \frac{2\pi n}{T}\right)\right|^2, \qquad K = 1/N_0'T$$

and λ is a Lagrange multiplier to be determined from the constraint on the average transmitted power. Observe that these functions are independent over the range $-\pi/T \leqq \omega \leqq \pi/T$ and therefore consider the variation of I with respect to, say, $S_j^2(\omega)$

$$\delta_j I = \int_{-\pi/T}^{\pi/T} \left\{ \frac{KH_j^2(\omega)}{K \sum_n S_n^2(\omega) H_n^2(\omega) + 1} \delta S_j^2(\omega) + \lambda \delta S_j^2(\omega) \right\} d\omega. \quad (20)$$

When $S_j^2(\omega) \neq 0$, setting (20) to zero implies

$$\frac{KH_j^2(\omega)}{K \sum_n S_n^2(\omega) H_n^2(\omega) + 0} + \lambda = 0 \quad (21)$$

for all $\omega \in [-\pi/T, \pi/T]$. Now, assume that $H_n^2(\omega) > H_m^2(\omega)$ for $|n| < |m|$ and $\omega \in [-\pi/T, \pi/T]$. When this condition is satisfied it is not possible to solve the system of equations given in (21) unless

$$S_n^2(\omega) = 0 \qquad \text{for all } n \neq j$$

in which case we get

$$KH_j^2(\omega) = -\lambda[KS_j^2(\omega)H_j^2(\omega) + 1]. \quad (22)$$

Substituting this into (20) indicates that the largest value is obtained when $S_j^2(\omega) = S_0^2(\omega) = S^2(\omega)$ and furthermore λ must be negative.

So far we can conclude that for channels possessing monotonically decreasing amplitude characteristics, i.e., $H_m^2(\omega) > H_n^2(\omega)$, $-\pi/T \leqq \omega \leqq \pi/T$, when $|n| > |m|$, the optimum transmitting filter cuts off at the Nyquist frequency π/T. The optimum system allows no transmission outside the band $|\omega| > \pi/T$. The restrictions imposed on the channels are mild and are expected to be satisfied in most situations of interest. However, removing these restrictions makes the problem slightly more complicated, and it is left up to the reader to reason how it can be solved.

Making use of this partial solution permits writing the mean-square error in the simplified form

$$-\ln\left(\frac{\text{MSE}}{\sigma_a^2}\right) = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left[KH^2(\omega)S^2(\omega)+1\right]d\omega \tag{23}$$

and the functional now to be further maximized with respect to the inband structure of $S^2(\omega)$ reduces to

$$I[S(\omega)] = \int_{-\pi/T}^{\pi/T} \ln\left[KS^2(\omega)H^2(\omega)+1\right]d\omega + \lambda\int_{-\pi/T}^{\pi/T} S^2(\omega)d\omega. \tag{24}$$

As can be seen from (20) and (22), eq. (24) is maximized when

$$S^2(\omega) = \max\left[\frac{KH^2(\omega)-\lambda_0}{\lambda_1 KH^2(\omega)}, 0\right], \tag{25}$$

where $\lambda = -\lambda_0$.

To determine the Lagrange multiplier λ_0, two cases must be distinguished.

Case 1: $KH^2(\omega) - \lambda_0 > 0$, *for all* $|\omega| \leqq \pi/T$

In this case, we get

$$1 + KH^2S^2 = \mu H^2, \qquad \mu = K/\lambda_0,$$

which when substituted into (23) results in an explicit expression for the optimum MSE

$$-\ln\left(\frac{\text{MSE}}{\sigma_a^2}\right) = \ln\mu + \frac{T}{\pi}\int_0^{\pi/T} \ln H^2(\omega)d\omega. \tag{26}$$

The factor μ is determined from the average power constraint in the following manner. Use (25) and (18) to write

$$\begin{aligned} \text{P} &= \frac{\sigma_a^2}{T\pi}\int_0^{\pi/T}\left[\frac{\mu}{K} - \frac{1}{KH^2(\omega)}\right]d\omega \\ &= \frac{\sigma_a^2}{KT^2}(\mu - A), \end{aligned} \tag{27}$$

where

$$A = \frac{T}{\pi}\int_0^{\pi/T}\frac{1}{H^2(\omega)}d\omega.$$

Substituting the parameters $K = 1/TN_0'$ and $N_0' = N_0/2\sigma_a^2$ into (27) gives explicitly

$$\mu = \rho + A, \tag{28}$$

where

$$\rho = \frac{P}{\left(\frac{N_0}{2}\frac{1}{T}\right)} \equiv \frac{\text{Average transmitted signal power}}{\text{Average noise power in the Nyquist band}}.$$

Thus (26) and (28) provide a complete solution for this case.

Case 2: There exists a set of ω for which $KH^2(\omega) - \lambda_0 \leqq 0$

The optimization procedure in this case involves the standard water-pouring argument. To illustrate the nature of the solution we take the situation where $H^2(\omega)$ is strictly monotonically decreasing in the Nyquist band. This implies that there exists only one frequency ω_0 for which $KH^2(\omega) = \lambda_0$ and consequently we get

$$\mu = \frac{1}{H^2(\omega_0)}. \tag{29}$$

This gives one relation between the unknowns μ and ω_0 and another is obtained from the power constraint. Since the optimum filter characteristic is zero when $\omega > \omega_0$, the signal-to-noise ratio is

$$\rho = \frac{T}{\pi} \int_0^{\omega_0} \left[\frac{1}{H^2(\omega_0)} - \frac{1}{H^2(\omega)} \right] d\omega \tag{30}$$

and an explicit formula for the mean-square is

$$-\ln\left[\frac{\text{MSE}}{\sigma_a^2}\right] = \frac{T}{\pi} \int_0^{\omega_0} \ln H^2(\omega) d\omega - \frac{T}{\pi}\, \omega_0 \ln H^2(\omega_0). \tag{31}$$

We now briefly summarize how these optimized formulas are to be used:

(*i*) For a given transmitted average signal-to-noise ratio ρ, solve eq. (30) for ω_0.
(*ii*) If $\omega_0 < \pi/T$, use formula (31) to compute MSE.
(*iii*) If $\omega_0 \geqq \pi/T$, use formula (26) to compute MSE.

In Section VI we shall illustrate numerically the use of these formulas.

V. SIGNALING FASTER THAN THE NYQUIST RATE

Here we examine the behavior of the optimized mean-square error when the frequency support of an ideal unity-gain channel is smaller than the Nyquist rate $\frac{1}{2}T$. After deriving the optimized MSE for this situation, the possibility of further optimization relative to the signaling rate when the information rate per unit bandwidth is held fixed will be investigated. This has been an open question thus far and the issue is whether increasing the signaling rate beyond the Nyquist rate while decreasing the number of levels to maintain a fixed information rate is ever beneficial.

Consider a channel having the characteristic

$$H^2(\omega) = \begin{cases} 0, & \omega \in E_1 \\ 1, & \omega \in E_2' \end{cases} \tag{32}$$

where the sets E_1 and E_2 form a partition on the frequency interval $E = \{\omega : 0 \leqq \omega \leqq \pi/T\}$. By frequency support we mean the measure of the set E_2 denoted by $m(E_2)$.

For this channel it is easy to calculate explicitly the mean-square error. Observe from eq. (22) that for a piecewise constant channel the optimum transmitting filter is a constant when $\omega \in E_2$ and zero otherwise. The minimum MSE is then calculated from (23)

$$\begin{aligned} -\ln\left[\frac{\text{MSE}}{\sigma_a^2}\right] &= \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left[KS^2H^2(\omega) + 1\right]d\omega \\ &= \frac{T}{\pi}\int_{\omega\in E_2} \ln\left[KS^2 + 1\right]d\omega \\ &= \frac{T}{\pi}\, m(E_2)\ln\left[KS^2 + 1\right], \end{aligned} \tag{33}$$

where S^2 is a constant to be determined from the power constraint

$$P = \frac{\sigma_a^2}{\pi T}\int_{\omega\in E_2} S^2 d\omega = \frac{\sigma_a^2}{\pi T} S^2 m(E_2). \tag{34}$$

From this equation and the definition of $K = 1/(N_0'T)$ it can be checked that

$$\rho = KS^2 = \frac{\text{Average signal power}}{\text{Average noise power in a band} = m(E_2)}.$$

Substituting this into (33) gives the simple desired formula

$$\text{MSE} = \sigma_a^2(1 + \rho)^{-\alpha}, \tag{35}$$

where

$$\begin{aligned} \alpha &= \frac{m(E_2)T}{\pi} \\ &= \frac{\text{Channel bandwidth}}{2 \times \text{Signaling rate}} \leqq 1. \end{aligned}$$

For fixed ρ, MSE $\rightarrow \sigma_a^2$ when $\alpha = 0$ and, as expected, MSE $\rightarrow \sigma_a^2(1+\rho)^{-1}$ when $\alpha = 1$. It is curious that as long as $\alpha \neq 0$, MSE $\rightarrow 0$ as $\rho \rightarrow \infty$.

When the set E_2 is the interval $I = \{\omega : 0 \leqq \omega \leqq \pi/T_0 < \pi/T\}$, it is referred to as being less than the Nyquist band. Since there is no mathematical reason for making this distinction we shall refer to "signaling faster than the Nyquist rate" whenever $\alpha < 1$.

We now investigate whether it is ever advantageous to signal faster than the Nyquist rate. Clearly, for fixed $\sigma_a^2 = (L^2 - 1)/3$, or a fixed number of levels, (35) shows that MSE degrades rapidly with decreasing α. An interesting question, first raised by R. W. Lucky,[7] is the possibility of trading L with α to further minimize the mean-square error. This is the problem we address.

Let the source information rate be $R = \log_2 L/T$ bits/second. The available bandwidth is equal to $1/(2\pi)m(E_2)$ cycles/second. Thus the normalized information rate is

$$\begin{aligned}\mathcal{E} &= \frac{R}{\dfrac{1}{2\pi}\, m(E_2)} = \frac{2}{\left[\dfrac{Tm(E_2)}{\pi}\right]} \log_2 L \\ &= \frac{2}{\alpha} \log_2 L \, \frac{\text{bits}}{\text{cycle}}. \end{aligned} \tag{36}$$

Writing (35) in terms of this quantity gives

$$\text{MSE}(\alpha) = \frac{2^{\alpha\mathcal{E}} - 1}{3} \, \frac{1}{(1+\rho)^\alpha}. \tag{37}$$

Letting $C = \log_2(1 + \rho)$ be the ultimate attainable rate according to Shannon's theory, eq. (37) can be put into the form

$$\text{MSE}(\alpha) = (2^{-\alpha(C-\mathcal{E})} - 2^{-\alpha C})\tfrac{1}{3}. \tag{38}$$

Note that, since $L = 2^{\alpha\mathcal{E}/2}$ and the minimum allowable L is equal to 2, the parameter α must be in the range $(2/\mathcal{E}, 1)$.

The problem initially posed can now be stated as follows. Find a set of α's, $2/\mathcal{E} \leqq \alpha \leqq 1$, which minimize eq. (38). We begin by setting the derivative of (38) to zero,

$$\frac{d\text{MSE}(\alpha)}{d\alpha} = -(C - \mathcal{E})2^{-\alpha(C-\mathcal{E})} + C2^{-\alpha C} = 0,$$

from which we find a unique stationary point

$$\alpha = \frac{2}{\mathcal{E}} \log_2 \left[\frac{C}{C - \mathcal{E}}\right]^{\frac{1}{2}}. \tag{39}$$

Since $\text{MSE}(0) = 0$ and $\text{MSE}(\alpha) > 0$, the value of α found above must be a point where MSE attains a maximum. If this maximum is in the range ($0 \leqq \alpha \leqq 2/\mathcal{E}$), then the minimum value of MSE is attained at the boundary ($\alpha = 1$). The condition for this to be the case is determined from (39).

$$\log_2 \left[\frac{C}{C - \mathcal{E}} \right]^{\frac{1}{2}} \leqq 1. \tag{40}$$

From this we deduce that as long as $\mathcal{E} < \frac{3}{4}C$, $\alpha = 1$ minimizes MSE. For this region of $\mathcal{E}$ and C there is no advantage gained by signaling faster than the Nyquist rate. Since C is channel capacity, $\mathcal{E}_c = \frac{3}{4}C$ seems to be a critical rate. If the rate is greater than this critical rate, the maximum point lies within the allowable range of $\alpha(2/\mathcal{E}, 1)$ thus raising the possibility that either $\alpha = 2/\mathcal{E}$ or $\alpha = 1$ is a minimum point. Suppose $\alpha = 2/\mathcal{E}$ is a minimum point. Equation (38) gives for this case

$$\text{MSE}\left(\alpha = \frac{2}{\mathcal{E}}\right) = 2^{-2C/\mathcal{E}} \geqq 2^{-8/3} \approx 0.157. \tag{41}$$

Thus we have found a region for which signaling faster than the Nyquist rate appears to be beneficial. But at this high level of MSE, we are no longer justified in assuming that the feedback decisions are correct most of the time. In fact, what is more likely to happen is that errors begin to occur resulting in a larger value of MSE than predicted by the error-free model. We are therefore led to the conclusion that a minimum point other than at $\alpha = 1$ will render an MSE to be outside the range of practical utility. To emphasize this point further substitute (41) into (38) to get

$$\text{MSE}(\alpha = 1) = [\text{MSE}(2/\mathcal{E})]^{\mathcal{E}/2} \frac{2^{\mathcal{E}} - 1}{3}. \tag{42}$$

Thus $\alpha = 1/\mathcal{E}$ is a minimum point whenever

$$[\text{MSE}(2/\mathcal{E})]^{\mathcal{E}/2} \left(\frac{2^{\mathcal{E}} - 1}{3} \right) > \text{MSE}(2/\mathcal{E})$$

or

$$\text{MSE}(2/\mathcal{E}) > \left(\frac{3}{2^{\mathcal{E}} - 1} \right)^{2/\mathcal{E} - 2}.$$

As an example of the use of these inequalities, suppose that $\mathcal{E} = 3$ and $C = 3.5$; therefore, $\alpha = \frac{2}{3}$ is the minimum point and the achievable MSE = 0.198 which can be obtained with a binary system ($L = 2$). On the other hand, $M(\alpha = 1) = 2^{-3.5}(7/3) \cong 0.206$ which can be achieved with ($L = 2^{1.5} = 2.8$)!!

It is interesting to see what MSE can be achieved when only a linear equalizer is used. Using formula (15) and following the same reasoning as before yields an expression for the optimized mean-square error

$$(\mathrm{MSE})_{\mathrm{linear}} = \sigma_a^2 \frac{(1-\alpha)\rho + 1}{\rho + 1}. \tag{43}$$

In this case, unlike in the decision feedback case [eq. (35)], as $\rho \to \infty$, MSE $\to 1 - \alpha$, when $\alpha > 0$. Thus the mean-square error cannot be made vanishingly small as the signal-to-noise ratio increases without bound.

Expressing (43) in terms of the normalized rate $\mathcal{E}$ we get

$$(\mathrm{MSE})_{\mathrm{linear}} = \frac{2^{\alpha\mathcal{E}} - 1}{3}\left[\frac{\rho(1-\alpha) + 1}{\rho + 1}\right]. \tag{44}$$

Again we seek to minimize (44) with respect to α in the range $(2/\mathcal{E}, 1)$. It can be checked that (44) has at most one stationary point in the range $0 \leqq \alpha \leqq 1$. Since $\alpha = 0$ is a minimum point, the minimum in the range $(2/\mathcal{E}, 1)$ must lie on the boundary. Thus the condition for achieving a smaller MSE when signaling faster than the Nyquist rate $(\alpha < 1)$ is

$$\rho(1 - 2/\mathcal{E}) + 1 \leqq \frac{2^{\mathcal{E}} - 1}{3}$$

or

$$\rho = 2^C - 1 \leqq \frac{2^{\mathcal{E}} - 1}{3}\,\frac{1}{1 - 2/\mathcal{E}}. \tag{45}$$

Is it possible to find an $\mathcal{E} \leqq C$, $\mathcal{E} > 2$ which satisfies the inequality (45)? A straightforward analysis reveals that the answer is negative. In other words, the Nyquist rate is optimum provided the information rate is less than channel capacity.

VI. APPLICATION TO A CABLE CHANNEL†

This section will illustrate the use of the formulas developed in previous sections in a particular application. For this purpose we choose a cable channel having frequency characteristic

$$H(f) = \exp\{\sqrt{-2i\alpha f}\}. \tag{46}$$

We shall develop in detail the applicable formulas, provide asymptotic behaviors, and exhibit numerically the relevant parameter trade-offs. For comparison purposes, the applicable formulas for the optimum linear equalizer will also be developed.

† I am indebted to Dr. Robert Price for calling my attention to Ref. 8 where related work is reported. The paper is in Japanese; however, Dr. Price has an English translation.

We begin by first considering a suboptimum system where the transmitting filter is flat across the Nyquist band and zero outside. For this case, the minimum attainable mean-square error is [eq. (23)]

$$M_d = \exp\left\{-\frac{T}{\pi}\int_0^{\pi/T} \ln\left[S^2Ke^{-\sqrt{2\omega\alpha/\pi}}+1\right]d\omega\right\}, \tag{47}$$

where

$$M_d = \text{MSE}/\sigma_a^2, \qquad |H(\omega)|^2 = e^{-\sqrt{2\omega\alpha/\pi}}, \qquad \text{and} \quad K = 2\sigma_a^2/TN_0$$

are to be determined from the average power constraint. Since the transmitted power $P = (\sigma_a^2/T)S^2$, we find that $S^2K = 2PT/N_0 = \rho$, the transmitted signal-to-noise ratio, where the noise is measured in a band $= 1/T$. Substituting these constants into (47) and making some changes of variables result in

$$M_d = \exp\left\{-2\int_0^{\frac{1}{2}} \ln\left[\rho e^{-\sqrt{4\beta y}}+1\right]dy\right\}, \tag{48}$$

where $\beta = \alpha/T$. The parameter $\sqrt{2\beta}$ is seen to be proportional to the loss of the cable in dB at the Nyquist frequency $\frac{1}{2}T$.

We are interested in the behavior of (48) when ρ and β are varied. While it is not possible to express this integral in a closed form, it is possible to obtain a rapidly converging series in the two parameters of interest from which asymptotic behaviors can be deduced. Appendix B shows the details of the development. Different power series apply in different regions. The first series applies when $\ln \rho < \sqrt{2\beta}$ and the second $\ln \rho \geqq \sqrt{2\beta}$. The results are as follows

$1\colon \sqrt{2\beta} \geqq \ln \rho > 0$

$$-\ln M_d = \frac{(\ln \rho)^3 + \pi^2 \ln \rho}{6\beta} - \sqrt{\frac{2}{\beta}}\sum_{n=1}^{\infty}(-1)^{n+1}\frac{[\rho e^{-\sqrt{2\beta}}]^n}{n^2} + \frac{1}{\beta}\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^3}\left[\frac{1}{\rho^n} - (\rho e^{-\sqrt{2\beta}})^n\right]. \tag{49}$$

$2\colon \ln \rho \geqq \sqrt{2\beta} > 0$

$$-\ln M_d = \ln \rho - \frac{2}{3}\sqrt{2\beta} + \sqrt{\frac{1}{\beta}}\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^2}\left[\frac{e^{\sqrt{2\beta}}}{\rho}\right]^n + \frac{1}{\beta}\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^3}\left[\frac{1}{\rho^n} - \left(\frac{e^{\sqrt{2\beta}}}{\rho}\right)^n\right]. \tag{50}$$

It can be checked that (49) equals (50) when $\ln \rho = \sqrt{2\beta}$. The first asymptotic behavior is deduced from (50) when $\rho \to \infty$ and $\sqrt{2\beta}$ is held fixed. For this case we get

$$M_d \sim \frac{e^{\frac{2}{3}\sqrt{2\beta}}}{\rho}. \tag{51}$$

Another asymptotic behavior is deduced from (49) as $\beta \to \infty$ while ρ is held fixed. In this case

$$M_d \sim e^{-g(\rho)/\beta}, \tag{52}$$

where

$$g(\rho) = \frac{(\ln \rho)^3 + \pi^2 \ln \rho}{6} + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^3} \frac{1}{\rho^n}.$$

In order to make possible performance comparisons, we develop similar formulas and asymptotes for a system employing only a linear equalizer. The minimum mean-square error applicable in this situation is obtained from eq. (15). After substituting the cable characteristic we obtain

$$M_L = 2 \int_0^{\frac{1}{2}} [\rho e^{-\sqrt{4\beta y}} + 1]^{-1} dy, \tag{53}$$

where

$$M_L = \frac{(\text{MSE})_{\text{linear}}}{\sigma_a^2}.$$

Here, as in the decision feedback case, rapidly converging series can be developed from which asymptotic formulas are deduced. The detailed calculations are also given in Appendix B. The desired results are

1: $\sqrt{2\beta} \geqq \ln \rho > 0$

$$M_L = 1 - \frac{3(\ln \rho)^2 + \pi^2}{6\beta} + \sqrt{\frac{2}{\beta}} \ln [1 + \rho e^{-\sqrt{2\beta}}] + \frac{1}{\beta} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^2} \left[\frac{1}{\rho^n} + (\rho e^{-\sqrt{2\beta}})^n \right]. \tag{54}$$

2: $\ln \rho > \sqrt{2\beta} > 0$

$$M_L = \sqrt{\frac{2}{\beta}} \ln \left[1 + \frac{e^{\sqrt{2\beta}}}{\rho} \right] + \frac{1}{\beta} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^2} \left[\frac{1}{\rho^n} - \left(\frac{e^{\sqrt{2\beta}}}{\rho} \right)^n \right]. \tag{55}$$

The asymptotic formulas are readily deduced from (54) and (55).

When $\rho \to \infty$ and β is fixed we get

$$M_L \sim Q(\beta)\frac{1}{\rho}, \tag{56}$$

where

$$Q(\beta) = e^{\sqrt{2\beta}}\left[\sqrt{\frac{2}{\beta}} - \frac{1}{\beta} + \frac{e^{-\sqrt{2\beta}}}{\beta}\right].$$

On the other hand, when $\beta \to \infty$ and ρ is kept fixed,

$$M_L \sim 1 - \frac{3(\ln\rho)^2 + \pi^2}{6\beta} + \frac{D}{\beta}, \tag{57}$$

where

$$D = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^2}\frac{1}{\rho^n}.$$

At this point it is possible to make some judicious performance comparisons between the two schemes. The first observation is that, in order to get a small mean-square error, ρ must be large and greater than $e^{\sqrt{2\beta}}$. In this case, (51) and (56) apply. On the other hand, when $\sqrt{2\beta} > \ln\rho$, and $\beta \to \infty$, performance deteriorates rapidly as can be seen from (52) and (57). Suppose now that a large signal-to-noise ratio is available and we wish to obtain the same mean-square error in both schemes. How do the signaling rates compare?

Equating (51) and (56) shows that $\beta_d/\beta_L \sim 9/4$ when these quantities are large. In other words, asymptotically, the signaling speed of the cable may be increased by more than a factor of two with the use of decision feedback. Clearly when β is small no significant advantage can be obtained from using decision feedback equalization.

To exhibit these phenomena further, we have used numerical integration to evaluate (49) and (53) and checked the accuracy by summing terms in the various power series. The results of these calculations are exhibited graphically in Figs. 5 through 9. A striking feature in all these curves is the manner MSE degrades as β increases. The linear MSE exhibits a sharp threshold while the MSE for the decision feedback equalizer degrades much more gracefully.

Next we wish to examine the possible payoffs when the inband characteristics of the transmitter filter are optimized. To do this explicitly, we follow the procedure outlined in Section IV. Equation (30) must first be evaluated for the cable characteristics. (We omit all straightforward integrations and algebraic manipulations.) Re-

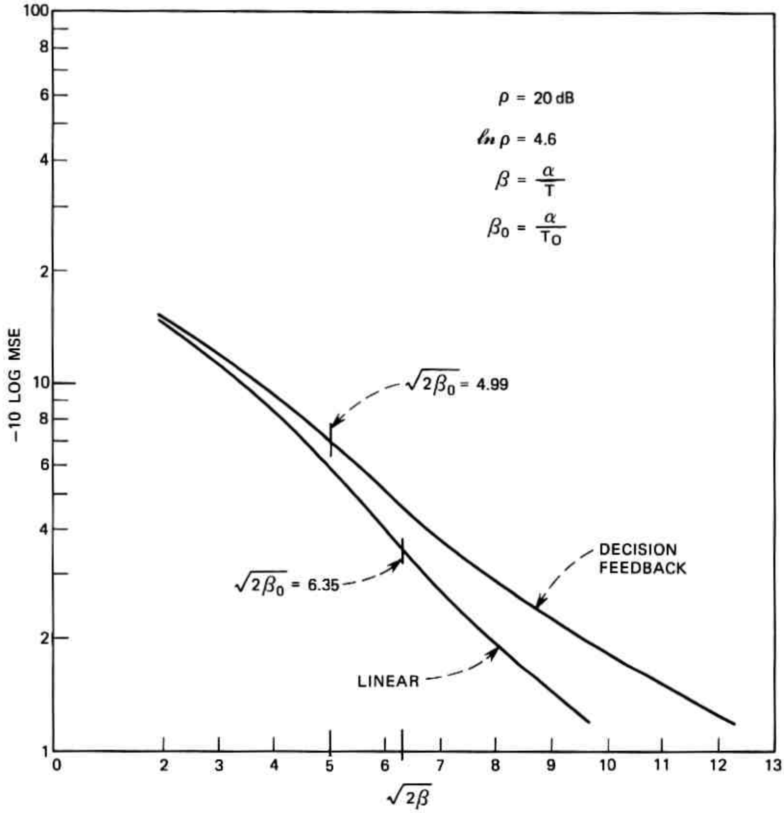

Fig. 5—MSE in dB vs $\sqrt{2\beta}$ for $\rho = 20$ dB.

writing eq. (30),

$$\rho = \frac{P}{\left(\dfrac{N_0}{2}\dfrac{1}{T}\right)} = \frac{T}{\pi}\int_0^{\omega_0}\left[\frac{1}{H^2(\omega_0)} - \frac{1}{H^2(\omega)}\right]d\omega$$

$$= \frac{\beta_0}{\beta}\left[e^{\sqrt{2\beta_0}} - \frac{F(2\beta_0)}{\beta_0}\right], \tag{58}$$

where

$$\beta_0 = \alpha/T_0, \qquad \omega_0 = \pi/T_0, \qquad \beta = \alpha/T,$$
$$H^2(\omega) = e^{-\sqrt{2\omega\alpha/\pi}},$$

and

$$F(x) = e^{\sqrt{x}}(\sqrt{x} - 1) + 1 \sim x/2 \text{ when } x \text{ is small.}$$

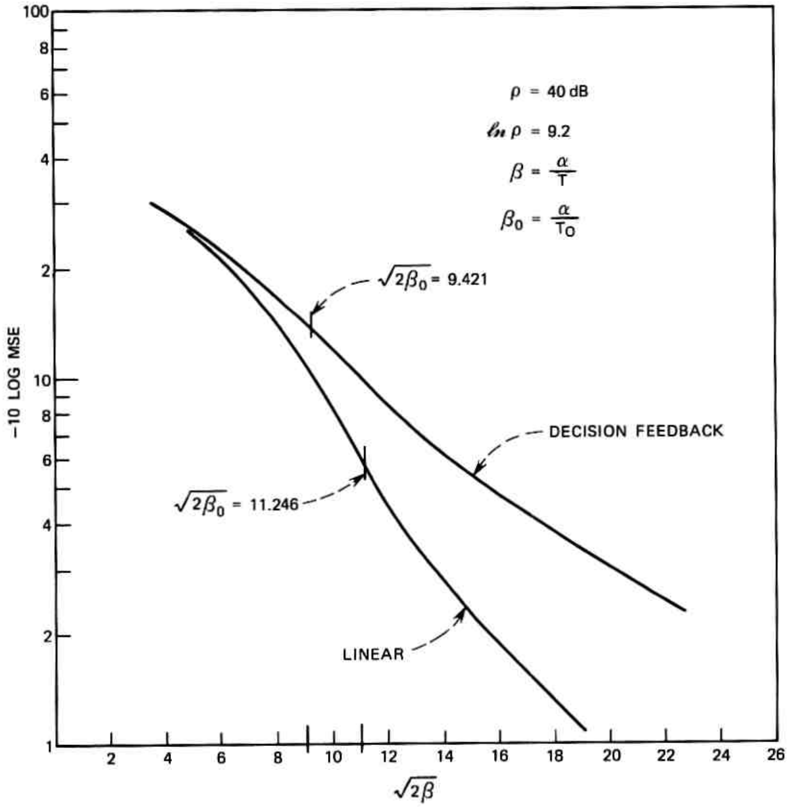

Fig. 6—MSE in dB vs $\sqrt{2\beta}$ for $\rho = 40$ dB.

The explicit evaluation of MSE is as follows: For a given ρ, β solve (58) for β_0. If $\beta_0 > \beta$, calculate MSE from eq. (26),

$$-\ln\left[\frac{\text{MSE}}{\sigma_a^2}\right] = \ln M_d = \ln \mu + \frac{T}{\pi}\int_0^{\pi/T} \ln H^2(\omega)d\omega.$$

An explicit evaluation gives

$$M_d = \frac{e^{\frac{4}{3}\sqrt{2\beta}}}{\rho + \dfrac{F(2\beta)}{\beta}}, \qquad \beta_0 \geq \beta. \tag{59}$$

On the other hand, if (58) yields a $\beta_0 < \beta$, use formula (31) to compute

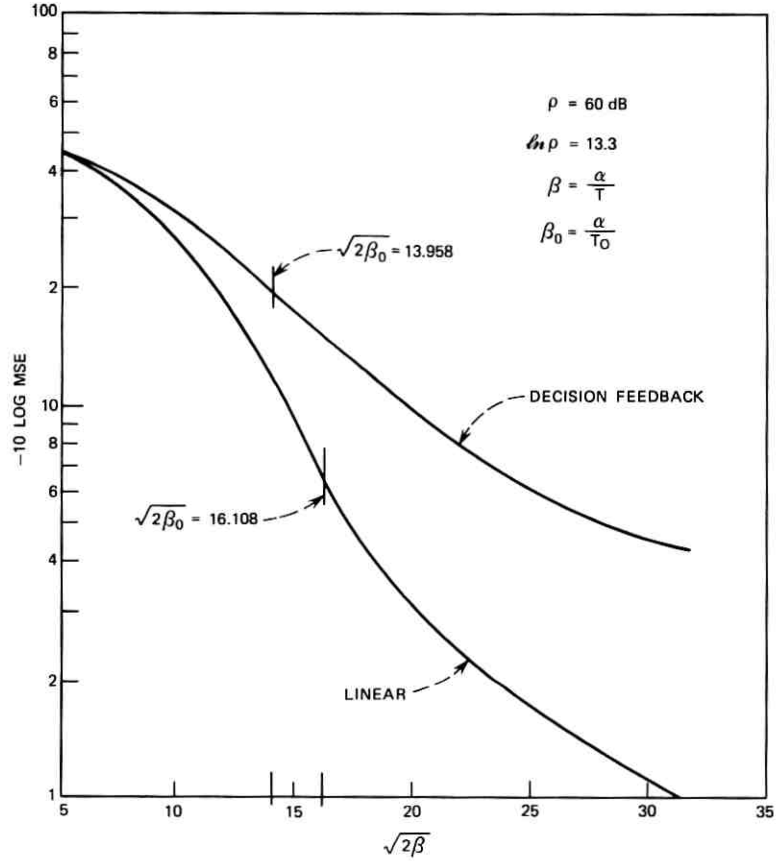

Fig. 7—MSE in dB vs $\sqrt{2\beta}$ for ρ = 60 dB.

MSE. The explicit evaluation for this case gives

$$M_d = e^{-\frac{1}{2}(\beta_0/\beta)\sqrt{2\beta_0}}, \qquad \beta_0 \geq \beta, \tag{60}$$

$$= \left[\frac{1}{\rho\,\dfrac{\beta}{\beta_0} + F(2\beta_0)/\beta_0}\right]^{\frac{1}{2}(\beta_0/\beta)},$$

where $e^{\sqrt{2\beta_0}}$ was obtained from (58). It can be checked that when $\beta_0 = \beta$ in (58), eq. (59) equals (60) as it must.

Let us now pause and examine what these optimized results are telling us. Suppose β is fixed in (58) and ρ is allowed to increase.

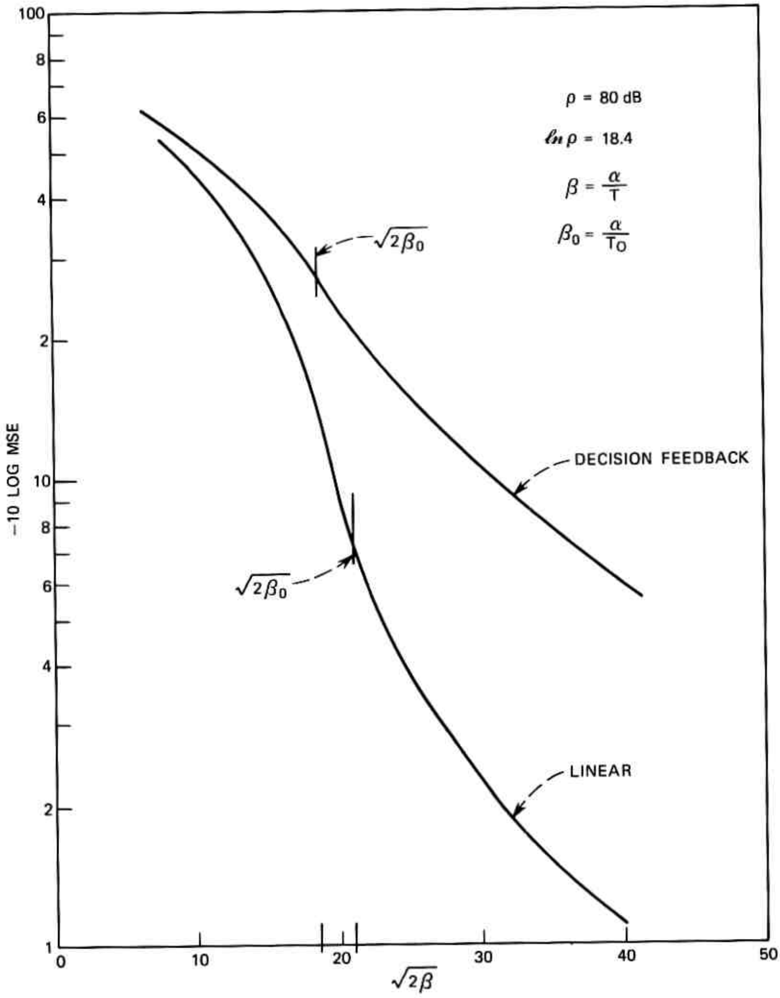

Fig. 8—MSE in dB vs $\sqrt{2\beta}$ for ρ = 80 dB.

Eventually a β_0 will be found which satisfies (58) and which ultimately will be greater than β. The physical implication of finding a β_0 which is less than β is that the transmitting filter cuts off before the Nyquist frequency $\frac{1}{2}T$. This will occur only when ρ is relatively small and thus results in a poor MSE. Practically, the region of interest is when ρ is large such that $\beta_0 \geqq \beta$, in which case the filter cuts off at the Nyquist frequency. In this region (59) applies and, upon comparing (59) with

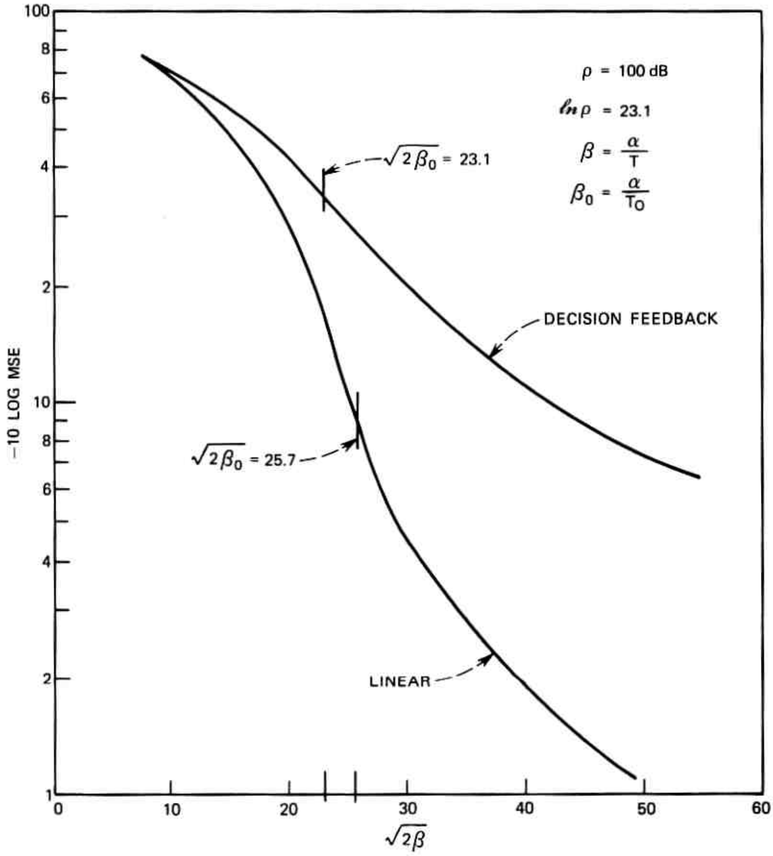

Fig. 9—MSE in dB vs $\sqrt{2\beta}$ for ρ = 100 dB.

the asymptotic suboptimized result, (51) shows that

$$\frac{e^{\frac{3}{2}\sqrt{2\beta}}}{\rho + \dfrac{F(2\beta)}{\beta}} \leqq \frac{e^{\frac{3}{2}\sqrt{2\beta}}}{\rho}\,.$$

Since $\min_\beta [F(\beta)/\beta] = 1$, the optimized result appears to be asymptotically equal to the suboptimized result. In other words, in the region where $\ln \rho > \sqrt{2\beta}$ and $\rho \to \infty$ no benefits are obtained from inband optimization. This is also evident from eq. (25) since when $\ln \rho$ is large relative to $\sqrt{2\beta}$ the structure of the optimum transmitting filter is a constant. The situation where $\beta_0 < \beta$ is slightly more complicated to compare. Here inband optimization should perhaps be beneficial. However, comparisons in this case between the optimized MSE and the suboptimized ones must be made on the basis of the same transmitted power rather than signal-to-noise ratio because the systems operate over different bandwidths.

Again for comparison purposes, we summarize the formulas that apply when the inband characteristics of the transmitting filter in a system using only linear equalization are optimized. Berger and Tufts[6] carried out such an optimization and the procedure is similar to the one carried out in Section V. Adopting our notation and the same definition of parameters as above we can obtain explicitly the following formulas applicable for a linear system.

Choose a ρ and a β and solve for β_0 in the equation below:

$$\rho = \frac{\beta_0}{\beta}\left[\frac{4}{\beta_0}F(\beta_0/2)e^{\sqrt{\beta_0 2}} - \frac{1}{\beta_0}F(2\beta_0)\right]. \tag{61}$$

If $\beta_0 > \beta$, calculate

$$M_L = \frac{\left[\frac{4}{\beta}F(\beta/2)\right]^2}{\rho + \frac{F(2\beta)}{\beta}}, \qquad \beta_0 > \beta. \tag{62}$$

If β_0 in (61) is $< \beta$, calculate

$$M_L = 1 - \frac{\beta_0}{\beta} + \frac{4}{\beta}F(\beta_0/2)e^{-\sqrt{\beta_0/2}}. \tag{63}$$

It is now possible to cross plot the formulas derived in this section *ad nauseum*. We shall show only two sets of graphs. Figures 10 and 11 show four curves of MSE/σ_a^2 in dB vs $\sqrt{2\beta}$ where $E = 2\alpha P/N_0$ and P is the transmitted power divided by the parameter $N_0/2\alpha$. In each case we plot the optimized results and the suboptimized results. The optimized decision feedback equalizer results were evaluated from equations (58), (59), and (60) and from equations (61), (62), and (63) for the linear equalizer. The nonoptimized results are given in (48) and (53) respectively. In all cases $E = \rho/\beta$ in dB. Marked on the curves is the value of $\sqrt{2\beta_0}$ where the transmitting filters cut off. We show two cases, $10 \log_{10} E = 60$ and $10 \log_{10} E = 100$. It appears that inband optimization does not provide a great deal of performance enhancement. As expected, inband optimization yields more improvement in the linear equalization scheme than in decision feedback.

We have also evaluated the optimized results as a function of the actual signal-to-noise ratio ρ, where the noise is measured in whatever

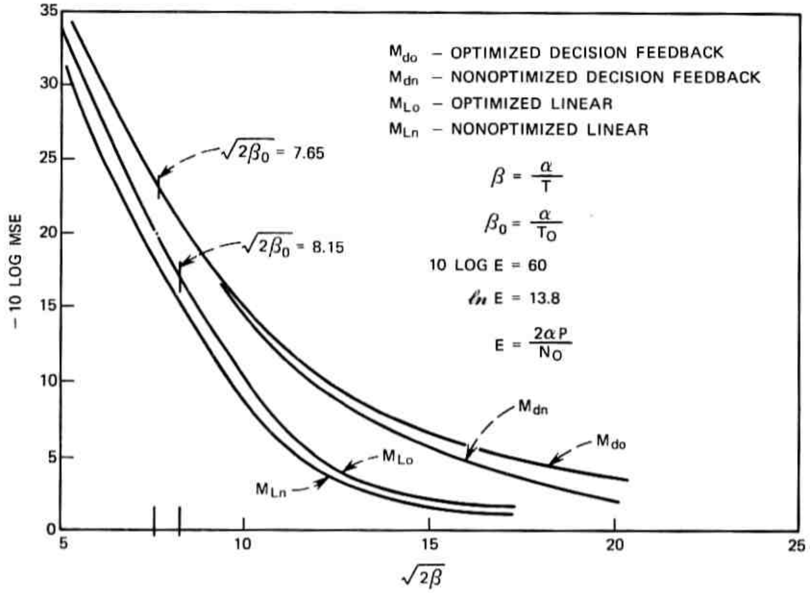

Fig. 10—MSE in dB vs $\sqrt{2\beta}$ for $E = 60$ dB.

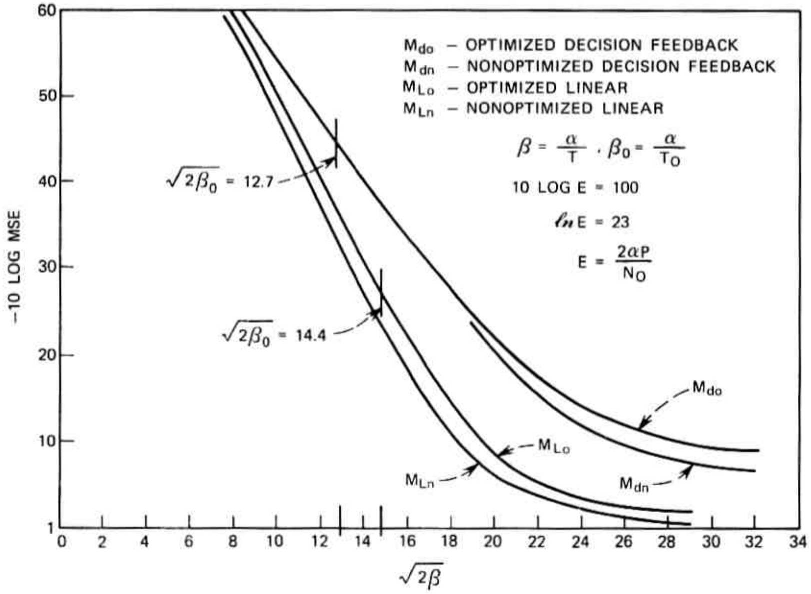

Fig. 11—MSE in dB vs $\sqrt{2\beta}$ for $E = 100$ dB.

band happens to be optimum. We found insignificant differences between these and the suboptimized results shown in Figs. 5 through 10.

In concluding this section we wish to stress that in practice error propagation problems may negate the indicated theoretical results for this channel. When the MSE is large, errors will result. In addition, the tap gains of the feedback filter may become quite large causing those errors which do result to propagate.

VII. ACKNOWLEDGMENTS

During the course of this work I have spoken to many people about this problem and received valuable suggestions. I wish to name these people in alphabetical order: L. H. Brandenburg, D. D. Falconer, G. J. Foschini, R. D. Gitlin, S. Halfin, R. W. Lucky, F. R. Magee, J. E. Mazo, S. O. Rice, and M. Segal. Finally, a great deal of credit is due to R. R. Anderson who did all the numerical work in this paper.

APPENDIX A

Solution of the Wiener-Hopf Equations

We wish to solve the set of linear equations

$$R_k = \sum_{n=-\infty}^{0} M_{n-k} S_n, \qquad k = 0, -1, -2 \cdots, \tag{64}$$

where $\{R_k\}_{-\infty}^{\infty}$ and $M_{n-k} = R_{n-k} + N_0' \delta_{n-k}$ ($\delta_{n-k} = 1, n = k; \delta_{n-k} = 0, n \neq k$) are given.

Since $\{M_n\}_{-\infty}^{\infty}$ is a correlation sequence with positive Fourier coefficients it is well known that it can be represented as the discrete convolution of a sequence $\{M_n^-\}_{-\infty}^{0}$ and a sequence $\{M_n^+\}_0^{\infty}$, namely

$$M_n = \sum_{j=0}^{\infty} M_j^+ M_{n-j}^- \qquad \text{for all } n. \tag{65}$$

Let the sequence $\{X_n\}_{-\infty}^{\infty}$ be determined from

$$R_k = \sum_{j=0}^{\infty} M_j^+ X_{k-j}, \qquad \text{all } k. \tag{66}$$

Substituting (65) and (66) into (64) gives

$$\sum_{j=0}^{\infty} M_j^+ \left\{ X_{k-j} - \sum_{n=-\infty}^{0} S_n M_{n-k-j}^- \right\} = 0. \tag{67}$$

Clearly a solution of

$$X_k = \sum_{n=-\infty}^{0} S_n M_{n-k}^{-}, \qquad k \leqq 0, \tag{68}$$

is also a solution of (67).

Define the two-sided discrete Fourier transform of a sequence $\{X_n\}_{-\infty}^{\infty}$ by

$$X(\theta) = \sum_{n=-\infty}^{\infty} X_n e^{in\theta}.$$

Take the transform of both sides of (66) to obtain

$$R(\theta) = M^{+}(\theta) X(\theta). \tag{69}$$

The one-sided transform $\left(\sum_{n=-\infty}^{0} \right)$ of (68) is

$$X^{-}(\theta) = S^{-}(\theta) M^{-}(\theta) \tag{70}$$

and $X^{-}(\theta)$ is obtained from (69) as

$$X^{-}(\theta) = \left[\frac{R(\theta)}{M^{+}(\theta)} \right]_{-}, \tag{71}$$

where $[\cdot]_{-}$ stands for "projection to negative integers only." To obtain the projection, expand $[\cdot]$ in a two-sided Fourier series and retain only the part of the series containing negative/positive coefficients (including zero).

Thus the desired solution is

$$S^{-}(\theta) = \frac{1}{M^{-}(\theta)} \left[\frac{R(\theta)}{M^{+}(\theta)} \right]_{-}. \tag{72}$$

To proceed further, observe that since $M(\theta) = M^{+}(\theta) M^{-}(\theta)$ and $M(\theta) = R(\theta) + N_0'$ it is possible to calculate explicitly

$$\left[\frac{R(\theta)}{M^{+}(\theta)} \right]_{-} = M^{-}(\theta) + \frac{N_0'}{\gamma_0}, \tag{73}$$

where γ_0 is the dc coefficient of $M^{+}(\theta)$.

The final solution for $S^{-}(\theta)$ is therefore

$$S^{-}(\theta) = 1 - \frac{N_0'}{M^{-}(\theta)\gamma_0}. \tag{74}$$

As pointed out in the text, there are various methods available for calculating $M^{\pm}(\theta)$ from a known function $M(\theta)$. We briefly outline one such approach. Since $M(\theta) > 0$, $0 \leqq \theta \leqq 2\pi$, $\ln M(\theta)$ may be expanded in a two-sided Fourier series

$$\ln M(\theta) = \sum_{n=-\infty}^{0} \gamma_n^- e^{in\theta} + \sum_{n=0}^{\infty} \gamma_n^+ e^{in\theta}. \tag{75}$$

Knowing the sequence $\{\gamma_n^\pm\}_{-\infty}^{\infty}$ we can get immediately

$$M^+(\theta) = \exp\left\{ \sum_{n=0}^{\infty} \gamma_n^+ e^{in\theta} \right\} \tag{76}$$

and

$$M^-(\theta) = \exp\left\{ \sum_{n=-\infty}^{0} \gamma_n^- e^{in\theta} \right\}.$$

Notice that the dc term of $M^+(\theta)$ equals the dc term of $M^-(\theta)$.

APPENDIX B

Evaluation of Integrals

B.1 *Decision Feedback*

The detailed evaluation of

$$I = 2\int_0^{\frac{1}{2}} \ln\left[1 + \rho e^{-\sqrt{4\beta y}}\right] dy \tag{77}$$

is accomplished as follows: Change the variable of integration to $x = (\sqrt{4\beta y} - \ln \rho)$ which gives

$$I = \frac{1}{\beta}\int_{-\ln \rho}^{\sqrt{2\beta}-\ln \rho} \ln[1 + e^{-x}][x + \ln \rho]dx. \tag{78}$$

Assume $\sqrt{2\beta} > \ln \rho > 0$ and write (78) as

$$\begin{aligned} I &= \frac{1}{\beta}\left(\int_{-\ln \rho}^{0} + \int_{0}^{\sqrt{2\beta}-\ln \rho} \right) \\ &= \frac{1}{\beta}\int_0^{\ln \rho} x(\ln \rho - x)dx + \frac{1}{\beta}\int_0^{\ln \rho} (\ln \rho - x)\ln[1 + e^{-x}]dx \\ &\qquad + \frac{1}{\beta}\int_0^{\sqrt{2\beta}-\ln \rho} (x + \ln \rho)\ln[1 + e^{-x}]dx \qquad (79) \\ &= \frac{(\ln \rho)^3}{2\beta} - \frac{(\ln \rho)^3}{3\beta} + \frac{1}{\beta}\int_0^{\ln \rho} (\ln \rho - x)\ln[1 + e^{-x}]dx \\ &\qquad + \frac{1}{\beta}\int_0^{\sqrt{2\beta}-\ln \rho} (x + \ln \rho)\ln(1 + e^{-x})dx. \end{aligned}$$

Since $x > 0$ in the range of integration, expand

$$\ln [1 + e^{-x}] = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{e^{-nx}}{n}$$

and substitute into (79) to obtain

$$I = \frac{(\ln \rho)^3}{6\beta} + \frac{1}{\beta} \sum_{n=1}^{\infty} (-1)^{n+1} \left\{ \frac{\ln \rho}{n} A_n(\ln \rho) - \frac{B_n(\ln \rho)}{n} \right.$$
$$\left. + \frac{B_n(\sqrt{2\beta} - \ln p)}{n} + \frac{\ln \rho}{\beta} \frac{A_n(\sqrt{2\beta} - \ln \rho)}{n} \right\}, \quad (80)$$

where

$$A_n(\xi) = \frac{1}{n}(1 - e^{-n\xi})$$

and

$$B_n(\xi) = \frac{1 - e^{-n\xi} - n\xi e^{-n\xi}}{n^2}.$$

Collecting terms and recognizing that

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^2} = \frac{\pi^2}{12}$$

we finally get

$$I = \frac{(\ln \rho)^3 + \pi^2 \ln \rho}{6\beta} - \sqrt{\frac{2}{\beta}} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(\rho e^{-\sqrt{2\beta}})^n}{n^2}$$
$$+ \frac{1}{\beta} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^3} \left[\frac{1}{\rho^n} - (\rho e^{-\sqrt{2\beta}})^n \right]. \quad (81)$$

When $\ln \rho > \sqrt{2\beta} > 0$, (77) can be expressed in the form

$$I = \frac{1}{\beta} \int_{-\ln \rho}^{0} (x + \ln \rho) \ln [1 + e^{-x}] dx$$
$$+ \frac{1}{\beta} \int_{0}^{-(\ln \rho - \sqrt{2\beta})} (x + \ln \rho) \ln [1 + e^{-x}] dx$$
$$= \frac{1}{\beta} \int_{0}^{\ln \rho} (\ln \rho - x)[x + \ln(1 + e^{-x})] dx$$
$$- \frac{1}{\beta} \int_{0}^{\ln \rho - \sqrt{2\beta}} (\ln \rho - x)[x + \ln (1 + e^{-x})] dx. \quad (82)$$

At this stage $\ln (1 + e^{-x})$ is again expanded in a power series and when the terms are collected we obtain

$$I = \ln \rho - \frac{2}{3} \sqrt{2\beta} + \sqrt{\frac{2}{\beta}} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{e^{n\sqrt{2\beta}}}{n^2 \rho^n}$$
$$+ \frac{1}{\beta} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n^3 \rho^n} [1 - e^{n\sqrt{2\beta}}]. \quad (83)$$

B.2 *Linear Equation*

The integral we wish to evaluate here is

$$I = 2\int_0^{\frac{1}{2}} (1 + \rho e^{-\sqrt{4\beta y}})^{-1} dy. \tag{84}$$

We follow the identical procedure as in the previous case. First change the variable of integration to obtain

$$I = \frac{1}{\beta}\int_{-\ln\rho}^{\sqrt{2\beta}-\ln\rho} \left(\frac{x+\ln\rho}{1+e^{-x}}\right) dx. \tag{85}$$

Assume that $\sqrt{2\beta} > \ln\rho > 0$ and write

$$\begin{aligned} I &= \frac{1}{\beta}\left(\int_{-\ln\rho}^{0} + \int_0^{\sqrt{2\beta}-\ln\rho}\right) \\ &= \frac{1}{\beta}\int_0^{\sqrt{2\beta}-\ln\rho} \frac{(x+\ln\rho)}{1+e^{-x}}\, dx + \frac{1}{\beta}\int_0^{\ln\rho} \frac{e^{-x}(\ln\rho - x)}{1+e^{-x}}\, dx \\ &= \frac{1}{2\beta}(\sqrt{2\beta} - \ln\rho)(\sqrt{2\beta} + \ln\rho) \\ &\qquad + \frac{1}{\beta}\sum_{n=1}^{\infty} (-1)^n [B_n(\sqrt{2\beta} - \ln\rho) + B_n(\ln\rho)] \\ &\qquad + \frac{\ln\rho}{\beta}\sum_{n=1}^{\infty} (-1)^n [A_n(\sqrt{2\beta} - \ln\rho) - A_n(\ln\rho)] \\ &= 1 - \frac{(\ln\rho)^2}{2\beta} - \frac{\pi^2}{6\beta} + \sqrt{2/\beta}\,\ln[1 + \rho e^{-\sqrt{2\beta}}] \\ &\qquad + \frac{1}{\beta}\sum_{n=1}^{\infty} (-1)^{n+1}\frac{1}{n^2}\left(\frac{1}{\rho^n} + (\rho e^{-\sqrt{2\beta}})^n\right). \end{aligned} \tag{86}$$

When $\ln\rho > \sqrt{2\beta} > 0$ we get

$$\begin{aligned} I &= \frac{1}{\beta}\int_0^{\ln\rho} \frac{e^{-x}(\ln\rho - x)}{1+e^{-x}}\, dx - \frac{1}{\beta}\int_0^{\ln\rho-\sqrt{2\beta}} \frac{e^{-x}(\ln\rho - x)}{1+e^{-x}}\, dx \\ &= \frac{1}{\beta}\sum_{n=1}^{\infty} (-1)^{n+1}\{\ln\rho A_n(\ln\rho) - B_n(\ln\rho) \\ &\qquad + \ln\rho A_n(\ln\rho - \sqrt{2\beta}) + B_n(\ln\rho - \sqrt{2\beta})\} \end{aligned} \tag{87}$$

and after collecting terms we finally obtain

$$I = \sqrt{\frac{2}{\beta}}\ln\left[1 + \frac{e^{\sqrt{2\beta}}}{\rho}\right] + \frac{1}{\beta}\sum_{n=1}^{\infty} (-1)^{n+1}\frac{1}{n^2}\left[\frac{1}{\rho^n} - \left(\frac{e^{\sqrt{2\beta}}}{\rho}\right)^n\right].$$

REFERENCES

1. George, D. A., Coll, D. C., Kay, A. R., and Bowen, R. R., "Channel Equalization for Data Transmission," The Engineering Journal (Canada), *53*, May 1970.
2. Keeler, R. J., "Construction and Evaluation of a Decision Feedback Equalizer," Rec. IEEE Int. Conf. Commun., Montreal, Canada, June 14–16, 1971.
3. Price, Robert, "Nonlinearly Feedback-Equalized PAM vs Capacity for Noisy Linear Channels," Rec. IEEE Int. Conf. Commun., Philadelphia, Pa., June 19–21, 1972.
4. Monsen, P., "Feedback Equalization for Fading Dispersive Channels," IEEE Trans. Info. Theory, *IT-17*, January 1971.
5. Doob, J. L., *Stochastic Processes*, New York: John Wiley & Sons, Inc., May 1967, pp. 159–161.
6. Berger, T., and Tufts, D. W., "Optimum Pulse Amplitude Modulation, Part: I: Transmitter-Receiver Design and Bounds from Information Theory," IEEE Trans. Info. Theory, *IT-13*, April 1967.
7. Lucky, R. W., "Decision Feedback and Faster-than-Nyquist Transmission," Paper 7.6, Abstracts 1970 Int. Symp. Info. Theory, Noordwijk, The Netherlands, June 15–19, 1970.
8. Miyakawa, H., and Harashima, H., "Capacity of Channels with Matched Transmission Technique, for Peak Transmitting Power Limitation," Nat. Conv. Rec. Inst. Elec. Commun. Eng. Japan, No. 1268, August 1969, p. 1269.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Traffic Measurement Biases Induced by Partial Sampling

By A. DESCLOUX

(Manuscript received March 22, 1973)

Under equilibrium conditions, the sample average of the delays encountered by all the calls submitted during a given time interval is an unbiased estimate of the mean of the delay distribution. If some of the delays are not observed, the resulting sample average need no longer be an unbiased estimator of the corresponding population mean. This is the case when, for instance, only a limited number of delays can be timed simultaneously. The purpose of this paper is to investigate these biases for queuing systems when only one clock is available and thus one delay only can be measured at a time. It is shown that, regardless of the order of service, the expected value of the observed average delays is always smaller than the mean waiting time for all calls.

Although the average delay on all calls is independent of the order of service, the measurement biases resulting when only one delay can be measured at once depend on the queue discipline. In particular, we shall show that the average delay for all calls is always larger than the average delay of the observed calls even if these calls are always served last (observed-call served-last).

I. INTRODUCTION

The following remarks due to J. F. C. Kingman appear in the *Proceedings of the Symposium on Congestion Theory* held at the University of North Carolina in 1964 (Ref. 1, pp. 314–315): "To illustrate the pitfalls of inference from congestion systems, let me tell a (more or less true) story. It was desired to estimate the mean waiting time in a particular queuing system, and for technical reasons, only one customer could be timed at once. Thus the waiting time ω_1 of a customer was measured. When he entered service, the next customer to arrive was observed and his waiting time ω_2 was noted. This procedure continued, the waiting times ω_3, ω_4, $\cdots$ being measured, and, for large n,

$$n^{-1}(\omega_1 + \omega_2 + \cdots + \omega_n)$$

was used as an estimate of the waiting time. It is, however, strongly biased and inconsistent, because of the selection of the customers to be observed. The mean waiting time is overestimated by a factor which becomes arbitrarily large as the traffic intensity approaches one." We stress that, according to this sampling procedure, a customer is observed if and only if it arrives when the clock is free.

Another instance of biases induced by the measurement procedure is reported by Oberer and Riesz.[2] These authors have investigated the possibility of estimating blocking probabilities in telephone networks by means of test calls generated by a single source repeatedly calling a dedicated number. Their study shows that the proportion of blocked test-calls does not yield a suitable estimate of the grade of service as it is markedly biased downwards. As expected the bias becomes larger as the intervals between consecutive (nonoverlapping) test-calls becomes smaller. It is also established in Ref. 2 that the relative test-call biases increase as the blocking probability decreases.

It is worth noting that the biases studied in Ref. 2, as well as here, are of a different sign than those referred to by Kingman. Much more important, however, is the fact that measurement techniques which, superficially, appear to be adequate may prove to be very unreliable. It is thus becoming increasingly clear that great care is required in the design of performance measurements for stochastic service systems so that unanticipated biases are not encountered.

The purpose of this paper is to investigate the effects of partial sampling on the estimate of the mean (overall) waiting time obtained by averaging measured delays. The biases induced by such limitations will be studied here for M/G/1 and GI/M/s when at most one call can be observed at once and the estimation procedure is as described by Kingman. We shall see that, in these systems, the equilibrium average delay of the observed calls is always smaller than the equilibrium average delay for all calls.

It is well known that the average delay for all calls is the same for all queue disciplines which are independent of the lengths of the individual calls (no other type of queue disciplines will be considered here). As we shall see, this is not true of the mean *measured* delay when only one delay can be recorded at a time. In this case, both the unconditional and the conditional average delays† are (as expected)

† As customary, unconditional and conditional delays pertain to arbitrary and delayed calls respectively.

smallest when the observed calls are served first and largest when they are served last. (In particular, the second extreme case occurs in systems with first-come last-served queue discipline.) Furthermore, in view of the general inequality mentioned at the end of the preceding paragraph, the upper bound for the unconditional average delay of the observed calls (which is reached when the observed calls are served last) is always a strict lower bound for the unconditional average delay for all calls!

The preceding result pertains to unconditional delays and does not always hold for the average delay of those sampled calls which encounter a delay. Thus for M/M/s, the conditional average delay for all delayed calls is equal to the conditional average delay of the observed delayed calls so long as these are always served last. (Note that for M/M/s, the average delay of the delayed calls is equal to the average length of the busy period, and that the waiting-time distribution of the observed calls coincides with the busy-period distribution for the observed-served-last measurement procedure.) In contrast, for the M/Γ_k/1 queue, the upper bound for the conditional average delay of the observed delay calls is larger than the average conditional delay for all delayed calls when $k > 1$, the inequality being reversed whenever $0 < k < 1$. (Γ_k is used here to designate the gamma distribution with mean 1 and variance k^{-1}. Thus M/Γ_1/s is identical to M/M/s. When k is an integer, Γ_k is the Erlangian distribution often designated E_k.)

Expressions for the moments of the equilibrium delay-distribution of the observed calls are given for M/G/1 and first-come first-served queue discipline. The equilibrium delay-distribution of the observed calls is also derived for M/M/s with order-of-arrival service. (Corresponding results for the "observed-call served-first" and the "observed-call served-last" measurement procedures are immediate.) These formulas are used to show that the biases induced by partial sampling can be quite substantial.

When the average service-time is unity, an assumption made throughout, the average delay, EW, for the single-server queue M/G/1 is given by the formula (Ref. 3, pp. 46–50):

$$EW = \alpha m_2/2(1 - \alpha),$$

where α is the server occupancy and m_2 is the second moment about 0 of the service-time distribution. Since m_2 can be arbitrarily large, no bound can be placed on the value of EW. But when only one delay can be timed at once, we shall see that the expectation of the observed delays cannot exceed $1/(1 - \alpha)$. Therefore for any prescribed value of α, it is always possible to find service-time distributions for which the ratio of the average delay for all calls to the average delay of the observed calls exceeds any given bound.

To simplify the exposition we restrict ourselves to full-access delay systems with recurrent inputs in which delays are measured by means of a single clock. Some of the results obtained below can, however, be extended to more general situations.

(In the sequel, W, with or without affix, is used to designate the waiting time of an arbitrary call while W_*, with or without affix, is used as the generic symbol for the observed delays when only one clock is available.)

II. A GENERAL DELAY FORMULA

Consider the queuing system GI/G/s and suppose that the arrival and service-time distributions are such that equilibrium can be reached. (To avoid trivial qualifications we assume throughout that the mean interarrival time is finite. For the same reason, the underlying distributions are also supposed to be such that simultaneous occurrences of events need not be considered.) The purpose of this section is to derive a formula relating the average delay of the observed calls to the equilibrium probability, Φ, that an observed call has immediate access to a server. To this end we prove first that

$$\Phi = (1 - B)(1 + A)/[(1 - B)(1 + A) + B], \tag{1}$$

where B is the equilibrium blocking probability for all calls and A is the expectation of the number of unobserved calls originating during the waiting time of an arbitrary observed delayed call.

It follows from (1) that the probability that an observed call is blocked is always (strictly) smaller than the overall probability of delay (so long as $B \neq 0$ or $B \neq 1$, two trivial cases that we exclude from our considerations). This, of course, is a consequence of the fact that all nonblocked calls are observed whereas, with one clock only, some delays may not be recorded.

We turn now to the proof of (1). Consider an infinite sequence of consecutive calls and for the ith call ($i = 1, 2, \cdots$) let

$$X_i = \begin{cases} 0 \text{ if the } i\text{th call is delayed,} \\ 1 \text{ if the } i\text{th call is not delayed,} \end{cases}$$

$$Y_i = \begin{cases} 0 \text{ if the } i\text{th call is not observed,} \\ 0 \text{ if the } i\text{th call is observed and not delayed,} \\ 1 \text{ if the } i\text{th call is observed and delayed.} \end{cases}$$

Let $\epsilon > 0$. Then assuming that the system is in equilibrium when the first call arrives, we have, by the integral stationarity theorem (Ref. 4, p. 419),

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1 + \epsilon) + \cdots + (X_n + Y_n + \epsilon)} = \frac{EX_1}{E(X_1 + Y_1) + \epsilon} \tag{2}$$

and

$$\lim_{n\to\infty} \frac{(X_1 + Y_1) + \cdots + (X_n + Y_n)}{(X_1 + \epsilon) + \cdots + (X_n + \epsilon)} = \frac{E(X_1 + Y_1)}{EX_1 + \epsilon}, \tag{3}$$

with probability 1. However, since (Ref. 4, p. 421)

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = EX_1 = \Pr[X_1 = 1] > 0,$$

with probability 1, there is, for almost all realizations of the process, an integer n such that the ratios

$$\frac{X_1 + \cdots + X_m}{(X_1 + Y_1) + \cdots + (X_m + Y_m)}, \qquad m \geqq n,$$

are well defined. Hence, by (2) and (3), we have

$$\frac{EX_1}{E(X_1 + Y_1) + \epsilon} \leqq \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1) + \cdots + (X_n + Y_n)} \leqq \frac{EX_1 + \epsilon}{E(X_1 + Y_1)},$$

with probability 1 and, letting ϵ tend to 0,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1) + \cdots + (X_n + Y_n)} = \frac{EX_1}{E(X_1 + Y_1)}, \tag{4}$$

with probability 1.

[The preceding derivation makes use of the fact that the stationarity—and hence the integral stationarity—of the processes $\{X_i, i = 1, \cdots\}$ and $\{X_i + Y_i, i = 1, \cdots\}$ follows from the property that the random variables X_i and Y_i, $i = 1, \cdots$, whose means are finite, are "translates" defined on the stationarity queuing process (Ref. 4, p. 417 ff.).

[The formulas in Ref. 4, p. 419, Theorem A, involve conditional expectations with respect to fields of invariant events. Under the present circumstances these expressions can be simplified. Indeed let us specify the state of the system, $\mathcal{T}_t$ at time t, by means of the vector whose components are the arrival time of the last request placed before t and the elapsed portions of the service-times in progress at time t. Then the only invariant sets (Refs. 4 and 5) of the process $\{\mathcal{T}_t, \ -\infty < t < \infty\}$ are the whole space and the null-set. This property, in turn, implies that the conditional expectations of the random variables X_i and $X_i + Y_i$ relative to their invariant fields can be replaced by the unconditional expectations EX_i and $E(X_i + Y_i)$, respectively. These and other similar substitutions are made here without formal justification.]

Consider now an infinite sequence of observed calls and let

$$Z_i = \begin{cases} 0 \text{ if the } i\text{th observed call is delayed,} \\ 1 \text{ if the } i\text{th observed call is not delayed.} \end{cases}$$

Then (Ref. 4, p. 421)

$$\lim_{n\to\infty} \frac{Z_1 + Z_2 + \cdots + Z_n}{n} = EZ_1 = \Phi. \tag{5}$$

Since (4) and (5) are both equal to the proportion of observed calls with zero delay over an interval of infinite length, we have

$$\Phi = EX_1/E(X_1 + Y_1). \tag{6}$$

We note that EX_1 is equal to the probability that a call (observed or not) is not delayed. Hence

$$EX_1 = 1 - B, \tag{7}$$

and to complete the proof of (1) we have to show that

$$EY_1 = B/(1 + A). \tag{8}$$

To this end consider again a stationary sequence of observed calls and let A_i be the number of unobserved calls placed during the waiting time of the ith call that is both observed and delayed. Then we have (Ref. 4, pp. 419–421)

$$\lim_{n\to\infty} \frac{n}{n + A_1 + \cdots + A_n} = \frac{1}{1 + EA_1} = \frac{1}{1 + A}, \tag{9}$$

with probability 1.

Furthermore, by the integral stationarity theorem (Ref. 4, p. 419) and a simple ϵ-argument of the type used in the proof of (2), we have:

$$\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{(1 - X_1) + \cdots + (1 - X_n)} = \frac{EY_1}{E(1 - X_1)}, \tag{10}$$

with probability 1.

Since the left-hand sides of (9) and (10) are both equal to the proportion of delayed calls that are observed, we have

$$EY_1 = E(1 - X_1)/(1 + A) = B/(1 + A).$$

This completes the proof of (1).

Our next step will now be to relate A to the average delay, EW_*, of the observed calls. Let W_{i*} be the delay of the ith observed call and let U_i be the interval between the end of the ith and the beginning of the $(i + 1)$st measurement. Let also I_n be the interval between the arrival epochs of the nth and $(n + 1)$st call (in the whole sequence of calls, observed or not). Then we have:

$$(W_{1*} + U_1) + \cdots + (W_{n*} + U_n) = I_1 + \cdots + I_{K_n}, \qquad (11)$$

where K_n, a random variable, is equal to the number of calls placed during the interval that starts with an observed call and ends just before the beginning of the $(n + 1)$st measurement. By the stationarity theorem, we have (Ref. 4, p. 421):

$$\lim_{n\to\infty} \frac{(W_{1*} + U_1) + \cdots + (W_{n*} + U_n)}{n} = E(W_{1*} + U_1), \qquad (12)$$

with probability 1, and, by the strong law of large numbers (note that the I_n's are, by assumption, independent random variables with finite means and that $K_n \geqq n$),

$$\lim_{n\to\infty} \frac{I_1 + I_2 + \cdots + I_{K_n}}{K_n} = \alpha^{-1}, \qquad (13)$$

with probability 1, where α^{-1} is the expected interarrival-time.

Furthermore,

$$K_n = [Z_1 + (1 - Z_1)(1 + A_1)] + \cdots + [Z_n + (1 - Z_n)(1 + A_n)],$$

so that

$$\lim_{n\to\infty} \frac{K_n}{n} = E[Z_1 + (1 - Z_1)(1 + A_1)] = \Phi + (1 - \Phi)(1 + A), \quad (14)$$

with probability 1.

Combining (11)–(14) we find that

$$\alpha E(W_{1*} + U_1) = 1 + A(1 - \Phi). \qquad (15)$$

In particular, when the input is Poissonian, $EU_1 = \alpha^{-1}$ and (15) reduces to

$$\alpha EW_{1*} = A(1 - \Phi). \qquad (16)$$

Thus, taking (1) into account, we find that:

$$EW_* \equiv EW_{1*} = (\Phi + B - 1)/\alpha(1 - B). \tag{17}$$

It should be noted that the preceding relation is valid regardless of the order of service.

When the calls are served in order of arrival, (16) is an immediate consequence of the fact that the waiting time of any given call is not affected by the stream of requests placed after its arrival epoch. This is also true when the observed calls are always served first and (16) can then be written down with equal ease.

We note that (17) can be obtained quickly whenever the epochs at which measurements begin or terminate constitute a renewal process. In such cases, the expected number of observations in time t is (asymptotically)

$$(EW_* + \alpha^{-1})^{-1} \cdot t + 0(1), \ t \text{ large}, \tag{18}$$

and the expected number of arrival points at which the system is empty in time t is

$$\alpha(1 - B) \cdot t + 0(1), \ t \text{ large}. \tag{19}$$

The long-term proportion of observed calls with no delay is given by the ratio of (19) to (18) with probability 1 (cf. Ref. 6, p. 264, alternative form of Theorem IV):

$$\alpha(1 - B)(EW_* + \alpha^{-1}). \tag{20}$$

Since the probability Φ that an arbitrary call is not delayed is independent of the past, a simple application of the strong law of large numbers show that (20) may be equated to Φ and (17) therefore holds. An instance where the preceding argument can be applied is the M/G/1 system with observed calls always served last. With this order of service, there is exactly one call in the system at the termination of each measurement. These epochs constitute a renewal process since they also coincide with the beginnings of the service-times of the observed calls. With an obvious change, the previous argument remains true for M/M/s with observed-calls served-last.

III. TWO EXTREME CASES

In this section we show that if only one clock is available then the expected average delay of the observed calls is largest when the observed calls are served last and smallest when the observed calls are served first. These relations are not statistical: They are satisfied by all the realizations of the process over any finite or infinite time interval regardless of the arrival and service-time distributions. (To avoid ambiguities, we assume that the timing device is free at the beginning of the realizations.)

We note first that under any measurement procedure, all the calls which arrive when all the servers are busy, but no request is waiting, are observed. These calls are the only delayed calls that are observed when the observed calls are served last. Therefore (*i*) the number of observed delayed calls takes its smallest value for the observed-served-last procedure and (*ii*) during the measurement of a delay, D, under this particular procedure any observed delay under any alternate single-clock measurement procedure cannot exceed D. Combining these two facts and taking into account that all calls with zero delay are observed we may conclude that the observed average delay takes always its largest value when the observed calls are served last, as is the case for the first-come last-served queue discipline.

When the observed calls are served first, we note that (*i*) the number of observed delays over any busy period (initial or not) is never smaller than for any other single-clock measurement procedure and (*ii*) to each observed delayed call there corresponds, under any other single-clock measurement procedure, one call whose delay is at least as large and this correspondence involves all the observed delays under the alternate procedure. All calls with zero delay are again observed and the average delay of the observed calls takes therefore its smallest value when the observed calls are served first.

Clearly the conditional average delays of the observed calls do have the same property.

IV. Bounds for the Average Delay of the Observed Calls in M/G/1

The object of this section is to determine the upper and lower bounds for the average delay, EW_*, of the observed calls in M/G/1. These bounds, as we have seen, are reached when the observed calls are served last and first respectively. Under the present conditions, formula (17) may be written as follows:

$$EW_* = (\Phi + \alpha - 1)/\alpha(1 - \alpha). \tag{21}$$

For a given server occupancy α, EW_* is a monotone increasing function of $\Phi = \Phi(\alpha)$. Since $\Phi < 1$, (21) implies that $EW_* < (1 - \alpha)$. Hence for $\alpha < 1$, EW_* is always bounded (but EW is not).

It will be convenient to define here the service backlog, at a given instant t, as the sum of the service-times of all waiting requests plus the residual of the service-time of the request being served. [When calls are served in order of arrival, the service backlog is equal to the virtual waiting time (Ref. 3, p. 59 ff.).]

Now let $F(\cdot)$ be the stationary cumulative distribution of the service backlog at the end of a measurement. The probability, $\Phi(\alpha)$, that an observed call does not suffer a delay is simply the Laplace-Stieltjes transform of $F(\cdot)$ evaluated at α since it is equal to the probability that no call originates during a time interval whose length is that of the service backlog:

$$\Phi(\alpha) = \int_0^\infty e^{-\alpha t} dF(t).$$

Writing $\sigma(\cdot)$ for the Laplace-Stieltjes transform of the service-time distribution we have the following inequality:

$$\Phi(\alpha) \leqq \sigma(\alpha). \tag{22}$$

This inequality is a consequence of the fact that, at the conclusion of a measurement, the service backlog may be represented as the sum of two independent random variables, one of which is the full service-time of the request whose delay has just come to an end while the other is equal to the sum of the service-times of all the waiting requests. Writing $R(\cdot)$ for the c.d.f. of the latter and $S(\cdot)$ for the service-time distribution, we have:

$$\begin{aligned}\Phi(\alpha) &= \int_0^\infty e^{-\alpha t} d\int_0^t R(t-v)dS(v)\\ &= \sigma(\alpha)\int_0^\infty e^{-\alpha t} dR(t) \leqq \sigma(\alpha).\end{aligned}$$

When the observed calls are served last,

$$R(t) = \begin{cases} 1 & \text{for} \quad t \geqq 0, \\ 0 & \text{for} \quad t < 0, \end{cases}$$

and

$$\Phi(\alpha) = \sigma(\alpha).$$

We can therefore conclude that

$$EW_* \leqq \frac{\sigma(\alpha) + \alpha - 1}{\alpha(1-\alpha)}. \tag{23}$$

We are now in a position to prove that the average delay, EW_*, is always smaller than the average delay for all calls (observed or not). Since the service-time is unity, we have:

$$\Phi(\alpha) \leqq \sigma(\alpha) = \int_0^\infty e^{-\alpha t} dS(t) < 1 - \alpha + \frac{\alpha^2 m_2}{2},$$

where m_2 is the second moment of S about the origin. Hence, substituting $1 - \alpha + \alpha^2 m_2/2$ for $\sigma(\alpha)$ in (23) we find that, irrespective of the service order:

$$EW_* < \frac{\alpha m_2}{2(1 - \alpha)} = EW. \tag{24}$$

By (23) and the equality in (24) we also have:

$$EW/EW_* > \frac{\alpha m_2}{2},$$

so that, for any given α, we can always find a service-time distribution such that EW/EW_* exceeds any preassigned value.

We now derive an absolute lower bound for EW_*. As shown above, this bound can be found by assuming that the observed calls are served first. Our first step here is to determine A.

For the observed-served-first procedure, the circumstances under which a positive delay can be observed are as follows: at some time the clock is not in use and a service-time begins, and during this service-time a new call arrives. Thus A is the conditional expectation of the number of arrivals minus 1 during an arbitrary service-time given that at least one call is placed during a service-time. A, therefore, is given by the formula

$$\begin{aligned} A &= \int_0^\infty \sum_{n=1}^\infty (n - 1) \frac{(\alpha t)^n}{n!} e^{-\alpha t} dS(t) \Big/ \int_0^\infty (1 - e^{-\alpha t}) dS(t) \\ &= \int_0^\infty [\alpha t - 1 + e^{-\alpha t}] dS(t) \Big/ \int_0^\infty (1 - e^{-\alpha t}) dS(t) \\ &= [\alpha - 1 + \sigma(\alpha)]/[1 - \sigma(\alpha)]. \end{aligned}$$

By means of (1) and (21), it is now readily shown that, for the observed-calls-served-first procedure:

$$\Phi(\alpha) = \frac{(1 - \alpha)}{2 - \alpha - \sigma(\alpha)}$$

and

$$EW_* = \frac{\sigma(\alpha) + \alpha - 1}{\alpha[2 - \alpha - \sigma(\alpha)]}.$$

Summing up, we have the following inequalities for Φ and EW_*, regardless of the measurement procedure:

$$\frac{1-\alpha}{2 - \alpha - \sigma(\alpha)} \leqq \Phi(\alpha) \leqq \sigma(\alpha), \tag{25}$$

and

$$\frac{\sigma(\alpha) + \alpha - 1}{\alpha[2 - \alpha - \sigma(\alpha)]} \leqq EW_* \leqq \frac{\sigma(\alpha) + \alpha - 1}{\alpha(1-\alpha)}. \tag{26}$$

For exponential service-times, (25) and (26) reduce to

$$(1 - \alpha^2)/(1 + \alpha - \alpha^2) \leqq \Phi_1(\alpha) \leqq 1/(1+\alpha),$$
$$\alpha/(1 + \alpha - \alpha^2) \leqq EW_{*1} \leqq \alpha/(1 - \alpha^2).$$

(The subscript 1 is added to Φ and EW_* to indicate that the service-times are exponentially distributed.)

V. THE SINGLE-SERVER QUEUE M/G/1 WITH ORDER-OF-ARRIVAL SERVICE

In this section we consider the M/G/1 queue under the assumption that the calls are served in order of arrival. Our principal aim here is to determine $\Phi = \Phi(\alpha)$ and then, by means of (17), EW_*. To this end let p_n be the probability that there are n calls in the system immediately after the conclusion of a measurement. Then, relating the state probabilities at two consecutive conclusions of delay measurements, we find that

$$\begin{aligned} p_1 &= \sum_{n=1}^{\infty} p_n \int_0^\infty e^{-\alpha t} dS^{(n)}(t) + \sum_{n=1}^{\infty} p_n \int_0^\infty \alpha t e^{-\alpha t} dS^{(n)}(t), \\ p_k &= \sum_{n=1}^{\infty} p_n \int_0^\infty \frac{(\alpha t)^k}{k!} e^{-\alpha t} dS^{(n)}(t), \qquad k > 1, \end{aligned} \tag{27}$$

where $S^{(n)}$ is the nth convolution of the service-time distribution, S, with itself.

Now let

$$G(x) \equiv \sum_1^\infty p_n x^n.$$

Equations (27) yield:

$$
\begin{aligned}
G(x) = \sum_{m=1}^{\infty} p_m x^m &= \sum_{m=1}^{\infty} x^m \sum_{n=1}^{\infty} p_n \int_0^{\infty} \frac{(\alpha t)^m}{m!} e^{-\alpha t} dS^{(n)}(t) \\
&\qquad\qquad + x \sum_{n=1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} dS^{(n)}(t) \\
&= \sum_{n=1}^{\infty} p_n \int_0^{\infty} \left[\sum_{m=1}^{\infty} \frac{(\alpha t x)^m}{m!} e^{-\alpha t} \right] dS^{(n)}(t) + x \sum_{n=1}^{\infty} p_n \sigma^n(\alpha) \\
&= \sum_{n=1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} (e^{\alpha t x} - 1) dS^{(n)}(t) + x \sum_{n=1}^{\infty} p_n \sigma^n(\alpha) \\
&= \sum_{n=1}^{\infty} p_n \sigma^n[\alpha(1 - x)] - (1-x) \sum_{n=1}^{\infty} p_n \sigma^n(\alpha) \\
&= G\{\sigma[\alpha(1 - x)]\} - (1 - x) G[\sigma(\alpha)].
\end{aligned}
$$

Summing up, we have the relation

$$
G(x) = G\{\sigma[\alpha(1 - x)]\} - (1 - x) G[\sigma(\alpha)]. \tag{28}
$$

Note also that

$$
\Phi(\alpha) = \sum_{1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} dS^{(n)}(t) = G[\sigma(\alpha)].
$$

Let $x_0 \equiv \sigma(\alpha)$ and $x_n \equiv \sigma[\alpha(1 - x_{n-1})]$, $n = 1, 2, \cdots$.

Since $0 \leqq \alpha < 1$, we have $x_0 < x_1 < \cdots \leqq 1$ and $\lim_{n\to\infty} x_n$ does therefore exist. With this notation, we obtain, from (28):

$$
\begin{aligned}
\Phi(\alpha) &= G(x_1) - (1 - x_0)\Phi(\alpha), \\
G(x_1) &= G(x_2) - (1 - x_1)\Phi(\alpha), \\
&\vdots \\
G(x_{n-1}) &= G(x_n) - (1 - x_{n-1})\Phi(\alpha).
\end{aligned} \tag{29}
$$

Adding up these relations, we find that:

$$
\Phi(\alpha)\left[1 + \sum_{m=0}^{n-1} (1 - x_m)\right] = G(x_n),
$$

and, by passing to the limit,

$$
\Phi(\alpha)\left[1 + \sum_{m=0}^{\infty} (1 - x_m)\right] = 1. \tag{30}
$$

(Note that $\lim_{n\to\infty} G(x_n)$ exists since the x_n constitute a positive monotone-increasing sequence bounded by 1. By letting $n \to \infty$ in the last of the relations (29) it follows immediately that $\lim_{n\to\infty} x_n = 1$ so long as $\Phi(\alpha) \neq 0$. This last condition is however clearly satisfied whenever $\alpha < 1$.)

In particular, when the service-times are negative exponential, we have: $S(t) = 1 - e^{-t}$, $t \geqq 0$, $\sigma(s) = (1+s)^{-1}$, and $(1 - x_m) = \alpha^{m+1}/(1 + \alpha + \cdots + \alpha^{m+1})$. Hence, by (30),

$$\Phi(\alpha) = \left[1 + \sum_{m=0}^{\infty} \frac{\alpha^{m+1}}{1 + \alpha + \cdots + \alpha^{m+1}}\right]^{-1}. \tag{31}$$

We examine briefly the case where the service-times have a gamma distribution with parameter k (the subscript k is added to the symbols considered earlier in order to stress their dependence on k). We have, in this case:

$$\sigma_k(\alpha) = [k/(k+\alpha)]^k.$$

Then $\sigma_k(\alpha)$ is a strictly decreasing function of $k(>0)$ as can be seen by taking the derivative of $\ln \sigma_k^{-1}(\alpha) = \ln(1 + \alpha/k)^k$ and using the inequality $\ln(1-x) > x/(1+x)$, $x > 0$ (Ref. 7, p. 68). This monotonicity property of σ_k implies that

$$x_{k0} > x_{k+h,0}, \qquad x_{k1} > x_{k+h,1}, \cdots, \qquad k > 0, \qquad h > 0,$$

and we have, therefore:

$$\sum_{m=0}^{\infty} (1 - x_{km}) < \sum_{m=0}^{\infty} (1 - x_{k+h,m}), \qquad h > 0,$$

so that

$$\Phi_k(\alpha) > \Phi_{k+h}(\alpha),$$

and from (21)

$$EW_{*k} > EW_{*,k+h}, \qquad k > 0.$$

For $k = 1$ (exponential service-time) the conditional average delay for the delayed calls under the observed-last-served procedure is equal to the average length of the busy period. To see this one need only note that each positive observed delay begins with an arrival that occurs when there is exactly one customer in the system and ends when, for the first time thereafter, there is no waiting customer. Hence, for $k = 1$, the conditional delay distribution of the observed calls is the same as the busy-period distribution (Ref. 8, p. 32). Since the average length of the busy period and the average of the conditional waiting times of all the delayed calls are both equal to $(1-\alpha)^{-1}$ (Ref. 3, p. 63), we have

$$\frac{EW_1}{\alpha} = \frac{\overline{EW}_{*1}}{1 - \sigma_1(\alpha)} = \frac{\sigma_1(\alpha) + \alpha - 1}{[1 - \sigma_1(\alpha)]\alpha(1-\alpha)} = \frac{1}{1-\alpha}, \tag{32}$$

where $\overline{EW}_{*1}$ designates the average delay when the observed calls are served last.

Clearly, the inequalities (22) and (23) imply that

$$\frac{EW_{*k}}{1 - \Phi_k} \leqq \frac{\sigma_k(\alpha) + \alpha - 1}{[1 - \sigma_k(\alpha)]\alpha(1 - \alpha)}. \tag{33}$$

We note that if the service-times have a gamma distribution with transform $\sigma_k(s) = (k/k + s)^k$, then the conditional average delay on all delayed calls is given by the formula (Ref. 3, p. 50):

$$\frac{EW_k}{\alpha} = \left[\frac{k + 1}{k}\right]\frac{1}{2(1 - \alpha)}. \tag{34}$$

Substituting $(k/k + \alpha)^k$ for $\sigma_k(\alpha)$ in (33) we obtain the following upper bound for the conditional average delay of the observed delayed calls (this bound, as we know, is reached when the observed calls are served last):

$$\frac{[k/(k + \alpha)]^k + \alpha - 1}{\{1 - [k/(k + \alpha)]^k\}\alpha(1 - \alpha)}. \tag{35}$$

Subtracting (34) from (35) we find that the difference is of the same sign as

$$\alpha - \{1 - [k/(k + \alpha)]^k\}[1 + (k + 1)\alpha/2]. \tag{36}$$

The two factors in the second term of (36) are both increasing functions of k and since (36) vanishes for $k = 1$ [a fact that we already know from (32)] we may conclude that (35) is smaller than (34) for $0 < k < 1$, and greater than (34) for $k > 1$. This proves that, for $k < 1$, the conditional average delay for all delayed calls is still larger than the conditional average delay for all observed delayed calls even if these calls are served last. For $k > 1$, the conditional average delay of the observed delayed calls for the observed-served-last procedure is larger than the conditional average delay of all the delayed calls.

Expressions for the higher moments of the observed calls delay-distribution are also readily obtained. Let F be the equilibrium cumulative distribution of the virtual waiting time, V, at the conclusion of the measurement of a delay (this delay, of course, may be equal to zero). The delay distribution, K, of the calls whose delays are observed, can be readily expressed in terms of F. Indeed we have:

$$K(w) = \Pr[W_* \leqq w] = \alpha\int_0^w \left\{\int_t^\infty \exp - \alpha(y - t)\cdot dF(y)\right\}dt + \int_0^\infty e^{-\alpha y}dF(y). \tag{37}$$

TABLE I—MEANS AND STANDARD DEVIATIONS OF THE DELAY DISTRIBUTIONS FOR ALL CALLS AND FOR ALL OBSERVED CALLS (1 CLOCK) IN M/M/1—FIRST-COME FIRST-SERVED

α	EW_1	SW_1	EW_{*1}	SW_{*1}
0.1	0.11111	0.48432	0.09259	0.42264
0.2	0.25000	0.75000	0.17840	0.58463
0.3	0.42857	1.0202	0.26684	0.72360
0.4	0.66667	1.3333	0.36669	0.87308
0.5	1.0000	1.7321	0.48958	1.0590
0.6	1.5000	2.2913	0.65554	1.3188
0.7	2.3333	3.1798	0.90754	1.7310
0.8	4.0000	4.8990	1.3659	2.5191
0.9	9.0000	9.9499	2.5808	4.7519
0.92	11.500	12.460	3.1398	5.8266
0.94	15.667	16.637	4.0284	7.5793
0.96	24.000	24.980	5.6974	10.987
0.98	49.000	49.990	10.243	20.776
0.99	99.000	99.995	18.393	39.434

The problem of finding the distribution K of the observed delays is thus reduced to the problem of finding F. The distribution F satisfies the following integral equation:

$$F(t) = \sum_{n=1}^{\infty} \int_0^t dS^{(n)}(u) \int_0^{\infty} e^{-\alpha y} \frac{(\alpha y)^n}{n!} dF(y) + \int_0^t dS(u) \int_0^{\infty} e^{-\alpha y} dF(y), \quad (38)$$

where $S^{(n)}$ stands for the nth convolution of S with itself. Equation (38) follows immediately upon noticing that, at the conclusion of a measurement, the only calls in the system are (*i*) the call whose delay has just been measured and (*ii*) all the calls which arrived during the measurement interval (we note that the preceding argument makes essential use of the assumption that calls are served in order of arrival).

Let φ and σ be respectively the Laplace-Stieltjes transform of F and S. Then transforming (38) we obtain

$$\begin{aligned}\varphi(s) &= \sum_{n=1}^{\infty} \sigma^n(s) \int_0^{\infty} e^{-\alpha y} \frac{(\alpha y)^n}{n!} dF(y) + \sigma(s) \int_0^{\infty} e^{-\alpha y} dF(y) \\ &= \varphi\{\alpha[1 - \sigma(s)]\} + \Phi(\alpha)[\sigma(s) - 1]. \quad (39)\end{aligned}$$

This formula may be used to derive recurrence relations for the moments of V. Let $n!\mu_n = EV^n$ and

$$n!\nu_n = (-1)^n \frac{d}{ds^n} \sigma(s) \bigg|_{s=0} = \int_0^{\infty} t^n dS(t),$$

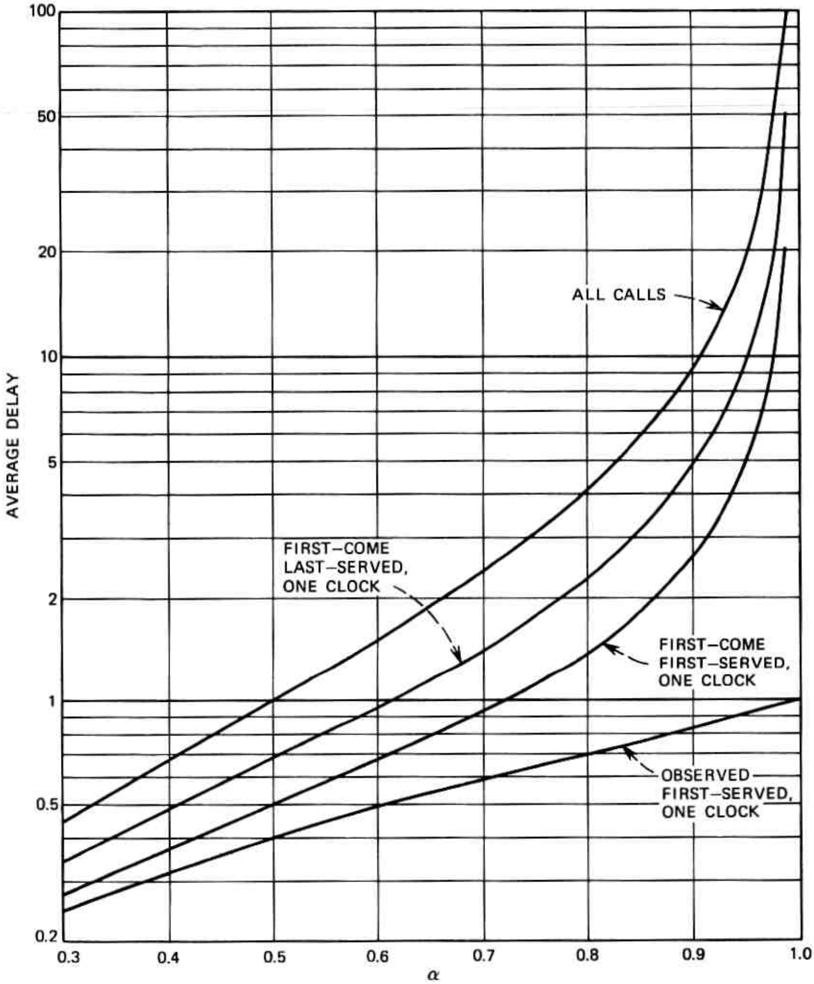

Fig. 1—Average delay for M/M/1 vs occupancy.

so that $n!\nu_n$ is the nth moment of the service-time distribution. Using Faa di Bruno's formula for the derivative of a composite function (Ref. 9, p. 36) we find that:

$$\mu_n = \sum \frac{k!}{k_1!\cdots k_n!} \mu_k \alpha^k \nu_1^{k_1} \cdots \nu_n^{k_n} + \Phi(\alpha)\nu_n, \tag{40}$$

with $k = k_1 + \cdots + k_n$ and the sum over all solutions in non-negative

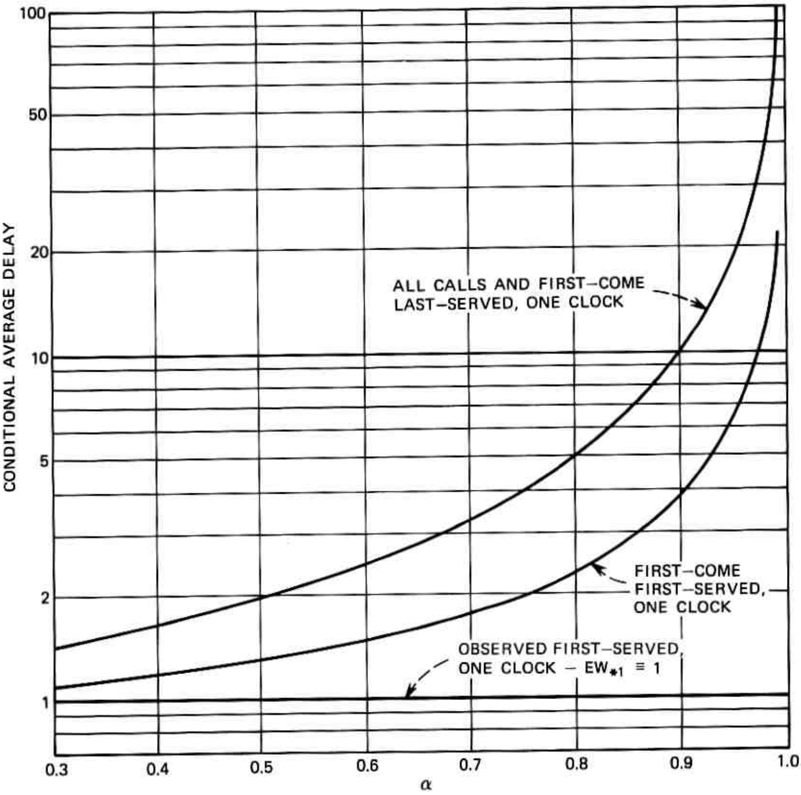

Fig. 2—Conditional average delay for M/M/1 vs occupancy.

integers of $k_1 + 2k_2 + \cdots + nk_n = n$ (note that $\nu_1 = 1$ since the average service-time is assumed here to be equal to 1).

In particular, for $n = 1, 2$, and 3 we have:

$$\mu_1 = EV = \Phi(\alpha)/(1-\alpha),$$

$$\mu_2 = \frac{EV^2}{2!} = \Phi(\alpha)\nu_2/(1-\alpha)(1-\alpha^2),$$

$$\mu_3 = \frac{EV^3}{3!} = \Phi(\alpha)[2\alpha^2\nu_2^2 + \nu_3(1-\alpha^2)]/(1-\alpha)(1-\alpha^2)(1-\alpha^3),$$

where $\Phi(\alpha)$ is the probability that an observed call is not delayed.

When the service-time distribution is exponential (with mean 1) we have $\nu_i = 1, i = 0, 1, \cdots$, and (40) becomes:

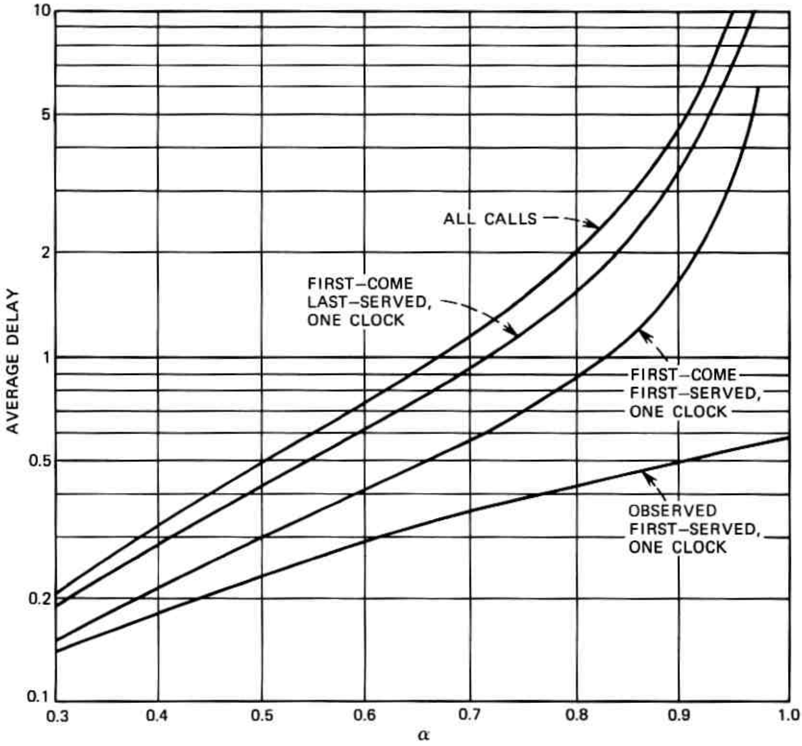

Fig. 3—Average delay for M/D/1 vs occupancy.

$$\mu_n = \sum \frac{k!}{k_1! \cdots k_n!} \mu_k \alpha^k + \Phi_1(\alpha)$$

$$= \sum_{k=1}^{n} \binom{n-1}{k-1} \mu_k \alpha^k + \Phi_1(\alpha), \qquad n \geqq 1.$$

Equation (37) may be used to express the moments of K in terms of the moments of V. We have:

$$EW_*^n = \int_0^\infty w^n dK(w) = \alpha \int_0^\infty w^n \int_w^\infty \exp - \alpha(y - w) dF(y) dw,$$

and, upon integrating by parts, we obtain

$$EW_* = EV - \frac{1}{\alpha}[1 - \Phi(\alpha)],$$

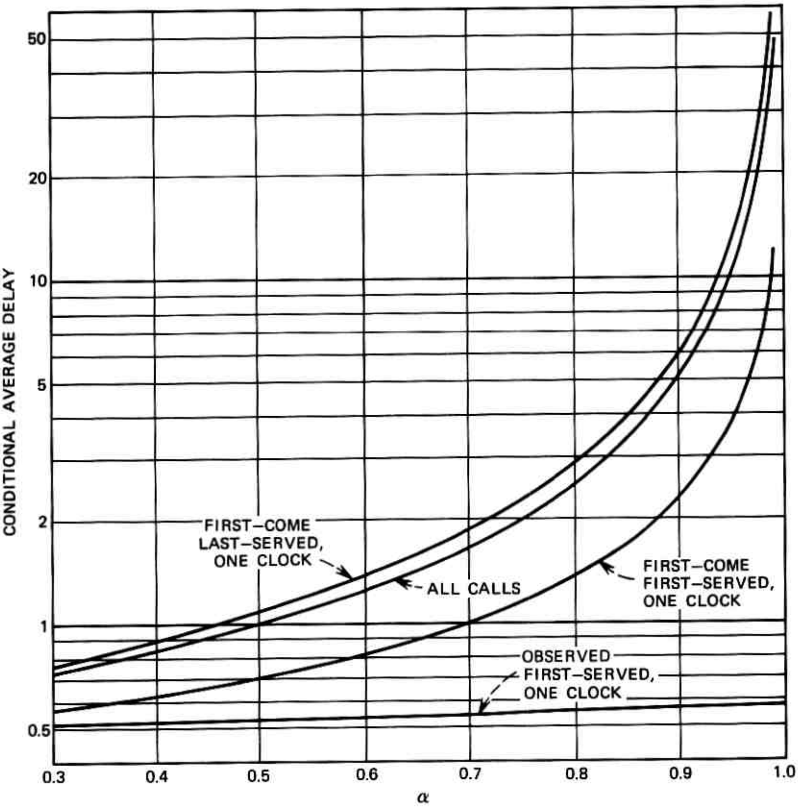

Fig. 4—Conditional average delay for M/D/1 vs occupancy.

and

$$EW_*^{n+1} = EV^{n+1} - \frac{n+1}{\alpha} EW_*^n, \qquad n > 0.$$

Thus, in particular, we have

$$EW_* = \frac{\Phi(\alpha) + \alpha - 1}{\alpha(1-\alpha)},$$

$$EW_*^2 = \frac{2\alpha^2\Phi(\alpha)\nu_2 - 2[\Phi(\alpha) + \alpha - 1](1-\alpha^2)}{\alpha^2(1-\alpha)(1-\alpha^2)}.$$

The moments of W_* depend only on the moments of the service-time distribution and on $\Phi(\alpha)$.

As a numerical illustration of the biases induced when only one clock is available, the means and the standard deviations of W_1 and

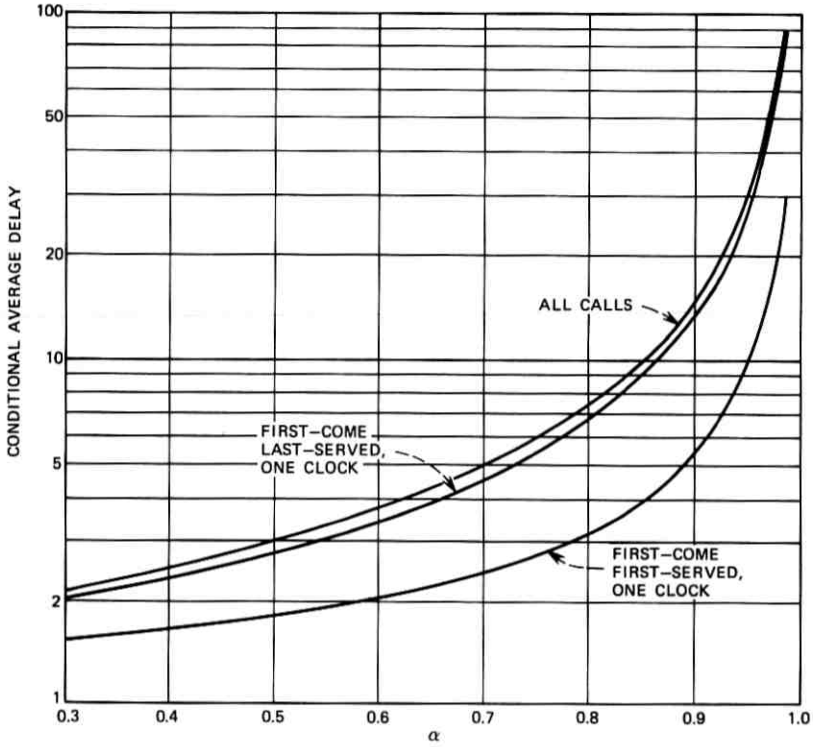

Fig. 5—Conditional average delay for $M/\Gamma_{\frac{1}{4}}/1$ vs occupancy.

W_{*1} are given in Table I. (The standard deviations of W_1 and W_{*1} are designated by SW_1 and SW_{*1} respectively.) For further quantitative results, see Figs. 1–6.

VI. THE SINGLE–SERVER QUEUE M/M/1

In this section we consider a single-server delay system and assume that: (*i*) calls arrive in a Poisson process of intensity α; (*ii*) the service-times are independent random variables with the same negative exponential distribution; and (*iii*) calls are served in order of arrival. We again suppose that only one delay can be measured at a time. Our purpose here is to derive the delay distribution of the observed calls.

Let $F(\cdot)$ be the equilibrium cumulative distribution of the virtual waiting time at the conclusion of the measurement of a delay; the measured delay may of course be equal to zero. From (38), the distri-

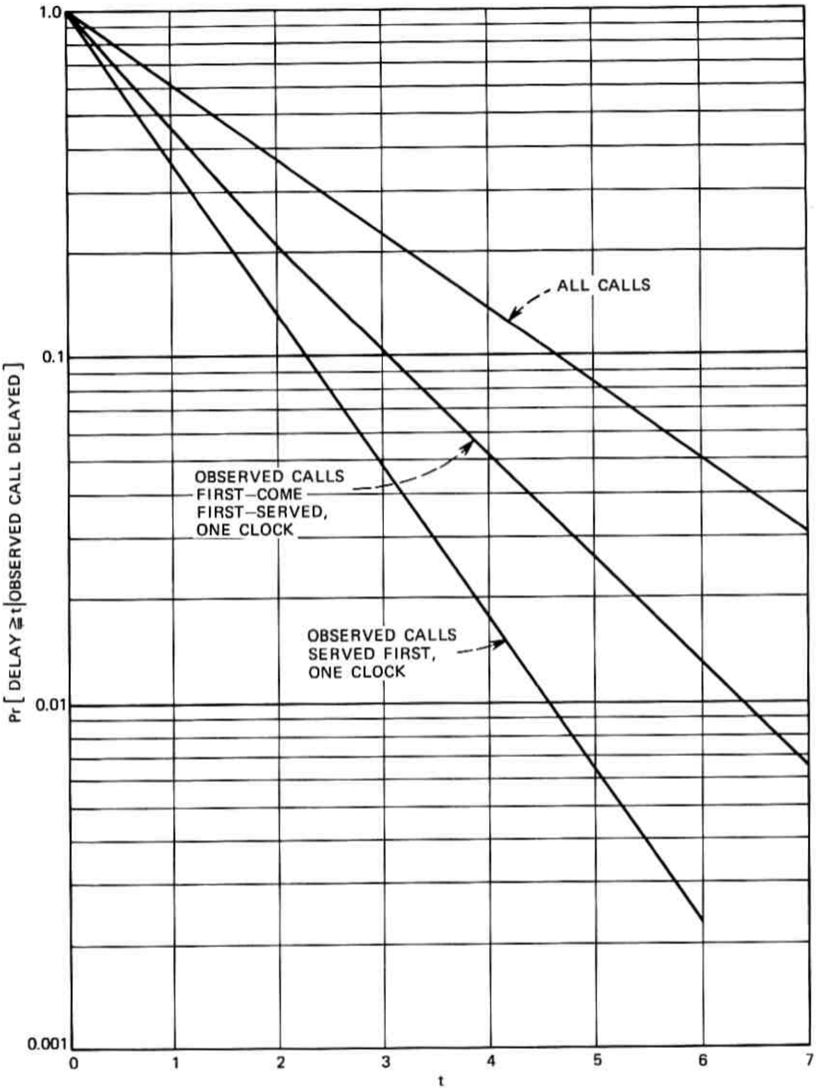

Fig. 6—Conditional delay distributions for M/M/1 − α = 0.5 first-come first-served.

bution $F(\cdot)$ satisfies the following integral equation:

$$F(t) = \int_0^t \left\{ \int_0^\infty \sum_{n=1}^\infty e^{-\alpha y} \frac{(\alpha y)^n}{n!} \frac{u^{n-1}}{(n-1)!} e^{-u} dF(y) \right\} du + \int_0^t e^{-u} du \cdot \int_0^\infty e^{-\alpha y} dF(y).$$

This relation implies that $F(\cdot)$ is continuous and that the virtual waiting time, at the conclusion of a measurement, has a density function $f(\cdot)$ [at $t = 0$, the latter is defined as the right-hand derivative of $F(\cdot)$]. We have therefore

$$f(t) = e^{-t}\int_0^\infty f(y)e^{-\alpha y}\sum_{n=1}^\infty \frac{(\alpha y)^n}{n!}\,\frac{t^{n-1}}{(n-1)!}\cdot dy + e^{-t}\int_0^\infty f(y)e^{-\alpha y}dy$$

$$= \alpha^{\frac{1}{2}}e^{-t}t^{-\frac{1}{2}}\int_0^\infty f(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy + e^{-t}\int_0^\infty f(y)e^{-\alpha y}dy, \qquad (41)$$

where $I_1(\cdot)$ is the modified Bessel function of order 1 (Ref. 7, p. 374).

The preceding relation implies that $f(\cdot)$ is of the form

$$f(t) = f_1(t) + ce^{-t},$$

where c is a constant. Substitution of this expression in (41) yields, on taking relation 29.3.81, p. 1026, of Ref. 7 into account:

$$f_1(t) = \alpha^{\frac{1}{2}}e^{-t}t^{\frac{1}{2}}\int_0^\infty f_1(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy + \frac{c\alpha e^{-t}}{(1+\alpha)^2}\,e^{\alpha t/(1+\alpha)}. \qquad (42)$$

Thus $f_1(\cdot)$ is of the form

$$f_1(t) = f_2(t) + \frac{c\alpha}{(1+\alpha)^2}\exp - t/(1+\alpha),$$

and substituting this expression in (42) we find that $f_2(\cdot)$ is of the form

$$f_2(t) = f_3(t) + \frac{c\alpha^2}{(1+\alpha+\alpha^2)^2}\exp - t/(1+\alpha+\alpha^2).$$

Proceeding in this manner, we define successively $f_4(\cdot)$, $f_5(\cdot)$, $\cdots$, and it is readily shown, by induction, that:

$$f_m(t) = f_{m+1}(t) + \frac{c\alpha^m}{(1+\alpha+\alpha^2+\cdots+\alpha^m)^2}$$
$$\cdot\exp - t/(1+\alpha+\cdots+\alpha^m), \qquad m = 0, 1, 2, \cdots;$$
$$f_0(\cdot) = f(\cdot). \qquad (43)$$

Passing to the limit, we obtain, in this manner:

$$f(t) = f_\infty(t) + c\sum_{m=0}^\infty \frac{\alpha^m}{(1+\alpha+\cdots+\alpha^m)^2}$$
$$\cdot\exp - t/(1+\alpha+\cdots+\alpha^m),$$

where $f_\infty(\cdot) = \lim_{m\to\infty} f_m(\cdot)$ satisfies the integral equation:

$$f_\infty(t) = \alpha^{\frac{1}{2}}e^{-t}t^{-\frac{1}{2}}\int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy. \qquad (44)$$

Note that, by virtue of (43), $f_m(t) > f_{m+1}(t)$ for all t and that $\lim_{m\to\infty} f_m(t)$ does therefore exist and is non-negative since $f_m > 0$ for all m.

We shall prove now that the only non-negative solution of (44) is $f_\infty(t) \equiv 0$.

Let

$$\theta(s) \equiv \int_0^\infty f_\infty(t)e^{-st}dt.$$

Then, transforming the previous relation, we have by Ref. 7, p. 1026, equation 29.3.81:

$$\begin{aligned}\theta(s) &= \alpha^{\frac{1}{2}}\int_0^\infty e^{-st}e^{-t}t^{-\frac{1}{2}}\int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy\cdot dt\\ &= \alpha^{\frac{1}{2}}\int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}(\alpha y)^{-\frac{1}{2}}(e^{\alpha y/1+s}-1)dy\\ &= \int_0^\infty f_\infty(y)\exp-\alpha y\left(1-\frac{1}{1+s}\right)dy\\ &\qquad -\int_0^\infty f_\infty(y)\exp(-\alpha y)dy\\ &= \theta\left(\frac{\alpha s}{1+s}\right)-\theta(\alpha).\end{aligned}$$

Setting s equal to zero in the preceding relation, we find that $\theta(\alpha) = 0$ which implies that $f_\infty(t) \equiv 0$ and we have, therefore, for exponential service-times:

$$\begin{aligned}f(t) &= c\sum_{m=0}^\infty \frac{\alpha^m}{(1+\alpha+\cdots+\alpha^m)^2}\\ &\qquad\cdot\exp-t/(1+\alpha+\cdots+\alpha^m), \qquad t \geqq 0,\\ 1-F(t) &= c\sum_{m=0}^\infty \frac{\alpha^m}{1+\alpha+\cdots+\alpha^m}\\ &\qquad\cdot\exp-t/(1+\alpha+\cdots+\alpha^m), \qquad t \geqq 0,\\ f(t) &= F(t) = 0, \qquad t < 0,\end{aligned} \tag{45}$$

where the constant c is determined by the condition $F(0) = 0$.

By means of (37) and (45), it is readily seen that the conditional delay distribution is given by the following formula:

$$\Pr[W_* \geqq t|\text{observed call delayed}] = c'\sum_{m=0}^\infty \frac{\alpha^{m+1}}{1+\alpha+\cdots+\alpha^{m+1}}$$
$$\cdot\exp-t/(1+\alpha+\cdots+\alpha^m), \qquad t > 0,$$

where c' is determined by the requirement that the preceding expression be equal to 1 for $t = 0$.

The effect of partial sampling on the delay distribution is illustrated in Fig. 6.

VII. THE MULTISERVER QUEUE M/M/s

In this section we consider a full-access multiserver delay system with Poisson arrivals and exponential service-times. Our purpose here is to determine the probability $\Phi_1^{(s)}$ that an observed call is served without delay and the expected delay $EW_{*1}^{(s)}$ of the observed calls $(\Phi_1^{(1)} = \Phi_1, W_{*1}^{(1)} = W_{*1})$. This is easily done. Indeed, under the present assumptions, $A_1^{(s)}$ the expected number of nonobserved calls during the measurement of a positive delay is:

$$A_1^{(s)} = \frac{\alpha EW_{*1}}{s(1 - \Phi_1)}, \tag{46}$$

where EW_{*1} and Φ_1 pertain to the single-server queue with load α/s. Equation (46) is an immediate consequence of the fact that in an s-server system with demand rate α and service rate 1, the conditional average delay of an observed call, $EW_{*1}^{(s)}/(1 - \Phi_1^{(s)})$, is equal to the conditional average delay of an observed call in a single-server queue with offered load α/s and service rate s, i.e., $EW_{*1}/s(1 - \Phi_1)$.

Hence, by (46) and (1) we have

$$\Phi_1^{(s)} = \frac{(1 - B)[(\alpha/s)EW_{*1} + 1 - \Phi_1]}{(1 - B)[(\alpha/s)EW_{*1} + 1 - \Phi_1] + B(1 - \Phi_1)}$$

so that, by (17),

$$EW_{*1}^{(s)} = \frac{B \cdot EW_{*1}}{s(1 - \Phi_1) + \alpha(1 - B)EW_{*1}}.$$

We note that

$$\frac{EW_{*1}^{(s)}}{EW_1^{(s)}} = \frac{(1 - \alpha/s)EW_{*1}}{1 - \Phi_1 + (1 - B)(\alpha/s)EW_{*1}}.$$

For α/s fixed, the blocking probability B is strictly decreasing and tends to 0 as s increases. Hence

$$\frac{EW_{*1}^{(s)}}{EW_1^{(s)}} > \frac{EW_{*1}^{(s+m)}}{EW_1^{(s+m)}}, \qquad m > 0,$$

and

$$\left[\frac{EW_*}{EW}\right]_\infty \equiv \lim_{s \to \infty} \frac{EW_{*1}^{(s)}}{EW_1^{(s)}} = \frac{(1 - \alpha/s)EW_{*1}}{1 - \Phi_1 + (\alpha/s)EW_{*1}}.$$

We stress that the preceding relations are valid regardless of the order of service.

Numerical values are given in Table II. They show, in particular, that, for a given server occupancy, the magnitudes of the relative biases become larger as the number of servers, s, increases but remain bounded.

VIII. AN INEQUALITY FOR GI/M/s

For the M/G/1 queue we have seen that the average delay on all calls, EW, is always larger than the expected delay EW_* even if the observed calls are served last. It will be shown here that the same relation also holds for the multiserver queue GI/M/s.

When the observed calls are served last, the waiting times of the observed delayed calls have the same distribution as the busy period whenever the service-times are exponential. Writing $\overline{EW}^{(s)}_{*1}$ for the unconditional average delay for the observed-served-last procedure we have therefore:

$$\overline{EW}^{(s)}_{*1} = (1 - \Phi)/s(1 - b), \tag{47}$$

where b is the root of smallest absolute value of the equation (Ref. 10, p. 225 ff.)

$$z = \mathbf{A}^*(1 - z)$$

and $\mathbf{A}^*$ is the Laplace-Stieltjes transform of the interarrival distribution $\mathbf{A}$. We note here that b is also the blocking probability in the associated single-server queue GI/M/1 with $\mathbf{A}$ as interarrival distribution and exponential service-time distribution with mean $1/s$.

By means of (1) we can rewrite (47) as follows:

$$\overline{EW}^{(s)}_{*1} = B/s(1 - b)[(1 - B)(1 + A) + B], \tag{48}$$

where B is the probability of delay in the GI/M/s queue.

When the observed calls are served last, $1 + A$ is equal to the expected number of calls served during a busy period of GI/M/s which, in turn, is equal to the expected number of calls served during a busy period of the associated single-server queue GI/M/1. Hence, we have (Ref. 10, p. 286):

$$1 + A = (1 - b)^{-1},$$

and on taking this relation into account, (48) reduces to

$$\overline{EW}^{(s)}_{*1} = B/s[1 - B + B(1 - b)].$$

TABLE II—AVERAGE DELAYS IN M/M/s—FIRST-COME FIRST-SERVED

α/s	EW_{*1}	EW_{*1}/EW_1	$EW_{*1}^{(2)}$	$EW_{*1}^{(2)}/EW_1^{(2)}$	$EW_{*1}^{(4)}$	$EW_{*1}^{(4)}/EW_1^{(4)}$	$EW_{*1}^{(8)}$	$EW_{*1}^{(8)}/EW_1^{(8)}$	$[EW_{*1}/EW_1]_\infty$
0.1	0.093	0.83	0.008	0.82					0.82
0.2	0.178	0.71	0.029	0.70	0.002	0.69			0.69
0.3	0.227	0.62	0.059	0.60	0.008	0.58	0.0004	0.57	0.57
0.4	0.337	0.55	0.099	0.52	0.018	0.49	0.0019	0.48	0.48
0.5	0.490	0.49	0.151	0.45	0.037	0.42	0.0059	0.40	0.39
0.6	0.656	0.44	0.224	0.40	0.065	0.36	0.0146	0.34	0.31
0.7	0.908	0.39	0.336	0.35	0.111	0.31	0.0316	0.28	0.24
0.8	1.37	0.34	0.541	0.30	0.199	0.27	0.0665	0.23	0.16
0.9	2.58	0.29	1.09	0.26	0.438	0.22	0.166	0.19	0.086
0.95	4.71	0.25	2.06	0.22	0.866	0.19	0.349	0.17	0.045

But the average delay for all calls is given by the formula (Ref. 11, p. 383):

$$EW_1^{(s)} = B/s(1 - b)$$

so that

$$\overline{EW}_{*1}^{(s)} < EW_1^{(s)}.$$

REFERENCES

1. *Proceedings of the Symposium on Congestion Theory*, W. L. Smith and W. E. Wilkinson, eds., Chapel Hill, N. C.: The University of North Carolina Press, 1965.
2. Oberer, E., and Riesz, G. W., "Test Calling: Experience, Theory, Prospect," Proc. Seventh Int. Teletraffic Congress, Stockholm, June 13–20, 1973.
3. Riordan, J., *Stochastic Service Systems*, New York: Wiley, 1962.
4. Loève, M., *Probability Theory*, third edition, Princeton, N. J.: D. Van Nostrand, 1963.
5. Doob, J. L., *Stochastic Processes*, New York: Wiley, 1953.
6. Smith, W. L., "Renewal Theory and Its Ramifications," J. Roy. Stat. Soc., Ser. B, *20*, No. 2, 1958, pp. 243–302.
7. *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun, eds., National Bureau of Standards, Applied Mathematics Series, 55, 1965.
8. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
9. Riordan, J., *An Introduction to Combinatorial Analysis*, New York: Wiley, 1958.
10. Cohen, J. W., *The Single-Server Queue*, New York: American Elsevier Publishing Company, Inc., 1969.
11. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Edinburgh and London: Oliver and Boyd, 1960.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Use of a Gate to Reduce the Variance of Delays in Queues With Random Service

By R. D. COLEMAN

(Manuscript received April 18, 1973)

We consider an N-server queuing system with Poisson arrivals and exponential service, in which arriving customers must pass through a gate into a waiting room before becoming eligible for service. Customers who find the gate closed wait outside until the gate opens; customers inside the waiting room are served at random. When the last customer inside acquires a server, the gate admits all those outside and then closes again. If no customer is waiting outside when the gate opens, the gate remains open until there is a queue of k waiting customers.

Service offered by this system is intermediary between random service and order-of-arrival service. As long as the gate is open and fewer than $N + k$ *customers are in the system, service is purely random. The parameter k can be regarded as a threshold at which the queue is judged too long to permit random service to continue.*

Our main results are (i) the Laplace-Stieltjes transform of the equilibrium distribution of the waiting time of an arbitrary customer and (ii) a comparison of the second moments of the waiting time for different values of k with those of the waiting time under random service and order-of-arrival service. The service is shown to be "nearly random" at low loads and "not quite order-of-arrival" at high loads; for higher values of k this transition occurs at higher traffic intensities.

I. INTRODUCTION

Switching systems, particularly electromechanical switching systems, are often constructed so that if several customers are awaiting service simultaneously, they will receive service in what is essentially random order, i.e., a server which becomes idle will choose its next customer at random from the queue of customers awaiting service. Such an arrangement may be satisfactory when the traffic intensity is low, but as the intensity increases, a progressively greater number of customers will have to wait an undesirably long time, and the quality of service may become unacceptable.

There are available, however, methods of providing service other than "random service"; the most obvious one is service in order of arrival. The quality of service, which depends in part on the variance of the waiting time, will still diminish as the traffic intensity increases, but not as quickly as when random service is used. In fact, order-of-arrival service is the "best" discipline (at least when the order of service does not depend on individual service times) in the sense that, for a given traffic intensity, and hence mean waiting time, the variance of the waiting time is smallest.[1] Unfortunately, it may not be worthwhile (or even possible) to build a system which offers service in order of arrival. One is therefore led to consider an intermediary queue discipline, one for which the variance of the equilibrium waiting time lies between that of random service and that of order-of-arrival service.

Suppose we have an M/M/N queuing system with customers arriving at rate λ and requiring a mean service time μ^{-1}. Suppose further that customers must first pass through a gate into a "corral," or "waiting room," before becoming eligible for service. So long as there are not more than N customers in the system, the gate remains open; an arriving customer enters the corral immediately and, if some server is idle, begins service. As soon as a customer has to wait (having found N customers in the system), the gate closes until that customer enters service. The gate then opens to admit all those who have accumulated outside the corral, and immediately closes again, remaining closed until all those inside have acquired a server, and so on. Thus the customers are admitted in bunches to the corral, and once inside, they are served at random. It is assumed that the corral has an unlimited capacity. If there is no customer waiting when the gate opens, the gate remains open until there is again a customer who has to wait.

A gated queuing system merits consideration, not only because of the anticipated improvement over random service, but also because it can overcome design problems in certain telephone equipment which might otherwise lead to poor service for some customers. For example, in some switching equipment, each server hunts in a fixed sequence over a group of customer-sources; when a customer is found, the server stops to provide service and then resumes hunting from that point. Such a hunting procedure may be desirable from a hardware viewpoint, but it can result in unequal service among customers.

Gates have been successfully used to improve service in the manner described above, that is, by temporarily blocking subsequent bids for service until all those customers present have been served.

Some discussion of the gated queuing system is given in a 1953 paper by Wilkinson.[2] Theoretical results were summarized in 1958 by Beckwith,[3] although he gives few details as to how the results are derived. The model we consider in this paper is more general than the one described above. We shall assume that as soon as there are k customers (rather than one) waiting, the gate closes until all k customers have entered service. Thus the parameter k can be thought of as a threshold at which the queue is judged to be too long to permit purely random service to continue; the system enters the "gating mode," and the gate admits customers in bunches as described previously. If the gate opens and there is no customer waiting outside, then the system leaves the gating mode and the gate remains open until there is again a queue of k waiting customers.

In Section II we obtain a recurrence relation for the probability-generating function of the nth bunch-size and, after that, the generating function and moments of the equilibrium bunch-size distribution. Using these results we determine in Section III the Laplace-Stieltjes transform of the distribution of the equilibrium waiting time of an arbitrary customer. In Section IV we obtain the first two moments of the equilibrium waiting time, and we make some comparisons of the second moments of gated service for different values of the parameter k with those of random service and order-of-arrival service.

In several places in the text, we specialize a general result to the case $k = 1$, since that is the simplest of our gated systems. This allows us to verify that our results then agree with Beckwith's, and it often reduces a complicated expression to one that is more easily comprehended.

We can assume, without loss of generality, that the system is initially empty. Since we assume that the traffic intensity, $\rho = \lambda/N\mu$, is less than unity, the number of customers will always return to zero in a finite length of time, regardless of how many customers may be present at the beginning. Consequently, our equilibrium results do not depend on the initial conditions.

Several other authors have considered systems which operate in a manner similar to ours, but they all take $k = 1$ and assume that the underlying system is an M/G/1 queue. The "generations" in the M/G/1 queue, as described by Kendall[4] and later studied by Neuts,[5] correspond to the bunches in our model. Nair and Neuts[6,7] subsequently consider the waiting time distribution for the M/G/1 queue under the assumption that the queue discipline was either longest-processing-time-first or shortest-processing-time-first.

II. DISTRIBUTION OF THE EQUILIBRIUM BUNCH-SIZE

Let X_n be the number of customers in the nth bunch; that is, there are $N + X_n$ customers in the system the instant after the gate closes for the nth time. Thus X_n is the number of customers waiting outside if that number is positive, and is k if there was no one waiting, when the gate was opened for the $(n-1)$st time. Let T_n be the time it takes to serve X_n customers. Then, denoting by $p^d(T_n = t)$ the density of T_n at t, we can express the distribution of X_n in terms of that of X_{n-1} by

$$P(X_n = j) = \sum_{i=1}^{\infty} P(X_{n-1} = i) \int_0^{\infty} P(X_n = j \mid X_{n-1} = i,\, T_{n-1} = t) \cdot p^d(T_{n-1} = t \mid X_{n-1} = i)dt. \quad (1)$$

But, given that $X_{n-1} = i$, T_{n-1} has a gamma distribution with parameters i and $N\mu$; and

$$P(X_n = j \mid X_{n-1} = i,\, T_{n-1} = t) = P(X_n = j \mid T_{n-1} = t) = \begin{cases} \dfrac{(\lambda t)^j}{j!} e^{-\lambda t}, & j \neq k \\[2ex] \dfrac{(\lambda t)^k}{k!} e^{-\lambda t} + e^{-\lambda t}, & j = k. \end{cases}$$

The extra term $e^{-\lambda t}$ in this expression is the probability that no one is waiting when the gate opens; when this happens, the next bunch-size is necessarily k. Substituting these expressions into eq. (1) gives

$$P(X_n = j) = \sum_{i=1}^{\infty} P(X_{n-1} = i) \left(\frac{1}{1+\rho}\right)^i \times \left[\left(\frac{\rho}{1+\rho}\right)^j \binom{j+i-1}{j} + \delta_{jk}\right],$$

where $\rho = \lambda/N\mu$ and δ_{jk} is the Kronecker delta. Now, letting

$$P_n(u) = \sum_{j=1}^{\infty} P(X_n = j)u^j$$

be the probability-generating function of the distribution of X_n, we have

$$P_n(u) = P_{n-1}\left(\frac{1}{1+\rho(1-u)}\right) + (u^k - 1)P_{n-1}\left(\frac{1}{1+\rho}\right). \quad (2)$$

Starting with $P_1(u) = u^k$, we can determine successively the $P_n(u)$. When $k = 1$, eq. (2) agrees with Beckwith's eq. (1).

We wish to obtain the distribution of the equilibrium bunch-size $X = \lim_{n\to\infty} X_n$. To see that an equilibrium distribution exists, we observe that the bunch-sizes X_n form a Markov chain which is irreducible and aperiodic. Since we are assuming $\rho < 1$, the number of customers in the system cannot grow without bound, so the states are not all transient or null. Therefore all states are ergodic, and there is a unique equilibrium distribution.[8] The distribution of $X = \lim_{n\to\infty} X_n$ has a probability-generating function $P(u) = \lim_{n\to\infty} P_n(u)$, which satisfies

$$P(u) = P\left(\frac{1}{1+\rho(1-u)}\right) + (u^k - 1)P\left(\frac{1}{1+\rho}\right).$$

This can be written in the form

$$P[r(u)] - P(u) = F(u)$$

by setting

$$r(u) = 1/[1 + \rho(1-u)]$$

and

$$F(u) = (1 - u^k)P[1/(1+\rho)].$$

The solution to this functional equation is[9]

$$P(u) = \eta - \sum_{n=0}^{\infty} F[r_n(u)],$$

where r_n is the nth iterate of r and η is a constant. To evaluate this expression, we first need to find $r_n(u)$. We have $r_0(u) = u$ and

$$r_n = \frac{1}{1 + \rho - \rho r_{n-1}}. \quad (3)$$

Letting y_n be such that

$$r_n = \frac{y_{n+1}}{y_n} + 1 + \frac{1}{\rho}$$

transforms eq. (3) into a linear homogeneous difference equation, which is easily solved. Thus we determine that

$$r_n(u) = \frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}}.$$

The solution to our functional equation is therefore

$$P(u) = \eta - \sum_{n=0}^{\infty} \left\{1 - \left[\frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}}\right]^k\right\} P\left(\frac{1}{1+\rho}\right).$$

Since $P(1) = 1$, we must have $\eta = 1$; since no bunch-size can be zero, $P(0) = 0$. This determines $P[1/(1+\rho)]$, so that finally

$$\begin{aligned} P(u) &= 1 - \frac{\displaystyle\sum_{n=0}^{\infty} \left\{1 - \left[\frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}}\right]^k\right\}}{\displaystyle\sum_{n=0}^{\infty} \left\{1 - \left[\frac{1 - \rho^n}{1 - \rho^{n+1}}\right]^k\right\}} \\ &= 1 - \frac{h(u)}{h(0)}, \end{aligned} \tag{4}$$

where $h(u)$ is the numerator in the second term of $P(u)$ above. By

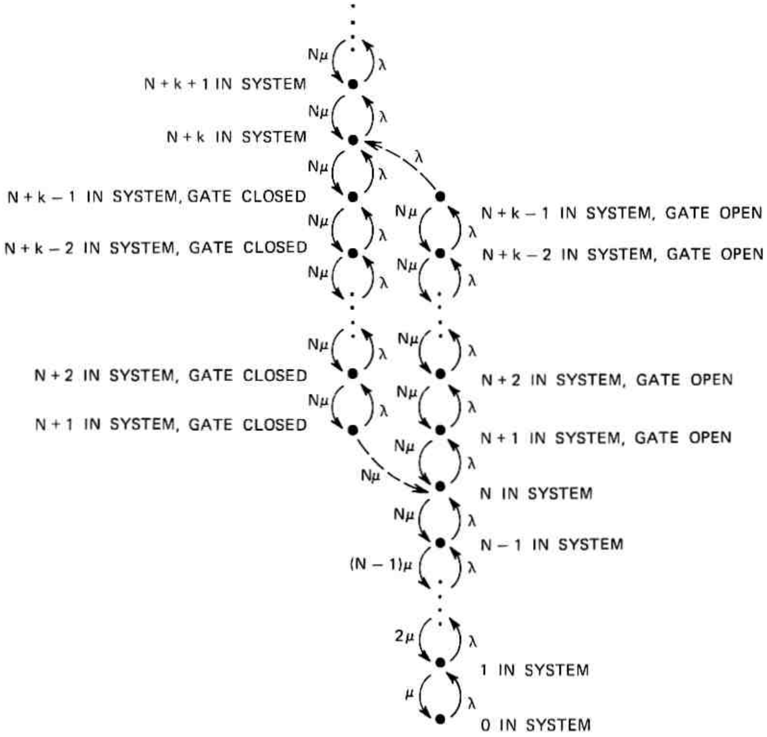
Nμ
λ
N + k + 1 IN SYSTEM
Nμ
λ
N + k IN SYSTEM
λ
Nμ
λ
N + k − 1 IN SYSTEM, GATE CLOSED
Nμ
λ
N + k − 1 IN SYSTEM, GATE OPEN
Nμ
λ
N + k − 2 IN SYSTEM, GATE CLOSED
Nμ
λ
N + k − 2 IN SYSTEM, GATE OPEN
Nμ
λ
Nμ
λ
Nμ
λ
Nμ
λ
N + 2 IN SYSTEM, GATE CLOSED
N + 2 IN SYSTEM, GATE OPEN
Nμ
λ
Nμ
λ
N + 1 IN SYSTEM, GATE CLOSED
N + 1 IN SYSTEM, GATE OPEN
Nμ
λ
Nμ
N IN SYSTEM
Nμ
λ
N − 1 IN SYSTEM
(N − 1)μ
λ
2μ
λ
1 IN SYSTEM
μ
λ
0 IN SYSTEM

Fig. 1—Possible states of the system and the transition rates between states.

setting $k = 1$, we can reduce eq. (4) to Beckwith's result, his eq. (2):

$$P(u) = \frac{1}{g(1)}\left[\frac{u-1}{1-\rho} + g(1) - g\left(\frac{1-u}{\frac{1}{\rho} - u}\right)\right],$$

where

$$g(u) = u \sum_{n=0}^{\infty} \frac{\rho^n}{1 - \rho^{n+1}u}.$$

The moments of X are obtained by differentiating $P(u)$. We find that

$$E(X) = P'(1) = \frac{k}{(1-\rho)h(0)}$$

and

$$\text{Var}\,(X) = \frac{k}{h(0)(1-\rho)}\left[\frac{k}{1+\rho} + \frac{\rho}{1-\rho} - \frac{k}{h(0)(1-\rho)}\right].$$

III. DISTRIBUTION OF THE EQUILIBRIUM WAITING TIME

Let W be the waiting time to the point of entering service of an arbitrary customer when the system is in equilibrium. The distribution of W depends on which state the system is in when the customer arrives; i.e., it depends on the number of customers already in the system and on whether the gate is open or closed. These states, together with the transition rates from one state to another, have been enumerated in Fig. 1. Let

$p_j = P[j$ customers in the system$], \qquad j \geqq 0,$

$p_j^c = P[j$ customers in the system, gate closed$],$
$j = N+1, N+2, \cdots, N+k-1,$

$p_j^o = P[j$ customers in the system, gate open$],$
$j = N+1, N+2, \cdots, N+k-1.$

Obviously, $p_j^c + p_j^o = p_j$. It is also clear that when $j \leqq N$, the gate is open, and that when $j \geqq N + k$, the gate is closed. The values of the p_j are just the equilibrium state probabilities of an M/M/N queue, which are

$$p_j = \frac{(\lambda/\mu)^j}{j!}\, p_0 = \frac{(\rho N)^j}{j!}\, p_0, \qquad j = 0, 1, \cdots, N$$

$$p_{N+j} = \rho p_{N+j-1} = \rho^j p_N = \rho^j \frac{(\rho N)^N}{N!}\, p_0, \qquad j > 0$$

$$p_0 = \left[\sum_{j=0}^{N-1} \frac{(\rho N)^j}{j!} + \frac{(\rho N)^N}{N!(1-\rho)}\right]^{-1}.$$

To find p_j^o, we equate the rates at which the system leaves and enters the state $\{j$ customers in the system, gate open$\}$. From Fig. 1, we see that

$$(1+\rho)p_{N+j}^o = \rho p_{N+j-1}^o + p_{N+j+1}^o, \qquad j = 1, 2, \cdots, k-1,$$

with $p_{N+k}^o = 0$ and $p_N^o = p_N$, where p_N is known from the above. The solution to this equation is

$$p_{N+j}^o = \frac{\rho^j - \rho^k}{1-\rho^k}\, p_N, \qquad j = 0, 1, \cdots, k-1.$$

It then follows that

$$\begin{aligned} p_{N+j}^c &= p_{N+j} - p_{N+j}^o \\ &= \frac{1-\rho^j}{1-\rho^k}\, \rho^k p_N, \qquad j = 1, 2, \cdots, k-1. \end{aligned}$$

To find the distribution of the waiting time W, we shall consider what happens when an arriving customer encounters one of the following conditions:

H_1: $< N$ customers in the system,
H_2: the gate is closed,
$H_{3\cdot j}$: $N + j$ customers in the system, and the gate is open,
$j = 0, 1, \cdots, k-1.$

From the above computations,

$$P(H_1) = 1 - \sum_{j=0}^{\infty} \rho^j p_N = 1 - \frac{1}{1-\rho}\, p_N, \tag{5}$$

$$P(H_2) = \sum_{j=1}^{k-1} p_{N+j}^c + \sum_{j=k}^{\infty} p_{N+j} = \frac{k\rho^k}{1-\rho^k}\, p_N, \tag{6}$$

and

$$P(H_{3\cdot j}) = \frac{\rho^j - \rho^k}{1-\rho^k}\, p_N, \qquad j = 0, 1, \cdots, k-1. \tag{7}$$

An arriving customer who encounters H_1 immediately gains access to a server, so

$$P(W = t \mid H_1) = \begin{cases} 1, & t = 0 \\ 0, & t > 0 \end{cases}. \tag{8}$$

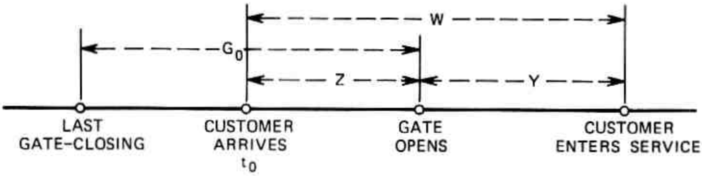

Fig. 2—Relations among the variables used.

An arriving customer who encounters H_2 will have to spend some time, Z, outside the gate waiting for the gate to open (see Fig. 2); once he gets inside, he will then have to spend some additional time, Y, waiting to be chosen for service from the bunch he entered with. The total amount of time the customer spends waiting to enter service is therefore $W = Z + Y$, where Z can be regarded as the residual lifetime of an interval G_0 during which the gate is closed. Thus we can write

$$\begin{aligned} p^d(W = t \mid H_2) &= \int_{x=0}^{\infty} \int_{y=0}^{\infty} p^d(Z + Y = t,\, G_0 = y,\, Z = x)dydx \\ &= \int_{x=0}^{\infty} \int_{y=0}^{\infty} p^d(Y = t - x \mid G_0 = y,\, Z = x) \\ &\qquad\qquad \cdot p^d(Z = x \mid G_0 = y)p^d(G_0 = y)dydx. \qquad (9) \end{aligned}$$

We first need the distribution of G_0. Disregarding our arbitrary customer temporarily, let G be the equilibrium length of the time interval during which the gate is closed. If we are given that j customers entered the waiting room when the gate last opened, then the gate will remain closed for j service times. Since the equilibrium probability that j customers entered when the gate last opened is $P(X = j)$, we have

$$p^d(G = y) = \sum_{j=1}^{\infty} P(X = j)b^{*j}(y),$$

where $b(y) = N\mu \exp(-N\mu y)$ and the asterisk denotes convolution. The mean of this distribution is easily found to be

$$E(G) = \frac{E(X)}{N\mu} = \frac{k}{N\mu(1 - \rho)h(0)}.$$

Let t_0 be the epoch at which the arbitrary customer arrives, and let

G_0 be the length of the G-interval containing t_0. Then (see Ref. 10)

$$\begin{aligned} p^d(G_0 = y) &= \frac{y}{E(G)}\, p^d(G = y) \\ &= \frac{1}{k}\, N\mu y(1 - \rho)h(0) \sum_{j=1}^{\infty} P(X = j) b^{*j}(y). \qquad (10) \end{aligned}$$

Next, the distribution of Z, given $G_0 = y$, is uniform on $(0, y)$, so

$$p^d(Z = x | G_0 = y) = \begin{cases} \dfrac{1}{y}, & 0 \leqq x \leqq y \\ 0, & x > y. \end{cases} \qquad (11)$$

We also have

$$\begin{aligned} p^d(Y = t - x | G_0 = y, Z = x) &= p^d(Y = t - x | G_0 = y) \\ &= \sum_{n=1}^{\infty} p^d(Y = t - x | n \text{ arrivals in } (0, y), G_0 = y) \\ &\qquad \cdot p^d(n - 1 \text{ other arrivals in } (0, y) | G_0 = y) \\ &= \sum_{n=1}^{\infty} \frac{1}{n} \left(\sum_{l=1}^{n} b^{*l}(t - x) \right) \cdot \frac{(\lambda y)^{n-1}}{(n-1)!} e^{-\lambda y}, \qquad (12) \end{aligned}$$

where the first factor in each summand reflects the fact that the arbitrarily chosen customer has probability $1/n$ of waiting one service time, $1/n$ of waiting two service times, $\cdots$, and $1/n$ of waiting n service times. Substituting eqs. (10), (11), and (12) into eq. (9), we obtain

$$\begin{aligned} p^d(W = t | H_2) &\\ = \frac{1}{k} N\mu(1 - \rho)h(0) &\sum_{j=1}^{\infty} \sum_{n=1}^{\infty} P(X = j) \sum_{l=1}^{n} \int_{x=0}^{t} b^{*l}(t - x) \\ &\cdot \int_{y=x}^{\infty} b^{*j}(y) e^{-\lambda y} \frac{(\lambda y)^{n-1}}{n!}\, dy dx. \qquad (13) \end{aligned}$$

Notice that the range of integration of the inner integral was reduced from $(0, \infty)$ to (x, ∞) because of eq. (11), and that of the outer integral was reduced from $(0, \infty)$ to $(0, t)$ because $b^{*l}(t - x) = 0$ for $x > t$. By setting $k = 1$, $N\mu = 1$, and $\lambda = \rho$ in eq. (13), we can obtain Beckwith's expression for the corresponding density in his model, the density of W given that the arbitrary customer finds more than N customers in the system.

We can calculate the Laplace-Stieltjes transform, $\psi(s)$, of the distribution (13), but the algebra is long and tedious. The results are given below, one in terms of $P(X = j)$ and the other explicitly:

$$
\begin{aligned}
&\psi(N\mu s)\\
&= \frac{(1-\rho)h(0)}{\rho k s^2}\biggl\{P(X=1)\log\frac{(1+s)(1+s+s\rho)}{(1+s+\rho)(1+s)-\rho}\\
&\quad + \sum_{j=2}^{\infty}\frac{1}{j-1}P(X=j)\biggl[1-\left(\frac{1}{1+s}\right)^{j-1}-\left(\frac{1+s}{1+s+s\rho}\right)^{j-1}\\
&\qquad\qquad + \left(\frac{1+s}{(1+s)(1+\rho+s)-\rho}\right)^{j-1}\biggr]\biggr\} \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1-\rho}{k\rho s^2}\sum_{n=0}^{\infty}\biggl[\frac{k\rho^n(1-\rho)^2}{(1-\rho^{n+1})^2}\biggl\{\left(\frac{1-\rho^n}{1-\rho^{n+1}}\right)^{k-1}\\
&\quad \cdot\log\left(\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho^{n+2})}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right)\\
&\quad + \sum_{j=2}^{k}\binom{k-1}{j-1}\left(\frac{1-\rho^{n-1}}{1-\rho^n}\right)^{k-j}\left(\frac{\rho^{n-1}(1-\rho)^2}{(1-\rho^n)(1-\rho^{n+1})}\right)^{j-1}\\
&\quad \cdot\sum_{m=1}^{j-1}\frac{1}{m}\biggl[\left(\frac{1-\rho^{n+1}}{1-\rho}\right)^m-\left(\frac{(1-\rho^{n+1})(1+s)}{1+s-\rho-s\rho^{n+1}}\right)^m\\
&\quad -\left(\frac{(1-\rho^{n+1})(1+s+s\rho)}{1+s-\rho-s\rho^{n+2}}\right)^m\\
&\quad +\left(\frac{(1-\rho^{n+1})[(1+s)^2+s\rho]}{(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}}\right)^m\biggr]\biggr\}\\
&\quad -\frac{1}{1-\rho^{n+1}}\biggl\{(1-\rho)-\frac{(1+s-\rho-s\rho^n)^k}{(1+s-\rho-s\rho^{n+1})^{k-1}}\\
&\quad -\frac{1}{1+s}\cdot\frac{(1+s-\rho-s\rho^{n+1})^k}{(1+s-\rho-s\rho^{n+2})^{k-1}}\\
&\qquad\qquad +\frac{1}{1+s}\cdot\frac{[(1+s)^2-\rho-s(1+s+\rho)\rho^n]^k}{[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]^{k-1}}\biggr\}\biggr]. \qquad (15)
\end{aligned}
$$

When k is set equal to 1 in eq. (15), the sum from 2 to k is trivially zero, and the expression in the last pair of braces collapses to zero; the transform then reduces to

$$
\begin{aligned}
\psi(N\mu s) &= \frac{(1-\rho)^3}{\rho s^2}\sum_{n=0}^{\infty}\frac{\rho^n}{(1-\rho^{n+1})^2}\\
&\quad\times\log\left[\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho^{n+2})}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right],\\
&\qquad\qquad (k=1). \qquad (16)
\end{aligned}
$$

The final portion of the waiting time distribution needed is $f_j(t) = p^d(W = t|H_{3\cdot j})$, $j = 0, 1, \cdots, k-1$. Let W_j be the waiting time of a customer who arrives to find the gate open and $N+j$ in the system. Then W_j has the density $f_j(t)$, and, letting $E_\lambda(t)$ denote the exponential density $\lambda e^{-\lambda t}$,

$$\begin{cases} f_j = \dfrac{\lambda}{\lambda+N\mu} E_{\lambda+N\mu} * f_{j+1} + \dfrac{N\mu}{\lambda+N\mu} \\ \qquad \cdot \left[\dfrac{1}{j+1} E_{\lambda+N\mu} + \dfrac{j}{j+1} E_{\lambda+N\mu} * f_{j-1} \right], \\ \qquad\qquad\qquad\qquad j = 0, 1, \cdots, k-2 \\ f_{k-1} = \dfrac{1}{k} E_{N\mu} + \dfrac{1}{k} E_{N\mu}^{*2} + \cdots + \dfrac{1}{k} E_{N\mu}^{*k}. \end{cases}$$

The reasoning behind these equations is that, if an arriving customer finds the gate open and $j(< k-1)$ customers waiting, then either the next event is an arrival, in which case he waits until that event plus an additional time distributed as W_{j+1}, or else the next event is a departure, in which case he waits until that event, after which either he is served immediately [with probability $1/(j+1)$] or someone else is chosen and our customer must wait an additional time distributed as W_{j-1} [with probability $j/(j+1)$]. If, on the other hand, the arriving customer finds the gate open and $k-1$ customers waiting, then the gate shuts behind him and he waits either 1, 2, $\cdots$, or k service times, each having probability $1/k$. Taking Laplace-Stieltjes transforms of the equations, and denoting the transform of $f_j(t)$ by $\phi_j(s)$, we obtain

$$\begin{cases} \rho j \phi_j(N\mu s) = (1+\rho+s) j \phi_{j-1}(N\mu s) - (j-1)\phi_{j-2}(N\mu s) - 1, \\ \qquad\qquad\qquad\qquad j = 1, \cdots, k-1, \\ \phi_{k-1}(N\mu s) = \dfrac{1}{ks}\left[1 - \left(\dfrac{1}{1+s}\right)^k\right]. \end{cases} \tag{17}$$

For any particular k, this set of equations can be solved explicitly by successive substitution. Finally, using eqs. (5), (6), (7), and (8), we can represent the Laplace-Stieltjes transform of the distribution of the waiting time W as

$$\phi(N\mu s) = 1 - \frac{1}{1-\rho} p_N + p_N \sum_{j=0}^{k-1} \frac{\rho^j - \rho^k}{1-\rho^k} \phi_j(N\mu s) + \frac{k\rho^k}{1-\rho^k} p_N \psi(N\mu s), \tag{18}$$

where $\phi_j(N\mu s)$ is given by eq. (17) and $\psi(N\mu s)$ is given by eq. (14)

or eq. (15). When $k = 1$, eq. (18), with the help of eq. (16), can be written explicitly as

$$\phi(N\mu s) = 1 - \frac{1}{1-\rho}\, p_N + \frac{1}{1+s}\, p_N + \frac{(1-\rho)^2}{s^2}\, p_N \sum_{n=0}^{\infty} \frac{\rho^n}{(1-\rho^{n+1})^2}$$

$$\cdot \log\left[\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho)^{n+2}}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right], \qquad (k = 1).$$

IV. MOMENTS OF THE EQUILIBRIUM WAITING TIME

Since the mean waiting time does not depend on the queue discipline (see Ref. 11), the mean is the same as for a simple queue with service in order of arrival, i.e.,

$$E(W) = \sum_{j=0}^{\infty} p_N \rho^j \frac{j+1}{N\mu} = \frac{p_N}{N\mu(1-\rho)^2}.$$

The second-moment computations are fairly lengthy, finally yielding

$$E(W^2) = \frac{2p_N}{N\mu(1-\rho^k)}\,[M'(\rho) + M(\rho) - \rho^{k-1}M'(1) - \rho^{k-1}M(1)]$$

$$+ \frac{kp_N\rho^{k-1}(1-\rho)}{(N\mu)^2(1-\rho^k)}\left\{\frac{(k-1)(k-2)}{6}\left[\frac{2+2\rho+3\rho^2}{1-\rho^3}\right]\right.$$

$$\left. + (k-1)\left[\frac{2+3\rho+3\rho^2-\rho^4}{(1-\rho^2)(1-\rho^3)}\right] + \frac{2+4\rho+5\rho^2+2\rho^3+\rho^4}{(1-\rho)(1-\rho^2)(1-\rho^3)}\right\}, \qquad 19)$$

where M is a function defined by

$$M(x) = -\sum_{j=0}^{k-1} x^j \phi_j'(0) = \sum_{j=0}^{k-1} x^j E(W \mid H_{3\cdot j}).$$

The variance of W is obtained by subtracting the square of the mean of W:

$$\operatorname{Var}(W) = E(W^2) - \frac{p_N^2}{(N\mu)^2(1-\rho)^4}.$$

For comparison purposes, we also need the second moments for order-of-arrival service and for random service. When service is in the order of arrival, we obtain[11]

$$E(W^2) = E(W^2 \mid W > 0)P(W > 0) = \frac{2p_N}{(N\mu)^2(1-\rho)^3}. \qquad (20)$$

When service is at random, the second moment can be written as[11]

$$E(W^2) = \frac{2p_N}{(N\mu)^2(1-\rho)^3}\cdot\frac{2}{2-\rho}. \qquad (21)$$

Observe that the second moments depend on N only through the factor $p_N/(N\mu)^2$, assuming the value of ρ is fixed. This is true also for the second moment in eq. (19), since each of the M-terms contains a factor $(N\mu)^{-1}$. It is therefore convenient to consider the ratio of $E(W^2)$ for the gated system [eq. (19)] to the second moment for the order-of-arrival system [eq. (20)], i.e.,

$$R(k,\rho) = \frac{N\mu(1-\rho)^3}{1-\rho^k}[\rho M'(\rho) + M(\rho) - \rho^{k-1}M'(1) - \rho^{k-1}M(1)]$$
$$+ \frac{k\rho^{k-1}(1-\rho)^4}{2(1-\rho^k)}\left\{\frac{(k-1)(k-2)}{6}\left[\frac{2+2\rho+3\rho^2}{1-\rho^3}\right]\right.$$
$$\left. + (k-1)\left[\frac{2+3\rho+3\rho^2-\rho^4}{(1-\rho^2)(1-\rho^3)}\right] + \frac{2+4\rho+5\rho^2+2\rho^3+\rho^4}{(1-\rho)(1-\rho^2)(1-\rho^3)}\right\}. \quad (22)$$

Because this ratio is independent of N, it provides a useful tool for examining the effect of the value of k on the second moment of the waiting time W. Thus we shall be interested in determining its properties.

The function $R(k,\rho)$ has been plotted in Fig. 3 on the interval $0 < \rho < 1$ for a variety of values of k. We observe that $R(k,\rho)$ is bounded from below by unity, increases as k increases, and is bounded from above by $2/(2-\rho)$, the ratio of eq. (21) and eq. (20), which corresponds to $k = \infty$. This general behavior is, of course, just what we expected *a priori*. In order to demonstrate analytically that

$$1 \leqq R(k,\rho) \leqq \frac{2}{2-\rho}, \quad (23)$$

we first introduce the inequality

$$\frac{1}{N\mu}\frac{j+2}{2} \leqq E(W|H_{3\cdot j}) \leqq \frac{1}{N\mu}\frac{j+2}{2}\frac{2}{2-\rho},$$
$$j = 0, 1, \cdots, k-1. \quad (24)$$

The left half of this inequality is demonstrated by considering what happens when an arriving customer finds j others waiting, but no more arrivals are permitted to enter the system: the customer's expected waiting time will decrease to $(j+2)/2N\mu$, since the customer will have to wait either 1, 2, $\cdots$, or $j+1$ service times, each with probability $(j+1)^{-1}$. The right half of the inequality is demonstrated by considering what happens when there is no gate at all to block future arrivals when a threshold k is reached: the mean waiting time $E(W|H_{3\cdot j})$ would then increase to the corresponding mean in a system

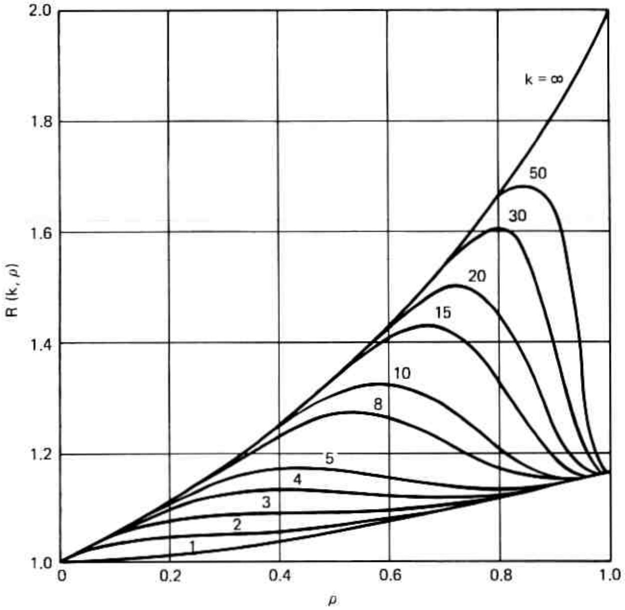

Fig. 3—Ratio of second moments.

employing purely random service. In a random service system, a customer who arrives to find j others waiting has an expected delay of

$$\frac{1}{N\mu}\cdot\frac{j+2}{2}\cdot\frac{2}{2-\rho}, \qquad j = 0, 1, \cdots, k-1,$$

a fact which is derived in the appendix.

Using the left half of eq. (24), we can substitute $(j+2)/2N\mu$ for $E(W \mid H_{3\cdot j})$ in $R(k, \rho)$ and combine terms, obtaining

$$R(k, \rho) \geqq 1 + \frac{k\rho^{k+1}(1-\rho)}{12(1-\rho^k)}\left\{2 + \frac{3k(1-\rho)(1-\rho^2)}{(1+\rho)(1+\rho+\rho^2)} + \frac{k^2(1-\rho)^2}{1+\rho+\rho^2}\right\},$$

from which it is obvious that $R(k, \rho) \geqq 1$. Similarly, using the right half of eq. (24), we can substitute $(j+2)/N\mu(2-\rho)$ in $R(k, \rho)$. Combining terms, we obtain

$$R(k, \rho) \leqq \frac{2}{2-\rho} - \frac{k\rho^k(1-\rho)}{(2-\rho)(1-\rho^k)}\left\{\frac{k^2(1-\rho)^2(2+3\rho^2)}{12(1+\rho+\rho^2)} + \frac{k(1-\rho)(2+2\rho+5\rho^2+\rho^4)}{4(1+\rho)(1+\rho+\rho^2)} + \frac{2+7\rho+10\rho^2+8\rho^3+3\rho^4}{6(1+\rho)(1+\rho+\rho^2)}\right\},$$

from which it is clear that $R(k, \rho) \leqq 2/(2 - \rho)$. Thus, eq. (23) is established.

Perhaps the most striking feature of Fig. 3 is that all the curves approach the same value, 7/6, as $\rho \rightarrow 1$. It is easy to demonstrate, by using eq. (22), that this must occur. Since the conditional means $E(W | H_{3.j})$ remain bounded as $\rho \rightarrow 1$, $M(\rho) \rightarrow M(1)$ and both are finite. Thus the first term of $R(k, \rho)$ approaches zero as $\rho \rightarrow 1$. In the second term, the factor $(1 - \rho)^4$ in front causes all but the last term in braces to approach zero. Thus,

$$R(k, 1) = \lim_{\rho \rightarrow 1} \frac{k\rho^{k-1}}{2(1 + \rho + \cdots + \rho^{k-1})} \cdot \frac{2 + 4\rho + 5\rho^2 + 2\rho^3 + \rho^4}{(1 + \rho)(1 + \rho + \rho^2)} = \frac{7}{6}.$$

In order to gain some insight as to why the curves meet at a common point at $\rho = 1$, we will find it helpful to consider a supplementary variable, the fraction of time the system spends in the gating mode. When the system is in equilibrium, this is simply the probability that an arriving customer finds the gate closed, and is given by eq. (6):

$$P(\text{system is in gating mode}) = \frac{k\rho^k}{1 - \rho^k} p_N.$$

This quantity has been plotted in Fig. 4 as a function of ρ, for the arbitrarily chosen value $N = 7$. It can easily be seen (and is intuitively obvious) that when ρ is very close to 1, the system spends almost all its time in the gating mode. But when the system is in the gating mode, the system's operation is independent of the value of k; it is only when the gate is open that the threshold value k can have any effect. Thus, as ρ approaches 1, the system becomes independent of k; so we can expect the curves in Fig. 3 to be independent of k at $\rho = 1$.

Another feature of the curves in Fig. 3 is that the slope at zero for $k \geqq 2$ is the same as the slope of $2/(2 - \rho)$; but the slope for $k = 1$ is zero. The reason for this becomes clear when we realize that when $k = 1$, order-of-arrival service is guaranteed until there are $N + 3$ customers in the system, while $k \geqq 2$ makes it possible for passing to occur as soon as there are $N + 2$ customers in the system. For small ρ, $N + 3$ in the system is much less likely than $N + 2$.

The ratio in eq. (22) is convenient because it is independent of N. The variance of a distribution, however, is also frequently of interest. It is clear from Fig. 3 that the second moments, and therefore the variances, increase with increasing k. There is, therefore, a similar family of curves, one family for each value of N, obtained by computing the ratio of the variance with threshold k to the variance with order-of-

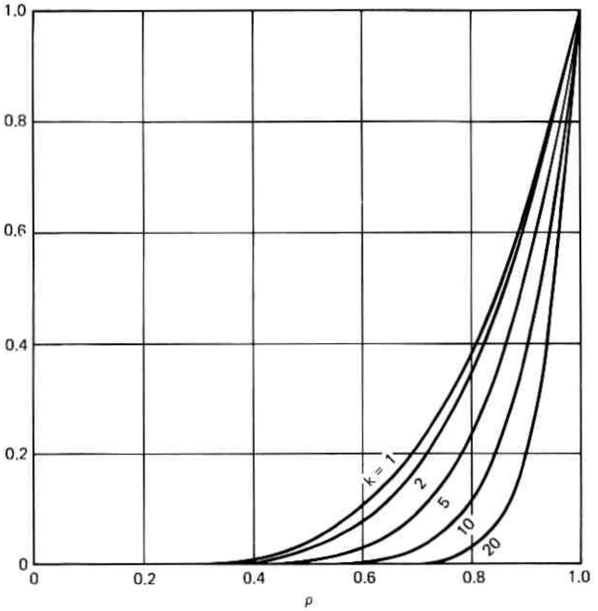

Fig. 4—Equilibrium probability that gate is closed ($N = 7$).

arrival service. In Fig. 5 we have plotted some of these curves, again for $N = 7$. It is obvious that these curves have the same general shape as those in Fig. 3. The main differences are (*i*) that the ratio of the variances when $k = \infty$ (random service) goes to 3 as $\rho \rightarrow 1$, while the ratio of the second moments when $k = \infty$ goes to 2, and (*ii*) that the ratios of the variances for $k < \infty$ go to $8/6$ as $\rho \rightarrow 1$, while the ratios of the second moments go to $7/6$. These facts are easily verified analytically by letting $\rho \rightarrow 1$ in the actual expressions for the ratios.

Since the second moments (and variances) increase as a function of k, there arises the question of why one might consider a threshold value k greater than 1. Clearly, if the queuing systems were otherwise equivalent, one would prefer $k = 1$ to any larger value of k. But it is equally possible that a queuing system would be more costly to operate when it is in the gating mode, since more bookkeeping is necessary: each waiting customer must be classified as "inside" or "outside" the waiting room, and these labels must be changed when the gate operates. One would then prefer a system in which the gate were used as little as possible (large k), consistent with an acceptable quality of service. The resulting tradeoff between cost and quality of service

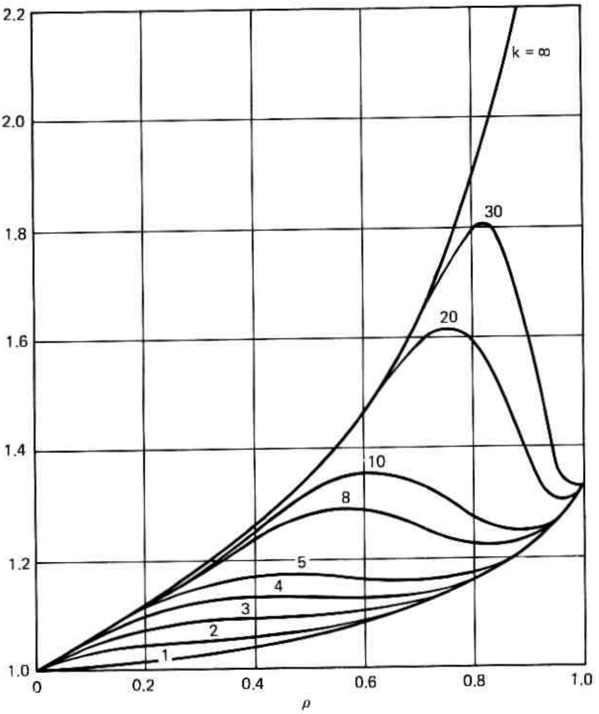

Fig. 5—Ratio of variances ($N = 7$).

can be resolved only by examination of the particular application at hand.

V. CONCLUDING REMARKS

In summary, our main result is the specification of the distribution of the equilibrium waiting time of an arbitrary customer in a queuing system whose service discipline is a compromise between service in order of arrival and random service. Our model contains a parameter, k, which determines how "close" the discipline is to order-of-arrival service or to random service. We have seen that the variance of the waiting time is bounded from below by the variance for order-of-arrival service, and that as k increases, the variance increases, approaching the variance of the waiting time when random service is employed. We also found a convenient quantity, $R(k, \rho)$, which is independent of the number of servers, and which, together with Fig. 3, allowed us to examine the effect of the threshold k on the waiting time.

Figure 3 shows how, for fixed $k > 1$, the service changes from "nearly random" to "not quite order-of-arrival" with increasing load, and how this transition occurs at higher loads as k increases.

There are, of course, other variables which can be derived from the system we have described; for example, in studies of equipment life, it might be useful to know the distribution of the number of gate-closings that occur in an interval of length t. It did not seem worthwhile to pursue such questions in the present study, which deals with gating from the viewpoint of traffic performance.

APPENDIX

Suppose we have an M/M/N queuing system employing random service, with arrival rate λ and mean holding time μ^{-1}. The mean waiting time to the point of entering service of an arbitrary customer in such a system is

$$m = \frac{p_N}{N\mu(1 - \rho)^2},$$

where $\rho = \lambda/N\mu < 1$ and p_N is the known equilibrium probability that there are exactly N customers in the system. We wish to determine the mean waiting time, m_j, of a customer who arrives to find $N + j$ other customers in the system. The m_j's satisfy

$$m_j = \frac{1}{\lambda + N\mu} + \frac{\lambda}{\lambda + N\mu} m_{j+1} + \frac{N\mu}{\lambda + N\mu} \cdot \frac{j}{j + 1} m_{j-1}, \qquad j \geqq 0. \quad (25)$$

The rationale for this equation is that a customer must wait at least until the next change of state; the mean of this initial delay is $(\lambda + N\mu)^{-1}$. If the change of state is caused by an arrival [which occurs with probability $\lambda/(\lambda + N\mu)$], then the customer will have to wait an additional period of time whose mean is m_{j+1}. If, on the other hand, the change of state is caused by a departure [which occurs with probability $N\mu/(\lambda + N\mu)$], then with probability $j/(j + 1)$ our customer will not be chosen from the group of $j + 1$ customers, and he will have to wait an additional period of time whose mean is m_{j-1}.

We now introduce the function

$$H(x) = \sum_{j=0}^{\infty} m_j x^j.$$

This series converges for $x \leqq \rho$, since the mean waiting time of an arbitrary customer is

$$m = p_N H(\rho) = \frac{p_N}{N\mu(1 - \rho)^2}. \quad (26)$$

Multiplying eq. (25) by $(1 + \rho)(j + 1)x^j$ and summing on j, we can obtain a first-order differential equation:

$$H'(x) + \frac{1 + \rho - x}{(1 - x)(x - \rho)} H(x) = \frac{1}{N\mu} \frac{1}{(1 - x)^3(x - \rho)}.$$

The solution to this equation is

$$H(x) = C(1 - x)^\rho \left(\frac{1 - x}{x - \rho}\right)^{1/(1-\rho)} + \frac{2 - x}{N\mu(1 - x)^2(2 - \rho)}.$$

Using eq. (26) as the boundary condition, we see that C must be zero in order that $H(x)$ remain finite as $x \rightarrow \rho$. Thus

$$H(x) = \frac{2 - x}{N\mu(1 - x)^2(2 - \rho)}.$$

The power series expansion of this function is found to be

$$H(x) = \sum_{j=0}^{\infty} \frac{j + 2}{N\mu(2 - \rho)} x^j;$$

therefore, the means we desire are given by

$$m_j = \frac{j + 2}{N\mu(2 - \rho)}.$$

REFERENCES

1. Kingman, J. F. C., "The Effect of Queue Discipline on Waiting Time Variance," Proc. Cambridge Phil. Soc., *58* (1962), pp. 163–164.
2. Wilkinson, R. I., "Working Curves for Delayed Exponential Calls Served in Random Order," B.S.T.J., *32*, No. 2 (March 1953), pp. 360–383.
3. Beckwith, D. A., "Delay at a Simple Gate," unpublished work, May 16, 1958.
4. Kendall, D. G., "Some Problems in the Theory of Queues," J. Roy. Stat. Soc., Series B, *13*, No. 2 (1951), pp. 151–185.
5. Neuts, M. F., "The Queue with Poisson Input and General Service Times, Treated as a Branching Process," Duke Math. J., *36*, No. 2 (June 1969), pp. 215–231.
6. Nair, S. S., and Neuts, M. F., "A Priority Rule Based on the Ranking of the Service Times for the M/G/1 Queue," Operations Res., *17*, No. 3 (May-June 1969), pp. 466–477.
7. Nair, S. S., and Neuts, M. F., "An Exact Comparison of the Waiting Times Under Three Priority Rules," Operations Res., *19*, No. 2 (March-April 1971), pp. 414–423.
8. Feller, W., *An Introduction to Probability Theory and its Applications, Volume I*, New York: John Wiley and Sons, 1957, p. 355.
9. Kuczma, M., *Functional Equations in a Single Variable*, New York: Hafner Publishing Company, 1968, pp. 46–58.
10. Feller, W., *An Introduction to Probability Theory and its Applications, Volume II*, New York: John Wiley and Sons, 1966, p. 356.
11. Riordan, J., *Stochastic Service Systems*, New York: John Wiley and Sons, 1962, pp. 101–105.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Losses and Impulse Response of a Parabolic Index Fiber With Random Bends

By D. MARCUSE

(Manuscript received March 29, 1973)

The coupling coefficients of the modes of a parabolic index fiber with randomly curved axis are derived and are used to compute its excess losses and impulse response. It is found that bends with a period comparable to the natural ray oscillation period in the parabolic index medium are catastrophic. The average radius of curvature R_c of a guide composed of circular sections with an average length of 1 cm must not decrease below approximately $R_c = 1$ m. Mode coupling by random bends has the tendency to reduce the width of the impulse response function. However, this improvement is accompanied by losses. Reducing the width of the impulse response for coupled mode operation to half its uncoupled width causes 0.7 dB additional loss, a ten-fold reduction of the pulse width costs 18 dB.

I. INTRODUCTION

Optical fibers with parabolic refractive index profile[1,2] have less pulse delay distortion than conventional fibers with piecewise constant, discontinuous index distribution.[3] The width of the impulse response increases in direct proportion to the length of the fiber. It is well known that an improvement of the impulse response results if the modes are coupled among each other.[4,5] In the presence of mode coupling the width of the impulse response increases only proportionally to the square root of the length of the fiber.

Pulse propagation in multimode parabolic index fibers is studied in this paper by means of converting the coupled power equations to a partial differential equation.[5–7] Random changes of the direction of the waveguide axis are considered as the coupling mechanism.

The problem is simplified by assuming that the modes of the parabolic index fiber are essentially the same as the modes of an infinitely extended square-law medium. The fiber boundary is included in the description by requiring that modes interacting with it suffer high losses so that an effective cutoff exists. Modes below the cutoff value propagate as if they were in an infinitely extended medium. At cutoff we demand that the modes do not carry power. The effect of the waveguide wall is thus taken into account as a boundary condition that has to be satisfied by the solutions of the partial differential equation.

This formalism provides information about the width of the impulse response function and the losses associated with the coupling mechanism. The achievable improvement in pulse width due to mode coupling can thus be expressed in terms of the associated loss penalty. We conclude that mode coupling is capable of improving the already favorable impulse response of the parabolic index fiber. However, this improvement of the width of the impulse response is accompanied by excess losses. The product of the square of the pulse width ratio (width of the impulse response of coupled modes to the uncoupled pulse width) times the loss penalty is independent of the waveguide parameters and the statistics of the axis deformation. There is thus no hope of reducing the loss penalty of delay distortion improvement by optimizing the waveguide parameters.

II. MODES AND COUPLING COEFFICIENTS

We use the modes of the infinite square-law medium. The refractive index distribution is assumed to be of the form

$$n^2 = n_0^2 \left(1 - 2\bar{\Delta} \frac{r^2}{a^2}\right). \tag{1}$$

The fiber radius is at $r = a$. However, the modes are assumed to be unaffected by the fiber boundary if their mode number remains below a cutoff value. We use linearly polarized modes and obtain for the y component of the electric field[8]

$$E_{pq} = \frac{2\left(\sqrt{\frac{\mu_0}{\epsilon_0}}P\right)^{\frac{1}{2}} H_p\left(\sqrt{2}\,\frac{x}{w}\right) H_q\left(\sqrt{2}\,\frac{y}{w}\right) e^{-r^2/w^2}}{(n_0 \pi 2^{p+q} p!\,q!)^{\frac{1}{2}} w} e^{-i\beta z}. \tag{2}$$

There is also an electric field component in axial direction. But, for small refractive index changes, it is negligible. The functions H_p and H_q are Hermite polynomials of order p and q, the radius r is defined by

$$r^2 = x^2 + y^2, \tag{3}$$

and the mode radius w is

$$w = \left(\frac{\sqrt{2}a}{n_0 \text{k} \sqrt{\bar{\Delta}}}\right)^{\frac{1}{2}} \tag{4}$$

with the free-space propagation constant

$$\mathrm{k} = \omega\sqrt{\epsilon_0\mu_0}. \tag{5}$$

The propagation constant of the mode is given by the expression[8]

$$\beta = n_0\mathrm{k}\left[1 - \frac{2\sqrt{2\bar{\Delta}}}{n_0\mathrm{k}a}(p + q + 1)\right]^{\frac{1}{2}}. \tag{6}$$

The orthogonality of the modes and their normalization follows from the following equation:

$$\frac{\beta}{2\mathrm{k}}\sqrt{\frac{\epsilon_0}{\mu_0}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{pq}E^*_{p'q'}dxdy = P\delta_{pp'}\delta_{qq'}. \tag{7}$$

The asterisk indicates complex conjugation.

The cutoff condition for the guided mode has been derived in Ref. 9. The permitted maximum values of p and q are defined by the relation

$$(p + q)_c = \sqrt{\frac{\bar{\Delta}}{2}}\, n_0\mathrm{k}a = \frac{a^2}{w^2}. \tag{8}$$

Any deviation of the parabolic index fiber from its perfect geometry can be expressed by a change of its refractive index distribution.

Changes in the direction of the waveguide axis can be expressed by the following index distribution:

$$n^2 = n_0^2\left\{1 - 2\frac{\bar{\Delta}}{a^2}\left[(x - f(z))^2 + y^2\right]\right\}. \tag{9}$$

We consider waveguides bends in only one plane for simplicity. The results thus obtained can easily be extended to the general case. The appropriate coupling coefficient for this type of index change is [10,11]

$$K_{pq,p'q'} = \frac{\omega\epsilon_0}{4P(\beta_{pq} - \beta_{p'q'})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\partial n^2}{\partial z}\,\mathcal{E}^*_{pq}\mathcal{E}_{p'q'}dxdy. \tag{10}$$

Bends of the waveguide axis couple even modes to their immediate odd neighbors and odd modes to their immediate even neighbors. Only the following coupling coefficients are different from zero:

$$K_{pq,p\pm1,q} = \frac{n_0\mathrm{k}w\bar{\Delta}}{a^2}\sqrt{p}\,\frac{\frac{df}{dz}}{\beta_p - \beta_{p\pm1}} = i\,\frac{n_0\mathrm{k}w\bar{\Delta}}{a^2}\sqrt{p}f(z). \tag{11}$$

Because we restricted the waveguide curvature to the $x - z$ plane there is no coupling between the modes with different values of q.

All coupling coefficients with different q values vanish. The derivative of $f(z)$ was replaced by the function itself with the help of the relation

$$\frac{1}{\beta-\beta'}\frac{df}{dz} \to if(z). \tag{12}$$

This replacement is permissible since it is the spatial frequency $\beta - \beta'$ of $f(z)$ that is responsible for the coupling process.

III. COUPLED POWER EQUATIONS

Pulse propagation in multimode optical fibers can be described by the following set of coupled equations for the average power P_ν carried by the modes:[5]

$$\frac{\partial P_\mu}{\partial z} + \frac{1}{v_\mu}\frac{\partial P_\mu}{\partial t} = -\alpha_\mu P_\mu + \sum_{\nu=1}^{N} h_{\mu\nu}(P_\nu - P_\mu). \tag{13}$$

The single label ν indicates both p and q. The power coupling coefficients $h_{\nu\mu}$ are defined by [5,6]

$$h_{\mu\nu} = h_{\nu\mu} = |\hat{K}_{\mu\nu}|^2 F(\beta_\mu - \beta_\nu). \tag{14}$$

The coupling coefficient (11) enters the power coupling coefficients via the definition

$$K_{\mu\nu} = \hat{K}_{\mu\nu} f(z). \tag{15}$$

The power spectrum $F(\beta_\mu - \beta_\nu)$ is defined by the equation

$$F(\theta) = \left\langle \left| \frac{1}{\sqrt{L}} \int_0^L f(z) e^{-i\theta z} dz \right|^2 \right\rangle. \tag{16}$$

The symbol $\langle\ \rangle$ indicates an ensemble average.

Coupling between the modes of the parabolic index fiber is an ideal application for a diffusion theory of the power coupling process.[7] Since only nearest neighbors couple directly to each other, power redistributes itself by jumping from mode to mode in the same way as particles diffuse through real space. If the mode number is very large we can consider the set of discrete modes as a quasi-continuum and change the equation system (13) into a partial differential equation. To accomplish this transformation we consider the following expression using $h_{\mu\nu} = h_{\nu\mu}$:

$$\sum_{\nu=1}^{N} h_{\mu\nu}(P_\nu - P_\mu) = h_{\mu+1,\mu}(P_{\mu+1} - P_\mu) - h_{\mu,\mu-1}(P_\mu - P_{\mu-1}). \tag{17}$$

Considering μ as a continuous variable $\mu = \theta$ we use the approximation

$$h_{\mu+1,\mu}(P_{\mu+1} - P_\mu) - h_{\mu,\mu-1}(P_\mu - P_{\mu-1})$$
$$= \Delta\theta \left\{ h(\theta + \Delta\theta) \left(\frac{\partial P}{\partial \theta} \right)_{\theta+\Delta\theta} - h(\theta) \left(\frac{\partial P}{\partial \theta} \right)_\theta \right\}$$
$$= (\Delta\theta)^2 \frac{\partial}{\partial \theta} \left(h \frac{\partial P}{\partial \theta} \right). \tag{18}$$

With $\Delta\theta = 1$ we can write (13) as a partial differential equation

$$\frac{\partial P}{\partial z} + \frac{1}{v} \frac{\partial P}{\partial t} = - \alpha P + \frac{\partial}{\partial \theta} \left(h \frac{\partial P}{\partial \theta} \right). \tag{19}$$

The propagation constant (6) can be approximated as follows:

$$\beta = n_0 \mathrm{k} - \frac{\sqrt{2\bar{\Delta}}}{a} (\theta + q + 1) \tag{20}$$

with $p = \theta$. The difference of the propagation constants of adjacent modes,

$$\Delta\beta = \beta(\theta + \Delta\theta) - \beta(\theta) = - \frac{\sqrt{2\bar{\Delta}}}{a} = - \Omega, \tag{21}$$

is independent of θ. The power spectrum entering (14) contributes to the coupling process only at one fixed spatial frequency and is independent of the variable θ. The power coupling coefficient (14) can be expressed with the help of (4), (11), and (21) as follows ($p = \theta$):

$$h(\theta) = \frac{\sqrt{2} n_0 \mathrm{k} \bar{\Delta}^{\frac{3}{2}}}{a^3} F(\Omega) \theta. \tag{22}$$

With (22) the partial differential equation (19) assumes the form

$$\frac{\partial P}{\partial z} + \frac{1}{v} \frac{\partial P}{\partial t} = - \alpha P + K \left[\theta \frac{\partial^2 P}{\partial \theta^2} + \frac{\partial P}{\partial \theta} \right] \tag{23}$$

with

$$K = \frac{\sqrt{2} n_0 \mathrm{k} \bar{\Delta}^{\frac{3}{2}} F(\Omega).}{a^3} \tag{24}$$

We assume that the attenuation coefficient α in (19) is constant and describe the high loss, that we must attribute to modes interacting with the waveguide boundary, by means of the boundary condition

$$P(z, t, \theta) = 0 \qquad \text{for} \qquad \theta = \theta_c. \tag{25}$$

The cutoff value $\theta = \theta_c$ follows from (8):

$$\theta_c = \frac{a^2}{w^2} - q = \frac{n_0 \mathrm{k} a \sqrt{\Delta}}{\sqrt{2}} - q. \tag{26}$$

The slope $\partial P/\partial\theta$ determines the rate of power diffusion. Since no power can be lost at $\theta = 0$ we must also require

$$\frac{\partial P}{\partial \theta} = 0 \qquad \text{at} \qquad \theta = 0. \tag{27}$$

IV. STEADY-STATE POWER LOSS

We begin the discussion of the solutions of (23) by neglecting the fact that each guided mode has a slightly different group velocity and consider $v(\theta) = \text{const}$. We construct a solution of (23) by introducing the trial solution

$$P(z, t, \theta) = e^{-(\sigma+\alpha)z} G(\theta). \tag{28}$$

Substitution of (28) into (23) yields the ordinary differential equation

$$\theta \frac{d^2G}{d\theta^2} + \frac{dG}{d\theta} + \frac{\sigma}{K} G = 0. \tag{29}$$

The normalized solutions of this equation that satisfy the boundary conditions (25) and (27) are

$$G_\nu(\theta) = \frac{1}{\sqrt{\theta_c}} \frac{J_0\left(u_\nu \sqrt{\frac{\theta}{\theta_c}}\right)}{J_1(u_\nu)} \tag{30}$$

with

$$u_\nu = 2\sqrt{\frac{\sigma_\nu}{K}\theta_c}. \tag{31}$$

The parameters u_ν are determined as the roots of the equation

$$J_0(u_\nu) = 0. \tag{32}$$

The functions $G_\nu(\theta)$ are mutually orthogonal.

$$\int_0^{\theta_c} G_\nu(\theta) G_\mu(\theta) d\theta = \delta_{\nu\mu}. \tag{33}$$

The general solution of the power equation (23) is obtained as the superposition of the trial solutions

$$P(z, t, \theta) = e^{-\alpha z} \sum_{\nu=1}^{\infty} c_\nu G_\nu(\theta) e^{-\sigma_\nu z}. \tag{34}$$

The expansion coefficient c_ν can be determined from the power distribution at $z = 0$ with the help of the orthogonality condition (33),

$$c_\nu = \int_0^{\theta_c} G_\nu(\theta) P(0, t, \theta) d\theta. \tag{35}$$

The eigenvalues σ_ν are obtained from (31) and (32),

$$\sigma_\nu = \frac{K u_\nu^2}{4\theta_c}. \tag{36}$$

The eigenvalues increase with the increasing values of the roots u_ν. It is thus apparent that only the first term in the series (34) needs to be considered for large values of z. The steady-state power distribution is thus described by the equation

$$P(z, t, \theta) = c_1 e^{-(\alpha+\sigma_1)z} G_1(\theta) \quad \text{for} \quad z \rightarrow \infty. \tag{37}$$

After an initial transient has decayed, the power distribution (versus mode number θ) assumes the steady state (37). The power loss in the steady state is the sum of the constant loss α, that was assumed to be the same for every mode, plus the loss value σ_1 that stems from mode coupling due to waveguide curvature. With

$$u_1 = 2.405 \tag{38}$$

we obtain the steady-state curvature loss from (24), (26), and (36),

$$\sigma_1 = \frac{2.045 \cdot n_0 \mathrm{k} \bar{\Delta}^{3/2}}{\left(\frac{n_0 \mathrm{k} a \sqrt{\bar{\Delta}}}{\sqrt{2}} - q\right) a^3} F(\Omega). \tag{39}$$

Because of our assumption that the waveguide is curved only in one plane the steady state losses depend on the mode number q. For small values of q we find low curvature loss

$$\sigma_1 = \frac{2.89 \bar{\Delta}}{a^4} F(\Omega) \quad \text{for} \quad q = 0. \tag{40}$$

With increasing values of q the losses increase until they reach infinitely high values.

In any actual cases it is unrealistic to assume that the waveguide would be bent in only one plane. Bends in the perpendicular plane couple modes with different q values. The total steady-state loss is thus a weighted average of the losses (39). The weight factor is the number of modes for each value of q. According to (8) we have

$$p = N(q) = \sqrt{\frac{\bar{\Delta}}{2}}\, n_0 \mathrm{k} a - q \tag{41}$$

different modes for each value of q. The average loss that results from coupling all the modes by random bends in both planes is thus (see appendix)

$$\bar{\sigma}_1 = \frac{4}{(n_0\text{k}a)^2\bar{\Delta}}\int_0^{n_0 \text{k} a\sqrt{\bar{\Delta}/2}} \sigma_1 N(q)dq = \frac{5.8\bar{\Delta}}{a^4}F(\Omega). \tag{42}$$

Comparison with the loss coefficient (40) for the mode group with $q = 0$ shows that the total fiber loss (42) is just twice as large. The loss coefficient (40) is representative of a slab waveguide. The actual loss of the round fiber can thus be deduced from the slab waveguide model. That the loss coefficient of the fiber is twice as large as the slab waveguide loss might be expected, since $F(\Omega)$ stands for the amplitude of the power spectrum for bends in only one plane. The fiber is assumed to be bent in both planes with equal power spectra. The effect of both bends add, doubling the loss coefficient.

In a fiber cable the fibers may (or may not) be twisted around each other. Such twists could introduce an almost sinusoidal deformation of the fiber axis. The length Λ of the period of sinusoidal deformations that should be avoided follows from (21), $\Lambda = 2\pi/\Omega$. For numerical estimates we are using a fiber with radius $a = 4.85 \times 10^{-3}$ cm and $\bar{\Delta} = 0.014$. The critical period for this fiber is $\Lambda = 0.18$ cm. If such a period should be built into the fiber by the method of cable construction we can estimate the losses that would be caused by a given amplitude. $F(\Omega)$ has the dimension of cm^3. It can be interpreted as the ratio of the square of the amplitude of the sinusoidal deformation and the spatial bandwidth that may be caused by random phase changes. With $\Omega = 34.5$ cm^{-1} let us assume a spatial bandwidth of $\Delta\Omega = 3$ cm^{-1}. An excess loss of $\bar{\sigma}_1 = 10$ dB/km $= 2.3 \times 10^{-5}$ cm^{-1} would require $F(\Omega) = 1.57 \times 10^{-13}$ cm^3. The square of the amplitude A is given by the product of $F(\Omega)$ with the spatial bandwidth. We thus obtain the amplitude of the sinusoidal deformation of the fiber axis that causes an excess loss of 10 dB/km: $A = [F(\Omega)\Delta\Omega]^{\frac{1}{2}} = 6.9 \times 10^{-7}$ cm $= 69$ Å. Sinusoidal axis deformations at the critical wavelength are thus seen to be extremely dangerous.

V. LOSSES FOR A STATISTICAL MODEL

Even though it is only one spatial frequency of the power spectrum $F(\beta_\mu - \beta_\nu)$ that determines the steady state loss (42), it is hard to guess the amplitude $F(\Omega)$ that might be expected at this spatial frequency Ω. In order to gain insight into the expected steady-state curvature losses it is necessary to consider statistical models. We consider a model consisting of waveguide sections with constant curvature whose magnitude and sign varies randomly.

The power spectrum can be written as follows:

$$F(\Omega) = \frac{1}{L}\left\langle\left|\int_0^L f(z)e^{-i\Omega z}dz\right|^2\right\rangle$$
$$= \frac{1}{\Omega^4 L}\left\langle\left|\int_0^L \frac{d^2f}{dz^2}\, e^{-i\Omega z}\right|^2\right\rangle. \tag{43}$$

The step from the function $f(z)$ to its second derivative involved two partial integrations. The end points of the integration range do not contribute if we assume that the randomly disturbed guide is connected to two perfectly straight waveguide sections so that $f(z)$ and its first two derivatives vanish at $z = 0$ and at $z = L$.

For waveguides that are only slightly bent we can consider the second derivative of $f(z)$ as the curvature function $1/R_c(z)$. We denote with $C(u)$ the autocorrelation function of the curvature function,

$$C(u) = \left\langle \frac{1}{R_c(z)R_c(z+u)} \right\rangle. \tag{44}$$

It is well known that the power spectrum of a function is equal to the Fourier transform of its autocorrelation function.[12] We may thus write

$$F(\Omega) = \frac{1}{\Omega^4}\int_{-\infty}^{\infty} C(u)e^{-i\Omega u}du. \tag{45}$$

The autocorrelation function of waveguide sections with piecewise constant curvature and fixed length D is

$$C(u) = \begin{cases} \dfrac{D - |u|}{D}\,\kappa^2 & |u| \leqq D \\ 0 & |u| > D. \end{cases} \tag{46}$$

The parameter κ^2 is the variance (square of the rms value) of the curvature $1/R_c$. Substitution of (46) into (45) results in

$$F = \frac{2\kappa^2}{D\Omega^6}(1 - \cos \Omega D). \tag{47}$$

An average over this expression, that allows us to consider D as an averaged quantity, leaves us with

$$F(\Omega) = \frac{2\kappa^2}{D\Omega^6}. \tag{48}$$

From (21), (42), and (48) we obtain the following expression for the steady-state loss of our statistical waveguide model:

$$\sigma^{(1)} = \frac{1.4\kappa^2 a^2}{D\bar{\Delta}^2}. \tag{49}$$

As a numerical example, we consider a parabolic index waveguide with the following parameters:

$$\left.\begin{aligned} \lambda &= 1\ \mu\text{m, free-space wavelength (k} = 6.28 \times 10^4\ \text{cm}^{-1}) \\ a &= 4.85 \times 10^{-3}\ \text{cm, waveguide radius} \\ n_o &= 1.56 \\ \bar{\Delta} &= 0.014 \\ w &= 7.69 \times 10^{-4}\ \text{cm, mode radius} \\ \frac{a}{w} &= 6.3 \end{aligned}\right\}. \tag{50}$$

With these data we have (κ in cm^{-1}, D in cm)

$$\sigma^{(1)} = 0.17\,\frac{\kappa^2}{D}\ \text{cm}^{-1}. \tag{51}$$

We may now ask for the rms value of the curvature that is required to cause a steady-state loss of 2.3×10^{-5} cm^{-1} = 10 dB/km with an average length of the waveguide sections of $D = 1$ cm. We find from (51) $\kappa = 0.0116$ cm^{-1} or $1/\kappa = 86$ cm. Our result tells us that a waveguide composed of individual sections of constant curvature of average length $D = 1$ cm with $R_c \approx 1$ m radius of curvature has 10 dB/km additional loss. For our derivation we assumed that the high-order modes suffer very high losses since their fields reach into the vicinity of the waveguide wall. Whether the interaction with the outer waveguide boundary causes high losses depends on the construction of the waveguide. If the outer surface is rough or coated with an absorbing material to reduce crosstalk, the losses are high and our estimate applies.

VI. PULSE DELAY DISTORTION

It was shown in Ref. 5 that the width of the impulse response of a multimode fiber with coupled modes is given by the equation

$$\Delta t = 4\sqrt{\rho L}. \tag{52}$$

L is the length of the waveguide and ρ is the second-order perturbation of the eigenvalue σ_1 defined by (36). For the discrete case we write

$G_\nu(\theta) = G_p{}^{(\nu)}$ and obtain ρ in the form[5]

$$\rho = \sum_{\nu=2}^{N} \frac{\left[\sum_{p=1}^{N}\left(\frac{1}{v_p} - \frac{1}{v_o}\right) G_p{}^{(1)} G_p{}^{(\nu)}\right]^2}{\sigma_\nu - \sigma_1}. \tag{53}$$

The average group velocity v_o actually does not contribute to (53) on account of the orthogonality of the vectors $G^{(\nu)}$. With the assumption of a continuum of modes we obtain instead of (53)

$$\rho = \sum_{\nu=2}^{\infty}\left\{\frac{1}{\sigma_\nu - \sigma_1}\left[\int_0^{\theta_c}\left(\frac{1}{v(\theta)} - \frac{1}{v_o}\right) G_1(\theta) G_\nu(\theta) d\theta\right]^2\right\}. \tag{54}$$

The inverse group velocity is obtained by approximating (6),

$$\beta \approx n_o\mathrm{k} - \frac{\sqrt{2\bar{\Delta}}}{a}(p+q) - \frac{\bar{\Delta}}{n_o\mathrm{k}a^2}(p+q)^2, \tag{55}$$

and taking the derivative. With $v_o = c/n_o$ we obtain (with $p = \theta$ and c = light velocity in vacuum)

$$\frac{1}{v(\theta)} - \frac{1}{v_o} = \frac{1}{c}\frac{d\beta}{d\mathrm{k}} - \frac{n_o}{c} = \frac{\bar{\Delta}}{cn_o\mathrm{k}^2a^2}(\theta^2 + 2\theta q + q^2). \tag{56}$$

The term with q^2 does not contribute to the following integral because of the orthogonality relation (33):

$$\int_0^{\theta_c}\left(\frac{1}{v(\theta)} - \frac{1}{v_o}\right) G_1(\theta) G_\nu(\theta) d\theta$$
$$= \frac{16\bar{\Delta}\theta_c}{cn_o\mathrm{k}^2a^2}\frac{u_\nu u_1}{(u_\nu^2 - u_1^2)^2}\left\{\theta_c\left[1 - \frac{12(u_\nu^2 + u_1^2)}{(u_\nu^2 - u_1^2)^2}\right] + q\right\}. \tag{57}$$

Each mode group with a given value of q has a different spread of the group velocities of its uncoupled modes. However, we have seen that the waveguide losses could be obtained from the simpler slab waveguide model. This simplification is expected to apply also to the pulse distortion problem. The slab model is obtained by setting $q = 0$. The spread of inverse group velocities is largest for $q = 0$ since the allowed θ range is largest in this case. However, even though the spread of the group velocities is reduced for increasing values of q, the mode groups with different q values arrive at different times. This delay distortion is reduced by coupling of the different mode groups by means of waveguide bends in the perpendicular plane (perpendicular to the plane coupling the modes with different p values). The mode group with $q = 0$ can be excited by shining light into the fiber that is collimated in one plane but spreads in the plane of the bends in such a way that all modes with $q = 0$ and p values in the range $0 < p < n_o \mathrm{k} a(\bar{\Delta}/2)^{\frac{1}{2}}$ are excited. The bends in one plane do not cause coupling to modes with different q values but reduce the delay distortion of the modes with different p values. From this physical picture we see that the delay distortion problem is reduced to studying delay distortion in a slab waveguide. Bends of the fiber in the perpendicular plane couple the modes with different q but fixed p values. Their velocity spread is the same as that considered in the first problem. We thus expect to obtain the correct result by considering the delay distortion reduction for the mode group with $q = 0$.

With $q = 0$ we obtain from (26) and (57)

$$\int_0^{\theta_c} \left(\frac{1}{v(\theta)} - \frac{1}{v_o} \right) G_1(\theta) G_\nu(\theta) d\theta = \frac{8 n_o \bar{\Delta}^2}{c} \frac{u_\nu u_1}{(u_\nu^2 - u_1^2)^2} \left[1 - \frac{12(u_\nu^2 + u_1^2)}{(u_\nu^2 - u_1^2)^2} \right]. \tag{58}$$

Using (24), (36), (54), and (58) we have

$$\rho = \frac{128 n_o^2 a^4 \bar{\Delta}^3}{c^2 F(\Omega)} \left\{ \sum_{\nu=2}^{\infty} \frac{u_\nu^2 u_1^2}{(u_\nu^2 - u_1^2)^5} \left[1 - \frac{12(u_\nu^2 + u_1^2)}{(u_\nu^2 - u_1^2)^2} \right]^2 \right\} = 2.26 \times 10^{-4} \frac{n_o^2 a^4 \bar{\Delta}^3}{c^2 F(\Omega)}. \tag{59}$$

The width of the impulse response follows from (52),

$$\Delta t = 0.06 \frac{n_o a^2}{c} \bar{\Delta}^{\frac{3}{2}} \left(\frac{L}{F(\Omega)} \right)^{\frac{1}{2}}. \tag{60}$$

From Ref. 9 we obtain the width of the impulse response for uncoupled modes

$$\Delta\tau = \frac{n_o L}{2c} \bar{\Delta}^2. \tag{61}$$

The improvement that is caused by mode coupling is the ratio of the widths of these two impulse responses

$$R = \frac{\Delta t}{\Delta \tau} = \frac{0.12 a^2}{[F(\Omega) L \bar{\Delta}]^{\frac{1}{2}}}. \tag{62}$$

For the statistical model of a sequence of circularly bent waveguide sections we obtain with (48) and (21)

$$R = \frac{0.24 \bar{\Delta}}{\kappa a} \left(\frac{D}{L} \right)^{\frac{1}{2}}. \tag{63}$$

We may ask for the average radius of curvature that is required to cause a ten-fold improvement of the width of the impulse response due to mode coupling. Using $L = 1$ km and an average length of the bent sections of $D = 1$ cm and the numbers in (50) we obtain for $R = 0.1$, $\kappa = 0.022$ cm^{-1} or an average radius of curvature $R = 1/\kappa = 45$ cm. This relatively small radius of curvature may cause very substantial excess loss according to the loss example in Section V.

In order to relate the excess loss to the delay distortion improvement we consider the loss penalty that must be paid for a given amount of improvement in the width of the impulse response. Both the loss formula (40) and the improvement factor R, (62), contain the power spectrum of the distortion function $f(z)$. By taking the square of R and multiplying it with the loss per length L, $\sigma_1 L$, we obtain[†]

$$R^2\sigma_1 L = 0.042 = 0.18 \text{ dB}. \tag{64}$$

This important formula is independent of any of the waveguide parameters and of the statistics of the axis deformation. This means that the loss penalty, $\sigma_1 L$, for parabolic index fibers depends only on the delay distortion improvement that one wants to achieve. For $R = 1$ we have a loss of $\sigma_1 L = 0.18$ dB. Clearly, the range of applicability of (64) is exceeded in this case since $R = 1$ means that there is no improvement at all. For $R = 0.5$ we pay a loss penalty of 0.7 dB, $R = 0.1$ increases the loss to 18 dB. The already favorable delay distortion of the parabolic index fiber can be improved by intentional curvature of the waveguide axis.

VII. DISCUSSION

We have studied the performance of the parabolic index fiber with randomly curved axis. The curvature of the waveguide axis has the tendency to force a light beam inside of the fiber towards the fiber boundary. In terms of wave optics this means that the wave field begins to interact with the boundary of the fiber. If this boundary is perfectly smooth no particular harm may be done except that the impulse response of the fiber is likely to deteriorate. However, the interfaces between two dielectric regions tend to be rough. Surface roughness leads to scattering losses. We have thus assumed that the interaction of the mode fields with the fiber boundary causes significant losses to high-order modes. On this basis we were able to calculate the fiber loss caused by random bends of the waveguide axis. For bends that approach a sinusoidal shape, with a period comparable to

[†] We use σ_1 of (40) instead of σ_1 of (42) since R was computed for $q = 0$.

the ray oscillation period in the parabolic index medium, the excess losses are extremely high. Bending of the waveguide axis with a period equal to the ray oscillation period must be avoided. For a statistical model, based on the assumption that the waveguide is composed of a sequence of circularly bent sections with random length and random radius of curvature, the waveguide losses have been predicted. We conclude that average radii of curvature of approximately 1 m can be allowed if an excess loss of 10 dB/km can be tolerated. The waveguide sections were assumed to have an average length of 1 cm.

It is possible to reduce the width of the impulse response of a parabolic index fiber by coupling its modes by random bends of the fiber axis. The impulse response of parabolic index fibers is already quite favorable compared to the impulse response of the conventional optical fiber with a discontinuous but piecewise constant index distribution. Our analysis shows that additional reduction of pulse delay distortion is accompanied by losses. A reduction of the pulse width to half its uncoupled width increases the loss by 0.7 dB, a ten-fold pulse width reduction increases the fiber loss by 18 dB.

VIII. ACKNOWLEDGMENT

Several helpful discussions with E. A. J. Marcatili are gratefully acknowledged.

APPENDIX

The averaging process used to obtain (42) can be justified as follows. Each mode group characterized by the mode number q comprising all modes with

$$0 < p < \left(\sqrt{\frac{\bar{\Delta}}{2}}\, n_o \mathrm{k} a - q\right)$$

has the loss coefficient $\sigma_1(q)$. By definition this can be written

$$\sigma_1(q) = \frac{\Delta P(q)}{P(q)}.$$

$\Delta P(q)$ is the power lost from the mode group per unit length and $P(q)$ is the power carried by these modes. If we assume, for simplicity, that each mode carries the power P we can write $P(q) = N(q)P$ so that we have

$$\sigma_1(q) = \frac{\Delta P(q)}{N(q)P},$$

with $N(q)$ indicating the number of modes in the group. The total loss is

$$\bar{\sigma}_1 = \frac{\sum_q \Delta P(q)}{\sum_q P(q)} = \frac{\sum_q \sigma_1(q) N(q)}{\sum_q N(q)}.$$

Replacing the sum by an integral yields formula (42).

REFERENCES

1. Uchida, M., Furukawa, M., Kitano, I., Koizumi, K., and Matsumura, H., "A Light Focusing Fibre Guide," IEEE J. Quantum Elec., (Digest of Technical Papers), *QE-5*, No. 6 (June 1969), p. 331.
2. Pearson, A. D., French, W. G., and Rawson, E. G., "Preparation of a Light Focusing Glass Rod by Ion Exchange Techniques," Appl. Phys. Ltrs., *15*, No. 2 (July 15, 1969), pp. 76–77.
3. Kawakami, S., and Nishizawa, J., "An Optical Waveguide with Optimum Distribution of the Refractive Index with Reference to Waveform Distortion," IEEE Trans. Microwave Theory and Techniques, *MTT-16*, No. 10 (October 1968), pp. 814–818.
4. Personick, S. D., "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
5. Marcuse, D., "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J. *51*, No. 6 (July-August 1972), pp. 1199–1232.
6. Marcuse, D., "Derivation of Coupled Power Equations," B.S.T.J., *51*, No. 1 (January 1972), pp. 229–237.
7. Gloge, D., "Optical Power Flow in Multimode Fibers," B.S.T.J., *51*, No. 8, (October 1972), pp. 1767–1783.
8. Marcuse, D., *Light Transmission Optics*, New York: Van Nostrand Reinhold Company, 1972, p. 270.
9. Marcuse, D., "The Impulse Response of an Optical Fiber with Parabolic Index Profile," B.S.T.J., *52*, No. 7 (September 1973), pp. 1169–1174.
10. Snyder, A. W., "Mode Propagation in a Nonuniform Cylindrical Medium," IEEE Trans. Microwave Theory and Techniques, *MTT-19*, No. 4 (April 1971), pp. 402–403.
11. Marcuse, D., "Coupling Coefficients for Imperfect Asymmetric Slab Waveguides," B.S.T.J., *52*, No. 1 (January 1973), pp. 63–82.
12. Reference 8, pp. 369–370.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Effect of Misalignments on Coupling Efficiency of Single-Mode Optical Fiber Butt Joints

By J. S. COOK, W. L. MAMMEL, and R. J. GROW

(Manuscript received April 9, 1973)

Analysis and computations made here, corroborated by experiment, determine the effects of axial displacement and angular misalignment on the power coupled between butt-joined, single-mode optical fibers. The absolute accuracy with which fibers must be joined on-centers is reduced for fibers with relatively smaller core; the angular accuracy is increased.

I. INTRODUCTION

The lowest-order mode in a clad optical fiber, the hybrid HE_{11} mode, is the only propagating mode for core sizes less than a few wavelengths in diameter ($v \leqq 2.4$).[1] Hence, small-core, single-mode glass fibers are attractive for transmitting optical signals because of their potentially low dispersive effects. A possible disadvantage lies in the difficulty of joining small-core fibers end-to-end.[2,3] It has been suggested[4] that butt joining may be made less critical by reducing the size of the fiber in the vicinity of the joint. The computation and experimental measurements disclosed here evaluate the advantage to be gained by such a procedure. It is found that as the fiber gets smaller, the accuracy with which the ends of the fibers to be joined must meet on-centers is indeed reduced. At the same time, however, as might be expected, the required angular alignment of the fibers becomes more critical.

Both calculations and experiments have been made under the assumption that the cladding is of sufficient extent that the role played by possible conversion of the zero-order mode to cladding modes need not be considered.[4] The calculation was made by matching the fields of the zero-order modes at the joint.

II. ANALYSIS

The power loss caused by displacement at a joint is readily found by first determining the ratio of power accepted by the displaced fiber to that presented by the sending fiber, that is, the power-coupling ratio. This is determined by assuming that the incident field E_i separates at the plane interface (which is perpendicular to the axis of the receiving fiber) into the desired zero-order mode that propagates in the receiving fiber, and into modes orthogonal to this propagating mode. The field of the propagating mode is represented by $B\cdot E_p$, since the form E_p is known but not the amplitude B. The field of the orthogonal modes is represented by E_o.

$$E_i = B\cdot E_p + E_o.$$

Multiplication by E_p and integration over the entire interface gives

$$\int E_iE_pdA = B\int E_p^2dA + 0.$$

The desired power ratio, c, is then represented by B^2.

$$c = \left[\int E_iE_pdA \bigg/ \int E_p^2dA\right]^2.$$

The power-coupling ratio between fiber ends with parallel but translated axes is c_1. In order to show the effect of axial displacement of a given magnitude, independent of core size, the ratio c_1 is computed

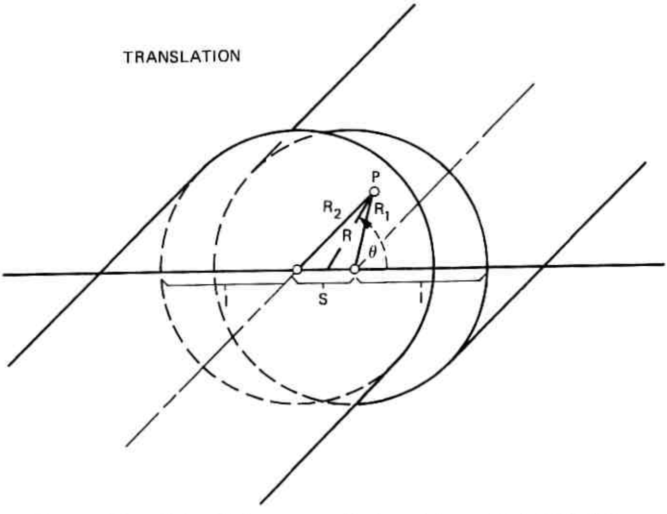

Fig. 1—Butt-joined fiber cores displaced by axial translation.

as a function of v, the normalized core size,[1] for several values of the parameter d, the normalized displacement.

$$v = \frac{2\pi a}{\lambda}(n_c^2 - n_o^2)^{\frac{1}{2}}, \tag{1}$$

$$d = \frac{2\pi s}{\lambda}(n_c^2 - n_o^2)^{\frac{1}{2}}, \tag{2}$$

where a is the core radius, s is the axial displacement distance, λ is the wavelength, and n_c and n_o are the core and cladding refractive indexes, respectively.

Figure 1 shows the cross section of a fiber displaced by axial translation. R and S are normalized to radius a. The origin is defined as the midpoint between core centers. R and θ are the polar coordinates of an arbitrary point P in the interface. R_1 is the distance from P to the center of the first fiber, and R_2 is the distance to the center of the second fiber.

$$R_1^2 = R^2 + \left(\frac{S}{2}\right)^2 - RS\cos\theta,$$

$$R_2^2 = R^2 + \left(\frac{S}{2}\right)^2 + RS\cos\theta.$$

Let A_{12} be the area of the interface of the displaced fibers, and A be the area of the cross section of the sending fiber; then

$$c_2(v) = \left[\int_{A_{12}} E(R_1)E(R_2)dA_{12} \bigg/ \int_A E^2(R)dA\right]^2.$$

The function E is defined[1] by

$$E(R) = \begin{cases} \dfrac{J_o(uR)}{J_o(u)}, & R \leqq 1, \\[2ex] \dfrac{K_o(wR)}{K_o(w)}, & R > 1, \end{cases}$$

where J_o and K_o are the regular and modified Bessel functions of order zero.

The values of u and w are determined from the eigenvalue equations,

$$v = (u^2 + w^2)^{\frac{1}{2}},$$

$$\frac{uJ_1(u)}{J_o(u)} = \frac{wK_1(w)}{K_o(w)}.$$

The integral in the denominator of the expression c_1 can be integrated analytically. For an infinite cladding,

$$\int_A E^2(R)dA = \pi\left[\frac{vJ_1(u)}{wJ_o(u)}\right]^2.$$

For a fiber of radius R_c, ($R_c > 1$),

$$\int_A E^2(R)dA = \pi\left(\frac{v}{w}\frac{J_1(u)}{J_o(u)}\right)^2 - \frac{\pi R_c^2}{K_o^2(w)}\left[K_1^2(wR_c) - K_o^2(wR_c)\right].$$

The integral in the numerator can be divided so that portions of the integration can be done analytically.

The power-coupling ratio between fibers with angular displacement of the fiber axes is c_2. In order to show the effect of a fixed angular displacement of the axes for different core sizes, the ratio c_2 is computed as a function of v for several values of the parameter b, the normalized displacement angle,

$$b = \frac{\sin\phi}{\sqrt{1 - n_o^2/n_c^2}} \approx \frac{\phi}{\sqrt{2\Delta}}, \tag{3}$$

where ϕ is the angle between the fiber axes, and Δ is the ratio of core-to-cladding index difference to the core index.

Figure 2 shows a cross section of the fiber joint in the plane of the axes of the fiber and auxiliary cross sections perpendicular to the axes of the fibers. R', θ', z' and x', y', z' are coordinate systems oriented with respect to the sending fiber while R, θ, z and x, y, z are coordinate systems oriented with respect to the receiving fiber; ϕ is the angle between the axes of the two fibers (and therefore between the planes perpendicular to the axes). From Fig. 2 it can be seen that

$$R' = R(1 - \sin^2\theta\sin^2\phi)^{\frac{1}{2}}.$$

The field of the sending fiber at the point P of the interface L_1 is

$$E_i(R) = E(R')\cos\beta z',$$

where β is the normalized propagation constant.

$$\beta \approx \frac{v}{\sqrt{2\Delta}}.$$

For the small angles under consideration the maximum deviation of R'/R from 1 is at most $2\cdot 10^{-2}$, so it is feasible to approximate R' by R. Therefore, $\beta z' = bvR\sin\theta$ and

$$c_2(v) = \left[\int_{A_{12}} E^2(R)\cos(bvR\sin\theta)dA \Big/ \int_A E^2(R)dA\right]^2.$$

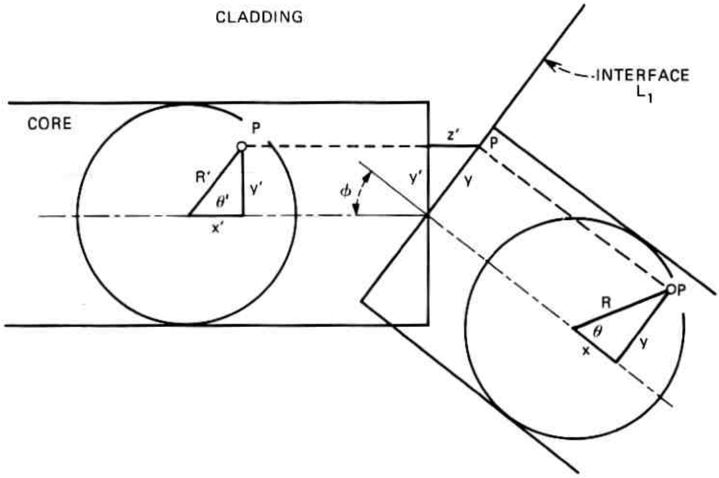

Fig. 2—Analytical model of butt-joined fiber cores displaced by angular misalignment, ϕ.

The integral in the numerator, since we have approximated R' by R, can be converted to a single integral.

$$\int_0^{R_c}\int_0^{2\pi} E^2(R)\cos\,(bvR\sin\theta)Rd\theta dR = 2\pi\int_0^{R_c} E^2(R)J_o(bvR)RdR.$$

The integral in the denominator is the same as before and can be integrated analytically.

III. EXPERIMENT

Experimental verification was carried out at microwave frequencies because equipment was readily available and dimensional control more certain than at optical frequencies. Polyfoam served as the core, and air as the cladding. The experimental arrangement is shown in Fig. 3. The polyfoam rod had dielectric constant of 1.06; hence, $(n_c^2 - n_o^2)^{\frac{1}{2}} = 0.245$. Results of computations from the analysis and experimentally measured points are shown overlaid in Figs. 4 and 5. The vertical point dimension indicates the disparity between two systematic measurements. Disparity between the analytical and experimental results for very small core dimensions and larger coupling loss is not understood; but it is not critical to the conclusions.

Figure 6 shows an overlay of calculated power coupling as a function of v due to both angular and lateral misalignments. It can be argued that the total power lost at the junction is roughly equal to the sum

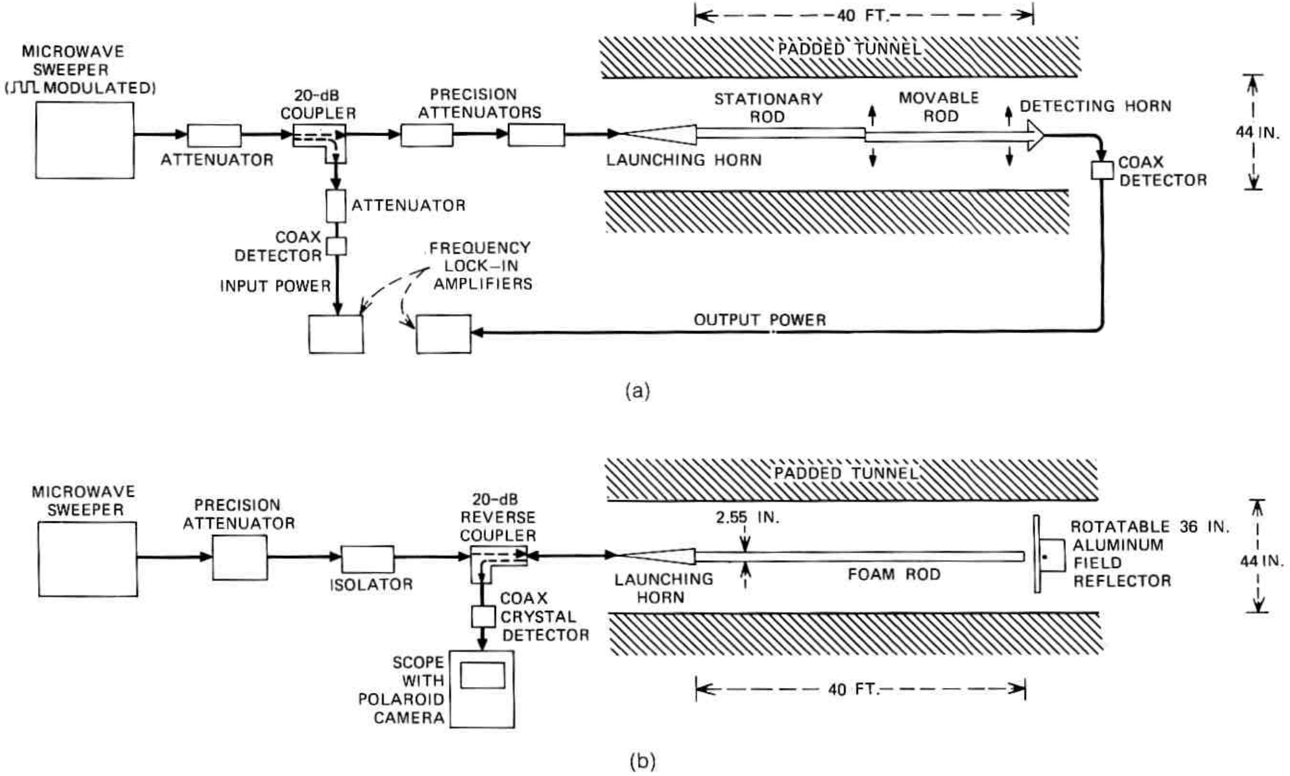

Fig. 3—Experimental arrangement used to corroborate analysis using microwave-guiding polyfoam rods. (a) Coupling vs axial displacement measurement configuration. (b) Coupling vs angular misalignment measurement configuration.

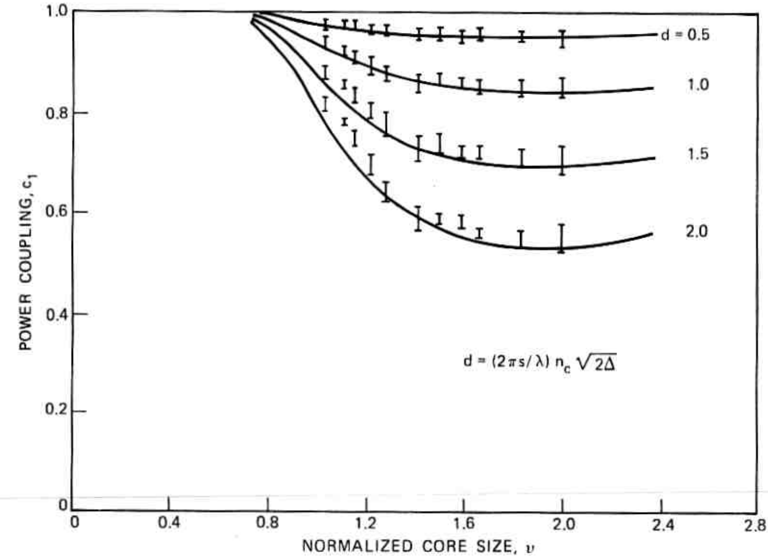

Fig. 4—Power coupling through translationally displaced joint as a function of normalized core size and axial translation, *d*.

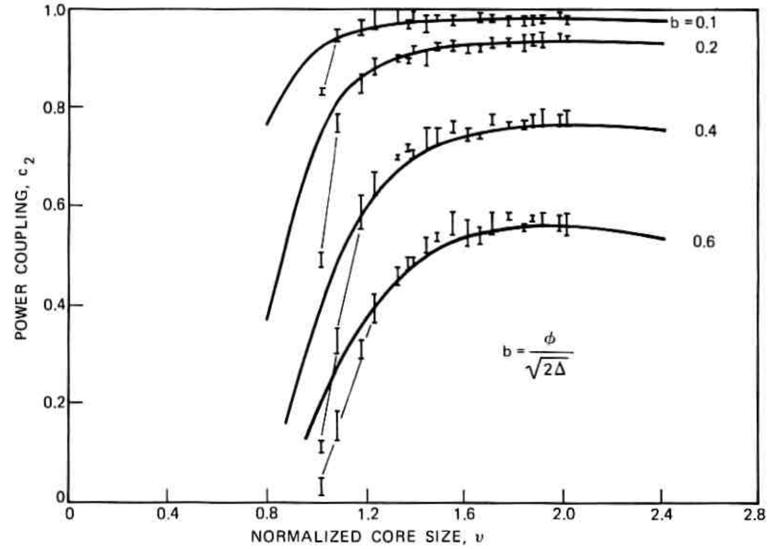

Fig. 5—Power coupling through angularly misaligned joint as a function of normalized core size and misalignment, *b*.

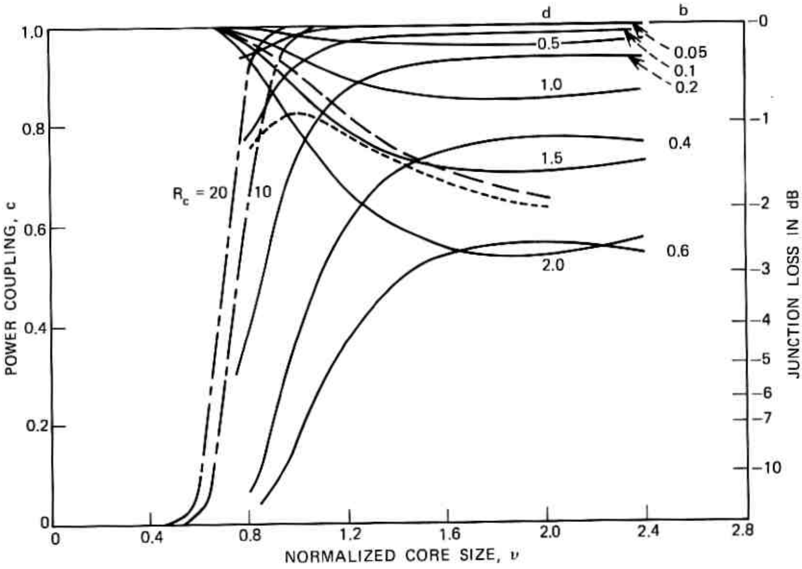

Fig. 6—Minimum power coupling through example joint as a function of normalized core size. Dashed line shows power coupled at maximum axial translation, dotted line shows composite minimum coupling.

of the lost power due to the two kinds of misalignment. Since many modes are involved in the scattered energy, and the displacement coordinates are different, the scattered modes must be essentially orthogonal, hence, power-additive. The dash-dot curves in Fig. 6 show the fraction of power lying within 10 and 20 radii of the core as labeled.

IV. CONCLUSIONS

In general, considering the nature of field spreading that accompanies the decrease in single-mode fiber core diameter, one concludes what one would expect. As the core size parameter, v, decreases below about 2, the fields spread, and axial displacement at the joints becomes less and less critical since more and more overlap of fields results from a given offset. At the same time, the effective "aperture" at the fiber end increases, the "antenna" becomes more directive, and the angular alignment becomes more critical.

More detailed conclusions must be drawn from the particular joining problem at hand. Suppose, for example, one wants to join single-mode fibers of 3-mil diameter, having a core loaded to produce an index difference at the interface of $\frac{1}{2}$ percent. If the refractive index of the fiber material is about 1.5, then

$$(n_c^2 - n_o^2)^{\frac{1}{2}} \approx n_c\sqrt{2\Delta} = 0.15.$$

If the fiber is designed to have $v = 2.0$ and the optical wavelength, λ, is 1 μm, the core radius may be found from eq. (1) to be

$$a \approx 2.1 \ \mu\text{m}.$$

Since 1.5 mils $\approx$ 38 μm, the cladding radius is about 18 times the core radius, so

$$R_c = 18.$$

Suppose the centering accuracy of the tool or fixture that will be used in aligning the fiber ends at a joint is about ± 1 μm. (It presumably will center the fiber with reference to its o.d.) And suppose the core is centered in the fiber with $\pm$1-percent accuracy, that is, the core center is never displaced from the fiber center by more than 1 percent of the fiber diameter. The net maximum displacement, then, will be

$$s_m \approx 1.75 \ \mu\text{m}$$

and, from (2),

$$d_m \approx 1.65.$$

If the fiber is drawn down to a smaller size at the end to decrease the lateral displacement sensitivity at the joint, it is reasonable to assume that the alignment tool maintains the same 1-μm accuracy and the core maintains the same $\pm$1-percent centering accuracy; the latter of which produces a net decrease in d_m. The dashed curve in Fig. 6 shows what happens to the minimum coupling, c_1, as v scales down with fiber size.

If at the same time one assumes that the net angular misalignment, ϕ, is less than about 0.01 radian, then from (3),

$$b \approx 0.1,$$

and the combined effect of angular and lateral displacement varies with v as shown by the dotted line in Fig. 6.

The total worst-case joint loss in this example, then, may be reduced by about a factor of two by drawing the fiber (core and cladding together) down to half its normal size at the ends for joining. At $v = 1$ the field extending outside the cladding is still negligible.

V. ACKNOWLEDGMENT

We thank E. A. J. Marcatili for providing the key to the analysis, so deftly drawn from the concepts of superposition and orthogonality.

REFERENCES

1. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, No. 10 (October 1971), pp. 2252–2258.
2. Bisbee, D. L., "Measurements of Loss Due to Offsets and End Separations of Optical Fibers," B.S.T.J., *50*, No. 10 (December 1971), pp. 3159–3168.
3. Somena, C. G., "Simple, Low-Loss Joints Between Single-Mode Optical Fibers," B.S.T.J., *52*, No. 4 (April 1973), pp. 583–596.
4. Dyott, R. B., Stern, J. R., and Stewart, J. H., "Fusion Junctions for Glass-Fiber Waveguides," Elec. Ltrs., *8*, No. 11 (June 1, 1972), pp. 290–292.

Contributors to This Issue

James E. Bennett, B.Sc. (Met.E.), 1959, Carnegie Institute of Technology; M.Sc. (Metallurgy), 1961, and Ph.D. (Metallurgy), 1965, Case Institute of Technology; General Electric Refractory Metals Laboratory, 1964–1966; Battelle Memorial Institute, 1966–1968; Bell Laboratories, 1968—. At Bell Laboratories in Columbus, Mr. Bennett has been conducting fundamental studies on the interaction of liquid mercury with contact metals, interdiffusion in connector materials, phase transformations in and processing behavior of Fe/Co/2–3% V alloys, and contact materials development. Member, AIME, IMS, Alpha Sigma Mu, Sigma Xi.

Min-Te Chao, B.S. 1961, National Taiwan University; M.A., 1965, and Ph.D., 1967, University of California at Berkeley; Bell Laboratories, 1968—. Mr. Chao's interests have been in the fields of large sample theory, sequential analysis, and, recently, application of Markov chains to error structures in digital data communication. Member, China Mathematical Society.

Roger D. Coleman, B.A., 1964, and Ph.D., 1968, The Johns Hopkins University; Bell Laboratories, 1968–1973. Mr. Coleman has worked on problems in economics, queuing theory, and traffic studies. Member, American Statistical Association, Institute of Mathematical Statistics.

John S. Cook, B.S. and M.S. (Electrical Engineering), 1952, Ohio State University; Bell Laboratories, 1952—. Mr. Cook has been engaged in several areas of electronics and electromagnetics research. At present, he is working on problems of communication with optical fibers. Senior member, IEEE; member, Optical Society of America.

A. Descloux, Math. Dipl., 1948, Swiss Federal Institute of Technology, Zürich; Ph.D. (Mathematical Statistics), 1961, University of North Carolina. After spending 1955–56 on the staff of the University of Washington where he taught mathematics and statistics, Mr. Descloux joined Bell Laboratories in 1956. At Bell Laboratories, he has been concerned chiefly with the application of probability theory to traffic problems. Member, Institute of Mathematical Statistics, American Mathematical Society.

ROBERT J. GROW, B.S. (Electrical Engineering), 1973, University of Utah; Bell Laboratories, Summer 1972. Mr. Grow carried out the experimental work associated with loss in single-mode optical waveguide junctions reported herein.

JOHN O. LIMB, B.E.E., 1963, and Ph.D., 1967, University of Western Australia; Research Laboratories, Australian Post Office, 1966–1967; Bell Laboratories, 1967—. Mr. Limb has worked on the coding of picture signals to reduce channel capacity requirements involving intraframe coding, frame-to-frame coding, and the coding of color signals. He currently heads the Visual Communication Research Department. Member, IEEE, Association for Research in Vision and Opthalmology, Optical Society of America.

WANDA L. MAMMEL, A.B. (Mathematics), 1943, Winthrop College; M.Sc. (Applied Mathematics), 1945, Brown University; Bell Laboratories, 1956—. Ms. Mammel is engaged in finding mathematical methods for the numerical solution of a variety of problems. In particular, she has applied linear programming techniques to problems of crystal plasticity. At present, she is working on problems in microwave propagation and optical waveguides.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of two books. Fellow, IEEE; member, Optical Society of America.

M. ROBERT PINNEL, B.Sc. (E.E.), 1966, M.Sc. (Met.E.), 1968, and Ph.D. (Materials Sci.), 1970, Drexel University; Bell Laboratories, 1970—. Mr. Pinnel has been engaged in research on the physical and mechanical characterization of numerous copper-based alloys used as spring materials, interdiffusion in electrical connector materials, the detailed characterization of the Fe/Co/2–3% V alloy system, and the interaction of liquid mercury with numerous metallic elements. His current interests are in the areas of semihard magnetic materials and multiphase metallic contact materials. Member, ASM, AIME, Tau Beta Pi, Phi Kappa Phi, Alpha Sigma Mu, Eta Kappa Nu.

Stephen O. Rice, B.S. (Electrical Engineering), 1929, and D.Sc. (Hon.), 1961, Oregon State College; Bell Laboratories, 1930–1972. Mr. Rice has been concerned with theoretical problems related to electromagnetic wave propagation, signal modulation, and noise. At the time of his retirement from Bell Laboratories, he was head of the Communications Analysis Research Department. In 1965, Mr. Rice received the Mervin J. Kelly Award from the Institute of Electrical and Electronic Engineers. Fellow, IEEE.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems, and is currently a supervisor in the Advanced Data Communications Department. During the academic year 1967–68 he was on leave as Professor of Electrical Engineering at the University of Florida. Member, Sigma Xi.