THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 52 May–June 1973 Number 5

Distortion Produced by Band Limitation of an FM Wave

By S. O. RICE

(Manuscript received November 21, 1972)

The bandwidth required to transmit an FM wave is related to how much distortion is allowed in the signal. Here expressions are developed for the distortion (interchannel interference) produced when an FDM-FM wave passes through an ideal filter. The signal is represented by a flat (PM) band of Gaussian noise. The formulas obtained hold only for small rms frequency deviation, but fortunately this is an important case in microwave communication systems. The theoretical expressions agree well with Monte Carlo results published recently by Anuff and Liou.

I. INTRODUCTION

When a frequency-modulated wave passes through a filter, distortion is produced in the signal by nonlinearity in the filter phase shift (usually the chief offender) and by the filter attenuation. Much effort has been spent in devising methods for computing this distortion.

A related problem is "What radio frequency bandwidth is required to transmit a given FM wave?" An approximate answer, known as "Carson's rule," states that the required bandwidth $2f_h$ is given by[1]

$$2f_h = 2B + 2D_{\max}, \tag{1}$$

where B is the bandwidth of the baseband signal and $D_{\max}$ is the maximum amount the instantaneous frequency deviates from the carrier frequency. Note that (1) implies a conventional FM system. This is the only type we shall consider in this paper. We shall not be concerned with single-sideband FM or other schemes for reducing the rf bandwidth.

Carson's rule has been revised recently by Anuff and Liou.[2] They make use of Monte Carlo calculations of the interchannel interference produced when an FM wave carrying a multichannel signal passes through an ideal filter. The ideal filter has zero attenuation and phase shift within the passband, and infinite attenuation outside the band. Monte Carlo calculations of interchannel interference in microwave systems have also been made by Grierson and McGee.[3]

Here we make a beginning on the analysis (in contrast to Monte Carlo) required to calculate the interchannel interference produced by an ideal filter.

The FM wave is $\cos[\omega_o t + \varphi(t)]$ where $\omega_o = 2\pi f_o$ and $\varphi(t)$ is a stationary, zero-mean Gaussian process with the two-sided power spectrum

$$W_\varphi(f) = \begin{cases} W_o, & |f| \leqq B \\ 0, & |f| > B. \end{cases} \tag{2}$$

In (2), W_o is a constant and B is the top baseband frequency. In order to represent an idle channel at frequency f_c, we take $W_\varphi(f) = 0$ in the narrow slots $f_c \leqq |f| \leqq f_c + \Delta f_c$, Δf_c being so small that $W_\varphi(f)$ can be replaced, without appreciable error, by W_o in the integrals appearing in the analysis.

The mean-square value of $\varphi(t)$ and the rms frequency deviation D are given by the ensemble averages

$$\begin{aligned} \langle \varphi^2(t) \rangle &= 2W_o B \text{ (rad)}^2 \\ D^2 = \langle (\varphi'(t)/2\pi)^2 \rangle &= 2W_o B^3/3 \text{ (Hz)}^2 \end{aligned} \tag{3}$$

where $\varphi'(t) = d\varphi(t)/dt$. This $\varphi(t)$ gives a convenient approximation to the preemphasized wave assumed by Anuff and Liou. A representative value of $D_{\max}$ in (1) is $4D$.

The ideal filter passband extends from $f_o - f_h$ to $f_o + f_h$. It is assumed that $2f_h/f_o \ll 1$ and that $nB < f_h < (n+1)B$ where n is a positive integer.

Our aim is to apply results from the theory of Volterra series to obtain an expression for the dominant portion of the interchannel interference when the normalized rms frequency deviation D/B becomes small.

For the moment, consider one-sided power spectra. Now the power spectrum of $\varphi(t)$ extends from 0 to B and has the value $2W_o$. The average signal power (FM) appearing in the channel $(f, f + \Delta f)$ when it is busy is

$$S = (2\pi f)^2 2W_o \Delta f \text{ (rad/s)}^2. \tag{4}$$

Let N be the average interchannel interference power which appears in the same channel. The value of N depends upon whether the channel is idle or busy. When the channel is idle, the interference can be heard as crosstalk noise. In our expressions for N/S, we assume that our particular channel is idle, that all the other channels are busy, and that N/S is the limit obtained as Δf tends to zero.

The nature of our results is illustrated by the following expression for N/S in the top baseband channel:

$$\begin{gathered} N/S = \left[\frac{3}{2} \frac{D^2}{B^2} \left(n + 1 - \frac{f_h}{B} \right) \right]^{2n} C_{on} + 0[(D/B)^{4n+2}], \\ C_{on} = \frac{1}{(2n)!} \sum_{k=1}^{n} \frac{(2n-2k)!(2k-1)!}{(k-1)!k!^2(k+1)!} \left(\frac{1}{(n-k)!} \right)^4. \end{gathered} \tag{5}$$

Here the integer n is determined by the filter semibandwidth f_h and the relation $nB < f_h < (n + 1)B$. The first three values of C_{on} are $C_{o1} = 1/4$, $C_{o2} = 5/96$, and $C_{o3} = 19/10368$. For large n, C_{on} tends to $2^{2n+1}/[n!^4 \pi n(n + 2)]$.*

Equations (5) are a special case, $f = B$, of (52) which gives N/S in a channel whose frequency f satisfies $f_h - nB \leqq f \leqq B$. When $0 \leqq f < f_h - nB$, N/S is of order $(D/B)^{4n+4}$ and the formulas corresponding to (52) do not appear to be known. However, comparison with Monte Carlo values plotted by Grierson and McGee[3] indicates that replacing n by $n + 1$ in (52) [n still given by $nB < f_h < (n + 1)B$] gives an expression for N/S which is not greatly in error when f is in $0 < f \leqq f_h - nB$. The simplest instance of (52) holds for $n = 1$, $B < f_h < 2B$, and f in the range $f_h - B \leqq f \leqq B$:

$$N/S = \frac{1}{4} \left(\frac{3}{2} \frac{D^2}{B^2} \right)^2 \left[\left(2 - \frac{f_h}{B} \right)^2 - \left(1 - \frac{f}{B} \right)^2 \right] + 0(D^6/B^6). \tag{6}$$

The explicit part of (6) decreases to zero as f decreases from B to $f_h - B$. For $0 \leqq f < f_h - B$, N/S is $0(D^8/B^8)$.

* I am indebted to a reviewer for the observation that the presence of the factor $n!^{-4}$ in C_{on} and the behavior of the curves in Fig. 1 strongly suggest that the formulas give useful results subject only to D/f_h (instead of the more restrictive D/B) being small.

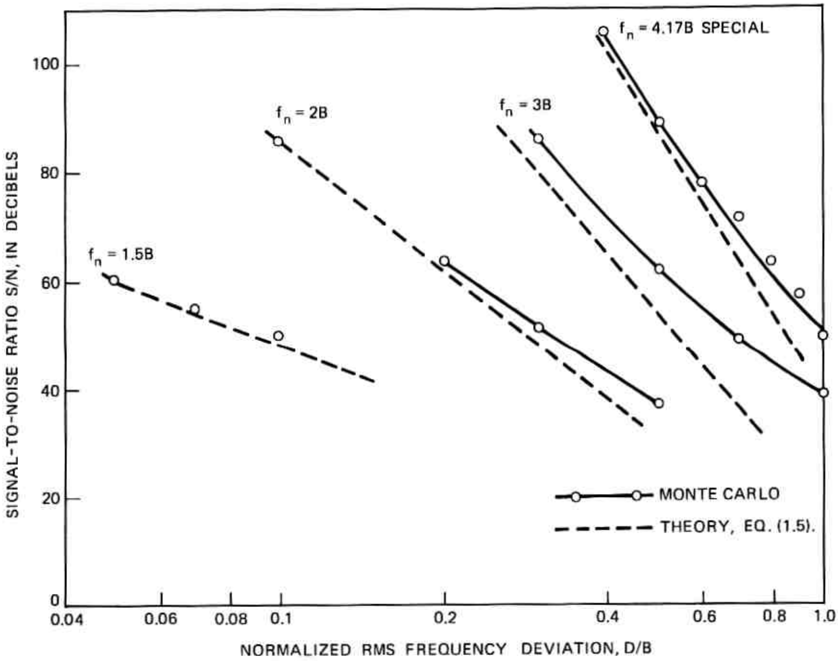


Fig. 1—Signal-to-noise ratio in top channel. The dashed lines show eq. (5) for flat baseband phase modulation. The Monte Carlo curve $4.17B$ is for flat baseband PM, and the curves $1.5B$, $2B$, and $3B$ are for the typical preemphasis used by Anuff and Liou.

It turns out that the explicit portions of (5) and (52) are obtained by considering modulation terms of order $2n + 1$ and of type $\cos 2\pi[(n+1)B - nB]t$.

The curves labeled $f_h = 1.5B$, $2B$, and $3B$ in Fig. 1 have been plotted to compare our eq. (5), based on the flat power spectrum (2) for $W_\varphi(f)$, with the Monte Carlo results given by Anuff and Liou for a typical preemphasis curve. The solid lines and dots show Monte Carlo values of S/N for the top baseband channel. The dashed lines are computed from our (5). It is seen that the slopes agree well for small D/B, but for $f_h = 3B$ a separation of about 6 dB appears. For $f_h = 4B$ (not shown) the separation increases to about 12 dB. Most of the separation appears to be due to the difference between (2) and the $W_\varphi(f)$ used by Anuff and Liou. This is indicated by later Monte Carlo computations made by Anuff for the $W_\varphi(f)$ of (2), and labeled $f_h = 4.17B$ in Fig. 1. There is still a separation of 2 or 3 dB. This may be due to the granularity of the Monte Carlo approximation to $W_\varphi(f)$ and also to the fact that the Monte Carlo filter is not quite ideal.

Section II contains a statement of results from the Volterra series theory needed in our analysis. In Section III, the simplest case, $B < f_h < 2B$, involving third-order modulation terms is discussed in some detail. Section IV and Appendices C and D deal with the general $nB < f_h < (n+1)B$ case. In Section V, formulas are given for the calculation of N/S. Appendices A and B contain material which provides some insight to the general work of Section IV. Appendix A discusses the case $\varphi(t) = A \cos \omega_a t$, and Appendix B treats a simple analog of the FM problem.

All of our work deals with the flat power spectrum $W_\varphi(f)$ defined by (2). The chief obstacle in going to a more general $W_\varphi(f)$ is the evaluation of the multiple integrals which occur in the analysis. Possibly $W_\varphi(f) = Af^\nu$ for $|f| < B$ and $\nu > -1$ could be handled by the procedure used here, but this extension has not been studied seriously.

II. RESULTS NEEDED FROM VOLTERRA SERIES THEORY

Because the carrier frequency f_o is at the center of the ideal filter passband, the even-order modulation products vanish. In the notation of Ref. 4, the Volterra series with the even terms equal to zero is

$$y(t) = \frac{1}{1!}\int_{-\infty}^{\infty} du_1\, g_1(u_1)x(t-u_1)$$
$$+ \frac{1}{3!}\int_{-\infty}^{\infty} du_1 \int_{-\infty}^{\infty} du_2 \int_{-\infty}^{\infty} du_3\, g_3(u_1, u_2, u_3) \prod_{k=1}^{3} x(t-u_k) + \cdots. \quad (7)$$

When $x(t)$ is a stationary, zero-mean Gaussian process with two-sided power spectrum $W_x(f)$, the Mircea-Sinnreich[5] series for the two-sided power spectrum $W_y(f)$ of $y(t)$ becomes [eqs. (14) and (160) of Ref. 4]:

$$W_y(f) = W_x(f)\bigg| G_1(f) + \frac{1}{1!2}\int_{-\infty}^{\infty} df_1' W_x(f_1')G_3(f, f_1', -f_1')$$
$$+ \frac{1}{2!2^2}\int_{-\infty}^{\infty} df_1' \int_{-\infty}^{\infty} df_2' W_x(f_1')W_x(f_2')G_5(f, f_1', -f_1', f_2', -f_2') + \cdots\bigg|^2$$
$$+ \frac{1}{3!}\int_{-\infty}^{\infty} df_1 \int_{-\infty}^{\infty} df_2 W_x(f_1)W_x(f_2)W_x(f - f_1 - f_2)$$
$$\times \bigg| G_3(f_1, f_2, f - f_1 - f_2)$$
$$+ \frac{1}{1!2}\int_{-\infty}^{\infty} df_1' W_x(f_1')G_5(f_1, f_2, f - f_1 - f_2, f_1', -f_1') + \cdots\bigg|^2$$
$$+ \frac{1}{5!}\int_{-\infty}^{\infty} df_1 \int_{-\infty}^{\infty} df_2 \int_{-\infty}^{\infty} df_3 \int_{-\infty}^{\infty} df_4 W_x(f_1)\cdots$$
$$W_x(f_4)W_x(f - f_1 - \cdots - f_4)|G_5(\cdots) + \cdots|^2 + \cdots. \quad (8)$$

Here, $G_m(f_1, f_2, \cdots, f_m)$ is the m-fold Fourier transform of $g_m(t_1, \cdots, t_m)$, i.e., the mth-order transfer function.

We shall need another result which can be derived from the analysis of Section VII of Ref. 4. Let $x(t)$ and $y(t)$ be as in (7) and (8), and let

$$y_L(t) = \int_{-\infty}^{\infty} du_1 \, g_1(u_1) x(t - u_1) \tag{9}$$

be the linear part of $y(t)$. Then the power spectrum of $y(t) - y_L(t)$ is given by

$$W_{y-y_L}(f) = [\text{Series for } W_y(f) \text{ with } G_1(f) \text{ replaced by } 0]. \tag{10}$$

This result can be established by using the series (152) of Ref. 4 for $\langle y(t + \tau) z(t) \rangle$ to evaluate the four ensemble averages appearing in the autocorrelation function of $y(t) - y_L(t)$.

In problems in which $\cos [2\pi f_o t + \varphi(t)]$ enters a filter with transfer function $K(f)$, the normalized transfer function

$$\Gamma(f) = K(f_0 + f)/K(f_0) \tag{11}$$

appears. For the ideal filter of our problem, $\Gamma(f) = 1$ when $-f_h < f < f_h$ and $\Gamma(f) = 0$ when $|f| > f_h$. Furthermore, the power spectrum $W_\theta(f)$ of the output phase angle $\theta(f)$ is given by the expression obtained from (8) by replacing $W_x(f)$ by $W_\varphi(f)$ and $G_1(f_1)$, $G_3(f_1, f_2, f_3), \cdots$ by [Mircea[6] and (52), (71), and (72) of Ref. 4]:

$$\begin{aligned}
G_{\theta 1}(f_1) &= \Gamma(f_1), \\
G_{\theta 3}(f_1, f_2, f_3) &= j^2[\Gamma(f_1 + f_2 + f_3) - \Gamma(f_1)\Gamma(f_2 + f_3) \\
&\quad - \Gamma(f_2)\Gamma(f_1 + f_3) - \Gamma(f_3)\Gamma(f_1 + f_2) \\
&\quad + 2\Gamma(f_1)\Gamma(f_2)\Gamma(f_3)], \\
G_{\theta 5}(f_1, \cdots, f_5) &= j^4[(12345) - 1! \sum_5{}' (1)(2345) - 1! \sum_{10}{}' (12)(345) \\
&\quad + 2! \sum_{10}{}' (1)(2)(345) + 2! \sum_{15}{}' (1)(23)(45) \\
&\quad - 3! \sum_{10}{}' (1)(2)(3)(45) + 4!(1)(2)(3)(4)(5)], \\
&\quad \vdots \\
G_{\theta m}(f_1, \cdots, f_m) &= j^{m-1} \sum_{\ell=1}^{m} (-1)^{\ell-1}(\ell - 1)! \sum_{(\nu;\ell,m)} \sum_N{}' \\
&\quad \times \Gamma(f_1 + \cdots + f_{\nu_1})\Gamma(f_{\nu_1+1} + \cdots + f_{\nu_1+\nu_2}) \cdots \\
&\quad \times \Gamma(f_{m-\nu_\ell+1} + \cdots + f_m).
\end{aligned} \tag{12}$$

The Γ's and f's have been omitted and the subscripts written within parentheses in $G_{\theta 5}$. In $G_{\theta m}$ the summation over ℓ and $(\nu; \ell, m)$ is essentially a summation over the partitions of m, ℓ being the number of parts and $\nu_1, \nu_2, \cdots, \nu_\ell$ the parts:

$$\begin{gathered} \nu_1 + \nu_2 + \cdots + \nu_\ell = m, \\ 1 \leqq \nu_1 \leqq \cdots \leqq \nu_\ell. \end{gathered} \tag{13}$$

The summation $\sum\limits_N{}'$ extends over the N (not to be confused with the N denoting noise power) nonidentical products that can be obtained by permuting the subscripts on the f's. The number of terms in the summation $\sum\limits_N{}'$ is

$$N = m!/\nu_1!\nu_2!\cdots\nu_\ell!r_1!r_2!\cdots r_k! \tag{14}$$

where r_1 is the number of equal ν's in the first run of equalities in the arrangement $\nu_1 \leqq \nu_2 \leqq \cdots \leqq \nu_\ell$, r_2 the number in the second run, etc. When the ν's are unequal, the r's do not appear. A more complete explanation of the notation is given in (24) to (29) of Ref. 4.

In our work, $G_{\theta(2n+1)}$ will be either 0 or -1 when $n \geqq 1$.

When $\varphi(t)$ is bandlimited to $|f| \leqq B$ and f_h exceeds B, the linear portion of $\theta(t)$ is equal to $\varphi(t)$. This can be seen formally by assuming $\varphi(t)$ to have a Fourier transform $F(f)$ which vanishes for $|f| > B$. Then, from (9) and $G_{\theta1}(f) = \Gamma(f) = 1$ for $|f| < f_h$, it follows that

$$\begin{aligned} \theta_L(t) &= \int_{-\infty}^{\infty} du\, g_{\theta1}(u)\, \varphi(t-u) \\ &= \int_{-\infty}^{\infty} df\, G_{\theta1}(f) F(f) e^{i2\pi ft} \\ &= \int_{-B}^{B} df F(f) e^{i2\pi ft} = \varphi(t). \end{aligned} \tag{15}$$

Most of our analysis will consist of using the combination of (8) and (10) to obtain expressions for $W_{\theta-\varphi}(f)$, the power spectrum of the difference $\theta(t) - \varphi(t)$ between the output and input phase angles.

III. $W_{\theta-\varphi}(f)$ WHEN $B < f_h < 2B$

In this section we take $B < f < 2B$, $f_h - B \leqq f \leqq B$, and assume D/B (and consequently W_oB) to be small. The power spectrum of the output phase angle θ is, from (8) with θ in place of y,

$$\begin{aligned} W_\theta(f) = W_\varphi(f) \Bigg| \Gamma(f) & \\ + \frac{1}{1!2} \int_{-B}^{B} df_1' W_\varphi(f_1') G_{\theta3}(f, f_1', -f_1') + 0(W_o^2B^2) \Bigg|^2 & \\ + \frac{1}{3!} \int_{-B}^{B} df_1 \int_{-B}^{B} df_2\, W_\varphi(f_1) W_\varphi(f_2) W_\varphi(f - f_1 - f_2) & \\ \times | G_{\theta3}(f_1, f_2, f - f_1 - f_2) + 0(W_oB) |^2 + 0(W_o^5B^4). & \end{aligned} \tag{16}$$

From (2), $W_\varphi(f_1')$ and $W_\varphi(f_i)$, $i = 1, 2$, can be replaced by W_o in the integrals. However, $W_\varphi(f - f_1 - f_2)$ will be retained for the present because it serves to make the integral vanish when $|f - f_1 - f_2| > B$. For completeness, we shall carry the first line in (16) along in the analysis even though it will vanish when we calculate the crosstalk noise in an idle channel represented by a slot in W_φ at $(f, f + \Delta f)$.

Since the linear portion of $\theta(t)$ is equal to $\varphi(t)$, the power spectrum $W_{\theta-\varphi}(f)$ of $\theta(t) - \varphi(t)$ is given by (16) with $\Gamma(f)$ in the first line replaced by zero:

$$W_{\theta-\varphi}(f) = W_\varphi(f)\left|\frac{1}{1!}\int_o^B df_1' W_o G_{\theta 3}(f, f_1', -f_1')\right|^2$$
$$+ \frac{1}{3!}\int_{-B}^{B} df_1 \int_{-B}^{B} df_2 W_o^2 W_\varphi(f - f_1 - f_2)|G_{\theta 3}(f_1, f_2, f - f_1 - f_2)|^2$$
$$+ 0(W_o^4 B^3). \quad (17)$$

In obtaining (17), we have used the fact that the integrand in the first line is an even function of f_1'.

Examination of (17) shows that the dominant terms in $W_{\theta-\varphi}(f)$ are $0(W_o^3 B^2)$ and hence correspond to third-order modulation. When f does not lie in an idle channel (i.e., $W_\varphi(f) \neq 0$), some of the third-order terms in $W_\theta(f)$ arise from the cross term $\Gamma(f)0(W_o^2 B^2)$ which requires a knowledge of $G_{\theta 5}$ for its evaluation. For this reason, we prefer to deal with $W_{\theta-\varphi}(f)$ [instead of $W_\theta(f)$] which requires only $G_{\theta 3}$ for the calculation of all its third-order terms.

When $0 \leqq f \leqq B$ and $0 \leqq f_1' \leqq B$, as in (17), all of the Γ's in

$$G_{\theta 3}(f, f_1', -f_1') = -\Gamma(f) + \Gamma(f)\Gamma(0) + \Gamma(f_1')\Gamma(f - f_1')$$
$$+ \Gamma(-f_1')\Gamma(f + f_1') - 2\Gamma(f)\Gamma(f_1')\Gamma(-f_1') \quad (18)$$

are unity except possibly $\Gamma(f + f_1')$ which is unity if $f + f_1' < f_h$ and zero if $f + f_1' > f_h$. Hence, $G_{\theta 3}(f, f_1', -f_1')$ is zero if $f_1' < f_h - f$ and is -1 if $f_h - f < f_1'$. It follows that

$$\frac{1}{1!}\int_o^B df_1' W_o G_{\theta 3}(f, f_1', -f_1')$$
$$= \begin{cases} -(B - f_h + f)W_o, & f \geqq f_h - B \\ 0, & f \leqq f_h - B. \end{cases} \quad (19)$$

The function $W_\varphi(f - f_1 - f_2)$ vanishes for $|f - f_1 - f_2| > B$, and the function

$$G_{\theta 3}(f_1, f_2, f - f_1 - f_2)$$
$$= -\Gamma(f) + \Gamma(f_1)\Gamma(f - f_1) + \Gamma(f_2)\Gamma(f - f_2)$$
$$+ \Gamma(f - f_1 - f_2)\Gamma(f_1 + f_2) - 2\Gamma(f_1)\Gamma(f_2)\Gamma(f - f_1 - f_2) \quad (20)$$

vanishes in part of the square $f_1 = \pm B$, $f_2 = \pm B$. The result is that, as will be shown, the region of integration for the double integral in (17) reduces to the shaded areas shown in Fig. 2. In Fig. 2, it is assumed that $f_h - B \leqq f \leqq B$. When $0 \leqq f \leqq f_h - B$, the double integral in (17) is zero because $G_{\theta 3}$ is zero.

In the present case, $B < f_h < 2B$, it is convenient to set

$$f_3 = f - f_1 - f_2 \tag{21}$$

so that the lines $f_3 = \pm B$, or $f_1 + f_2 = f \pm B$, mark boundaries outside of which $W_\varphi(f - f_1 - f_2)$ is zero. Equation (21) also enables us to write the boundaries $f_1 = f - f_h$ and $f_2 = f - f_h$ as $f_2 + f_3 = f_h$ and $f_1 + f_3 = f_h$, respectively, as shown in Fig. 2.

The expression (20) for $G_{\theta 3}(f_1, f_2, f - f_1 - f_2)$ is equal to -1 in the shaded areas of Fig. 2. This follows from the fact that all of the Γ's in (20) are unity except possibly $\Gamma(f - f_1)$, $\Gamma(f - f_2)$, and $\Gamma(f_1 + f_2)$, which are 0 when their arguments exceed f_h. The possibilities $f - f_1 < -f_h$ and $f - f_2 < -f_h$ are ruled out because $f > 0$, and $f_1 + f_2 < -f_h$ is discarded because it makes $f_3 > B$. Performing the integration over the shaded areas in Fig. 2 is equivalent to adding

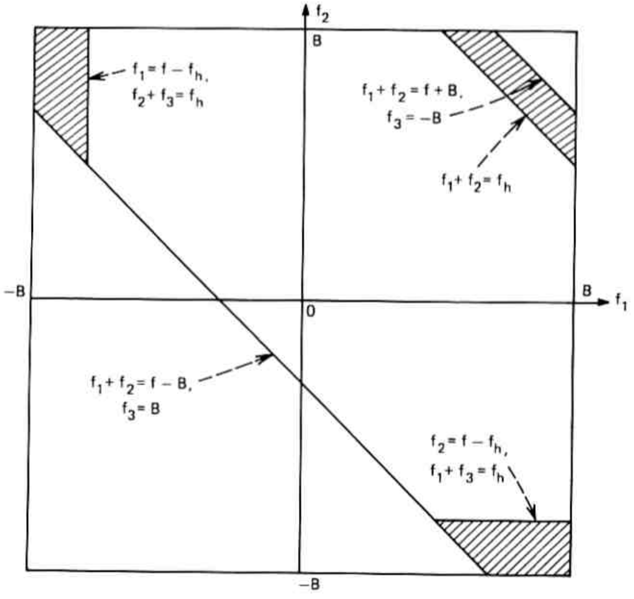


Fig. 2—The three areas of integration for the double integral in eq. (17) for $W_{\theta - \varphi}(f)$.

the areas and gives

$$\frac{1}{3!}\int_{-B}^{B} df_1 \int_{-B}^{B} df_2 W_o^2 W_\varphi(f - f_1 - f_2)\,|G_{\theta 3}(f_1, f_2, f - f_1 - f_2)|^2 = \tfrac{1}{4}[(2B - f_h)^2 - (B - f)^2]W_o^3 \quad (22)$$

when $f_h - B \leqq f \leqq B$. As mentioned earlier, the double integral vanishes when $f \leqq f_h - B$.

The main result of this section is obtained by substituting the values (19) and (22) of the integrals in the expression (17) for $W_{\theta-\varphi}(f)$:

$$W_{\theta-\varphi}(f) = W_\varphi(f)(B + f - f_h)^2 W_o^2 + \tfrac{1}{4}W_o^3[(2B - f_h)^2 - (B - f)^2] + 0(W_o^4B^3) \quad (23)$$

where f_h and f satisfy $B < f_h < 2B$ and $f_h - B \leqq f \leqq B$, respectively. When $0 \leqq f \leqq f_h - B$, the $G_{\theta 3}$'s are zero in the corresponding ranges of integration and it follows from (16) (with $\Gamma(f)$ replaced by zero) that

$$W_{\theta-\varphi}(f) = 0(W_o^5B^4). \quad (24)$$

It also appears that the third-order part of $W_{\theta-\varphi}(f)$ is constant when $B < f < f_h$.

Although it may not be obvious in Fig. 2, the areas of the three shaded regions are equal, and each contributes the same amount to $W_{\theta-\varphi}(f)$. There is an underlying symmetry which becomes evident when the boundaries of the three regions are written as follows:

$$\begin{array}{lll} f_1 = B & f_2 = B & f_3 = B \\ f_2 = B & f_3 = B & f_1 = B \\ f_3 = -B & f_1 = -B & f_2 = -B \\ f_1 + f_2 = f_h & f_2 + f_3 = f_h & f_3 + f_1 = f_h. \end{array} \quad (25)$$

Furthermore, the double integral in (17) can be written as

$$\frac{1}{3!}\int df_1 \int df_2 \int df_3\, W_\varphi(f_1)W_\varphi(f_2)W_\varphi(f_3)\delta(f - f_1 - f_2 - f_3) \times |G_{\theta 3}(f_1, f_2, f_3)|^2 \quad (26)$$

where, replacing $\delta(x)$ by the limit as $\epsilon \rightarrow 0$ of $h(x) = 1/\epsilon$ for $|x| < \epsilon/2$ and $h(x) = 0$ for $|x| > \epsilon/2$, the integration extends over three portions of a three-dimensional slab bounded by the planes $f_1 + f_2 + f_3 = f \pm \epsilon/2$. The three portions are cut out of the slab by the planes defined by eqs. (25). When the integration is accomplished by integrating with respect to f_3 first (the thickness of the slab is $\epsilon/3^{\frac{1}{2}}$ and f_3 is integrated over a length ϵ), the areas of integration for f_1 and f_2 are those shown in Fig. 2.

Thus the twofold integral is equal to the sum of three equal contributions where each contribution can be regarded as arising from a region near one of the corners of a three-dimensional cube. It turns out that the corresponding $2n$-fold integral encountered later is equal to the sum of $(2n+1)!/n!(n+1)!$ contributions arising from regions near $(2n+1)!/n!(n+1)!$ of the 2^{2n+1} corners of a $(2n+1)$-dimensional cube. The corners are those whose $(2n+1)$ coordinates consist of $(n+1)$ plus B's and n minus B's.

IV. $W_{\theta-\varphi}(f)$ WHEN $nB < f_h < (n+1)B$

For f_h and f such that $nB < f_h < (n+1)B$, $n = 1, 2, \cdots$, and $f_h - nB \leqq f \leqq B$, the dominant terms in $W_{\theta-\varphi}(f)$ are given by

$$\begin{aligned} W_{\theta-\varphi}(f) = {} & W_{\varphi}(f) \Bigg| \frac{W_o^n}{n!} \int_o^B df_1' \cdots \int_o^B df_n' \\ & \times G_{\theta(2n+1)}(f, f_1', -f_1', \cdots, f_n', -f_n') \Bigg|^2 \\ & + \sum_{k=1}^{n} \frac{W_o^{2k}}{(2k+1)!} \int_{-B}^{B} df_1 \cdots \int_{-B}^{B} df_{2k} \, W_{\varphi}(f - f_1 - \cdots - f_{2k}) \\ & \times \Bigg| \frac{W_o^{n-k}}{(n-k)!} \int_o^B df_1' \cdots \int_o^B df_{n-k}' \, G_{\theta(2n+1)}(f_1, \cdots, f_{2k}, f - f_1 \\ & - \cdots - f_{2k}, f_1', -f_1', \cdots, f_{n-k}', -f_{n-k}') \Bigg|^2 + 0(W_o^{2n+2} B^{2n+1}) \end{aligned} \tag{27}$$

where for $k = n$ it is understood that the quantity within the absolute value signs becomes $G_{\theta(2n+1)}(f_1, \cdots, f_{2n}, f - f_1 - \cdots f_{2n})$. No $G_{\theta m}$ for $m < 2n+1$ appears in (27) because, from Appendix D, all such terms vanish over the region of integration.

To aid in the evaluation of the integrals which arise in dealing with (27) we shall use[7]

$$\int dx_1 \cdots \int dx_m \, H(\sigma_m) = \frac{1}{(m-1)!} \int_K^L H(z) z^{m-1} dz \qquad K \leqq \sigma_m \leqq L \tag{28}$$

where $K \geqq 0$, $\sigma_m = x_1 + x_2 + \cdots + x_m$, and the integration on the left extends over the region specified by $x_i \geqq 0$, $i = 1, 2, \cdots, m$ and $K \leqq \sigma_m \leqq L$. The integrations with respect to the f_i's in our problem extend over regions where f_i is near $+B$ or $-B$; and we shall use (28) by making the change of variable $f_i = B - x_i$ or $f_i = -B + x_i$.

The $G_{\theta(2n+1)}$ in the second line of (27) is different from 0 (and, from Appendix D, equal to -1) only if

$$f + f_1' + \cdots + f_n' > f_h. \tag{29}$$

Setting $f_i' = B - x_i$, $i = 1, 2, \cdots, n$ carries this inequality into

$$x_1 + x_2 + \cdots + x_n < f + nB - f_h = P - Q \tag{30}$$

where we have introduced the parameters

$$\begin{aligned} P &= (n+1)B - f_h \\ Q &= B - f \end{aligned} \tag{31}$$

and have assumed $P > Q$. When $P < Q$, the inequality (29) and its analogues for the other terms in (27) cannot be satisfied. Consequently, all the $G_{\theta(2n+1)}$'s are zero and all the modulation terms of order $(2n+1)$ vanish from (27) when $P < Q$.

From (28) with $m = n$, $K = 0$, $L = P - Q$, and $H(z) = -1$, we get

$$W_\varphi(f)\left|\frac{W_o^n}{n!(n-1)!}\int_o^{P-Q}(-1)z^{n-1}dz\right|^2 = W_\varphi(f)W_o^{2n}(P-Q)^{2n}/n!^4 \tag{32}$$

for the first term on the right in (27).

Now consider the kth term in the sum in (27). For $G_{\theta(2n+1)}$ to be -1 instead of 0, the sum of $(n+1)$ of its arguments must exceed f_h (Appendix D). It can be shown that $(n-k)$ of the arguments must be $f_1', \cdots, f_{n-k}'$ and that the remaining $(k+1)$ arguments come from the set of $2k+1$ elements

$$f_1,\ f_2,\ \cdots,\ f_{2k},\ f - f_1 - \cdots - f_{2k}. \tag{33}$$

There are $(2k+1)!/(k+1)!k!$ different choices of $(k+1)$ items from the set (33). Let $f_1, f_2, \cdots, f_{k+1}$ represent the typical choice and

$$f_1 + f_2 + \cdots + f_{k+1} + f_1' + f_2' + \cdots + f_{n-k}' \tag{34}$$

be the typical sum of $(n+1)$ elements of $G_{\theta(2n+1)}$ which exceeds f_h. Each sum is associated with a region of integration, one boundary of which is obtained by setting (34) equal to f_h. For $k = 1$, there are three regions and, after the integrations with respect to the f_i''s have been performed, the regions become the ones shown in Fig. 2 with f_h replaced by $f_h - (n-1)B$. For k arbitrary, the regions correspond to the corners of a $(2k+1)$-dimensional cube, the corner coordinates consisting of $(k+1)$ plus B's and k minus B's. By virtue of the type of symmetry shown by (25) and (26) for the case $B < f_h < 2B$, each of the $(2k+1)!/(k+1)!$k! regions contributes the same amount to the kth term in (27).

The first step in evaluating the kth term $(k < n)$ is to perform the integrations with respect to $f_1', \cdots, f_{n-k}'$. Suppose that the values of the typical choice $f_1, \cdots, f_{k+1}$ are given. Then for $G_{\theta(2n+1)}$ to be equal to -1, it is necessary that

$$f_1' + \cdots + f_{n-k}' > f_h - f_1 - \cdots - f_{k+1}. \tag{35}$$

Setting $f_i' = B - x_i$, $i = 1, 2, \cdots, n - k$ carries (35) into

$$x_1 + x_2 + \cdots + x_{n-k} < (n - k)B \\ - f_h + f_1 + f_2 + \cdots + f_{k+1}. \quad (36)$$

Using (28) with $m = n - k$, $K = 0$, $H(o) = -1$, and L equal to the right side of (36) shows that the quantity inside the absolute value signs in the kth term is equal to

$$\frac{W_o^{n-k}}{(n-k)!(n-k-1)!}\int_0^L (-1)z^{n-k-1}dz \\ = -W_o^{n-k}L^{n-k}/(n-k)!^2 \quad (37)$$

where $L \geqq 0$.

Next, we integrate with respect to $f_1, f_2, \cdots, f_{k+1}$. The restriction that the right side of (36) be positive gives

$$f_1 + f_2 + \cdots + f_{k+1} > f_h - (n - k)B \quad (38)$$

and the fact that the argument of $W_\varphi(f - f_1 - \cdots - f_{2k})$ must exceed $-B$ gives

$$f_1 + f_2 + \cdots + f_{k+1} < B + f - f_{k+2} - \cdots - f_{2k}. \quad (39)$$

Setting $f_i = B - x_i$ for $i = 1, 2, \cdots, k + 1$ and $f_i = -B + x_i$ for $i = k + 2, \cdots, 2k$ carries (38), (39), and the L in (37) into

$$x_1 + x_2 + \cdots + x_{k+1} < (n + 1)B - f_h = P \\ x_i + x_2 + \cdots + x_{k+1} > B - f + x_{k+2} + \cdots + x_{2k} \\ = Q + x_{k+2} + \cdots + x_{2k} \quad (40) \\ L - P - x_1 - x_2 - \cdots\ x_{k+1}.$$

At this stage, the kth term in (27) is, for $k > 1$,

$$\frac{W_o^{2k}}{(2k+1)!}\,\frac{(2k+1)!}{(k+1)!k!}\int dx_{k+2}\cdots\int dx_{2k}\int dx_1\cdots\int dx_{k+1}\,W_o \\ \times W_o^{2n-2k}(P - x_1 - \cdots - x_{k+1})^{2n-2k}/(n-k)!^4 \quad (41)$$

where $(2k + 1)!/(k + 1)!k!$ is the number of equally contributing regions of integration. For fixed $x_{k+2}, \cdots, x_{2k}$, the integration with respect to $x_1, \cdots, x_{k+1}$ can be performed by using (28) with $m = k + 1$, $K = Q + x_{k+2} + \cdots + x_{2k}$, $L = P$, and $H(z) = (P - z)^{2n-2k}$. Expression (41) becomes

$$\frac{W_o^{2n+1}}{(k+1)!k!(n-k)!^4}\int dx_{k+2}\cdots\int dx_{2k} \\ \times \frac{1}{k!}\int_{Q+x_{k+2}+\cdots+x_{2k}}^{P}(P - z)^{2n-2k}z^k dz. \quad (42)$$

The integration in (42) extends over the region defined by $x_i \geqq 0$, $i = k + 2, \cdots, 2k$, and the inequality obtained by combining the two inequalities in (40):

$$x_{k+2} + \cdots + x_{2k} < P - Q. \tag{43}$$

Using (28) with $m = k - 1$, $K = 0$, $L = P - Q$, and

$$H(z) = \int_{Q+z}^{P} (P - y)^{2n-2k} y^k dy \tag{44}$$

leads to a double integral which can be reduced to a single integral by reversing the order of integration:

$$\begin{aligned}\frac{1}{(k-2)!} \int_{o}^{P-Q} H(z) z^{k-2} dz &= \frac{1}{(k-1)!} \int_{Q}^{P} y^k (P-y)^{2n-2k} (y - Q)^{k-1} dy \\ &= k \sum_{\ell=0}^{k} \frac{(2n-2k)!(2k-\ell-1)!}{\ell!(k-\ell)!(2n-\ell)!} Q^{\ell} (P-Q)^{2n-\ell}.\end{aligned} \tag{45}$$

When (45) is used in the expression (42) for the kth term in $W_{\theta-\varphi}(f)$, (42) becomes

$$\frac{W_o^{2n+1}(2n-2k)!}{(k+1)!k!(k-1)!(n-k)!^4} \sum_{\ell=0}^{k} \frac{(2k-\ell-1)! Q^{\ell}(P-Q)^{2n-\ell}}{\ell!(k-\ell)!(2n-\ell)!}. \tag{46}$$

It can be verified that (46) also holds for $k = 1$, even though $k > 1$ was assumed in the derivation. Adding the expression (32) to the sum of (46) from $k = 1$ to n and interchanging the order of summation gives the equation sought in this section:

$$\begin{aligned}W_{\theta-\varphi}(f) = W_{\varphi}(f) W_o^{2n} (P-Q)^{2n} n!^{-4} + W_o^{2n+1} \sum_{\ell=0}^{n} C_{\ell n} Q^{\ell} (P-Q)^{2n-\ell} \\ + 0(W_o^{2n+2} B^{2n+1})\end{aligned} \tag{47}$$

where n is a positive integer, $nB < f_h < (n+1)B$, $P > Q$, P and Q are given by (31), and

$$C_{\ell n} = \frac{1}{\ell!(2n-\ell)!} \sum_{k=\max(1,\ell)}^{n} \frac{(2k-\ell-1)!(2n-2k)!}{(k-\ell)!(k-1)!k!(k+1)!(n-k)!^4}. \tag{48}$$

When $P < Q$, i.e., $f < f_h - nB$, our analysis tells us only that $W_{\theta-\varphi}(f)$ is $0(W_o^{2n+3} B^{2n+2})$.

V. THE NOISE TO SIGNAL RATIO N/S

According to (4) the average signal power (FM) in the channel $(f, f + \Delta f)$ when it is busy is

$$S = (2\pi f)^2 (2W_o) \Delta f. \tag{49}$$

Likewise, the average interchannel interference noise power is

$$N = (2\pi f)^2[2W_{\theta-\varphi}(f)]\Delta f \tag{50}$$

and hence

$$N/S = W_{\theta-\varphi}(f)/W_o. \tag{51}$$

When the channel is idle, $W_\varphi(f)$ is 0 in $(f, f + \Delta f)$, and if all of the other channels are busy, (51) and (47) give

$$\begin{aligned} N/S &= W_o^{2n} \sum_{\ell=0}^{n} C_{\ell n} Q^\ell (P - Q)^{2n-\ell} + 0[(W_o B)^{2n+1}] \\ &= \left[\frac{3}{2}\frac{D^2}{B^2}\right]^{2n} \sum_{\ell=0}^{n} C_{\ell n}\left(1 - \frac{f}{B}\right)^\ell \left(n + \frac{f}{B} - \frac{f_h}{B}\right)^{2n-\ell} \\ &\qquad + 0[(D/B)^{4n+2}] \end{aligned} \tag{52}$$

provided $nB < f_h < (n + 1)B$, $f_h - nB \leqq f \leqq B$, and D/B is small. In going to the second line, we have used $W_o B = 3D^2/(2B^2)$ from (3) and the definitions (31) of P and Q. Equations (5) and (6) given as examples in Section I are obtained by setting $f = B$ and $n = 1$, respectively, in (52).

The first few values of $C_{\ell n} \times 10^n$ are listed below.

	$\ell = 0$	1	2	3	4
$n = 1$	2.5	5.0			
2	5.208	19.44	2.083		
3	1.832	9.375	3.906	0.193	
4	0.1994	1.226	1.182	0.2122	0.0060

VI. ACKNOWLEDGMENTS

I want to express my thanks to A. A. Anuff for his suggestions and helpful discussions regarding the comparison of the theory with the Monte Carlo computations. Furthermore, his Monte Carlo computation of the interchannel interference for the $W_\varphi(f)$ used in this paper gave a very welcome confirmation of the theoretical results. I am also indebted to Miss Judith Seery for computing values of N/S from the rather complicated formulas.

APPENDIX A

Sinusoidal Modulation

Some idea of how the FM distortion depends upon the radio frequency bandwidth when the deviation ratio, say A, is small can be obtained from the case $\varphi(t) = A \cos \omega_a t$, $\omega_a = 2\pi f_a$. The carrier frequency and ideal filter are the same as in Section I, but now n is such that $nf_a < f_h < (n+1)f_a$.

The input to the ideal filter is the real part of

$$\exp\left[j\omega_o t + j\varphi(t)\right] = \sum_{m=-\infty}^{\infty} c_m \exp\left(j\omega_o t + jm\omega_a t\right) \tag{53}$$

where $c_m = j^m J_m(A)$, $J_m(A)$ is the Bessel function of order m, and

$$\exp\left(jA\cos\omega_a t\right) = \sum_{m=-\infty}^{\infty} j^m J_m(A)\exp\left(jm\omega_a t\right). \tag{54}$$

Since $2f_h/f_o \ll 1$, the filter output is nearly equal to the real part of

$$\exp\left[j\omega_o t + j\theta(t)\right] = \sum_{m=-n}^{n} c_m \exp\left(j\omega_o t + jm\omega_a t\right). \tag{55}$$

Subtracting (54) from (55) and dividing by $\exp\left[j\omega_o t + j\varphi(t)\right]$ gives

$$e^{j\theta - j\varphi} = 1 - e^{-j\varphi}\left(\sum_{-\infty}^{-n-1} + \sum_{n+1}^{\infty}\right) c_m \exp\left(jm\omega_a t\right) \tag{56}$$

where the argument t has been omitted in $\theta(t)$ and $\varphi(t)$.

Replacing $\exp(-j\varphi)$ on the right by its series obtained from (53) and taking logarithms give the known first-order approximation

$$\theta - \varphi = -Im\left\{\sum_{\ell=-\infty}^{\infty}\left(\sum_{m=-\infty}^{-n-1} + \sum_{m=n+1}^{\infty}\right) c_\ell^* c_m \exp\left[j(m-\ell)\omega_a t\right]\right\} + \text{terms of order } \left[\sum_{\ell}\left(\sum_{m} + \sum_{m}\right)\right]^2. \tag{57}$$

Since A is small and $c_m = j^m J_m(A)$, we have for $m \geqq 0$

$$c_m = c_{-m} = (jA/2)^m/m! + 0(A^{m+2}). \tag{58}$$

The interchannel interference in a microwave system corresponds to the $\exp(j\omega_a t)$ and $\exp(-j\omega_a t)$ components in the expression (57) for $\theta - \varphi$. The largest contribution to these components comes from the values $m = n+1$, $\ell = -n$, and $m = -n-1$, $\ell = n$, respectively:

$$\begin{aligned}(\theta - \varphi)_{\omega_a} &= -Im\left[j^{-n+n+1}e^{j\omega_a t} + j^{-n+n+1}e^{-j\omega_a t}\right]\frac{(A/2)^{2n+1}}{n!(n+1)!} \\ &\qquad + 0(A^{2n+2}) \\ &= -\frac{2(A/2)^{2n+1}}{n!(n+1)!}\cos\omega_a t + 0(A^{2n+2}).\end{aligned} \tag{59}$$

It follows that the average power in the $\cos\omega_a t$ component of $\theta - \varphi$ is

$$P(A) = \frac{2(A/2)^{4n+2}}{n!^2(n+1)!^2} + 0(A^{4n+3}) \tag{60}$$

and dividing by the average power $A^2/2$ in $\varphi(t) = A \cos \omega_a t$ gives

$$\frac{P(A)}{\text{ave. power in } \varphi} = \frac{(A/2)^{4n}}{n!^2(n+1)!^2} + 0(A^{4n+1}). \tag{61}$$

If, instead of the pure sinusoidal signal $A \cos \omega_a t$, the signal $\varphi(t)$ were a very narrow band of Gaussian noise centered at the frequency f_a, its envelope R would fluctuate slowly according to a Rayleigh probability density

$$p(R) = \sigma^{-2} R \exp\left(-R^2/2\sigma^2\right) \tag{62}$$

where σ^2 is the average power in $\varphi(t)$. Replacing A in $P(A)$ by R and averaging with the help of (62) shows that the average of the total power in the components of $\theta - \varphi$ clustered around f_a is

$$\begin{aligned} N &= \int_o^\infty P(R)p(R)dR \\ &= \frac{(2n+1)!\sigma^{4n+2}}{n!^2(n+1)!^2\, 2^{2n}} + 0(\sigma^{4n+3}). \end{aligned} \tag{63}$$

This expression for N can be checked by using $W_x(f) = \sigma^2\delta(|f| - f_a)/2$ in place of $W_x(f) = W_o$, $|f| < B$, in the analysis of Sections II to V. Division by the average power $S = \sigma^2$ in $\varphi(t)$ gives

$$\frac{N}{S} = \frac{(2n+1)!\sigma^{4n}}{n!^2(n+1)!^2\, 2^{2n}} + 0(\sigma^{4n+1}). \tag{64}$$

In $\varphi(t) = A \cos \omega_a t$, A is the deviation ratio and in (64) σ is the rms frequency deviation ratio. The fact that N/S varies as σ^{4n} in (64) agrees with the case in which $\varphi(t)$ has a flat spectrum. However, (64) is larger by roughly the factor $(2n + 1)!$

APPENDIX B

Simple Analogue of Relation Between φ and θ

The relation between the FM input φ and output θ is somewhat similar to the relation between x and y given by

$$y = x + \frac{a}{(2n+1)!} x^{2n+1} \tag{65}$$

where a is real and x is a stationary, zero-mean Gaussian process with two-sided spectrum $W_x(f)$ and autocorrelation function $R_\tau \equiv R(\tau) = \langle x(t+\tau)x(t)\rangle$. We are given $W_x(f)$ and want to find $W_y(f)$ and $W_{y-x}(f)$.

Our aim here is to obtain some insight regarding the origin of the various terms in the series (17) and (27) for $W_{\theta-\varphi}(f)$.

From Volterra series theory, taking (65) to be the series, we get $G_1(f_1) = 1$, $G_{2n+1}(f_1, f_2, \cdots, f_{2n+1}) = a$, and $G_m = 0$ for all other values of m[Ref. 4, (22), (23)]. The Mircea-Sinnreich series [Ref. 4, (156), (160)] for $W_y(f)$ becomes

$$
\begin{aligned}
W_y(f) = {} & W_x(f)\left|1 + \frac{1}{n!2^n}\int_{-\infty}^{\infty} df'_1 \cdots \int_{-\infty}^{\infty} df'_n\, W_x(f'_1)\cdots W_x(f'_n)a\right|^2 \\
& + \frac{1}{3!}\int_{-\infty}^{\infty} df_1 \int_{-\infty}^{\infty} df_2\, W_x(f_1)W_x(f_2)W_x(f - f_1 - f_2) \\
& \times \left|\frac{1}{(n-1)!2^{n-1}}\int_{-\infty}^{\infty} df'_1 \cdots \int_{-\infty}^{\infty} df'_{n-1}\, W_x(f'_1)\cdots W_x(f'_{n-1})a\right|^2 \\
& + \cdots \\
& + \frac{1}{(2n+1)!}\int_{-\infty}^{\infty} df_1 \cdots \int_{-\infty}^{\infty} df_{2n}\, W_x(1)\cdots W_x(f_{2n}) \\
& \qquad \times W_x(f - f_1 - f_2 - \cdots - f_{2n})|a|^2. \qquad (66)
\end{aligned}
$$

According to (10), the power spectrum $W_{y-x}(f)$ of $y - x$ is equal to the expression obtained by replacing the 1 [i.e., $G_1(f)$] by zero in the first line of eq. (66) for $W_y(f)$.

In this particular example, $W_y(f)$ can be obtained as the Fourier transform of the autocorrelation function $\langle y(t)y(t+\tau)\rangle$. Let $y(t)$, $y(t+\tau)$, $x(t)$, $x(t+\tau)$ be denoted by y_1, y_2, x_1, x_2, respectively. Then

$$
\begin{aligned}
\langle y_1y_2\rangle = \langle x_1x_2\rangle + \frac{a}{(2n+1)!}\left[\langle x_1x_2^{2n+1}\rangle + \langle x_1^{2n+1}x_2\rangle\right] \\
+ \frac{a^2}{(2n+1)!^2}\langle x_1^{2n+1}x_2^{2n+1}\rangle. \qquad (67)
\end{aligned}
$$

From

$$\langle \exp\,(jux_1 + jvx_2)\rangle = \exp\left[-2^{-1}(u^2+v^2)R_0 - uvR_\tau\right]$$

we have the known results

$$
\begin{aligned}
\langle x_1x_2^{2n+1}\rangle = \langle x_1^{2n+1}x_2\rangle &= (2n+1)!R_\tau R_o^n/(n!2^n) \\
\langle x_1^{2n+1}x_2^{2n+1}\rangle &= \sum_{k=0}^{n}\frac{(2n+1)!^2R_\tau^{2k+1}(R_o/2)^{2n-2k}}{(2k+1)!(n-k)!^2} \qquad (68)
\end{aligned}
$$

Substituting in (67) and taking the Fourier transform:

$$
\begin{aligned}
W_y(f) &= \int_{-\infty}^{\infty} e^{-i2\pi f\tau}\langle y_1y_2\rangle d\tau \\
&= \int_{-\infty}^{\infty} d\tau e^{-i2\pi f\tau}\left[R_\tau + \frac{2aR_\tau R_o^n}{n!2^n} + \frac{a^2R_\tau(R_o/2)^{2n}}{n!^2}\right. \\
&\qquad \left. + \sum_{k=1}^{n}\frac{a^2R_\tau^{2k+1}(R_o/2)^{2n-2k}}{(2k+1)!(n-k)!^2}\right]. \qquad (69)
\end{aligned}
$$

The point being made in this appendix is that the terms in (69) have [after using $1 + 2\alpha + \alpha^2 = (1 + \alpha)^2$] a one-to-one correspondence with the terms in the Mircea-Sinnreich series (66). This can be seen with the help of

$$R_o = \int_{-\infty}^{\infty} W_x(f_i')df_i',$$

$$\int_{-\infty}^{\infty} e^{-i2\pi f\tau}R_\tau^m d\tau = \int_{-\infty}^{\infty} df_1 \cdots \int_{-\infty}^{\infty} df_{m-1}W_x(f_1)\cdots W_x(f_{m-1}) \times W_x(f - f_1 - \cdots - f_{m-1}). \quad (70)$$

APPENDIX C

Inequalities for Sums of Frequencies

Let $f_1, f_2, \cdots, f_{2n+1}$ denote the $(2n + 1)$ arguments of any one of the $G_{\theta(2n+1)}$'s occurring in the expression (27) for $W_{\theta-\varphi}(f)$. They satisfy the relations

$$\begin{aligned} |f_i| &\leqq B, \qquad i = 1, 2, \cdots, 2n + 1 \\ f_1 + f_2 + \cdots + f_{2n+1} &= f \end{aligned} \quad (71)$$

where f satisfies $0 < f \leqq B$ and is the frequency at which $W_{\theta-\varphi}(f)$ is being evaluated.

We shall call a set of r of the f_i's an "r-tuple" and the sum of the f_i's the "sum" of the r-tuple.

First we show that

$$f - nB \leqq \text{sum of any } (n + 1)\text{-tuple} \leqq f + nB. \quad (72)$$

Let $f_1 + f_2 + \cdots + f_{n+1}$ represent the typical $(n + 1)$-tuple sum. Then (72) follows upon using $|f_i| \leqq B$ on the right side of

$$f_1 + \cdots + f_{n+1} = f - f_{n+2} - \cdots - f_{2n+1}. \quad (73)$$

The next inequality is

$$-nB \leqq \text{sum of any } r\text{-tuple} \leqq nB \quad (74)$$

where $r = 1, 2, \cdots, n, n + 2, \cdots, (2n + 1)$. When $r \leqq n$, (74) follows from $|f_i| \leqq B$, and when $r \geqq n + 2$ it can be proved by using equations similar to (73).

The number of different $(n + 1)$-tuples is $(2n + 1)!/(n + 1)!n!$ If, for given set of values of the f_i's, any one of the $(n + 1)$-tuples, call it A, has a sum greater than nB, the sum of any one of the remaining $(n + 1)$-tuples satisfies

$$f - nB \leqq \text{sum of any } (n + 1)\text{-tuple except } A \leqq nB. \quad (75)$$

The left inequality follows from (72). To obtain the right inequality, note that all $(n + 1)$ elements (the f_i's) in A must be positive. Consider any other $(n + 1)$-tuple, say C. Then A contains k elements $1 \leqq k \leqq n$ which are not in C. Let $f_1, \cdots, f_{n+1}$ represent the elements of C so that the left side of (73) gives the sum of C. Then the right side of (73) contains k elements of A. Since the elements of A are positive, the right side of (73) is less than $f + (n - k)B$, and (75) follows from

$$f + (n - k)B \leqq f + (n - 1)B \leqq nB. \tag{76}$$

APPENDIX D

Values of $G_{\theta(2n+1)}(f_1, \cdots, f_{2n+1})$

The notation used in this appendix is the same as that in Appendix C.

Let $G_{\theta(2n+1)}$ stand for $G_{\theta(2n+1)}(f_1, f_2, \cdots, f_{2n+1})$ where the f_i's satisfy the relations (71). Here we show that

$$G_{\theta(2n+1)} = 0, \qquad f_h > (n + 1)B \tag{77}$$

where $n \geqq 1$ and f_h is the ideal filter semibandwidth. Furthermore, for a given set of $f_1, f_2, \cdots, f_{2n+1}$, it has been shown in Appendix C that there is at most only one $(n + 1)$-tuple, the sum of which exceeds nB. There may be none. When $nB < f_h < (n + 1)B$ with $n \geqq 1$ we shall show that

$$G_{\theta(2n+1)} = -1 \text{ if one } (n + 1)\text{-tuple sum} > f_h, \tag{78}$$

$$G_{\theta(2n+1)} = 0 \text{ if no } (n + 1)\text{-tuple sum} > f_h. \tag{79}$$

The inequalities (72) and (74) show that all of the Γ's in $G_{\theta(2n+1)}$ are unity (*i*) when $f_h > (n + 1)B$ or (*ii*) when no $(n + 1)$-tuple sum exceeds f_h where $nB < f_h < (n + 1)B$. Therefore, to prove (77) and (79), it is sufficient to show that $G_{\theta m}(f_1, \cdots, f_m)$, $m \geqq 2$, is zero when all of the Γ's in its expression (12) are equal to unity.

Consider the sum

$$\sum_{(\nu;\ell,m)} \sum_N{}' \Gamma(f_1 + \cdots + f_{\nu_1}) \cdots \Gamma(f_{m-\nu_\ell+1} + \cdots + f_m). \tag{80}$$

When all of the Γ's = 1, this sum is equal to the number of different ways m different objects $(f_1, f_2, \cdots, f_m)$ can be put in ℓ identical boxes with no box empty (the ℓ pairs of parentheses enclosing the arguments of the ℓ Γ's). From combinatorial theory, this number is $S(\mathrm{m}, \ell)$ the Stirling number of the second kind given by the generating equation[8], for $n \geqq 1$,

$$t^n = \sum_{k=1}^{n} S(n, k)t(t - 1) \cdots (t - k + 1). \tag{81}$$

To illustrate (80), let $m = 4$ and $\ell = 2$. Equations (13) show that the sum over $(\nu;\, \ell,\, m)$ in (80) now extends over the partitions of $m = 4$ which have $\ell = 2$ parts. There are two such partitions: $\nu_1 = 1$, $\nu_2 = 3$ and $\nu_1 = 2$, $\nu_2 = 2$. From (14), the corresponding values of N are $4!/1!3! = 4$ and $4!/2!2!2! = 3$, respectively. Hence the sum (80) is equal to $4 + 3 = 7$. This agrees with the known value $S(4, 2) = 7$. To return to the box problem, the 7 different ways of putting 4 different objects into 2 identical boxes with neither box empty is indicated by

$$\begin{gathered}(1)(234), \qquad (2)(134), \qquad (3)(124), \qquad (4)(123), \\ (12)(34), \qquad (13)(24), \qquad (14)(23).\end{gathered}$$

The expression (12) for $G_{\theta m}$ consists of the sum from $\ell = 1$ to $\ell = m$ of $j^{m-1}(-1)^{\ell-1}(\ell - 1)!$ times the sum (80). When all the Γ's are unity this gives

$$\begin{aligned} G_{\theta m} &= j^{m-1} \sum_{\ell=1}^{m} (-1)^{\ell-1}(\ell - 1)!S(m, \ell) \\ &= \begin{cases} 1, & m = 1 \\ 0, & m > 1 \end{cases} \end{aligned} \tag{82}$$

where the summation is accomplished by dividing (81) by t and then letting $t \to 0$. Setting $m = 2n + 1$ then gives (77) and (79).

Now we turn to (78). Let $f_{n+1} + f_{n+2} + \cdots + f_{2n+1}$ be the single $(n + 1)$-tuple whose sum exceeds f_h. Then $\Gamma(f_{n+1} + \cdots + f_{2n+1}) = 0$ and all the other Γ's in $G_{\theta(2n+1)}$ are unity. The problem is to determine the contribution of all of the terms in $G_{\theta(2n+1)}$ containing $\Gamma(f_{n+1} + \cdots + f_{2n+1})$. Subtracting this contribution from 0 will give the value of $G_{\theta(2n+1)}$.

Setting $m = 2n + 1$ in (12) shows that the terms in $G_{\theta(2n+1)}$ containing $\Gamma(f_{n+1} + \cdots + f_{2n+1})$ as a factor are those for which ℓ and the parts ν_i of the partition of $(2n + 1)$ into ℓ parts are such that

$$\begin{array}{lrl} \ell = 2, & \nu_1 = n, & \nu_2 = n + 1, \\ \ell = 3, & \nu_1 + \nu_2 = n, & \nu_3 = n + 1, \\ & \vdots & \\ \ell = n + 1, & \nu_1 + \cdots + \nu_n = n & \nu_{n+1} = n + 1. \end{array} \tag{83}$$

Therefore, with $k = \ell - 1$, the terms are the product of

$$j^{2n} \sum_{k=1}^{n} (-)^k k! \sum_{(\nu;\, k,\, n)} \sum_{N}{}' \Gamma(f_1 + \cdots + f_{\nu_1}) \cdots \Gamma(f_{n-\nu_k+1} + \cdots + f_n) \tag{84}$$

and $\Gamma(f_{n+1} + \cdots + f_{2n+1})$ where now

$$\begin{gathered} \nu_1 + \nu_2 + \cdots + \nu_k = n \\ \leqq \nu_1 \leqq \nu_2 \leqq \cdots \leqq \nu_k \\ N = n!/\nu_1! \cdots \nu_k! r_1! r_2! \cdots . \end{gathered}$$

When all of the Γ's in (84) are unity, (84) becomes

$$j^{2n} \sum_{k=1}^{n} (-)^k k! S(n, k) = j^{2n}(-1)^n = 1 \tag{85}$$

where the summation is performed by setting $t = -1$ in the generating equation (81). Subtracting the contribution (85) from 0 gives the value $G_{\theta(2n+1)} = -1$ stated in (78).

REFERENCES

1. Rowe, H. E., *Signals and Noise in Communication Systems*, Princeton, N. J.: Van Nostrand, 1965.
2. Anuff, A., and Liou, M. L., "A Note on Necessary Bandwidth in FM Systems," Proc. IEEE, *59* (October 1971), pp. 1522–1533.
3. Grierson, J. K., and McGee, W. F., "Microwave System Intermodulation Simulation," IEEE Int. Conf. Commun.—Conf. Record, Philadelphia, Pa., *4*, June 12–14, 1968, pp. 403–406.
4. 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.
5. Mircea, A., and Sinnreich, H., "Distortion Noise in Frequency-Dependent Nonlinear Networks," Proc. Inst. Elec. Eng., *116* (1969), pp. 1544–1648.
6. Mircea, A., "Harmonic Distortion and Intermodulation Noise in Linear FM Transmission Systems," Rev. Electrotech. Energet. (Romania), *12* (1967), pp. 359–371.
7. Edwards, J., *Treatise on the Integral Calculus*, vol. 2, London: MacMillan and Co., 1922, pp. 160–161.
8. Riordan, J., *An Introduction to Combinatorial Analysis*, New York: John Wiley & Sons, 1957.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

Computing Distortion in Analog FM Communication Systems

By A. J. RAINAL

(Manuscript received November 3, 1972)

This paper describes a method for computing baseband distortion in analog FM communication systems; the method is based on recent theoretical work available in the literature. The input baseband signal is taken to be a zero-mean, stationary Gaussian process having an arbitrary power spectral density. A variety of graphical results are presented in order to demonstrate the utility of this method of computing FM distortion. It is shown that the often-used noise loading test does not necessarily represent a worst-case test.

I. INTRODUCTION

Theoretically, analog FM signals generally possess an infinite bandwidth. Thus, when such signals are passed through a linear system having a finite bandwidth, some FM distortion must occur. The measurement of such distortion is costly and very time consuming. Accordingly, the development of methods for the computation of FM distortion is of practical interest.

The purpose of this paper is to describe how we used the theoretical results derived by A. Mircea,[1] E. Bedrosian[2] and S. O. Rice[3] to develop a computer program to compute the FM distortion resulting from linear time-invariant-filter structures. The input baseband modulation is taken to be a zero-mean, stationary Gaussian process having an arbitrary baseband power spectral density.

II. SERIES EXPANSION UNDERLYING THE COMPUTATION

Consider the analog FM communication system presented in Fig. 1a. The associated mathematical problem for studying FM distortion is illustrated in Fig. 1b. The problem is to deduce the double-sided power spectral density, $W_\theta(f)$, of the output random process, $\theta(t)$, given $\Gamma(f)$ and the double-sided power spectral density, $W_\phi(f)$, of the input Gaussian modulation. Once this mathematical problem is solved satisfactorily, we can then compute the FM distortion at baseband. In the FM case, $W_{\phi'}(f)$, the power spectral density of $\phi'(t)$, is given and $W_{\theta'}(f)$, the power spectral density of $\theta'(t)$, is desired. In the PM case, $W_{\phi}(f)$ is given and $W_{\theta}(f)$ is desired. However, the two problems are closely related because of the following relations:

$$W_{\phi'}(f) = \omega^2 W_{\phi}(f) \tag{1}$$

and

$$W_{\theta'}(f) = \omega^2 W_{\theta}(f). \tag{2}$$

In fact, an FM communication system can be designed by using only PM equipment, as is illustrated in Fig. 1a.

Using Rice's[3] notation, a series expansion of $W_{\theta}(f)$ is given by:

$$\begin{aligned} W_{\theta}(f) = {} & \theta_{\mathrm{dc}}^2 \delta(f) + \tfrac{1}{4} W_{\phi}(f) | U(f) + U^*(-f) |^2 \\ & + \frac{1}{8} \int_{-\infty}^{\infty} d\rho W_{\phi}(\rho) W_{\phi}(f-\rho) | T(\rho, f-\rho) - T^*(-\rho, -f+\rho) |^2 \\ & + \frac{1}{24} \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma W_{\phi}(\rho) W_{\phi}(\sigma) W_{\phi}(f-\rho-\sigma) \\ & \times | S(\rho, \sigma, f-\rho-\sigma) + S^*(-\rho, -\sigma, -f+\rho+\sigma) |^2 \\ & \qquad + 0(\phi^6 W_{\phi}) \end{aligned} \tag{3}$$

where

$$\begin{aligned} * &= \text{complex conjugate} \\ \theta_{\mathrm{dc}} &= \text{dc part of } \theta(t) \end{aligned}$$

$$\begin{aligned} T(\rho, f-\rho) = {} & S(\rho, f-\rho) + \int_{-\infty}^{\infty} d\sigma W_{\phi}(\sigma) [2S(\sigma, \rho) S(-\sigma, f-\rho) \\ & - S(\sigma, f-\sigma) - \Gamma(\sigma)\Gamma(-\sigma) S(\rho, f-\rho) \\ & + S(\rho+\sigma, f-\rho-\sigma)] \end{aligned}$$

$$\begin{aligned} U(f) = {} & \Gamma(f) + \int_{-\infty}^{\infty} d\rho W_{\phi}(\rho) \Gamma(\rho) S(-\rho, f) \\ & + \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma W_{\phi}(\rho) W_{\phi}(\sigma) \{ -\tfrac{1}{2} \Gamma(\rho+\sigma) S(-\rho-\sigma, f) \\ & + \Gamma(\sigma) [3S(-\sigma, \rho) S(-\rho, f) - S(\rho, f-\rho-\sigma) \\ & + S(\rho-\sigma, f-\rho)]\} \end{aligned}$$

$$S(\rho, \sigma) = \Gamma(\rho+\sigma) - \Gamma(\rho)\Gamma(\sigma)$$

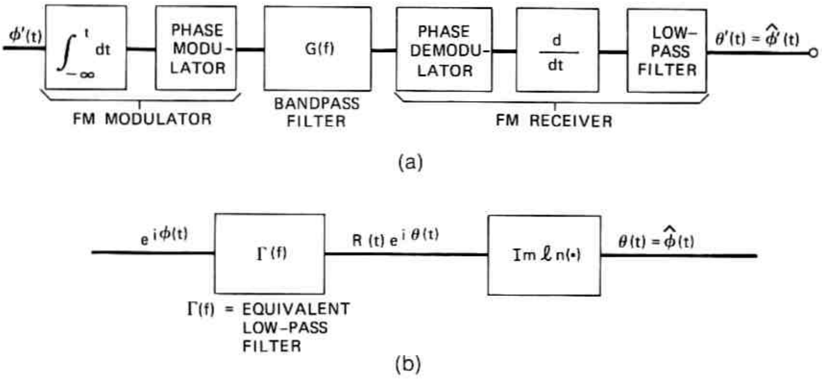


Fig. 1—(a) Analog FM communication system. $\hat{\phi}'(t)$ is a distorted version of $\phi'(t)$. (b) Associated mathematical model for studying FM distortion.

and

$$S(\rho, \sigma, \nu) = \Gamma(\rho + \sigma + \nu) - \Gamma(\rho + \sigma)\Gamma(\nu) - \Gamma(\rho + \nu)\Gamma(\sigma) \\ - \Gamma(\sigma + \nu)\Gamma(\rho) + 2\Gamma(\rho)\Gamma(\sigma)\Gamma(\nu).$$

We shall neglect the dc term, θ_{dc}^2, since we are mainly interested in the continuous part of $W_\theta(f)$. In addition, for the range of parameters of interest to us, we have found that the double integral associated with the $U(f)$ function can also be neglected.

Notice that $W_\phi(f_1) = 0$ does not imply that $W_\theta(f_1) = 0$. This is contrary to the case of a linear system. That is, even if we apply no input power in the frequency interval $(f_1, f_1 + df)$, we generally get some "intermodulation noise" at the output in this frequency interval. Actually, eq. (3) is a truncated form of an infinite series of functionals of $\Gamma(f)$ and $W_\phi(f)$. However, we shall see that it is possible to select system parameters which are of practical importance and which allow us to neglect all of the terms represented by $0(\phi^6 W_\phi)$. Accordingly, we shall define the signal power $S(f)df$ in the frequency interval $(f, f + df)$ at the output to be

$$S(f)df = \tfrac{1}{4}W_\phi(f)\,|\,U(f) + U^*(-f)\,|^2 df. \tag{4}$$

For the range of parameters which are of practical importance, it turns out that $S(f) \doteq \frac{1}{4}W_\phi(f)\,|\,\Gamma(f) + \Gamma^*(-f)\,|^2$. $S(f)$ represents the spectral contribution at the output which is free of intermodulation noise. We also define the FM distortion power $D(f)df$ appearing in the output frequency interval $(f, f + df)$ to be

$$\begin{aligned} D(f)df = \frac{1}{8}\int_{-\infty}^{\infty} d\rho W_{\phi}(\rho)W_{\phi}(f-\rho) \\ \times |T(\rho, f-\rho) - T^*(-\rho, -f+\rho)|^2 df \\ + \frac{1}{24}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma W_{\phi}(\rho)W_{\phi}(\sigma)W_{\phi}(f-\rho-\sigma) \\ \times |S(\rho, \sigma, f-\rho-\sigma) + S^*(-\rho, -\sigma, -f+\rho+\sigma)|^2 df. \end{aligned} \tag{5}$$

$D(f)$ represents the spectral contribution at the output which results from intermodulation noise.

The two quantities of prime interest in this section are the signal-to-distortion ratio, $S(f)/D(f)$, and the ratio of the average signal power, σ_S^2, to the average distortion power, σ_D^2, in the output baseband. The latter quantities are defined by

$$\sigma_S^2 = \begin{cases} 2\int_0^B S(f)df & \text{PM Case} \\ 2\int_0^B \omega^2 S(f)df & \text{FM Case} \end{cases} \tag{6}$$

$$\sigma_D^2 = \begin{cases} 2\int_0^B D(f)df & \text{PM Case} \\ 2\int_0^B \omega^2 D(f)df & \text{FM Case} \end{cases} \tag{7}$$

where B = baseband bandwidth of the modulation. When the ratio of $\sigma_S^2/\sigma_D^2 \geqq 10$, one is usually safe in disregarding the terms labeled $0(\phi^6 W_{\phi})$ in eq. (3).†

III. NUMERICAL METHODS EMPLOYED

An input power spectral density, $W_{\phi'}(f)$, which is often used when measuring FM distortion is the uniform spectrum, given by

$$W_{\phi'}(f) = \begin{cases} \dfrac{(2\pi D)^2}{2B}, & |f| \leqq B \\ 0, & |f| > B \end{cases} \tag{8}$$

† Equation (3) is a special case of a much more general equation which was recently reported by E. Bedrosian and S. O. Rice in the Proc. IEEE, *59*, No. 12, pp. 1688–1707, eq. (14) and Section IVc, December 1971.

where D = RMS frequency deviation and B = baseband bandwidth. From eq. (1), we have

$$W_{\phi}(f) = \begin{cases} \dfrac{D^2}{2Bf^2}, & |f| \leqq B \\ 0 \quad , & |f| > B \end{cases} \tag{9}$$

When such a uniform $W_{\phi'}(f)$ is used to measure FM distortion, the measurement is referred to as a "noise loading test." The noise loading test is used, for example, to estimate the FM distortion in microwave relay systems resulting when thousands of telephone channels are multiplexed to form a composite baseband signal.

A bandlimited form of $W_{\phi'}(f)$ is very convenient numerically, since it serves to convert the infinite limits of integration in eqs. (4) and (5) into finite limits of integration. However, if we attempt to evaluate equations (4) and (5) using a bandlimited $W_{\phi}(f)$ such as is given in eq. (9), we would run into difficulty whenever the argument of $W_{\phi}(\cdot)$ is equal to zero. In order to circumvent this apparent difficulty, we have selected an integration grid such that the argument of $W_{\phi}(\cdot)$ is never allowed to be zero. Equations (4) and (5) are then numerically evaluated by using a combination of Simpson's rule and the trapezoidal rule.

The particular integration grid used was obtained by setting

$$\rho = (2i + 1)\Delta \tag{10}$$

$$\sigma = (2l + 1)\Delta \tag{11}$$

$$f = (2n + 1)\Delta \tag{12}$$

$$\Delta = (2k + 1)^{-1} \tag{13}$$

$$B = 1.0 \tag{14}$$

where i, l, n, k are integers. In most of our numerical evaluations, $k = 20$.

IV. NUMERICAL RESULTS

4.1 *Test Cases*

In order to test the operation of the computer program, we evaluated $D(f)$ for the case when $W_{\phi'}(f)$ is uniform and $\Gamma(f)$ is the transfer

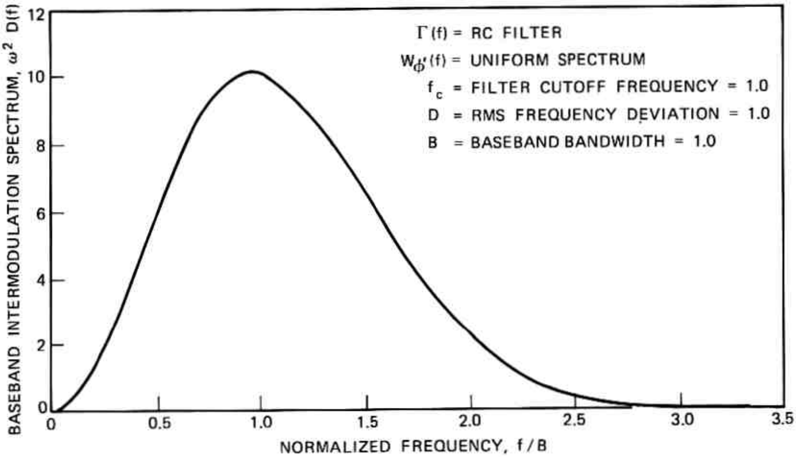


Fig. 2—Baseband intermodulation spectrum resulting from FM distortion.

function of a simple RC filter. That is, $W_\phi(f)$ is defined by eq. (9) and $\Gamma(f)$ is defined as

$$\Gamma(f) = \frac{1}{1 + i\left(\dfrac{f}{f_c}\right)} \tag{15}$$

where

$$f_c = 1.0$$

$$D = 1.0$$

$$B = 1.0.$$

Figure 2 shows the resulting computer plot of $\omega^2 D(f)$. Figure 3 shows a plot of $10 \log [S(f)/D(f)]$ for $f = 0.084B$, $0.36B$, and B, as a function of the RMS frequency deviation D. These results compare well with both the theoretical and experimental results which are presented in Refs. 2, 4, and 5.

As a final test case, we present the results for the case when $W_\phi(f)$ is still defined by eq. (9), but $\Gamma(f)$ now represents a 3-pole Butterworth filter with some mistuning, in which case

$$\Gamma(f) = \frac{1 + 2(i\xi_0) + 2(i\xi_0)^2 + (i\xi_0)^3}{1 + 2(i\xi) + 2(i\xi)^2 + (i\xi)^3} \tag{16}$$

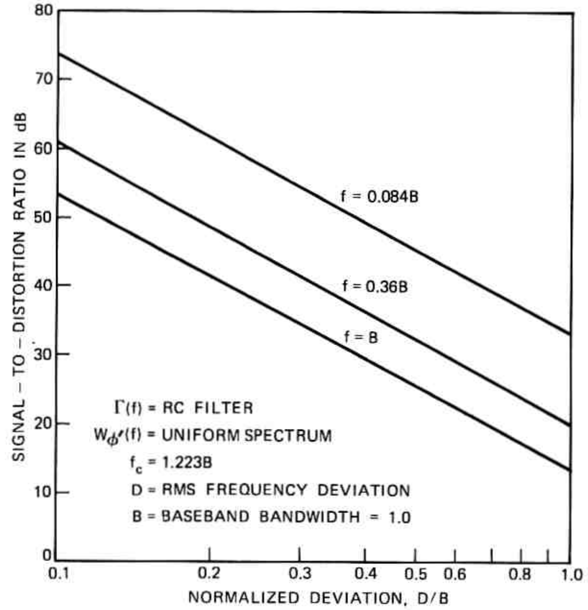
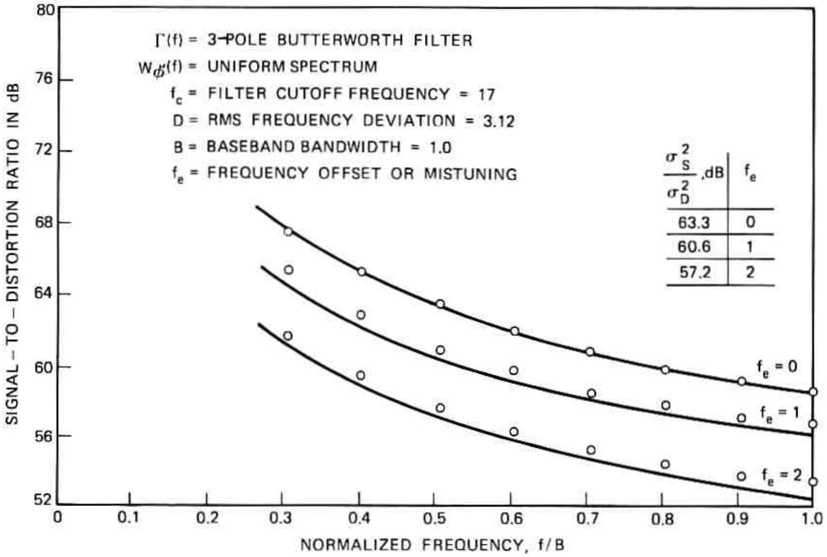


Fig. 3—Signal-to-distortion ratio at particular frequencies resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.



$\frac{\sigma_S^2}{\sigma_D^2}$, dB	f_e
63.3	0
60.6	1
57.2	2

Fig. 4—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$ for particular values of frequency offset or mistuning. The points are from the theoretical approximation given as eqs. (17) and (18).

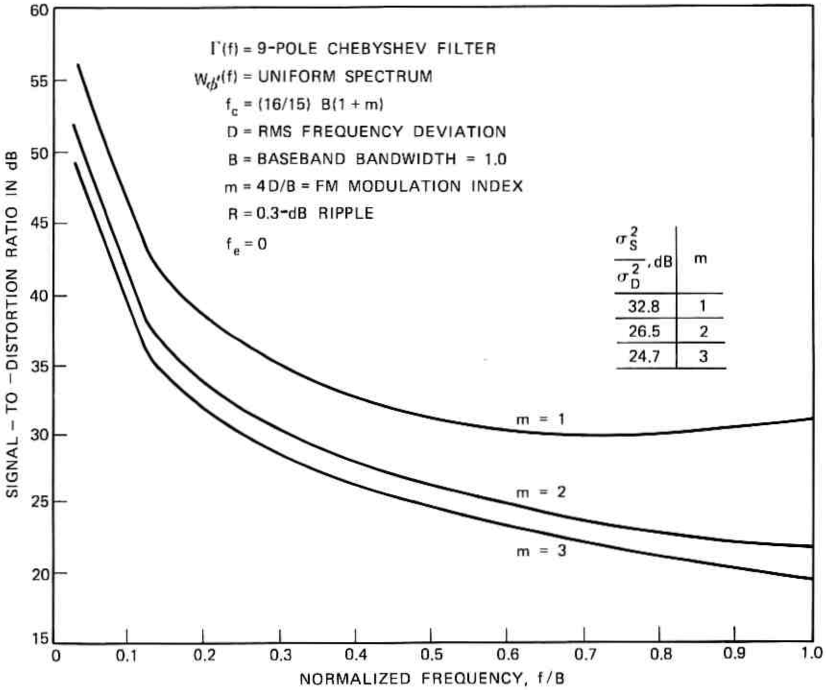


Fig. 5—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

where

$$\xi = \frac{f - f_e}{f_c}$$

$$\xi_0 = -\frac{f_e}{f_c}$$

$$B = 1.0$$

$$D = 3.12$$

$$f_c = 17.0$$

$$f_e = 0,\ 1,\ 2.$$

Figure 4 presents the computer plot for this case. The results compare very well with experimental and Monte Carlo results presented in Ref. 4.

A theoretical approximation for the above case, with $0 \leqq |f| \leqq B$,

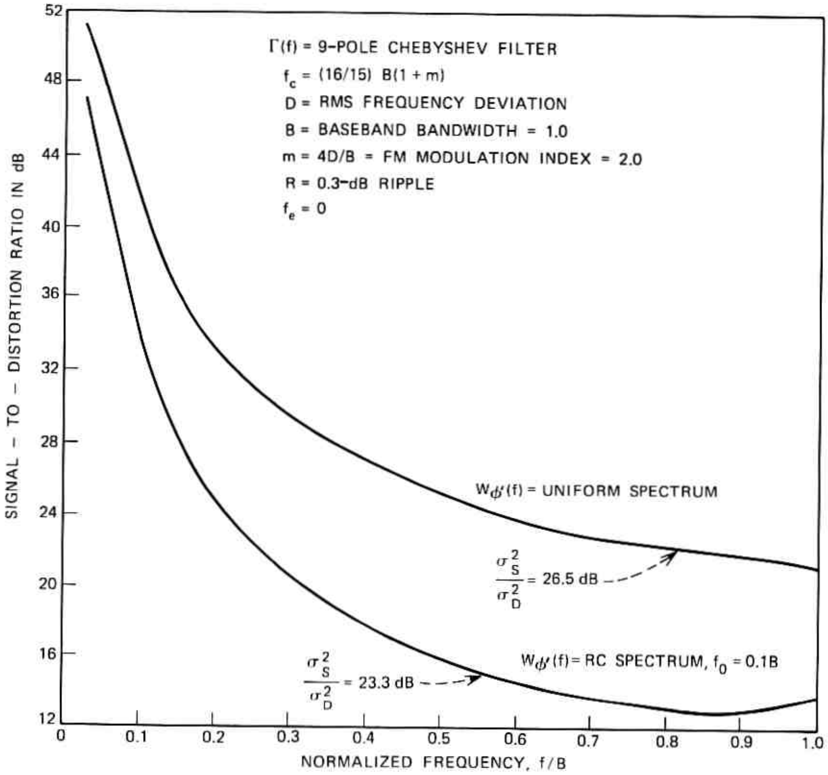


Fig. 6—Signal-to-distortion ratio resulting from $\Gamma(f)$ and two forms of $W_{\phi'}(f)$.

was developed by Rice[3] and is given by

$$D(f) \doteq \frac{D^4}{8B^2}(2B - |f|)(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 + \frac{D^6}{48B^3}(3B^2 - f^2)(\lambda_{3i})^2 \qquad (17)$$

and

$$S(f) \doteq W_\phi(f) \qquad (18)$$

where λ_{ni} is the imaginary part of λ_n and $\lambda_n/n!$ is the coefficient of f^n in the power series expansion of $\ln \Gamma(f)$

$$\lambda_{2i} = 2f_e f_c^{-3}$$

$$\lambda_{3i} = -2f_c^{-3}$$

$$\lambda_{4i} = 48f_e f_c^{-5}.$$

Some points which were computed from this theoretical approximation are indicated in Fig. 4.

Having verified the soundness of the computer program with the foregoing test cases, let us present some new results.

4.2 *New Results*

In this section, we shall present some new results which were obtained by using the above methods. These results will also help to demonstrate the general types of FM distortion problems that can be analyzed.

4.2.1 *n-pole Chebyshev filter*

Let $W_{\phi'}(f)$ be uniform as given by eq. (8) with $\Gamma(f)$ given by

$$\Gamma(f) = \frac{\prod_{k=1}^{n} (i\xi_0 - s_k)}{\prod_{k=1}^{n} (i\xi - s_k)} \tag{19}$$

where

$$s_k = -\sin\left[(2k-1)\frac{\pi}{2n}\right]\sinh\left[\frac{1}{n}\sinh^{-1}\left(\frac{1}{b}\right)\right] + i\cos\left[(2k-1)\frac{\pi}{2n}\right]\cosh\left[\frac{1}{n}\sinh^{-1}\left(\frac{1}{b}\right)\right], \qquad k = 1, \cdots, n$$

$$\xi = \frac{f - f_e}{f_c}, \qquad \xi_0 = -\frac{f_e}{f_c}$$

$$B = 1 = \text{baseband bandwidth}$$

$$2f_c = K2B(1+m) = K \text{ times Carson's rule} = \text{filter bandwidth}$$

$$m = \frac{4D}{B} = \text{FM modulation index}$$

$$R = 10\log(1+b^2) = \text{in-band ripple}$$

$$f_e = \text{offset frequency or mistuning.}$$

$\Gamma(f)$, defined by eq. (19), represents an n-pole Chebyshev filter. Some results for this case are presented in Figs. 5, 6 and 7. Also, a computer

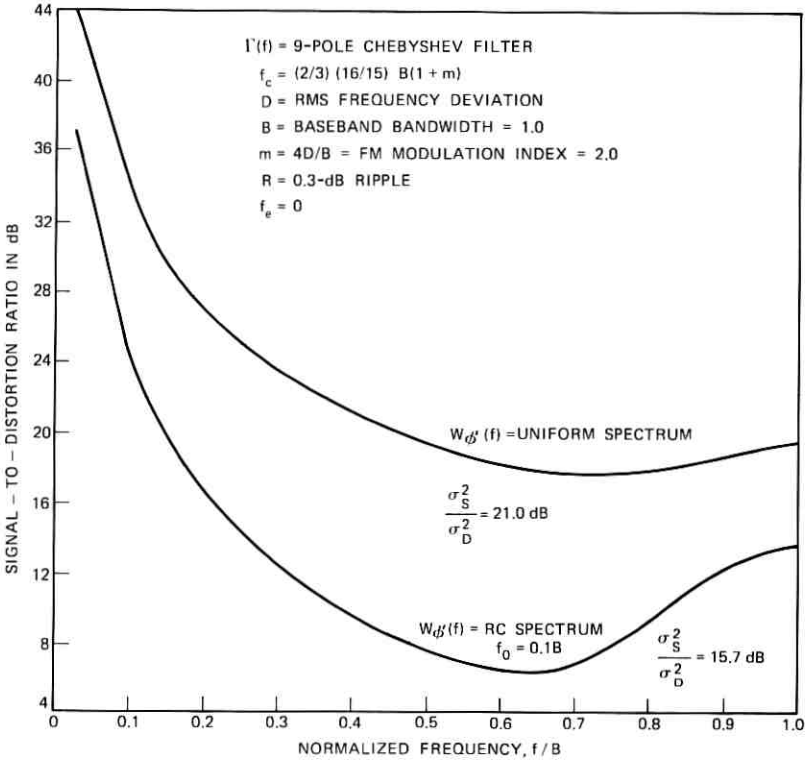


Fig. 7—Signal-to-distortion ratio resulting from $\Gamma(f)$ and two forms of $W_{\phi'}(f)$.

plot is presented in Figs. 6 and 7 for the case when $W_{\phi'}(f)$ is an RC spectrum[†] given by

$$W_{\phi'}(f) = \begin{cases} \dfrac{(2\pi D)^2}{2f_0 \tan^{-1}\left(\dfrac{B}{f_0}\right)} \left[1 + \left(\dfrac{f}{f_0}\right)^2\right]^{-1}, & |f| \leqq B \\ 0 \qquad\qquad\qquad\qquad\qquad , & |f| > B \end{cases} \tag{20}$$

where

$$f_0 = 3 \text{ dB bandwidth.}$$

An RC spectrum is often used to model a video signal. Notice that, as $f_0 \rightarrow \infty$, the RC spectrum approaches the uniform spectrum as given

[†] $W_{\phi'}(f)$ is the spectrum produced by passing bandlimited "white" noise through an RC filter.

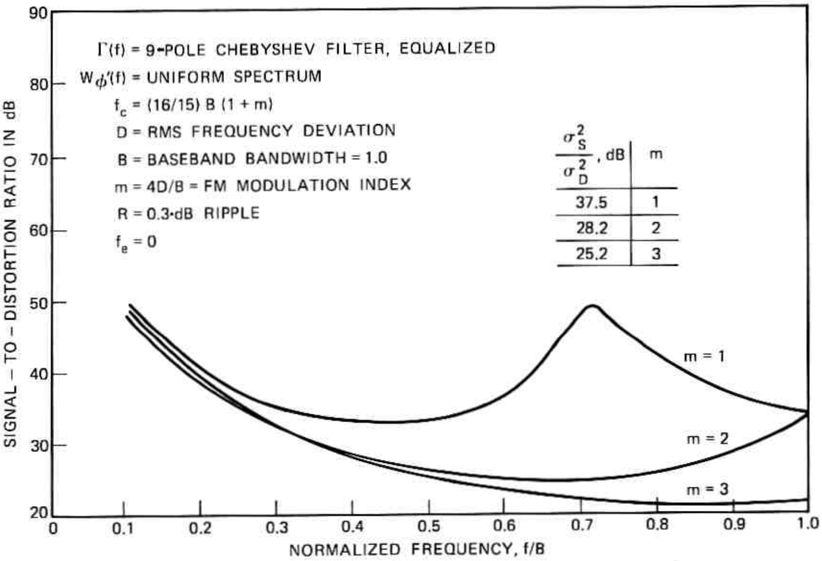


Fig. 8—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

by eq. (8). From Figs. 6 and 7, we see that more FM distortion results when $W_{\phi'}(f)$ is an RC or video spectrum than when $W_{\phi'}(f)$ is uniform.

4.2.2. *n-pole, equalized Chebyshev filter*

If the phase function associated with eq. (19) is taken to be zero,† $\Gamma(f)$ can be written as

$$\Gamma(f) = \left[\frac{1 + b^2 T_n^2(\xi_0)}{1 + b^2 T_n^2(\xi)}\right]^{\frac{1}{2}} \tag{21}$$

where $T_n(\xi)$ is a Chebyshev polynomial given by

$$T_n(\xi) = \begin{cases} \cos\left[n \cos^{-1}(\xi)\right], & |\xi| \leqq 1 \\ \cosh\left[n \cosh^{-1}(\xi)\right], & |\xi| > 1 \end{cases}.$$

$\Gamma(f)$, defined by eq. (21), represents an n-pole, equalized Chebyshev filter. Some results for this case are presented in Fig. 8. By comparing Figs. 5 and 8, we can determine the effect of equalization on FM distortion. In this case, equalization does not reduce the FM distortion significantly.

† Or linear in frequency since time delay is unimportant here.

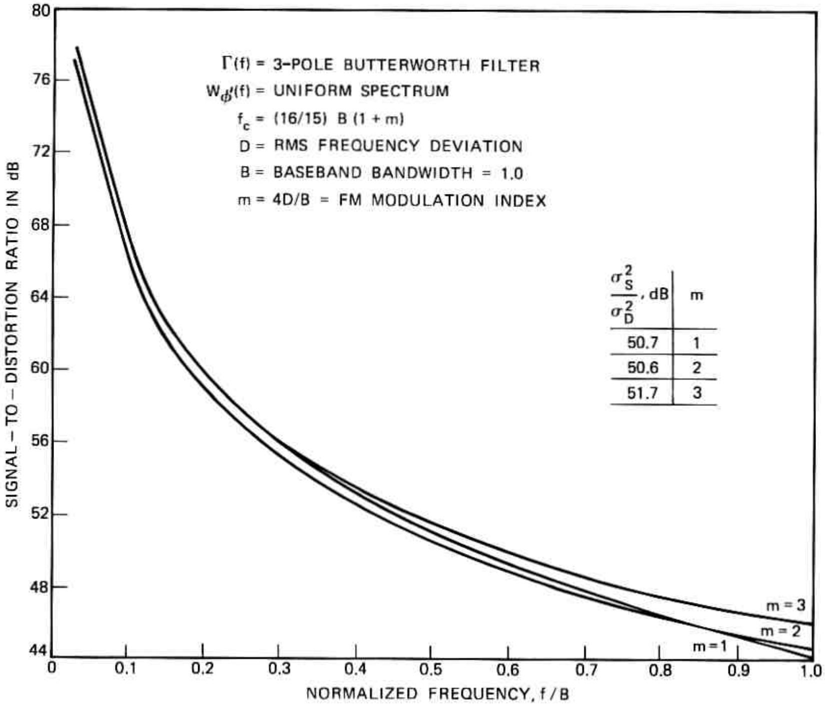


Fig. 9—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

4.2.3 *n-pole Butterworth filter*

Let $W_{\phi'}(f)$ be uniform as given by eq. (8) with $\Gamma(f)$ given by

$$\Gamma(f) = \frac{\prod_{k=1}^{n} (p_k)}{\prod_{k=1}^{n} (i\omega + p_k)} \tag{22}$$

where

$$p_k = (2\pi f_c) \exp\left[i\frac{\pi}{2}\left(\frac{2k-1}{n} - 1\right)\right], \qquad k = 1, 2, \cdots, n$$

$$B = 1 = \text{baseband bandwidth}$$

$$2f_c = K2B(1 + m)$$

$$m = \frac{4D}{B}.$$

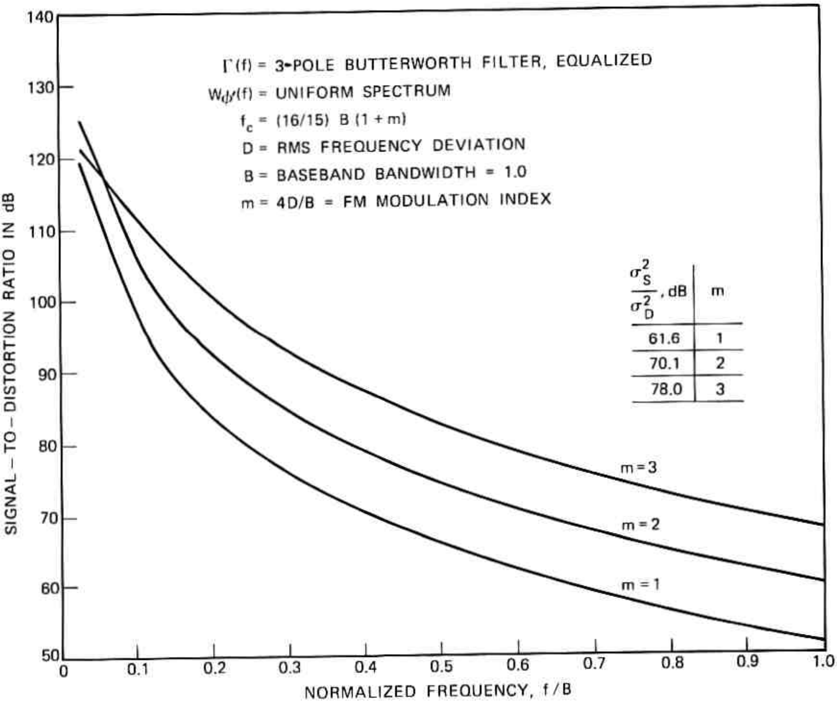


Fig. 10—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

$\Gamma(f)$, defined by eq. (22), represents an n-pole, Butterworth filter. Some results for this case are presented in Fig. 9. By comparing Figs. 5 and 9, we see that the 3-pole Butterworth filter produces much less FM distortion than does the 9-pole Chebyshev filter.

4.2.4 *n-pole, equalized Butterworth filter*

If the phase function associated with eq. (22) is taken to be zero, or linear in frequency, $\Gamma(f)$ can be written as

$$\Gamma(f) = \left[1 + \left(\frac{f}{f_c}\right)^{2n}\right]^{-\frac{1}{2}}. \tag{23}$$

$\Gamma(f)$, defined by eq. (23), represents an n-pole, equalized Butterworth filter. Some results for this case are presented in Fig. 10. By comparing Figs. 9 and 10, we see that equalization reduces the FM distortion significantly in this case.

4.2.5 *Echo, envelope delay, Butterworth filter*

Let $W_{\phi'}(f)$ be uniform as given by eq. (8) with $\Gamma(f)$ given by

$$\Gamma(f) = \underbrace{\left[\frac{1 + re^{i\omega T}}{1 + r}\right]}_{\text{(echo)}} \underbrace{[\exp\{i(b_2\omega^2 + b_3\omega^3)\}]}_{\text{(envelope delay)}} \times \underbrace{\left[\frac{1 + 2(i\xi_0) + 2(i\xi_0)^2 + (i\xi_0)^3}{1 + 2(i\xi) + 2(i\xi)^2 + (i\xi)^3}\right]}_{\text{(3-pole Butterworth filter)}} \tag{24}$$

where

$$\begin{aligned} r &= \text{amplitude of echo} \\ T &= \text{time delay of echo} \\ b_2 &= \text{linear envelope delay constant} \\ b_3 &= \text{quadratic envelope delay constant} \\ \xi &= \frac{f - f_e}{f_c} \\ \xi_0 &= -\frac{f_e}{f_c} \\ f_e &= \text{frequency offset or mistuning} \\ D &= \text{RMS frequency deviation} \\ B &= 1 = \text{baseband bandwidth} \\ 2f_c &= K2B(1 + m) = K \text{ times Carson's rule} \\ &= \text{filter bandwidth} \\ m &= \frac{4D}{B} = \text{FM modulation index.} \end{aligned}$$

Some results for this case are presented in Fig. 11.

Results are presented in Fig. 12 for the case when $\Gamma(f)$ is given by eq. (24) and $W_\phi(f)$, rather than $W_{\phi'}(f)$, is uniform and given by

$$W_\phi(f) = \begin{cases} \dfrac{3}{2}\dfrac{D^2}{B^3}, & |f| \leqq B \\ 0 \quad, & |f| > B. \end{cases} \tag{25}$$

This is the case of a noise loading test applied to a phase modulated

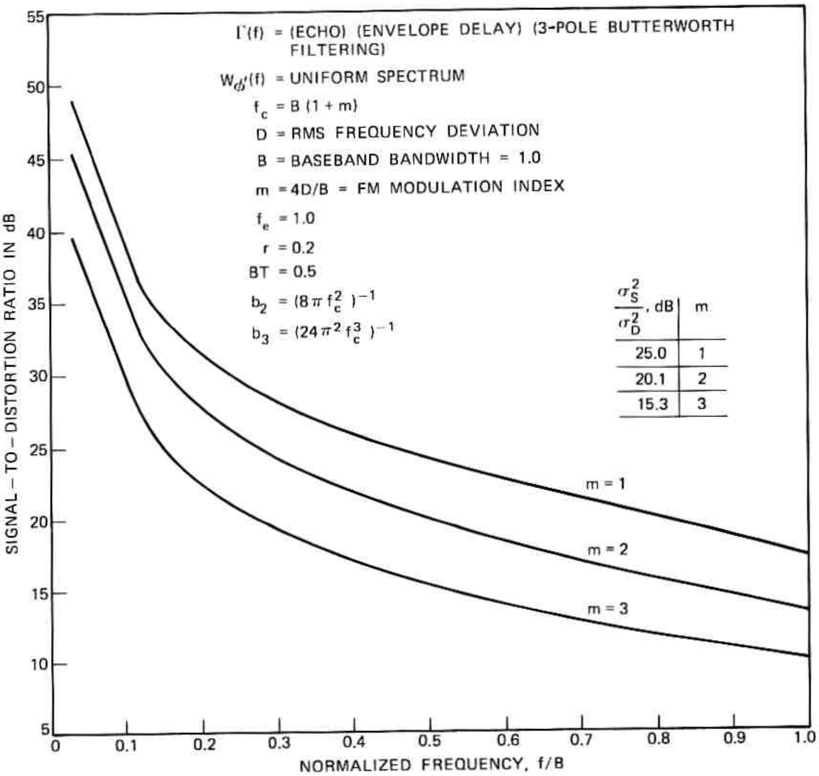


Fig. 11—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

system. From eq. (1) we have

$$W_{\phi'}(f) = \begin{cases} \dfrac{(2\pi D)^2 3 f^2}{2B^3}, & |f| \leqq B \\ 0 & |f| > B. \end{cases} \tag{26}$$

In this case, $W_{\phi'}(f)$ peaks at $f = B$ in contrast to the RC spectrum given by eq. (20), which peaks at $f = 0$. These results lead us to an interesting question. Given $W_{\phi'}(f)$ and $\Gamma(f)$, can we choose a predistortion characteristic such that the shape of $S(f)/D(f)$ is most suitable for a particular communication system? However, we have not investigated this question.

We can compare the distortion resulting from PM and FM systems by comparing Figs. 11 and 12. In fact, if the results in Fig. 12 apply to

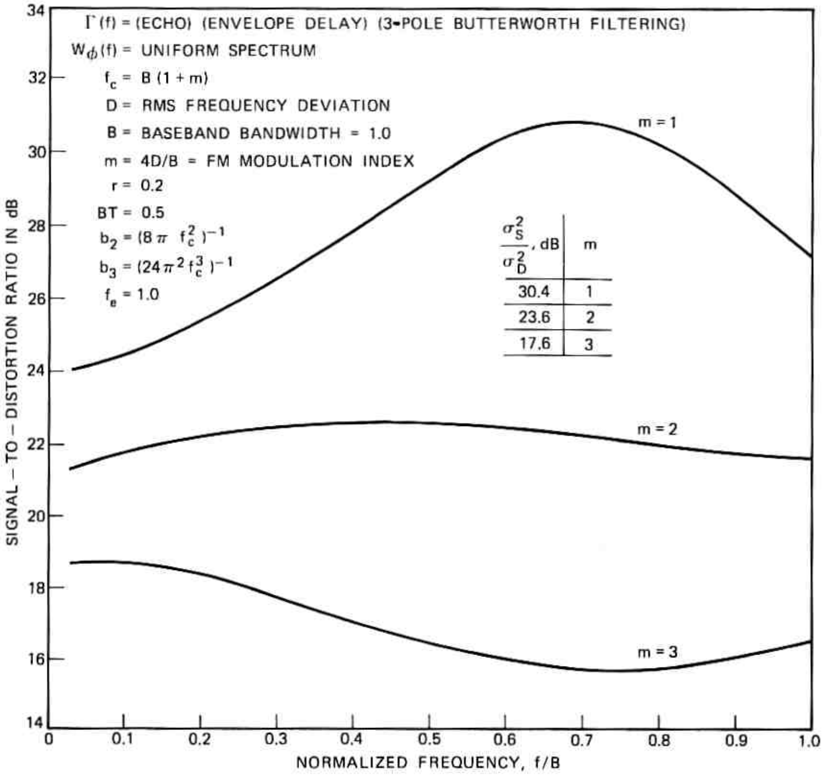


Fig. 12—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_\phi(f)$.

a Phase modulation system, then the results in Fig. 11 apply to the corresponding Frequency modulation system.

4.2.6 *Filter characteristic determined experimentally*

In all of the results presented above, the filter characteristic $\Gamma(f)$ was specified mathematically. However, many situations arise for which the measured amplitude and envelope delay characteristics are available in graphic form. Of course, one may try to fit a suitable analytical expression to these experimental points and proceed as above. However, there is no need to develop such an analytical expression. As our final case, we shall present some results which illustrate this point.

Let the amplitude and phase of $\Gamma(f)$ be given by the experimental

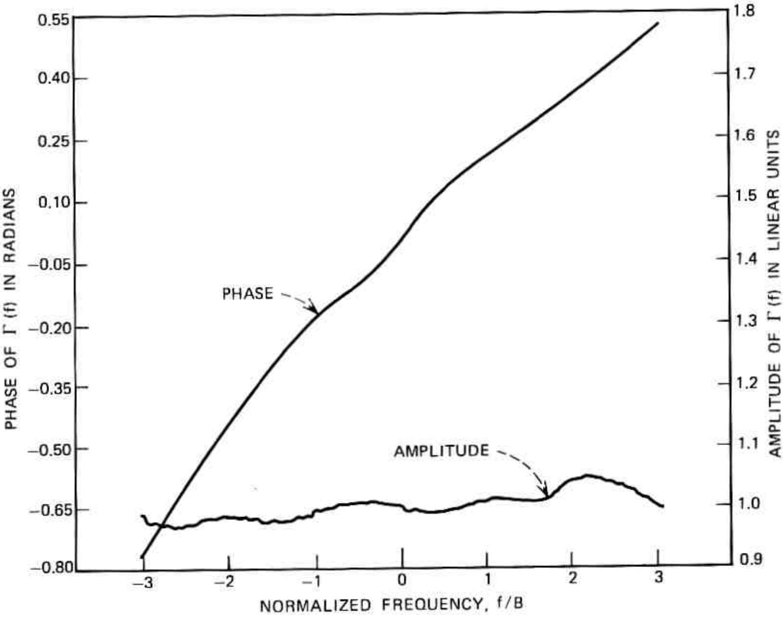


Fig. 13—Experimental amplitude and phase characteristic specifying $\Gamma(f)$.

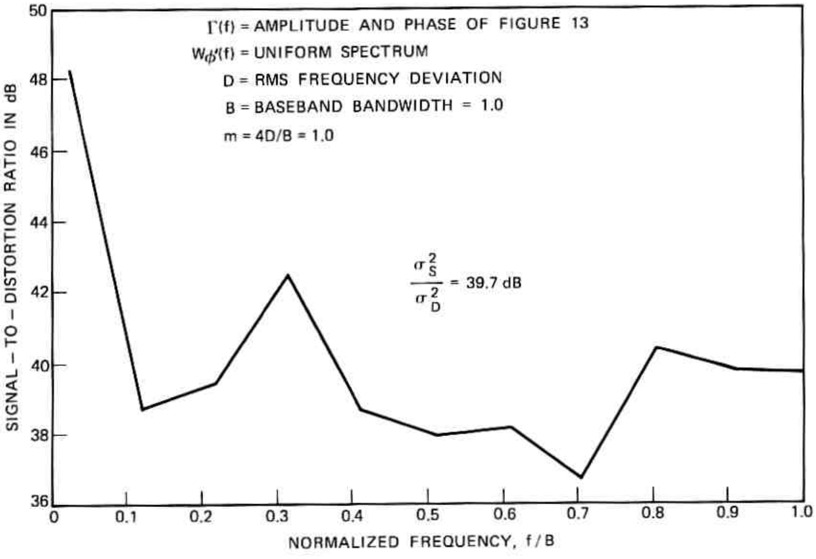


Fig. 14—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

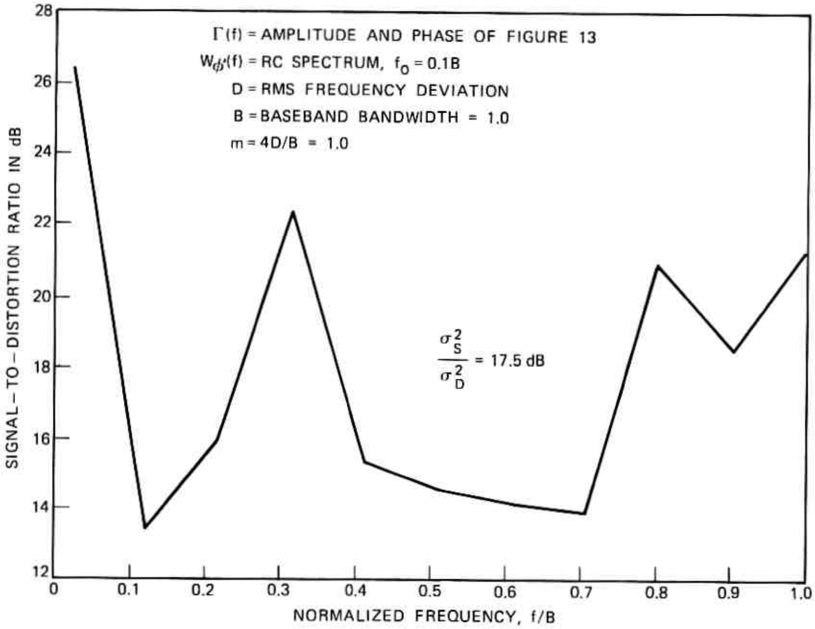


Fig. 15—Signal-to-distortion ratio resulting from $\Gamma(f)$ and $W_{\phi'}(f)$.

curves shown in Fig. 13. Let $W_{\phi'}(f)$ again be uniform and be given by eq. (8). Figure 14 presents some results for this case.

Now let $W_{\phi'}(f)$ be an RC spectrum given by eq. (20). Figure 15 presents some results for this case. We see that considerably more distortion is indicated in Fig. 15 as compared with Fig. 14.

Thus, the often-used noise loading test which makes use of a uniform spectrum may not represent a worst-case situation as far as FM distortion is concerned. A theoretical proof of this interesting fact is presented in the appendix.

V. CONCLUSIONS

Equations (4) through (7), together with a digital computer, can be used to compute the FM distortion resulting from passing FM waves through linear networks. To demonstrate the utility of the program, we have presented a variety of results in graphic form.

From this work, it is apparent that the often-used noise loading test does not necessarily represent a worst-case test. This was demonstrated for a system in which the modulating signal is a video signal. It is also apparent that predistortion may be useful in reducing FM distortion. This is in contrast to the use of conventional pre-emphasis, which is applied to combat RF noise.

Some important sources of FM distortion which were neglected in our analysis are AM-to-PM conversion, baseband and RF noise, and adjacent channel interference.

VI. ACKNOWLEDGMENTS

It gives me great pleasure to acknowledge helpful discussions with S. O. Rice. I also wish to express my thanks to J. N. Lahti and A. Yagoda for some useful comments, and to Miss R. Wright and F. L. Crawford for programming the digital computer.

APPENDIX

Theoretical Comparison of FM Distortion Resulting from Video and Uniform Spectra

The purpose of this appendix is to present a theoretical argument which shows that a video spectrum can lead to more FM distortion than a uniform spectrum. That is, a noise loading test which makes use of a uniform spectrum may not represent a worst-case test when the information source is a video signal.

Let the video signal $\phi'(t)$ be represented as a zero-mean, stationary Gaussian process having power spectral density $W_{\phi'}(f)$ given by

$$W_{\phi'}(f) = \begin{cases} \dfrac{(2\pi D)^2}{2f_0 \tan^{-1}\left(\dfrac{B}{f_0}\right)} \left[1 + \left(\dfrac{f}{f_0}\right)^2\right]^{-1}, & |f| \leqq B \\ 0 \qquad\qquad , & |f| > B \end{cases} \tag{27}$$

where

D = RMS frequency deviation

B = baseband bandwidth

f_0 = 3-dB bandwidth.

Let $\Gamma(f)$ be given by

$$\Gamma(f) = \exp\left[i(b_2\omega^2 + b_4\omega^4)\right] \tag{28}$$

where

b_2 = small linear envelope delay constant

b_4 = small cubic envelope delay constant.

By applying eq. (24) of Rice,[3] the leading term of $D(f)$ can be expressed as

$$D(f) = 2^{-1}(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 \int_{-\infty}^{\infty} d\rho W_\phi(\rho) W_\phi(f - \rho)\rho^2(f - \rho)^2 \tag{29}$$

where

$$\lambda_{2i} = (2!)(2\pi)^2 b_2$$

$$\lambda_{4i} = (4!)(2\pi)^4 b_4.$$

By taking $S(f) = W_\phi(f) = W_{\phi'}(f)/\omega^2$ and evaluating the integral for $D(f)$, we find that

$$R(f) = \frac{\left.\dfrac{D(f)}{S(f)}\right|_{f_0}}{\left.\dfrac{D(f)}{S(f)}\right|_{f_0=\infty}} = \frac{2}{[2 - |F|]\tan^{-1}\left(\dfrac{1}{F_0}\right)} \left[\frac{1 + \left(\dfrac{F}{F_0}\right)^2}{4 + \left(\dfrac{F}{F_0}\right)^2}\right]$$

$$\times \left\{\frac{F_0}{|F|} \ln \frac{1 + F_0^2}{(|F| - 1)^2 + F_0^2} + \tan^{-1}\left(\frac{1}{F_0}\right) + \tan^{-1}\left(\frac{1 - |F|}{F_0}\right)\right\} \tag{30}$$

where

$$F = \frac{f}{B}$$

$$F_0 = \frac{f_0}{B}$$

$$|f| \leqq B$$

$$-\frac{\pi}{2} \leqq \tan^{-1}(\cdot) \leqq \frac{\pi}{2}.$$

$R(f)$ represents the distortion-to-signal ratio for a video spectrum divided by the distortion-to-signal ratio for a uniform spectrum. If we can show that $R(f) > 1$ for particular values of f_0, the 3-dB bandwidth of $W_{\phi'}(f)$, then we can conclude that a video spectrum can produce more FM distortion than a uniform spectrum.

In order to show that $R(f)$ can be greater than unity, consider the important frequency range $0 < |F| < 1$. For this baseband frequency range, eq. (30) yields

$$\lim_{F_0 \to 0} R(f) = \frac{4}{2 - |F|} > 1. \tag{31}$$

A plot of $R(f)$ for various values of F_0 is shown in Fig. 16.

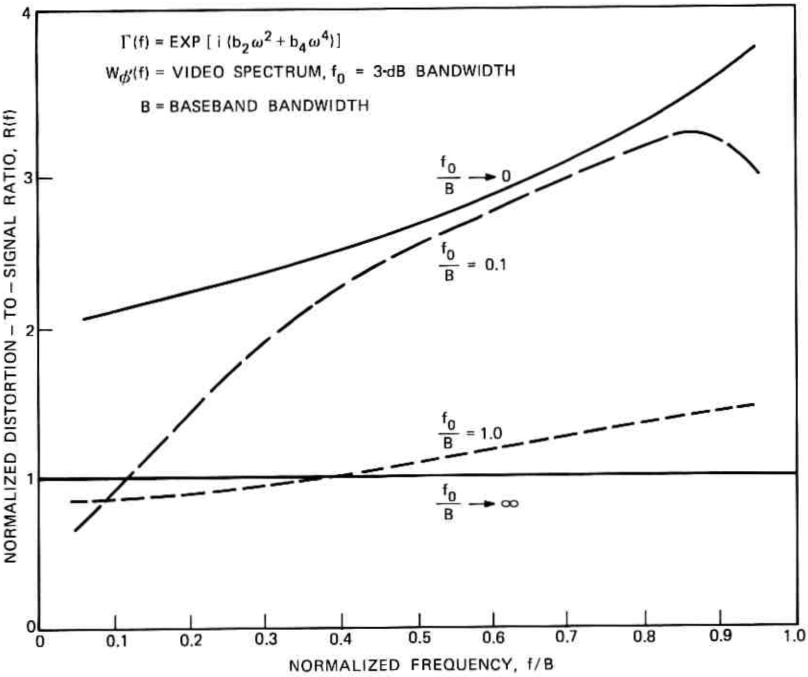


Fig. 16—$R(f)$ denotes the distortion-to-signal ratio for a video spectrum divided by distortion-to-signal ratio for a uniform spectrum.

Accordingly, we conclude that a video spectrum can produce more FM distortion in the baseband frequency range than a uniform spectrum.

REFERENCES

1. Mircea, A., "Harmonic Distortion and Intermodulation Noise in Linear FM Transmission Systems," Rev. Electrotech. Energet. (Romania), *12*, No. 3, 1967, pp. 359–371. Also see: Proc. IEEE (Correspondences) *54*, April 1966, pp. 705–706, and *54*, October 1966, pp. 1463–1466.
2. Bedrosian, E., and Rice, S. O., "Distortion and Crosstalk of Linearly Filtered Angle-Modulated Signals," Proc. IEEE, *56*, No. 1 (January 1968), pp. 2–13.
3. Rice, S. O., "Second and Third Order Modulation Terms in the Distortion Produced when Noise Modulated FM Waves are Filtered," B.S.T.J., *48*, No. 1 (January 1969), pp. 87–141.
4. Ruthroff, C. L., "Computation of FM Distortion in Linear Networks for Bandlimited Periodic Signals," B.S.T.J. *47*, No. 6 (July-August 1968), pp. 1043–1063.
5. Anuff, A., and Liou, M. L., "A Time-Domain Approach to Computing Distortion of Linearly Filtered FM Signals," IEEE Trans. Commun. Technol., *Com-19*, No. 2 (April 1971), pp. 218–221.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

A Volterra Series Description of Crosstalk Interference in Communications Systems

By JOEL GOLDMAN

(Manuscript received December 18, 1972)

This paper studies a general description of interchannel and intrachannel crosstalk interference created in a communications system. This description is in the form of a Volterra series expansion of the interference signal in terms of the signal which produced the interference. From it we are able to precisely define the "intelligible" part of the crosstalk. This description also provides us with quantitative measures of the amount of crosstalk created in some communications channel by signals in another channel, as well as a measure (intelligible crosstalk ratio) of the amount of intelligible crosstalk produced. We then consider a particular model for the generation of intelligible crosstalk [or direct adjacent channel interference (DACI)] between two neighboring angle-modulated channels in which the signal in one channel adds to the signal of the second channel, the sum is filtered, and the filter output then passes through an AM-PM conversion device. Using our definition, a simple expression for the intelligible crosstalk ratio for this model is derived in terms of the filter characteristic. We observe that this crosstalk ratio exhibits a number of properties usually associated with DACI.

I. INTRODUCTION

Crosstalk interference is an important consideration in transmission system engineering.[1] It is defined[2] as the disturbance created in one (desired) communications channel[†] by the signals in another (interfering) communications channel. Crosstalk is classified as due to interchannel or intrachannel effects and may be of either intelligible or unintelligible type. *Interchannel* crosstalk occurs between two different communications channels as, for example, when the transmitted signals of an interfering channel pass through the channel selectivity filters of

[†] Here "channels" refer to different communications paths (which are distinguished by, e.g., different frequency bands or different physical transmission media) together with the receivers associated with each of these paths.

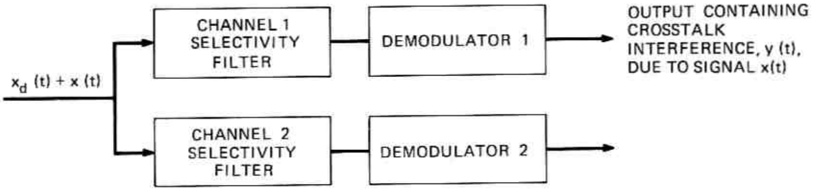

x_d (t) + x (t)
CHANNEL 1 SELECTIVITY FILTER
DEMODULATOR 1
OUTPUT CONTAINING CROSSTALK INTERFERENCE, y (t), DUE TO SIGNAL x(t)
CHANNEL 2 SELECTIVITY FILTER
DEMODULATOR 2

Fig. 1—Example of interchannel crosstalk. $x_d(t)$ is the desired signal in channel 1. $x(t)$, the desired signal in channel 2, creates interference in channel 1.

the desired channel and appear at its output (see Fig. 1). Another cause of interchannel crosstalk is electrical coupling between various transmission media, e.g., between wire pairs in a multipair cable. *Intrachannel* crosstalk occurs in a single communications channel and is due to nonlinearities in the receiver which act on the received signal to produce some disturbing signals in addition to the desired (linear) signal. Intrachannel crosstalk is also known as "intermodulation distortion."[3] If the signals in the channels are speech signals, crosstalk interference is described as *intelligible* or *unintelligible*, depending on whether the created interference is "understandable" or not. These terms are also applied to nonspeech signals, in which case intelligible means that the crosstalk is of "the same type as the desired signal."[2]

In this paper, we study a general mathematical technique which can be used to describe interchannel and intrachannel crosstalk created in a communications system. The description is in the form of a Volterra series expansion[4] of the interference signal in terms of the signal which produced the interference. This expansion furnishes some insight into which part of the total crosstalk interference is intelligible, and thus we will be able to precisely define what is meant by intelligible crosstalk. In this way, some of the subjectivity inherent in the earlier "definition" of intelligible crosstalk is removed. In addition, our description will provide quantitative measures of the amount of crosstalk created in some communications channel by signals in another channel, as well as a measure of the amount of intelligible crosstalk produced. The latter quantity will be called the *intelligible crosstalk ratio*. These measures may be valuable tools in systems design applications.

The Volterra series analysis of nonlinear systems with memory was first introduced by Wiener[5] and was further developed by Bedrosian and Rice.[4] In Section II, we discuss some definitions and results of this theory which will be needed in our analysis. A general description of crosstalk interference and a definition of intelligible crosstalk are given in Section III. We also define the intelligible crosstalk ratio in this section and compare it with previous measures of intelligible crosstalk. As an application of these results, we consider an example in Section IV of a model for the generation of intelligible crosstalk [or direct adjacent channel interference (DACI)[6]] between two neighboring angle-modulated channels in which the signal in one channel adds to the signal of the second channel, the sum is filtered, and the filter output then passes through an AM-PM conversion device. Using our definition, a simple expression for the intelligible crosstalk ratio for this model is derived in terms of the filter characteristic. We will see that this crosstalk ratio exhibits a number of properties usually associated with DACI. We conclude by calculating the crosstalk ratio for the case of a k-pole filter.

II. VOLTERRA SERIES ANALYSIS

In this section, we will discuss some definitions and results in the Volterra series analysis of nonlinear systems with memory. These results will be needed in the sequel. The reader is referred to Bedrosian and Rice[4] for a complete account of the theory of Volterra series as well as their application to the analysis of PM and other nonlinear systems.

For any two signals $y(t)$ and $x(t)$, possibly complex-valued, we will say that $y(t)$ *has a generalized Volterra series (GVS) expansion in terms of* $x(t)$ *with Volterra kernels (functions)* $\{g_n^{y,x}\}$ if and only if we can write:

$$y(t) = g_0^{y,x} + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{y,x}(u_1, \cdots, u_n) \cdot \prod_{r=1}^{n} x(t-u_r) \quad (1)$$

$$= g_0^{y,x} + \int_{-\infty}^{\infty} du_1 \, g_1^{y,x}(u_1) x(t-u_1) + \frac{1}{2!} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} du_1 \, du_2 \, g_2^{y,x}(u_1, u_2) x(t-u_1) x(t-u_2) + \frac{1}{3!} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} du_1 \, du_2 \, du_3 \, g_3^{y,x}(u_1, u_2, u_3) x(t-u_1) \cdot x(t-u_2) x(t-u_3) + \cdots$$

where the functions $g_n^{y,x}$, $n \geqq 1$, are symmetric functions of n variables and $g_0^{y,x}$ is a constant. For convenience, we denote this fact by the notation $y(t) = \text{GVS}[x(t); \{g_n^{y,x}\}]$. If $x(t)$ is the input to a system and $y(t)$ is its output, then the Volterra kernels $\{g_n^{y,x}\}$ completely characterize the system. If $g_0^{y,x} = a_0$ and $g_n^{y,x}(u_1, \cdots, u_n) = a_n \, \delta(u_1) \cdots \delta(u_n)$ for $n \geqq 1$ where $\delta(u)$ is the delta function, then

$$y(t) = \sum_{n=0}^{\infty} a_n \frac{[x(t)]^n}{n!}$$

represents the input-output relationship of a memoryless system.

The n-fold ($n \geqq 1$) Fourier transform of $g_n^{y,x}(u_1, \cdots, u_n)$ is denoted by:

$$G_n^{y,x}(f_1, \cdots, f_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{y,x}(u_1, \cdots, u_n) \cdot \exp\left[-j(\omega_1 u_1 + \cdots + \omega_n u_n)\right] \quad (2)$$

where $\omega_i = 2\pi f_i$, $i = 1, 2, \cdots$. Observe that if $g_0^{y,x} = 0$ and $g_n^{y,x} \equiv 0$ for $n \geqq 2$, then $g_1^{y,x}(u_1)$ is the familiar impulse response of a linear time-invariant system and $G_1^{y,x}(f_1)$ is its transfer function. By analogy, we will call $G_n^{y,x}(f_1, \cdots, f_n)$ the *nth order Volterra transfer function*. Since $\{g_n^{y,x}\}$ are symmetric functions, then so are $\{G_n^{y,x}\}$.

If $x(t)$ has Fourier transform $X(f)$, i.e.,

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt, \qquad \omega = 2\pi f,$$

then it is easy to see[4] that $y(t)$ and its Fourier transform $Y(f)$ are given by:

$$y(t) = g_0^{y,x} + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} df_1 \cdots df_n \, G_n^{y,x}(f_1, \cdots, f_n) \cdot e^{j(\omega_1 + \cdots + \omega_n)t} \prod_{r=1}^{n} X(f_r) \quad (3)$$

and

$$\begin{aligned} Y(f) = {} & g_0^{y,x}\delta(f) + \frac{1}{1!} G_1^{y,x}(f)X(f) \\ & + \frac{1}{2!}\int_{-\infty}^{\infty} df_1 \, G_2^{y,x}(f_1, f - f_1)X(f_1)X(f - f_1) \\ & + \frac{1}{3!}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} df_1 \, df_2 \, G_3^{y,x}(f_1, f_2, f - f_1 - f_2)X(f_1)X(f_2) \\ & \cdot X(f - f_1 - f_2) + \cdots . \end{aligned} \quad (4)$$

Next, suppose we apply a harmonic input of the form $\sum_{i=1}^{n} e^{j\omega_i t}$ to a system whose input and output are related by a generalized Volterra series expansion. Then the output of the system is an infinite series of harmonic terms. The following property, which is easy to demonstrate,[4] shows that the coefficients of these harmonic terms are the Volterra transfer functions of various orders.

Property 1: Suppose $y(t) = \text{GVS}[x(t); \{g_n^{y,x}\}]$. If $x(t) = \sum_{i=1}^{n} e^{j\omega_i t}$ where $\omega_i = 2\pi f_i$, $i = 1, \cdots, n$, and $\{f_i\}$ are incommensurable,[†] then for

[†] Frequencies $f_1, \cdots, f_n$ are said to be *incommensurable* if for any integers $m_1, \cdots, m_n$, not all zero, $m_1 f_1 + \cdots + m_n f_n \neq 0$.

$n \geqq 1$ and for $k \leqq n$:

$G_k^{y,x}(f_1, \cdots, f_k)$ = the coefficient of the $\exp[j(\omega_1 + \cdots + \omega_k)t]$ term in the expansion of $y(t)$.

Methods of measuring the Volterra kernels and transfer functions of a system having a Volterra series representation have been studied by George,[7] Schetzen,[8] and Lee and Schetzen.[9] These methods rely on the use of realizable input probing signals.

Bedrosian and Rice[4] have also shown the following:

Property 2: Suppose $y(t) = \text{GVS}[x(t); \{g_n^{y,x}\}]$. If $x(t) = P \cos \omega t$, $\omega = 2\pi f$, then

$$y(t) = g_0^{y,x} + \sum_{n=1}^{\infty} \sum_{k=0}^{n} \left(\frac{P}{2}\right)^n \frac{\exp[j(2k - n)\omega t]}{k!(n-k)!} G_{k,n-k}^{y,x}(f) \tag{5}$$

where $G_{k,n-k}^{y,x}(f)$ denotes $G_n^{y,x}(f_1, \cdots, f_n)$ with the first k of the f_i's equal to f and the remaining $n - k$ equal to $-f$. The leading terms in (5) are:

$$\begin{aligned} y(t) = & \left[\frac{P^2}{4} G_2^{y,x}(f, -f) + \cdots\right] \\ & + e^{j\omega t}\left[\frac{P}{2} G_1^{y,x}(f) + \frac{P^3}{16} G_3^{y,x}(f, f, -f) + \cdots\right] \\ & + e^{-j\omega t}\left[\frac{P}{2} G_1^{y,x}(-f) + \frac{P^3}{16} G_3^{y,x}(-f, -f, f) + \cdots\right] \\ & + e^{j2\omega t}\left[\frac{P^2}{8} G_2^{y,x}(f, f) + \cdots\right] \\ & + e^{-j2\omega t}\left[\frac{P^2}{8} G_2^{y,x}(-f, -f) + \cdots\right] \\ & + e^{j3\omega t}\left[\frac{P^3}{48} G_3^{y,x}(f, f, f) + \cdots\right] \\ & + e^{-j3\omega t}\left[\frac{P^3}{48} G_3^{y,x}(-f, -f, -f) + \cdots\right] + \cdots . \end{aligned} \tag{6}$$

When $x(t) = P \cos \omega_1 t + Q \cos \omega_2 t$, then $y(t)$ is a sum of complex exponentials, the $\exp[j(N\omega_1 + M\omega_2)t]$ component of $y(t)$ being, for $M \geqq 0$ and $N \geqq 0$,

$$e^{j(N\omega_1 + M\omega_2)t} \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} \frac{(P/2)^{2l+N}(Q/2)^{2k+M}}{(N+l)!\,l!\,(M+k)!\,k!} G_{N+l,l;M+k,k}^{y,x}(f_1, f_2) \tag{7}$$

where $\omega_i = 2\pi f_i$, $i = 1, 2$, and $G_{N+l,l;M+k,k}^{y,x}(f_1, f_2)$ denotes $G_n^{y,x}(f_1, \cdots, f_n)$

with $n = N+2l+M+2k$ and the first $N+l$ of the f_i's equal to f_1, the next l equal to $-f_1$, the next $M+k$ equal to f_2, and the last k equal to $-f_2$.

In Appendix A we show that the input-output pairs of a certain class of nonlinear systems can be related by a Volterra series expansion with certain Volterra kernels. The result is a slight generalization of one proved in Ref. 4.

Property 3: Suppose

$$y(t) = F\left[\int_{-\infty}^{\infty} g(u)\hat{h}[x(t-u)]du\right] \tag{8}$$

where F and $\hat{h}$ are functions of a complex variable having series expansions:

$$\hat{h}(z) = \sum_{\nu=0}^{\infty} \hat{h}_\nu \frac{z^\nu}{\nu!} \tag{9}$$

$$F(z) = \sum_{l=0}^{\infty} F_l \frac{(z-z_0)^l}{l!} \tag{10}$$

with

$$z_0 \triangleq \hat{h}_0 \int_{-\infty}^{\infty} g(u)du.$$

Let

$$G(f) \triangleq \int_{-\infty}^{\infty} g(u)e^{-j\omega u}du, \qquad \omega = 2\pi f.$$

Then $y(t) = \text{GVS}[x(t); \{g_n^{y,x}\}]$ with $g_0^{y,x} = F_0$ and

$$\begin{aligned} G_1^{y,x}(f_1) &= F_1\hat{h}_1 G(f_1) \\ G_2^{y,x}(f_1, f_2) &= F_1\hat{h}_2 G(f_1+f_2) + F_2\hat{h}_1^2 G(f_1)G(f_2) \\ G_3^{y,x}(f_1, f_2, f_3) &= F_1\hat{h}_3 G(f_1+f_2+f_3) + F_2\hat{h}_1\hat{h}_2[G(f_1)G(f_2+f_3) \\ &\quad + G(f_2)G(f_1+f_3) + G(f_3)G(f_1+f_2)] \\ &\qquad + F_3\hat{h}_1^3 G(f_1)G(f_2)G(f_3). \end{aligned}$$

Expressions for the higher-order Volterra transfer functions are given in eq. (49) of Ref. 4.

Finally, suppose that $y(t)$ and $x(t)$ are related by a Volterra series expansion, and that $y(t)$ is transformed by some function $\hat{F}(\cdot)$ to produce a signal $w(t) = \hat{F}[y(t)]$. Then, for a certain class of functions $\hat{F}(\cdot)$, the following result, which is proved in Appendix B, shows that $w(t)$ also has a Volterra series expansion in terms of $x(t)$ with specific kernels.

Property 4: Suppose $y(t) = \mathrm{GVS}[x(t); \{g_n^{y,x}\}]$ and $w(t) = \hat{F}[y(t)]$ where $\hat{F}$ is a function of a complex variable having series expansion:

$$\hat{F}(z) = \sum_{l=0}^{\infty} \hat{F}_l \frac{(z - g_0^{y,x})^l}{l!}.$$

Then $w(t) = \mathrm{GVS}[x(t); \{g_n^{w,x}\}]$ where $g_0^{w,x} = \hat{F}_0$ and

$$\begin{aligned} G_1^{w,x}(f_1) &= \hat{F}_1 G_1^{y,x}(f_1) \\ G_2^{w,x}(f_1, f_2) &= \hat{F}_1 G_2^{y,x}(f_1, f_2) + \hat{F}_2 G_1^{y,x}(f_1) G_1^{y,x}(f_2) \\ G_3^{w,x}(f_1, f_2, f_3) &= \hat{F}_1 G_3^{y,x}(f_1, f_2, f_3) + \hat{F}_2[G_1^{y,x}(f_1) G_2^{y,x}(f_2, f_3) \\ &\quad + G_1^{y,x}(f_2) G_2^{y,x}(f_1, f_3) + G_1^{y,x}(f_3) G_2^{y,x}(f_1, f_2)] \\ &\qquad + \hat{F}_3 G_1^{y,x}(f_1) G_1^{y,x}(f_2) G_1^{y,x}(f_3). \end{aligned} \tag{11}$$

Expressions for the higher-order kernels can be obtained from the method discussed in Appendix B.

III. MATHEMATICAL DESCRIPTION OF CROSSTALK INTERFERENCE

With the Volterra series analysis discussed in the last section, we can now give a mathematical description of interchannel and intrachannel crosstalk. Consider interchannel crosstalk first. Suppose $x(t)$ is some signal in one communications channel which enters a second channel as a signal $\hat{x}(t)$, where $\hat{x}(t)$ is $x(t)$ (possibly) transformed by some operation. Assume that the second channel contains some devices which operate on $\hat{x}(t)$ to produce a signal, $y(t)$, at the output of the channel. If the operations which transformed $x(t)$ into $\hat{x}(t)$ and $\hat{x}(t)$ into $y(t)$ consist of, for instance, nonlinear operations described by power series in cascade with time-invariant linear operations, then it is clear from Properties 3 and 4 that $y(t)$ will have a generalized Volterra series expansion in $x(t)$:

$$y(t) = g_0^{y,x} + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{y,x}(u_1, \cdots, u_n) \cdot \prod_{r=1}^{n} x(t - u_r). \tag{1}$$

That is, the crosstalk interference, $y(t)$, appearing at the output of the second channel, can be expressed in terms of the signal in the first channel, $x(t)$, which created it. The first term in the summation in (1) will be denoted by

$$y_L(t) = \int_{-\infty}^{\infty} du_1 \, g_1^{y,x}(u_1) x(t - u_1). \tag{12}$$

It is that part of $y(t)$ which is linear in $x(\cdot)$; $y_L(t)$ can be obtained by passing $x(\cdot)$ through a time-invariant linear filter with impulse response $g_1^{y,x}(\cdot)$. If $Y_L(f)$ is the Fourier transform of $y_L(t)$, then:

$$Y_L(f) = G_1^{y,x}(f)X(f). \tag{13}$$

The higher-order terms in the summation in eq. (1) represent greater nonlinear distortions of the signal $x(t)$ than do the lower-order terms. This can be seen from eq. (4), where we observe that each term in (1) has a spectrum which contributes to the spectrum of $y(t)$, the higher-order terms distorting the spectrum of $x(t)$ to a greater degree. The spectrum of the linear part of $y(t)$ given by (13), however, is simply $X(f)$ multiplied by a weight function. Because of this, we might expect $y_L(t)$ to be more intelligible than the other terms in (1), in the case when $x(t)$ is a speech-like signal. In fact, we will define $y_L(t)$ to be the *intelligible* part of the crosstalk, and $y(t) - y_L(t)$ will be called the *unintelligible* part.

The Volterra kernels $\{g_n^{y,x}\}$ and especially the Volterra transfer functions $\{G_n^{y,x}\}$ can be used as a *measure* of the degree of nonlinearity of each of the terms in (1). Moreover, since by Property 1 the Volterra transfer functions are the responses at certain frequencies to a harmonic sum input, they have further intuitive appeal as appropriate measures of system performance. In particular, as a measure of the intelligible crosstalk created in one channel by signals in the other channel, we will define the *intelligible crosstalk ratio at frequency f, R(f)*, to be

$$R(f) \triangleq \frac{|Y_L(f)|^2}{|X(f)|^2} = |G_1^{y,x}(f)|^2. \tag{14}$$

Previous authors followed two different approaches in defining intelligible crosstalk and intelligible crosstalk ratio. One idea, followed by Ruthroff,[6] Bennett,[10] Curtis,[11] and Hatch[12] was to assume that the signal, $x(t)$, in one channel is a constant amplitude sinusoid at frequency f and having power P_1. Then, for certain models, they were able to show that $y(t)$, the resulting interference in the second channel, contained a sinusoid at frequency f with power P_2. They defined the intelligible crosstalk ratio at frequency f to be P_2/P_1. Extending this idea a little further, one might let $x(t)$ be a sum of sinusoids at incommensurable frequencies $f_1, \cdots, f_n$ $(\omega_i \triangleq 2\pi f_i)$, i.e., $x(t) = \sum_{i=1}^{n} \sin \omega_i t$. If, for some problem, we can express the resulting interference $y(t)$ as a sum of sinusoids, with b the coefficient of $\sin \omega_1 t$ in this sum, then the intelligible crosstalk ratio at frequency f_1 would be taken to be $|b|^2$.

Our definition of intelligible crosstalk ratio in (14) is similar to this except that we use complex exponentials instead of sinusoids. But Property 2 shows that $G_1^{y,x}(f_1)$ is in fact the leading term of the coefficient of $\sin \omega_1 t$ (when $x(t)$ is a sum of sinusoids), and thus the two definitions may in some cases yield approximately the same numerical result. Lundquist[13] followed another approach. He assumed that $x(t)$ was arbitrary and, for a certain model, was able to express the interference $y(t)$ as a series of products of powers and derivatives of $x(t)$. He took the intelligible crosstalk to be that part of $y(t)$ which was "linear in $x(t)$." Expressing this part as a linear filtering operation on $x(t)$, having transfer function $H_L(f)$, he then defined the intelligible crosstalk ratio to be $|H_L(f)|^2$. The intelligible crosstalk ratio given in (14) is identical with that of Lundquist once the part of $y(t)$ linear in $x(t)$ is identified.

The preceding discussion is also applicable to the problem of intrachannel crosstalk. Earlier Volterra series techniques[4,14] had been applied to one such problem, namely, distortion in angle-modulated systems. In the intrachannel crosstalk problem, $x(t)$, the signal at the input of a channel, is transformed by some nonlinear devices into the output signal $y(t)$. If these devices consist of, for example, nonlinear operations described by power series in cascade with time-invariant linear filtering, then $y(t)$ has a generalized Volterra series expansion in terms of $x(t)$ as in (1). Assume that the *desired* output signal $y_0(t)$ in the absence of the (parasitic) nonlinear devices should be a time-invariant linear operation on $x(t)$ with impulse response $k(\cdot)$ and transfer function $K(\cdot)$, i.e.,

$$y_0(t) = \int_{-\infty}^{\infty} k(u_1)x(t - u_1)du_1. \tag{15}$$

Then the distortion or crosstalk at the channel output is

$$\begin{aligned} y_D(t) &= y(t) - y_0(t) \\ &= g_0^{y,x} + \int_{-\infty}^{\infty} du_1[g_1^{y,x}(u_1) - k(u_1)]x(t - u_1) \\ &\quad + \sum_{n=2}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{y,x}(u_1, \cdots, u_n) \\ &\qquad\qquad \cdot \prod_{r=1}^{n} x(t - u_r). \end{aligned} \tag{16}$$

The intelligible crosstalk is:

$$\int_{-\infty}^{\infty} du_1 \, \hat{g}_1^{y,x}(u_1)x(t - u_1) \tag{17}$$

with

$$\hat{g}_1^{y,x}(u_1) \triangleq g_1^{y,x}(u_1) - k(u_1). \tag{18}$$

The intelligible crosstalk ratio is:

$$R(f) = |\hat{G}_1^{y,x}(f)|^2 \tag{19}$$

where

$$\hat{G}_1^{y,x}(f) \triangleq G_1^{y,x}(f) - K(f). \tag{20}$$

The remainder of our preceding discussion for interchannel crosstalk is also valid for intrachannel crosstalk. The Volterra transfer functions may be used as measures of system performance. They are similar to (generalized) "intermodulation coefficients"[2] except that they are the response to complex exponentials and not to sinusoids.

IV. INTELLIGIBLE CROSSTALK RATIO FOR A PARTICULAR MODEL

In this section we look at a model for the generation of intelligible crosstalk [or direct adjacent channel interference (DACI)] between two neighboring angle-modulated channels in which the signal in one channel adds to the signal of the second channel, the sum is filtered, and the filter output then passes through an AM-PM conversion device. An example of such a situation occurs in the TD-2 microwave radio relay system[15,16] where the principal channel discrimination is provided by IF filters. The main AM-PM conversion in this system occurs in the transmitter amplifier. This model will illustrate the ideas and techniques of the previous sections. While we seek only the first Volterra transfer function (for intelligible crosstalk), the higher-order transfer functions can be found in a similar way.

Consider, in general, two neighboring phase-modulated† communications channels (labeled "1" and "2"). (See Fig. 2.) In channel 1, the received "desired" signal or carrier is taken to be:

$$v_{i1}(t) = \cos(\omega_1 t + \phi_1(t)) \tag{21}$$

where $\phi_1(t)$ is the phase modulation and the amplitude of $v_{i1}(t)$ has been normalized to unity. We assume that $v_{i1}(t)$ passes through a linear, time-invariant filter in channel 1 without distortion so that at the filter output the signal is:

$$v_{o1}(t) = \cos(\omega_1 t + \phi_1(t)). \tag{22}$$

In channel 2, the received "undesired" or interfering signal is assumed to be:

$$v_{i2}(t) = \kappa \cos(\omega_2 t + \phi(t)) \tag{23}$$

so that the signal (or carrier)-to-interference ratio is

$$\mu \triangleq \frac{1}{\kappa^2} \tag{24}$$

† Frequency-modulated channels can be treated in a similar way, and the results are the same.

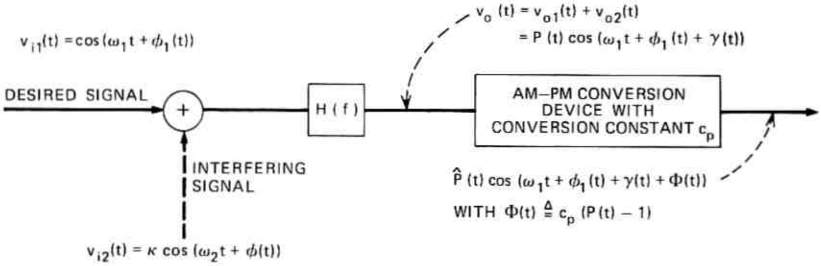

$v_o(t) = v_{o1}(t) + v_{o2}(t)$
$= P(t) \cos(\omega_1 t + \phi_1(t) + \gamma(t))$
$v_{i1}(t) = \cos(\omega_1 t + \phi_1(t))$
DESIRED SIGNAL
+
H(f)
AM–PM CONVERSION DEVICE WITH CONVERSION CONSTANT c_p
INTERFERING SIGNAL
$\hat{P}(t) \cos(\omega_1 t + \phi_1(t) + \gamma(t) + \Phi(t))$
WITH $\Phi(t) \triangleq c_p (P(t) - 1)$
$v_{i2}(t) = \kappa \cos(\omega_2 t + \phi(t))$

Fig. 2—Model for generation of intelligible crosstalk between two neighboring phase-modulated channels.

and in decibels:

$$\text{CIR} = 10 \log_{10} \mu \text{ (dB)}. \tag{25}$$

The signal $v_{i2}(t)$ is presumed to pass through the filter of channel 1 and produce the filter output:

$$\begin{aligned} v_{o2}(t) &= \int_{-\infty}^{\infty} du\, h(u) v_{i2}(t-u) \\ &= \kappa \int_{-\infty}^{\infty} du\, h(u) \cos[\omega_2(t-u) + \phi(t-u)] \end{aligned} \tag{26}$$

where $h(\cdot)$ is the filter impulse response. We will denote the filter's transfer function by:

$$H(f) = \int_{-\infty}^{\infty} h(u) e^{-j\omega u} du, \qquad \omega = 2\pi f. \tag{27}$$

Using the relation

$$\cos \alpha = \tfrac{1}{2}[e^{j\alpha} + e^{-j\alpha}] \tag{28}$$

and setting

$$A(t) \triangleq \int_{-\infty}^{\infty} du\, h(u) e^{-j\omega_2 u} e^{j\phi(t-u)}, \tag{29}$$

$$B(t) \triangleq \int_{-\infty}^{\infty} du\, h(u) e^{j\omega_2 u} e^{-j\phi(t-u)}, \tag{30}$$

we can rewrite (26) as:

$$v_{o2}(t) = \kappa[\tfrac{1}{2}A(t)e^{j\omega_2 t} + \tfrac{1}{2}B(t)e^{-j\omega_2 t}]. \tag{31}$$

It is easy to see that with

$$V(t) \triangleq [A(t)B(t)]^{\frac{1}{2}} \tag{32}$$

and

$$\phi_2(t) \triangleq \frac{1}{2j} \ln \frac{A(t)}{B(t)}, \tag{33}$$

$v_{o2}(t)$ equals:

$$v_{o2}(t) = \kappa V(t) \cos (\omega_2 t + \phi_2(t)). \tag{34}$$

We will assume that $|\kappa V(t)| < 1$. The output, $v_o(t)$, of the filter in channel 1 is:

$$\begin{aligned} v_o(t) &= v_{o1}(t) + v_{o2}(t) \\ &= \cos (\omega_1 t + \phi_1(t)) + \kappa V(t) \cos (\omega_2 t + \phi_2(t)) \\ &= [1 + \kappa^2 V^2(t) + 2\kappa V(t) \cos \theta(t)]^{\frac{1}{2}} \cos (\omega_1 t + \phi_1(t) + \gamma(t)) \\ &= P(t) \cos (\omega_1 t + \phi_1(t) + \gamma(t)) \end{aligned} \tag{35}$$

where

$$\begin{aligned} P(t) &\triangleq [1 + \kappa^2 V^2(t) + 2\kappa V(t) \cos \theta(t)]^{\frac{1}{2}}, \\ \theta(t) &\triangleq (\omega_2 - \omega_1)t + \phi_2(t) - \phi_1(t), \end{aligned} \tag{36}$$

and

$$\gamma(t) \triangleq \tan^{-1} \left[\frac{\kappa V(t) \sin \theta(t)}{1 + \kappa V(t) \cos \theta(t)} \right].$$

The amplitude function, $P(t)$, can be expanded in the power series:[17]

$$\begin{aligned} P(t) &= \sum_{n=0}^{\infty} C_n^{-\frac{1}{2}}(\cos \theta(t))(-1)^n(\kappa V(t))^n \\ &= 1 + \sum_{n=1}^{\infty} C_n^{-\frac{1}{2}}(\cos \theta(t))(-1)^n(\kappa V(t))^n \end{aligned}$$

where $\{C_n^{-\frac{1}{2}}(\cdot)\}$ are the Gegenbauer polynomials of degree n and order $-\frac{1}{2}$.

By definition,[18] if $a(t) \cos (\omega_c t + \psi(t))$ is the input to an AM-PM conversion device with conversion constant c_p (radians), then its output is $\hat{a}(t) \cos [\omega_c t + \psi(t) + c_p(a(t) - 1)]$. So if $v_o(t)$ passes through such a device, the undesired output phase in channel 1 is $\gamma(t) + \Phi(t)$ where

$$\Phi(t) \triangleq c_p(P(t) - 1) = c_p \sum_{n=1}^{\infty} C_n^{-\frac{1}{2}}(\cos \theta(t))(-1)^n(\kappa V(t))^n. \tag{37}$$

From Ref. 17, we also have:

$$C_n^{-\frac{1}{2}}(\cos \theta(t)) = \sum_{m=0}^{n} \frac{\Gamma(m - \frac{1}{2})\Gamma(n - m - \frac{1}{2})}{m!(n - m)![\Gamma(-\frac{1}{2})]^2} \cdot \cos [(n - 2m)\theta(t)]$$

where $\Gamma(\cdot)$ is the gamma function. Then,

$$\Phi(t) = c_p \sum_{n=1}^{\infty} \sum_{m=0}^{n} \frac{\Gamma(m - \frac{1}{2})\Gamma(n - m - \frac{1}{2})}{m!(n - m)![\Gamma(-\frac{1}{2})]^2} \cdot \cos [(n - 2m)\theta(t)] \cdot (-1)^n(\kappa V(t))^n. \tag{38}$$

Assuming that $f_2 - f_1$ ($f_i = \omega_i/2\pi$) is greater than the baseband frequencies of channel 1, we see from (36) that terms of the form $\cos [p\theta(t)]$, $p \neq 0$, do not contribute to the baseband interference in channel 1. In addition, it can be shown that $\gamma(t)$ is outside the baseband. Thus, retaining only the terms for n even and $m = n/2$ in (38), the undesired output phase or crosstalk interference is just:

$$\begin{aligned} y(t) &\triangleq c_p \sum_{n=1}^{\infty} \left[\frac{\Gamma(n - \frac{1}{2})}{\Gamma(-\frac{1}{2})n!} \right]^2 (\kappa V(t))^{2n} \\ &= c_p\left[F\left(-\tfrac{1}{2}, -\tfrac{1}{2}; 1; (\kappa V(t))^2 \right) - 1 \right] \end{aligned} \tag{39}$$

where $F(a, b; c; z)$ is the Gauss hypergeometric function[19] defined by:

$$F(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(a)} \sum_{n=0}^{\infty} \frac{\Gamma(a + n)\Gamma(b + n)}{\Gamma(c + n)} \frac{z^n}{n!}.$$

We next show that the crosstalk interference $y(t)$ has a generalized Volterra series in $\phi(t)$, the signal creating the interference, and we find the Volterra transfer function $G_1^{y,\phi}(f)$. We begin by rewriting (32) as:

$$\begin{aligned} V^2(t) &= \exp\,[\ln A(t) + \ln B(t)] \\ &= \exp\,[A_1(t) + B_1(t)] \end{aligned}$$

where

$$A_1(t) \triangleq \ln A(t) \quad \text{and} \quad B_1(t) \triangleq \ln B(t).$$

Recalling the definition of $A(t)$ in (29), we apply Property 3 [with $g(u) = h(u)e^{-j\omega_2 u}$, $\hat{h}(x) = e^{jx}$, and $F(z) = \ln z$] to get that $A_1(t) = \mathrm{GVS}[\phi(t); \{g_n^{A_1,\phi}\}]$ with

$$\begin{aligned} g_0^{A_1,\phi} &= F_0 = \ln z_0 = \ln H(f_2) \\ G_1^{A_1,\phi}(f) &= F_1\hat{h}_1 G(f) = jH(f_2 + f)/H(f_2). \end{aligned}$$

Similarly, for $B_1(t)$:

$$\begin{aligned} g_0^{B_1,\phi} &= \ln H(-f_2) \\ G_1^{B_1,\phi}(f) &= -jH(f - f_2)/H(-f_2). \end{aligned}$$

Setting $D(t) = A_1(t) + B_1(t)$, we have $D(t) = \mathrm{GVS}[\phi(t); \{g_n^{D,\phi}\}]$ and clearly $g_n^{D,\phi} = g_n^{A_1,\phi} + g_n^{B_1,\phi}$. Since $h(u)$ is real and $H(-f) = H^*(f)$ we have:

$$\begin{aligned} g_0^{D,\phi} &= \ln H(f_2) + \ln H(-f_2) = \ln |H(f_2)|^2 \\ G_1^{D,\phi}(f) &= j\,\frac{H(f_2 + f)}{H(f_2)} - j\,\frac{H(f - f_2)}{H(-f_2)}. \end{aligned}$$

Next we apply Property 4 to $V^2(t) = \exp\,[D(t)]$ with $\hat{F}(z) = e^z$ and $\hat{F}_0 = \hat{F}_1 = \cdots$ to get:

$$\begin{aligned} g_0^{V^2,\phi} &= \hat{F}_0 = \exp\,[g_0^{D,\phi}] = |H(f_2)|^2 \\ G_1^{V^2,\phi}(f) &= \hat{F}_1 G_1^{D,\phi}(f) = jH^*(f_2)H(f_2 + f) - jH(f_2)H^*(f_2 - f). \end{aligned}$$

Finally, we apply Property 4 to (39) with

$$\hat{F}(z) = c_p[F(-\tfrac{1}{2}, -\tfrac{1}{2}; 1; \kappa^2 z) - 1]$$

to get that $y(t) = \text{GVS}[\phi(t); \{g_n^{y,\phi}\}]$. Also

$$\begin{aligned}\hat{F}_0 &= \hat{F}(z) \text{ evaluated at } z = g_0^{V^2,\phi} \\ &= c_p[F(-\tfrac{1}{2}, -\tfrac{1}{2}; 1; |\kappa H(f_2)|^2) - 1]\end{aligned}$$

and

$$\hat{F}_1 = \frac{d}{dz}\,\hat{F}(z)\bigg|_{z=g_0^{V^2,\phi}} = \frac{c_p}{4}\,\kappa^2 F(\tfrac{1}{2}, \tfrac{1}{2}; 2; |\kappa H(f_2)|^2).$$

Hence,

$$\begin{aligned}g_0^{y,\phi} &= \hat{F}_0 = c_p[F(-\tfrac{1}{2}, -\tfrac{1}{2}; 1; |\kappa H(f_2)|^2) - 1] \\ G_1^{y,\phi}(f) &= \hat{F}_1 G_1^{V^2,\phi}(f) \\ &= \frac{c_p}{4}\,\kappa^2\, Q(\kappa|H(f_2)|)\cdot j[H^*(f_2)H(f_2+f) \\ &\qquad - H(f_2)H^*(f_2-f)] \qquad (40)\end{aligned}$$

where

$$\begin{aligned}Q(z) &\triangleq F(\tfrac{1}{2}, \tfrac{1}{2}; 2; z^2) \\ &= \frac{4}{\pi z^2}\,[E(z) - (1-z^2)K(z)] \qquad (41)\end{aligned}$$

and E and K denote complete elliptic integrals of modulus z (Ref. 19, pp. 47 and 358).

Then the intelligible crosstalk ratio equals:

$$\begin{aligned}R(f) &= |G_1^{y,\phi}(f)|^2 \\ &= \frac{c_p^2}{16}\,\kappa^4\, Q^2(\kappa|H(f_2)|)\bigg|H(f_2+f)H^*(f_2) \\ &\qquad - H^*(f_2-f)H(f_2)\bigg|^2 \qquad (42)\end{aligned}$$

where

$$\begin{aligned}Q(\kappa|H(f_2)|) &= 1 + \tfrac{1}{8}(\kappa|H(f_2)|)^2 \\ &\quad + \frac{3}{64}(\kappa|H(f_2)|)^4 + \frac{25}{1024}(\kappa|H(f_2)|)^6 + \cdots. \qquad (43)\end{aligned}$$

For a given value of κ (or CIR), we need only calculate the value of $Q(\kappa|H(f_2)|)$ once for any filter transfer function having attenuation $|H(f_2)|^2$. When $\kappa \leqq 1$ (or CIR $\geqq$ 0 dB) and $10 \log_{10} |H(f_2)|^2 \leqq -10$ dB, we can approximate, with very good accuracy, $Q(\kappa|H(f_2)|) \cong 1$, and then:

$$R(f) \cong \frac{c_p^2}{16}\,\kappa^4 |H(f_2+f)H^*(f_2) - H^*(f_2-f)H(f_2)|^2. \qquad (44)$$

If $C(f)$ and $\Theta(f)$ are the magnitude and phase of $H(f)$,

$$H(f) = C(f)e^{j\Theta(f)}, \tag{45}$$

then

$$|H(f_2 + f)H^*(f_2) - H^*(f_2 - f)H(f_2)|^2 = [C^2(f_2 + f) + C^2(f_2 - f)]C^2(f_2) - 2C(f_2 + f)C(f_2 - f)C^2(f_2) \cdot \cos[\Theta(f_2 + f) + \Theta(f_2 - f) - 2\Theta(f_2)]. \tag{46}$$

The last expression together with either (42) or (44) is well suited for computational purposes requiring only the values of the amplitude and phase of $H(\cdot)$ at frequencies f_2, $f_2 + f$, and $f_2 - f$.

One should note that in this analysis we have assumed that the filter gain at f_1 was unity. It is easy to see that, if the gain is not unity, the only difference in eqs. (42) to (46) is that $H(f)$ is replaced by the normalized transfer function $H(f)/|H(f_1)|$.

The expression for the intelligible crosstalk ratio given in (44) exhibits a number of properties usually associated with DACI.[1,11,16] For example, noting that $\mu = 1/\kappa^2$ and CIR $= 10 \log_{10} \mu$ (dB) and expressing the intelligible crosstalk ratio in decibels as $10 \log_{10} R(f)$ (dB), we see from (44) that if CIR decreases 1 dB then the crosstalk ratio increases 2 dB. We observe that the way in which we have defined $R(f)$ also makes $R(f)$ independent of the power of the input (phase). Moreover, by assuming that the amplitude of the desired signal in (21) is arbitrary (instead of unity), it is easy to check that, for fixed CIR, $R(f)$ is independent of the desired signal power.

V. EXAMPLE

The intelligible crosstalk ratio was calculated for the example considered by Lundquist[13] with CIR = 0 dB. The crosstalk ratio for other values of CIR can be found by adding 2 dB to the crosstalk ratio for each dB decrease in CIR. We assumed an AM-PM conversion constant of 5 degrees/dB[18] or $c_p = 5(0.1516) = 0.758$ radians, and a k-pole filter having transfer function:

$$H(f) = \frac{1}{\left[1 + j\left(\frac{f - f_1}{f_o}\right)\right]^k}. \tag{47}$$

Given the number of poles k, the frequency separation $\Delta f = f_2 - f_1$, and the value of the "attenuation at the adjacent channel" defined as $-10 \log_{10} |H(f_2)|^2$ (dB), we can determine f_o from (47). Equations (44) and (46) were used to compute $R(f)$ for various values of k, baseband frequency f, frequency separation Δf, and adjacent channel attenuation. The results are given in Figs. 3 to 5. Figure 3 shows the de-

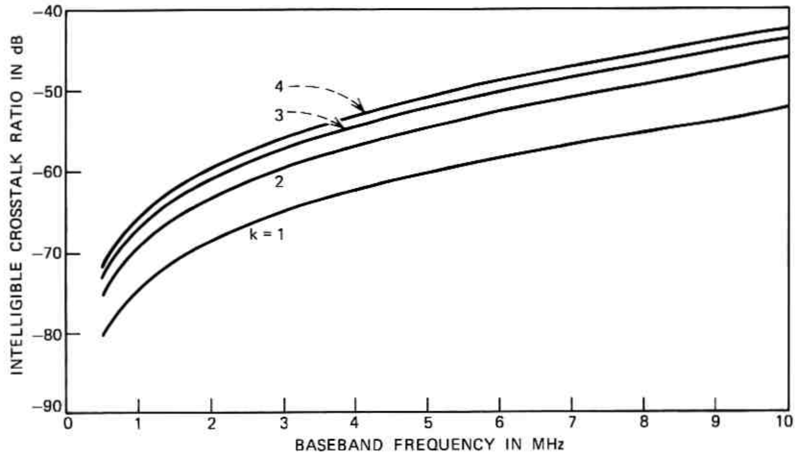


Fig. 3—Intelligible crosstalk ratio versus baseband frequency, for $\Delta f = 20$ MHz and adjacent channel attenuation = 20 dB.

pendence of the intelligible crosstalk ratio on the baseband frequency f, for fixed frequency separation $\Delta f = 20$ MHz and adjacent channel attenuation of 20 dB. We see from Fig. 3 that DACI is greater at higher

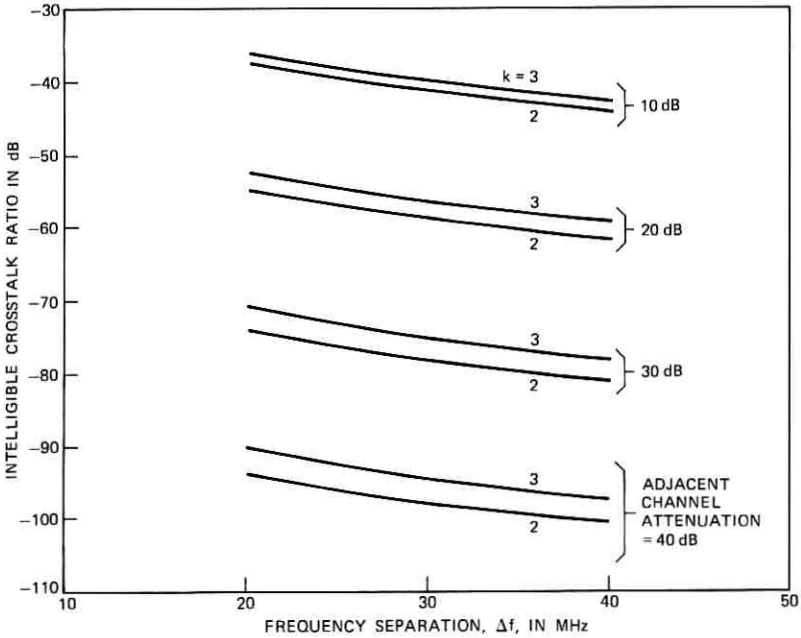


Fig. 4—Intelligible crosstalk ratio versus frequency separation, for baseband frequency = 5 MHz.

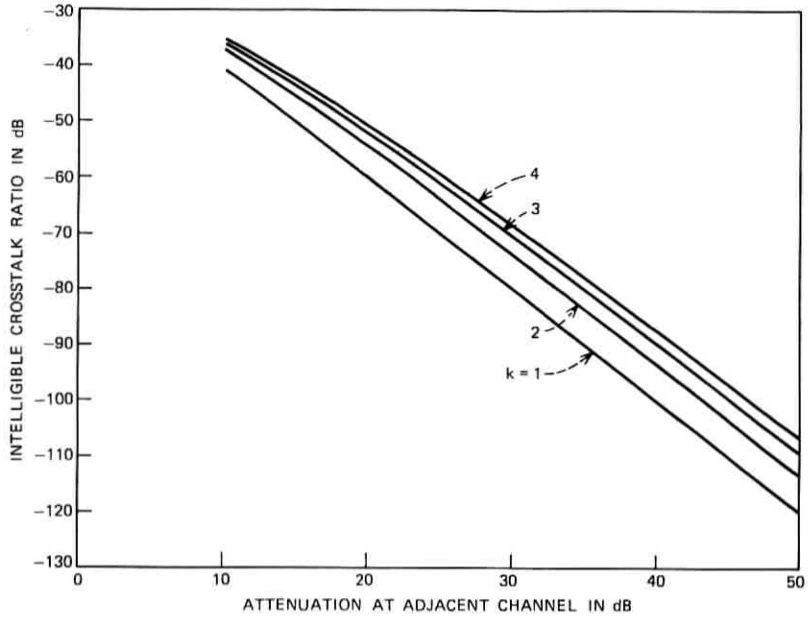


Fig. 5—Intelligible crosstalk ratio versus attenuation at adjacent channel, for Δf = 20 MHz and baseband frequency = 5 MHz.

baseband frequencies, increasing approximately 6 dB when f is doubled. For a fixed baseband frequency of 5 MHz, Fig. 4 shows the relation between $R(f)$ and the frequency separation Δf. We observe that there is not much variation of $R(f)$ with Δf for a given adjacent channel attenuation. In Fig. 5 we show the effect of increasing the adjacent channel attenuation for a fixed baseband frequency of 5 MHz and a frequency separation of 20 MHz. Here a 1-dB increase in attenuation produces about a 2-dB decrease in crosstalk ratio.

VI. CONCLUSION

By use of Volterra series analysis, we have presented a general mathematical description of the crosstalk interference created in a communications system. From this description, we were able to isolate the part of the crosstalk that was intelligible and to define the intelligible crosstalk ratio as a measure of the intelligible crosstalk created in the system. We then looked at a model in which intelligible crosstalk was generated between two neighboring PM channels. Using our results, we derived an expression for the intelligible crosstalk ratio for this model. This expression exhibited a number of properties usually associated with direct adjacent channel interference. The crosstalk ratio was computed for the case of a k-pole filter as a function of various parameters.

APPENDIX A

In this appendix we sketch the proof of Property 3. Following Bedrosian and Rice[4] we define the function[†]

$$\hat{H}(\xi) = \int_{-\infty}^{\infty} g(u)\hat{h}[\xi x(t-u)]du \tag{48}$$

so that from (9):

$$\hat{H}(0) = \hat{h}_0 \int_{-\infty}^{\infty} g(u)du = z_0. \tag{49}$$

From (10) we see that:

$$F[\hat{H}(0)] = F(z_0) = F_0. \tag{50}$$

Expanding the function $F[\hat{H}(\xi)]$ in a Maclaurin series we obtain:

$$F[\hat{H}(\xi)] = \sum_{n=0}^{\infty} \frac{\xi^n}{n!}\left[\frac{d^n}{d\xi^n}F[\hat{H}(\xi)]\right]_{\xi=0}. \tag{51}$$

Then

$$\begin{aligned} y(t) &= F[\hat{H}(1)] = F[\hat{H}(0)] + \sum_{n=1}^{\infty}\frac{1}{n!}\left[\frac{d^n}{d\xi^n}F[\hat{H}(\xi)]\right]_{\xi=0} \\ &= F_0 + \sum_{n=1}^{\infty}\frac{1}{n!}\left[\frac{d^n}{d\xi^n}F[\hat{H}(\xi)]\right]_{\xi=0}. \end{aligned} \tag{52}$$

Applying the results in eqs. (49), (114), and (115) of Ref. 4 we get:

$$y(t) = F_0 + \sum_{n=1}^{\infty}\frac{1}{n!}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} du_1\cdots du_n\, g_n^{y,x}(u_1,\cdots,u_n) \cdot \prod_{r=1}^{n} x(t-u_r) \tag{53}$$

for some kernels $\{g_n^{y,x}\}$ with

$$\begin{aligned} G_1^{y,x}(f) &= F_1\hat{h}_1 G(f), \\ G_2^{y,x}(f_1,f_2) &= F_1\hat{h}_2 G(f_1+f_2) + F_2\hat{h}_1^2 G(f_1)G(f_2), \end{aligned} \tag{54}$$

and

$$\begin{aligned} G_3^{y,x}(f_1,f_2,f_3) &= F_1\hat{h}_3 G(f_1+f_2+f_3) \\ &\quad + F_2\hat{h}_1\hat{h}_2[G(f_1)G(f_2+f_3) + G(f_2)G(f_1+f_3) + G(f_3)G(f_1+f_2)] \\ &\quad + F_3\hat{h}_1^3 G(f_1)G(f_2)G(f_3). \end{aligned}$$

[†] The dependence of $\hat{H}(\xi)$ on t will be suppressed.

The higher-order Volterra transfer functions are given by eq. (49) of Ref. 4. Thus, $y(t) = \text{GVS}[x(t); \{g_n^{y,x}\}]$ and $g_0^{y,x} = F_0$ which is the desired result.

APPENDIX B

Here we derive Property 4. Define the function $\hat{H}(\xi)$ by:[†]

$$\hat{H}(\xi) = g_0^{y,x} + \sum_{n=1}^{\infty} \frac{\xi^n}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{y,x}(u_1, \cdots, u_n) \cdot \prod_{r=1}^{n} x(t - u_r). \quad (55)$$

Then,

$$\hat{H}(0) = g_0^{y,x}$$

and the νth derivative of $\hat{H}(\xi)$ evaluated at $\xi = 0$ equals:

$$\hat{H}^{(\nu)}(0) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_\nu \, g_\nu^{y,x}(u_1, \cdots, u_\nu) \prod_{r=1}^{\nu} x(t - u_r), \quad \nu \geqq 1. \quad (56)$$

Next,

$$y(t) = \hat{H}(1)$$

and

$$w(t) = \hat{F}[y(t)] = \hat{F}[\hat{H}(1)]. \quad (57)$$

Note that with $\hat{F}^{(l)}(z)$ denoting the lth derivative of $\hat{F}(z)$:

$$\hat{F}^{(l)}[\hat{H}(0)] = \hat{F}^{(l)}[g_0^{y,x}] = \hat{F}_l, \qquad l \geqq 0. \quad (58)$$

Expanding $\hat{F}[\hat{H}(\xi)]$ in a Maclaurin series,

$$\hat{F}[\hat{H}(\xi)] = \sum_{n=0}^{\infty} \frac{\xi^n}{n!} \left[\frac{d^n}{d\xi^n} \hat{F}[\hat{H}(\xi)] \right]_{\xi=0}, \quad (59)$$

we get:

$$w(t) = \hat{F}[\hat{H}(1)] = \hat{F}[\hat{H}(0)] + \sum_{n=1}^{\infty} \frac{1}{n!} \left[\frac{d^n}{d\xi^n} \hat{F}[\hat{H}(\xi)] \right]_{\xi=0}$$

$$= \hat{F}_0 + \sum_{n=1}^{\infty} \frac{1}{n!} \left[\frac{d^n}{d\xi^n} \hat{F}[\hat{H}(\xi)] \right]_{\xi=0}. \quad (60)$$

Using the results in eqs. (98) and (112) through (115) of Ref. 4 we can write (60) as:

$$w(t) = \hat{F}_0 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} du_1 \cdots du_n \, g_n^{w,x}(u_1, \cdots, u_n) \cdot \prod_{r=1}^{n} x(t - u_r) \quad (61)$$

[†] The dependence of $\hat{H}(\xi)$ on t will be suppressed.

where, in particular,

$$\begin{aligned} g_1^{w,x}(u_1) &= \hat{F}_1\, g_1^{y,x}(u_1), \\ g_2^{w,x}(u_1, u_2) &= \hat{F}_1\, g_2^{w,x}(u_1, u_2) + \hat{F}_2[g_1^{y,x}(u_1) g_1^{y,x}(u_2)], \end{aligned} \tag{62}$$

and

$$\begin{aligned} g_3^{w,x}(u_1, u_2, u_3) &= \hat{F}_1\, g_3^{y,x}(u_1, u_2, u_3) \\ &\quad + \hat{F}_2[g_1^{y,x}(u_1) g_2^{y,x}(u_2, u_3) + g_1^{y,x}(u_2) g_2^{y,x}(u_1, u_3) + g_1^{y,x}(u_3) g_2^{y,x}(u_1, u_2)] \\ &\quad + \hat{F}_3[g_1^{y,x}(u_1) g_1^{y,x}(u_2) g_1^{y,x}(u_3)]. \end{aligned} \tag{63}$$

Therefore, $w(t) = \mathrm{GVS}[x(t);\ \{g_n^{w,x}\}]$ with $g_0^{w,x} = \hat{F}_0$ and the Volterra transfer functions given in the statement of Property 4.

REFERENCES

1. Kinzer, J. P., and Laidig, J. F., "Engineering Aspects of the TH Microwave Radio Relay System," B.S.T.J., *40*, No. 6 (November 1961), pp. 1486–1487.
2. Members of the Technical Staff of Bell Telephone Laboratories, *Transmission Systems for Communications*, 4th edition, 1970, pp. 279–304.
3. Ibid., pp. 237–278.
4. 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.
5. Wiener, N., *Nonlinear Problems in Random Theory*, Cambridge, Massachusetts: Technology Press of M.I.T., New York: Wiley & Sons, 1958.
6. Ruthroff, C. L., "A Mechanism for Direct Adjacent Channel Interference," Proc. IRE, *49* (June 1961), pp. 1091–1092.
7. George, D. A., "Continuous Nonlinear Systems," M.I.T. Res. Lab. Elec., Cambridge, Massachusetts, Tech. Rep. 355, July 24, 1959.
8. Schetzen, M., "Measurements of the Kernels of a Nonlinear System of Finite Order," Int. J. Control, *1* (March 1965), pp. 251–263.
9. Lee, Y. W., and Schetzen, M., "Measurement of the Wiener Kernels of a Nonlinear System by Cross-Correlation," Int. J. Control, *2*, 1965, pp. 237–254.
10. Bennett, W. R., unpublished work, 1960.
11. Curtis, H. E., unpublished work, 1961.
12. Hatch, R. W., unpublished work, 1963.
13. Lundquist, L., unpublished work, 1969.
14. Mircea, A., and Sinnreich, H., "Distortion Noise in Frequency-Dependent Nonlinear Networks," Proc. IEE, *116*, 1969, pp. 1644–1648.
15. Roetken, A. A., Smith, K. D., and Friis, R. W., "The TD-2 Microwave Radio Relay System," B.S.T.J., *30*, No. 4 (October 1951), pp. 1041–1077.
16. Curtis, H. E., Collins, T. R. D., and Jamison, B. C., "Interstitial Channels for Doubling TD-2 Radio System Capacity," B.S.T.J., *39*, No. 6 (November 1960), pp. 1505–1527.
17. Magnus, W., Oberhettinger, F., and Soni, R. P., *Formulas and Theorems for the Special Functions of Mathematical Physics*, New York: Springer-Verlag, 1966, pp. 218–227.
18. Cross, T. G., "Intermodulation Noise in FM Systems Due to Transmission Deviations and AM/PM Conversion," B.S.T.J., *45*, No. 10 (December 1966), pp. 1749–1773.
19. Magnus, Oberhettinger, and Soni, op. cit., pp. 37–65.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

The Potential in a Charge Coupled Device With No Mobile Minority Carriers And Zero Plate Separation

By J. McKENNA and N. L. SCHRYER

(Manuscript received November 1, 1972)

A two-dimensional analysis of the potential in charge coupled devices is presented. It is assumed that there are no mobile minority carriers, that the plate separation is zero, and that the plate voltage does not vary with time. The depletion layer approximation is used to linearize the equations, which are then solved exactly with the use of Fourier series. Both surface and buried channel devices are analyzed. These solutions can typically be evaluated on a computer in less than a tenth of the time it takes to obtain a solution by the method of finite differences. The solutions obtained here provide an important tool for the designer of charge coupled devices. In addition to describing the method of obtaining the solutions, we evaluate them to show the effects of a number of different design parameters, and compare the cost of these solutions with the cost of obtaining finite difference solutions.

I. INTRODUCTION AND SUMMARY

The recent invention[1,2] and development of charge coupled devices (CCD's) has led to renewed interest in the mathematical analysis of MIS-type structures. Ideally, one would like to solve the nonlinear equations describing the three-dimensional motion of charge as a function of the time-varying plate voltages. So far no one has succeeded in doing this for even the simplest geometries. For the most part, one-dimensional static models have been solved which yield only qualitative information about the behavior of such devices. A much more sophisticated, one-dimensional, time-varying model of a CCD has been developed and analyzed by Schryer and Strain.[3]

A static, two-dimensional model of a CCD has also been studied by Amelio[4] using finite difference techniques. He calculated the potential distribution in a two-dimensional model in the absence of mobile charge and with given static plate potentials. The results of this calculation are of great interest. The use of finite difference techniques in these cases has drawbacks, however. In even the relatively simple geometries considered so far, it is expensive to obtain reasonably accurate solutions for the potentials, and for more complicated devices, it soon becomes prohibitively expensive. Furthermore, as we shall show, even for simple geometries it is difficult to obtain accurate expressions for the fields from the finite difference solutions for the potentials.

In this paper, we show that when the plates on a CCD are close enough together so that they can be assumed to be abutted, and when the depletion layer approximation can be used,[5] the resulting linearized model can be solved analytically. These solutions can then be evaluated cheaply and quickly on a computer. This analysis will be valid for both surface and buried channel CCD's with an arbitrary number of plates. In a separate paper, we will show that these solutions can be used to obtain solutions for the potential in a CCD when there are gaps between the plates.[6]

In Section II we write down the equations describing the model and put them into appropriate dimensionless form. We then introduce the depletion layer approximation which linearizes the equations and discuss conditions under which this approximation is valid.

This paper has two main purposes: to show the behavior of the potentials and fields in a CCD and to demonstrate techniques by which these potentials and fields can be calculated cheaply and accurately. In Section III we present a discussion of how the solutions depend on the various parameters defining the devices.

In Section IV we derive the solution of the linearized potential equations. The reader interested only in the physical design of CCD's can skip the rest of the paper.

In Section V we discuss in some detail the solution by finite difference methods of the exact, nonlinear equations describing a surface CCD. Our purpose in doing this is twofold. We wish to show the difficulties involved in obtaining an accurate solution cheaply, especially if an accurate knowledge of the fields is required. Secondly, we want accurate solutions of the exact problem to compare with the analytic solutions of the linearized problem.

Finally, in Section VI we compare in detail some solutions of the exact problem obtained by finite differences with the corresponding analytic solutions of the linearized equations. It is shown that in many cases of interest the solutions of the linearized problem provide excellent approximations to the true potential and much more accurate approximations to the gradient of the potential than can be obtained from the finite difference solutions. Furthermore, the solutions of the linearized equations are at least an order of magnitude cheaper to obtain than are the finite difference solutions for any reasonable accuracy.

II. DERIVATION OF THE EQUATIONS

A surface CCD[1] consists of a layer of silicon covered with a thin insulating layer of silicon dioxide, and on top of the oxide layer, a sequence of closely spaced electrodes. Such a device is shown schematically in Fig. 1 with some typical dimensions indicated. Mobile charge trapped at the oxide-semiconductor interface is transferred from plate to plate by appropriately changing the potential of the plates. We consider the case where the substrate is n-type silicon and the mobile charges are injected holes. In this case, the plate potentials must be negative. Our analysis can be modified in an obvious way to describe the case where the substrate is p-type silicon and the mobile charges are electrons.

Some losses are introduced by the trapping of the mobile charges by surface states at the oxide-semiconductor interface. Smith and Boyle[7,8] have proposed a solution to this problem by inserting between the oxide and the substrate an additional layer of p-type silicon, thus forming a buried channel CCD. The p-layer is kept completely ionized, which

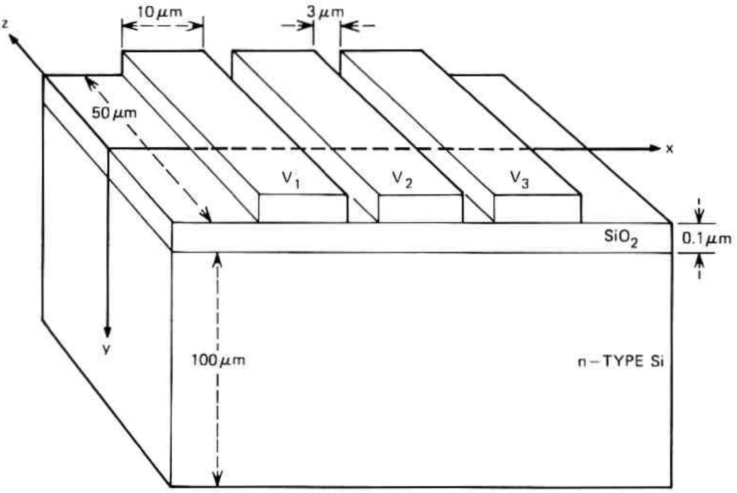


Fig. 1—A schematic diagram of a surface CCD.

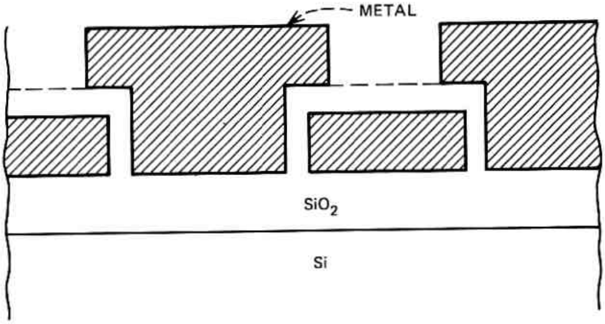


Fig. 2—A schematic diagram illustrating plate overlap structure.

causes the potential minimum to occur near the center of the p-layer. Thus the mobile positive charge stays at the potential minimum, safely away from the surface traps.

It is desirable to have the plates as close together as possible in order to minimize the transit time of the mobile charge between plates. The minimum plate separation presently obtainable from photolithography is ~3–5 μm, but a plate overlap structure,[8,9] as shown in Fig. 2, or undercut isolation[10] allows for plate separation of ~0.1 μm.

We propose to study the static potential in either a surface or buried channel CCD with plate overlap structure, in the absence of mobile charge. We begin by noting that since the length in the z-direction of each plate is much greater than its width in the x-direction, near the center of the plates ($z = 0$) the field is essentially two-dimensional. Hence we will treat the problem as two-dimensional. We assume that the plates are zero distance apart. Since in the overlap structure there should be little flux leakage between the plates, we feel this is a reasonable approximation. We make the additional assumption that the bottom substrate is infinitely thick. The field can penetrate into the substrate little beyond a depletion depth, and since for typical voltages the depletion depth ranges from 7 μm to 20 μm, and the thickness of the substrate in a typical device is 100 μm, this is a very reasonable approximation. Finally, we assume the structure is periodic in the x-direction, which in the usual mode of operation is an excellent approximation.

We begin by defining the boundary value problem describing a buried channel device. In all that follows, starred quantities have rationalized MKS dimensions; unstarred quantities, except for a few obvious physical parameters, are dimensionless. In the strip $0 \leqq x^* \leqq L^*$, let φ_1^* denote the potential in the oxide layer, $0 \leqq y^* \leqq h_1^*$; φ_2^* the potential in the p-layer, $h_1^* \leqq y^* \leqq h_2^*$; and φ_3^* the po-

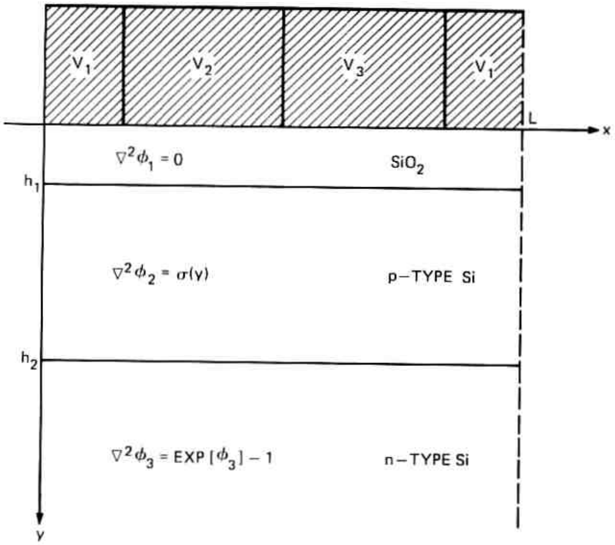

V1
V2
V3
V1
L
x
$\nabla^2\phi_1 = 0$
SiO2
h1
$\nabla^2\phi_2 = \sigma(y)$
p−TYPE Si
h2
$\nabla^2\phi_3 = \mathrm{EXP}\,[\phi_3] - 1$
n−TYPE Si
y

Fig. 3—A schematic diagram of one cell of a three-phase, buried channel CCD.

tential in the substrate, $h_2^* \leqq y^* < \infty$ (see Fig. 3). Then

$$\nabla_*^2 \varphi_1^* = 0, \tag{1}$$

$$\nabla_*^2 \varphi_2^* = \frac{e}{\epsilon_2} N_A^*(y^*) \tag{2}$$

$$\nabla_*^2 \varphi_3^* = -\frac{eN_D^*}{\epsilon_2}\left(1 - \exp\left(e\varphi_3^*/\mathrm{k}T\right)\right). \tag{3}$$

In eqs. (1) through (3), ∇_*^2 is the two-dimensional Laplacian; $-e$ is the charge of an electron; $N_A^*(y^*)$ is the acceptor number density in the p-region; N_D^* is the donor number density in the n-region; ϵ_1 and ϵ_2 are the permittivity of the oxide and silicon, respectively; k is Boltzman's constant; and T is the absolute temperature. The conditions under which eq. (3) can be expected to be valid are discussed in Ref. 11.

In the usual method of fabricating a CCD, the substrate donor number density, N_D^*, is a constant, independent of position. However, in a buried channel CCD, the p-layer is formed by diffusing the acceptor ions into the n-type Si, and so $N_A^*(y^*)$ is typically a function of $y^* - h_1^*$, the distance from the oxide surface. In many cases we have the representation[12]

$$N_A^*(y^*) = C_S^* \exp\left\{-\left(\frac{y^* - h_1^*}{h_2^* - h_1^*}\right)^2 \ell n \frac{C_S^*}{N_D^*}\right\} - N_D^* \tag{4}$$

where C_S^* is the number density of acceptor ions at the upper surface of the Si. The average value, $\bar{N}_A^*$, of $N_A^*(y)$ is easily shown to be

$$\begin{aligned}\bar{N}_A^* &= \frac{1}{h_2^* - h_1^*}\int_{h_1^*}^{h_2^*} N_A(y^*)dy^* \\ &= \frac{\sqrt{\pi}}{2}\frac{C_S^*}{\sqrt{\ell n(C_S^*/N_D^*)}}\,\text{erf}\,(\sqrt{\ell n(C_S^*/N_D^*)}) - N_D^* \end{aligned}\tag{5}$$

where erf (x) is the error function.[13]

Before writing down the boundary conditions, we introduce dimensionless variables. Define the Debye length

$$\lambda_D = (\epsilon_2 \text{k} T/e^2 N_D^*)^{\frac{1}{2}}.\tag{6}$$

Then normalize all lengths with respect to λ_D,

$$\begin{gathered} x = x^*/\lambda_D, \qquad y = y^*/\lambda_D, \qquad L = L^*/\lambda_D, \\ h_k = h_k^*/\lambda_D, \qquad (k = 1, 2), \end{gathered}\tag{7}$$

and define

$$\varphi_\alpha(x, y) = e\varphi_\alpha^*(x^*, y^*)/\text{k}T, \qquad (\alpha = 1, 2, 3),\tag{8}$$

$$\sigma(y) = N_A(y^*)/N_D^*, \qquad \eta = \epsilon_1/\epsilon_2.\tag{9}$$

Equations (1) through (3) become

$$\nabla^2\varphi_1 = 0,\tag{10}$$

$$\nabla^2\varphi_2 = \sigma(y),\tag{11}$$

$$\nabla^2\varphi_3 = e^{\varphi_3} - 1.\tag{12}$$

The boundary conditions can be written now as follows: For $0 \leqq x \leqq L$,

$$\varphi_1(x, 0) = V(x),\tag{13}$$

$$\varphi_1(x, h_1) = \varphi_2(x, h_1), \quad \eta\frac{\partial\varphi_1}{\partial y}(x, h_1) = \frac{\partial\varphi_2}{\partial y}(x, h_1) + Q(x),\tag{14}$$

$$\varphi_2(x, h_2) = \varphi_3(x, h_2), \quad \frac{\partial\varphi_2}{\partial y}(x, h_2) = \frac{\partial\varphi_3}{\partial y}(x, h_2),\tag{15}$$

$$\varphi_3(x, \infty) = 0,\tag{16}$$

and

$$\begin{gathered}\varphi(0, y) = \varphi(L, y) \\ \frac{\partial\varphi}{\partial x}(0, y) = \frac{\partial\varphi}{\partial x}(L, y), \qquad 0 \leqq y < \infty.\end{gathered}\tag{17}$$

In (13), $V(x)$ is a given, periodic function, assuming on each electrode the constant voltage of the electrode; and in (14), $Q(x)$ is a known, periodic surface charge density, which may include deliberately implanted charges.[14]

For future use, we record the expression for $\sigma(y)$ when $N_A(y^*)$ is given by (4). If

$$C_S = C_S^*/N_D^*, \tag{18}$$

then

$$\sigma(y) = C_S \exp\left\{-\left(\frac{y-h_1}{h_2-h_1}\right)^2 \ell n C_S\right\} - 1, \tag{19}$$

and

$$\bar{\sigma} = \frac{\sqrt{\pi}}{2}\,\frac{C_S \operatorname{erf}(\sqrt{\ell n C_S})}{\sqrt{\ell n C_S}} - 1. \tag{20}$$

The equations for the surface CCD are essentially the same, except that the p-layer is eliminated. In what follows, we will only give the analysis for the buried channel CCD. The results for the surface CCD can be obtained from those for the buried channel CCD by setting $\sigma = 0$, $h_1 = h_2$, and $\varphi_3 = \varphi_2$.

We now introduce the important depletion layer approximation.[5]

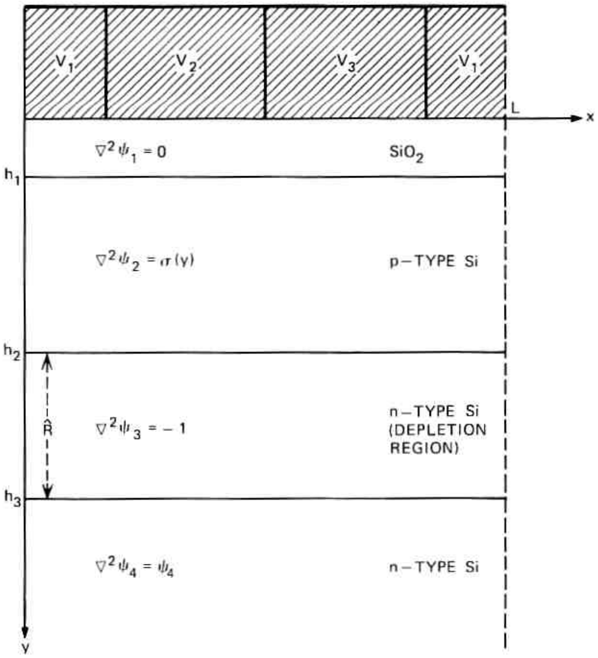


Fig. 4—A schematic diagram of the depletion layer approximation for one cell of a three-phase, buried channel CCD.

In most cases of interest, $\varphi_3(x, h_2) \ll -1$ for $0 \leqq x \leqq L$. For example, in a typical buried channel CCD, $\varphi_3(x, h_2) \sim -1000$ [and in a typical surface CCD, $\varphi_2(x, h_1) \sim -150$]. Thus, for $y - h_2$ small and positive, $e^{\varphi_3} \sim 0$. However, for $y \gg h_2$, $|\varphi_3| \ll 1$, and $e^{\varphi_3} \cong 1 + \varphi_3$. There is thus some curve $y = R(x)$ such that for $h_2 \leqq y \leqq R(x)$, $e^{\varphi_3} - 1 \approx -1$. The region $h_2 \leqq y \leqq R(x)$ is the depletion region. For $R(x) < y$, we have $e^{\varphi_3} - 1 \sim \varphi_3$. If $R(x)$ varies but little about its average value, $\hat{R}$, these remarks suggest that we replace eqs. (10) through (12) by the system of linear equations

$$\nabla^2\psi_1(x, y) = 0, \qquad 0 < y < h_1, \tag{21}$$

$$\nabla^2\psi_2(x, y) = \sigma(y), \qquad h_1 < y < h_2, \tag{22}$$

$$\nabla^2\psi_3(x, y) = -1, \qquad h_2 < y < h_3 = h_2 + \hat{R}, \tag{23}$$

$$\nabla^2\psi_4(x, y) = \psi_4(x, y), \qquad h_3 = h_2 + \hat{R} < y < \infty \tag{24}$$

where ψ_3 is the potential in $h_2 \leqq y \leqq h_3$ and ψ_4 is the potential in $h_3 \leqq y < \infty$. (See Fig. 4.) In addition to ψ_1, ψ_2, and ψ_3 satisfying boundary conditions (13) through (15), we have the boundary conditions, for $0 \leqq x \leqq L$,

$$\psi_3(x, h_3) = \psi_4(x, h_3), \qquad \frac{\partial\psi_3}{\partial y}(x, h_3) = \frac{\partial\psi_4}{\partial y}(x, h_3), \tag{25}$$

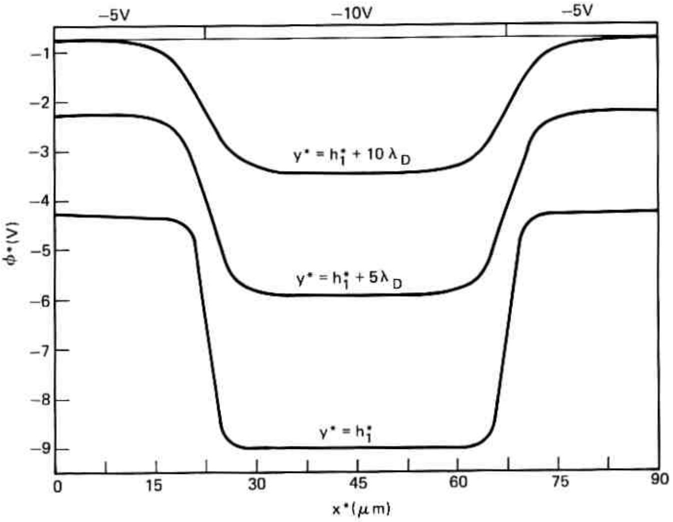


Fig. 5—The potential $\varphi^*(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2$ μm, 2.275 μm, and 4.35 μm. The 45-μm plates are alternately at −5 V and −10 V, and the oxide thickness is $h_1^* = 0.2$ μm.

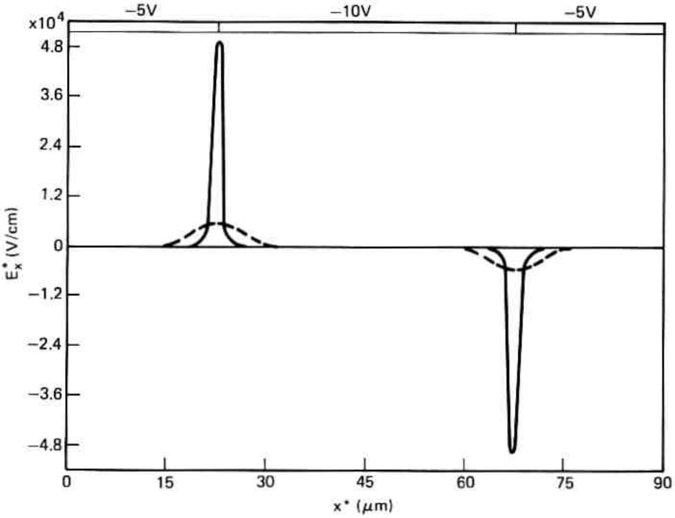


Fig. 6—The field $-(\partial\varphi^*/\partial x^*)(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2\ \mu$m. The 45-μm plates are alternately at -5 V and -10 V, and the oxide thickness is $h_1^* = 0.2\ \mu$m. The dashed curve is a plot of $-(\partial\varphi^*/\partial x^*)$ for the same device obtained from a finite difference calculation.

$$\psi_4(x, \infty) = 0, \tag{26}$$

and the ψ_α ($\alpha = 1, 2, 3, 4$) all satisfy (17).

It has been shown that for a one-dimensional version of this problem, the choice[5]

$$\hat{R} = -\left(1 + h_2 - h_1 + \frac{h_1}{\eta}\right) + \left[\left(h_2 - h_1 + \frac{h_1}{\eta}\right)^2 - 1 - 2V_o - \frac{2h_1}{\eta} Q_{SS} + 2\int_{h_1}^{h_2}\left(\xi - h_1 + \frac{h_1}{\eta}\right)\sigma(\xi)d\xi\right]^{\frac{1}{2}} \tag{27}$$

yields a solution which approximates the solution of the nonlinear problem very accurately in the region $h_1 \leqq y \leqq h_2$. Furthermore, the solution in this region is quite insensitive to the choice of $\hat{R}$. Since an accurate knowledge of the potential is only necessary in the p-layer for the buried channel CCD and near the oxide-semiconductor interface for the surface CCD, we feel this approximation is well justified. In this two-dimensional problem, we determine $\hat{R}$ from (27) by letting V_o and Q_{SS} be the averages of $V(x)$ and $Q(x)$:

$$V_o = \frac{1}{L}\int_0^L V(x)dx, \qquad Q_{SS} = \frac{1}{L}\int_0^L Q(x)dx. \tag{28}$$

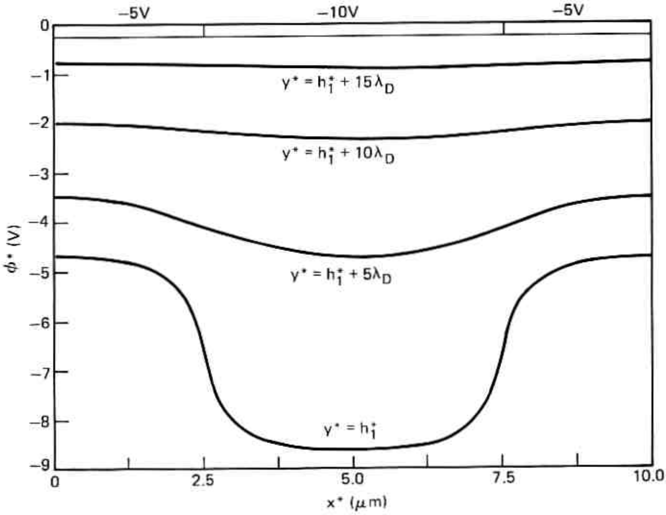


Fig. 7—The potential $\varphi^*(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2$ μm, 2.275 μm, 4.35 μm, and 6.425 μm. The 5-μm plates are alternately at −5 V and −10 V, and the oxide thickness is $h_1^* = 0.2$ μm.

In Section VI we will present a comparison of the solution of the linearized equations with the solution of the nonlinear equations obtained by finite difference methods. This will confirm for this example that the approximate solutions are accurate as claimed.

III. GENERAL BEHAVIOR OF THE POTENTIALS AND FIELDS

In this section we present graphical representations of the potentials and fields for both surface and buried channel CCD's for a number of design parameters. The graphs were obtained by evaluating the analytic expressions for the solutions, derived in Section IV, of eqs. (13) through (15), (17), and (21) through (26).

In all cases, we assume that the doping in the n-type substrate is $N_D = 10^{14}/\text{cm}^3$, that $\epsilon_2/\epsilon_0 = 12$, where ϵ_0 is the permittivity of free space, that $\epsilon_1/\epsilon_2 = \frac{1}{3}$, and that $Q(x) \equiv 0$, i.e., there is no trapped or implanted charge at the semiconductor-oxide interface. Then at $T = 300°\text{K}$, the value of λ_D defined in (6) is $\lambda_D = 4.15 \times 10^{-5}$ cm. Also, in all the examples presented here, we have used the factor $(\text{k}T/e) = 0.025$ V to convert dimensionless potentials to volts, and the factor $(\text{k}T/e\lambda_D) = 600$ V/cm to convert dimensionless fields to V/cm.

We consider first the effect of plate width in surface devices. The first pair of graphs illustrate a surface CCD with 45-μm plates, the second pair a surface CCD with 5-μm plates, and the third pair a surface CCD with 1.5-μm plates. The oxide layer in each of these CCD's is 0.2 μm thick, and the voltages on the plates are alternately -5 V and -10 V. These examples show the storage mode, and as a result there is no asymmetry to introduce a perferred direction of flow for the holes. However, they do exhibit the effects of plate width, and are easy to compare with finite difference calculations. In Figs. 5, 7, and 9, φ^* is plotted along the oxide-semiconductor interface ($y^* = h_1^*$) and along the lines $y^* = h_1^* + 5\lambda_D$, $h_1^* + 10\lambda_D$, and $h_1^* + 15\lambda_D$ inside the substrate. In Figs. 6, 8, and 10, $-(\partial\varphi^*/\partial x^*) = E_x^*$ is plotted along the oxide-semiconductor interface. The dashed curve is the field calculated by finite difference methods. The discrepancy between the two curves will be discussed in Section V.

In all three cases, the peak field available for moving positive charge from the left-hand plate to the center plate is about 4.8 × 10⁴ V/cm. However, in the 45-μm plate device, the field penetrates only about 7 μm under the plate from the edge, leaving most of the region under the plate field free. This would clearly be a very poor CCD. On the other hand, in the 1.5-μm device, there are substantial fields under the whole plate. These graphs show that if field penetration under the plates were the sole criterion, the narrower the plates the better. How-

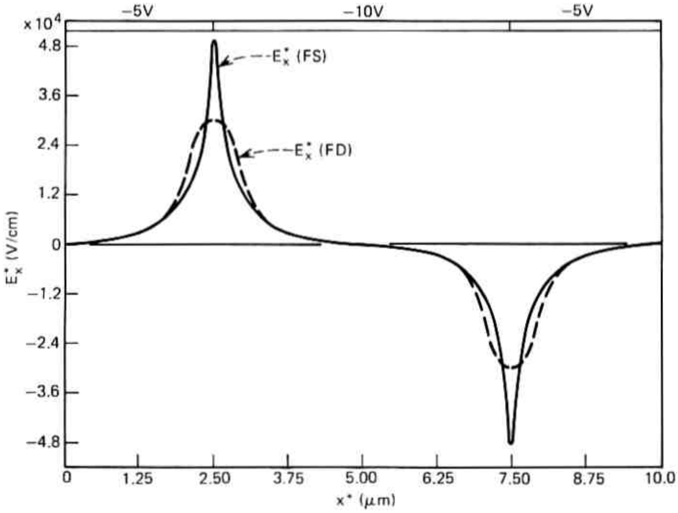


Fig. 8—The field $-(\partial\varphi^*/\partial x^*)(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2$ μm. The 5-μm plates are alternately at -5 V and -10 V, and the oxide thickness is $h_1^* = 0.2$ μm. The dashed curve is a plot of $-(\partial\varphi^*/\partial x^*)$ for the same device obtained from a finite difference calculation.

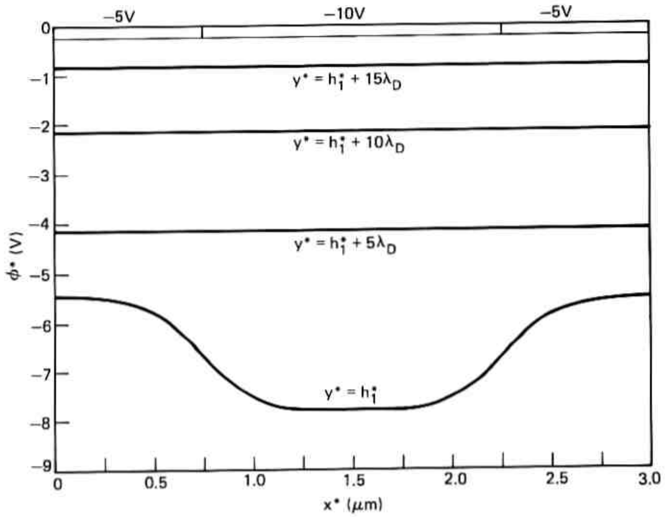


Fig. 9—The potential $\varphi^*(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2$ μm, 2.275 μm, 4.35 μm, and 6.425 μm. The 1.5-μm plates are alternately at -5 V and -10 V, and the oxide thickness is $h_1^* = 0.2$ μm.

ever, recent work by Tompsett[15] has shown that in surface CCD's the difference in losses of ones and fat zeros due to surface states becomes greater as the plate width decreases. His work shows that this puts a lower bound on plate widths in the neighborhood of 5 μm. However, Fig. 8 shows that for 5-μm plates there is still considerable field penetration under the plates.

Our calculations show that increasing (decreasing) the thickness of the oxide layer decreases (increases) the peak values of the fields, but does not materially affect the penetration of the fields under the plates.

We next consider buried channel CCD's. As for surface devices in general, the narrower the plates the better as far as field penetration is concerned. However, the plate width is apt to be determined by current photolithography tolerances, so this is a parameter not easily varied. In addition, if the plates are too narrow, the charge-carrying capacity of the CCD becomes very small.

Instead of considering the effects of plate width, we examine what happens for a given plate width if the thickness of the p-type layer is varied. We consider first a three-phase, buried channel CCD with 5-μm plates. The plates are at -5 V, -10 V, and -15 V, so charge is to be moved from under the -10 V plate to under the -15-V plate. The thickness of the oxide layer is $h_1^* = 0.1$ μm. The doping profile

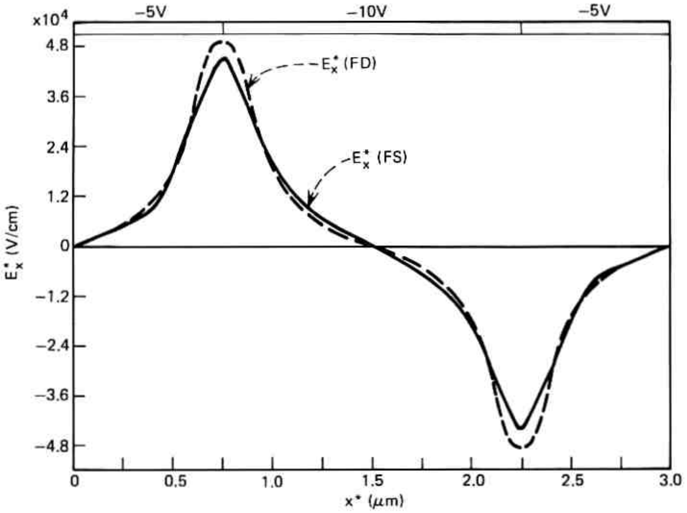


Fig. 10—The field $-(\partial\varphi^*/\partial x^*)(x^*, y^*)$ plotted as a function of x^* in a surface CCD for $y^* = 0.2$ μm. The 1.5-μm plates are alternately at -5 V and -10 V, and the oxide thickness is $h_1^* = 0.2$ μm. The dashed curve is a plot of $-(\partial\varphi^*/\partial x^*)$ for the same device obtained from a finite difference calculation.

in the p-type layer is assumed to be given by eq. (4) with $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$ ($C_S = 46$). This corresponds to an average number density of acceptor atoms of $2 \times 10^{15}/\text{cm}^3$. The remaining physical parameters are as described at the beginning of the section. In Fig. 11 we plot the channel field, $E_x^* = -(\partial\varphi^*/\partial x^*)$, (that is, the field at the potential minimum in the p-region) as a function of x^* for three different p-type layer thicknesses: $h_2^* - h_1^* = 0.1$ μm, 2 μm, and 4 μm. In Fig. 12 we plot the corresponding channel potentials, φ^*, (that is, the value of the potential at the potential minimum in the p-layer). The CCD with $h_2^* - h_1^* = 0.1$ μm is essentially a surface device. As the thickness of the p-layer is increased, the minimum value of the field under the center of the -10-V plate increases at first, while the peak value of the field decreases. Eventually, as the thickness of the p-layer is increased, the channel will be so far below the plates that the channel fields will start decreasing to zero. Thus, in terms of field penetration, there appears to be an optimal p-layer thickness. From Fig. 12, it is clear that as the p-layer gets thicker, the channel potential curve flattens out, and so the charge-carrying capacity of the CCD decreases.

We have also studied the effects of varying the doping in the p-layer (i.e., varying C_S^*). The behavior of the fields is relatively insensitive to changes in C_S^*.

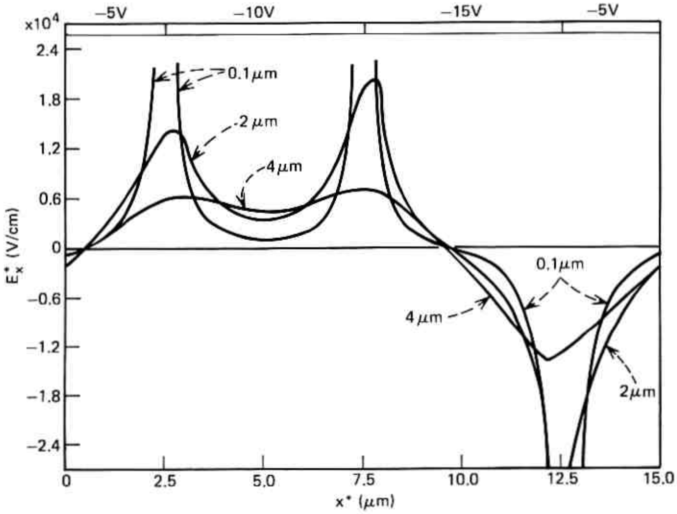


Fig. 11—The channel field $-(\partial\varphi^*/\partial x^*)$ plotted as a function of x^* along the channel for three buried channel CCD's. The 5-μm plates are at −5 V, −10 V, and −15 V, and the oxide thickness is $h_1^* = 0.1$ μm. $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are 0.1 μm, 2 μm, and 4 μm thick, respectively.

An operating buried channel CCD has been reported[16] in which the gaps between the plates have been filled with a resistive material so that the potential drop between the plates is essentially linear. The butted plate model can be easily adapted to describe this. In Figs. 13 and 14 we show the channel fields and potentials in such a CCD. This is a three-phase CCD with 10-μm plates and 5-μm gaps. The voltages on the plates are −5 V, −10 V, and −15 V and in the gaps the voltage varies linearly from one plate to the next. The thickness of the oxide layer is $h_1^* = 0.1$ μm. The doping profile in the p-type layer is given by (4) with $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$ ($C_S = 46$). The remaining physical parameters are as described at the beginning of this section. In Fig. 13 we plot the channel field, $E_x^* = -(\partial\varphi^*/\partial x^*)$, as a function of x^* for two different p-layer thicknesses, $h_2^* - h_1^* = 3$ μm and 7 μm. In Fig. 14 we plot the corresponding channel potentials, φ^*. Again, it is seen that the p-layer thickness is a sensitive parameter in terms of field penetration and charge-carrying capacity, and there is undoubtedly an optimal thickness. In Figs. 15 and 16 we plot the same quantities for a three-phase CCD which is identical to the one of Figs. 13 and 14, except that the gap spacing is zero. The devices seem to have essentially the same fields and charge-carrying capacities.

IV. ANALYTIC SOLUTION OF THE LINEARIZED EQUATIONS

In this section we briefly derive the solution of eqs. (21) through (24) subject to the boundary conditions (13) through (15), (17), (25), and (26). We assume as given the Fourier series expansions of $V(x)$ and $Q(x)$:

$$V(x) = \tfrac{1}{2}a_o + \sum_{n=1}^{\infty} (a_n \cos \lambda_n x + b_n \sin \lambda_n x), \tag{29}$$

$$Q(x) = \tfrac{1}{2}\zeta_o + \sum_{n=1}^{\infty} (\zeta_n \cos \lambda_n x + \xi_n \sin \lambda_n x) \tag{30}$$

where

$$\lambda_n = (2n\pi)/L. \tag{31}$$

Since in most cases of interest $V(x)$ and $Q(x)$ are either piecewise constant or linear functions, it is trivial to obtain the coefficients of these series.

Since $\psi(x, y)$ must be periodic in x with period L, we can expand the solution in each of the four regions in a series of the form

$$\psi(x, y) = \tfrac{1}{2}A_o(y) + \sum_{n=1}^{\infty} (A_n(y) \cos \lambda_n x + B_n(y) \sin \lambda_n x). \tag{32}$$

On substituting expressions of the form (32) for ψ into (21) through (24)

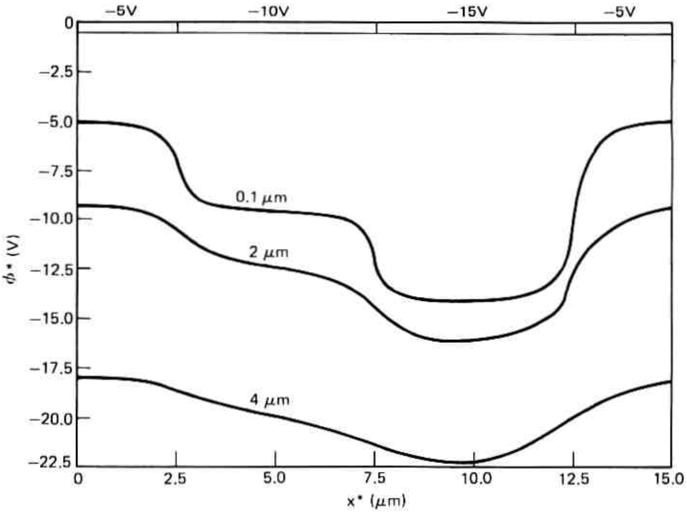


Fig. 12—The channel potential φ^* plotted as a function of x^* along the channel for three buried channel CCD's. The 5-μm plates are at -5 V, -10 V, and -15 V, and the oxide thickness is $h_1 = 0.1$ μm. $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are are 0.1 μm, 2 μm, and 4 μm thick, respectively.

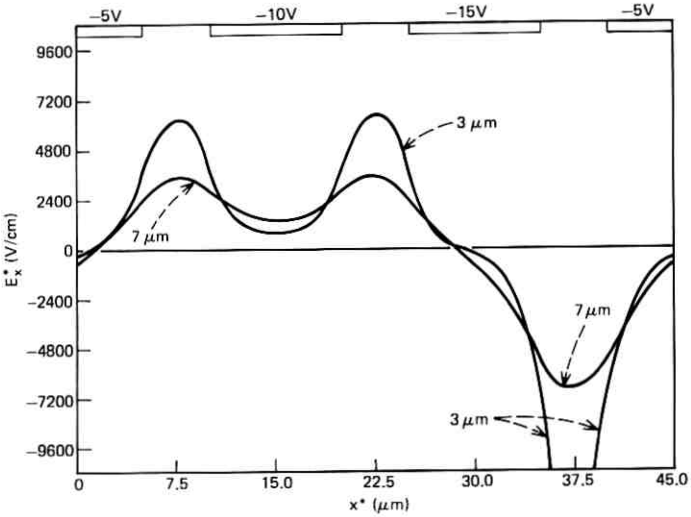


Fig. 13—The channel field $-(\partial\varphi^*/\partial x^*)$ plotted as a function of x^* along the channel for two buried channel CCD's. The 10-μm plates are at −5 V, −10 V, and −15 V and are separated by 5-μm gaps in which the potential varies linearly between plates. $h_1^* = 0.1$ μm, $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are 3 μm and 7 μm thick, respectively.

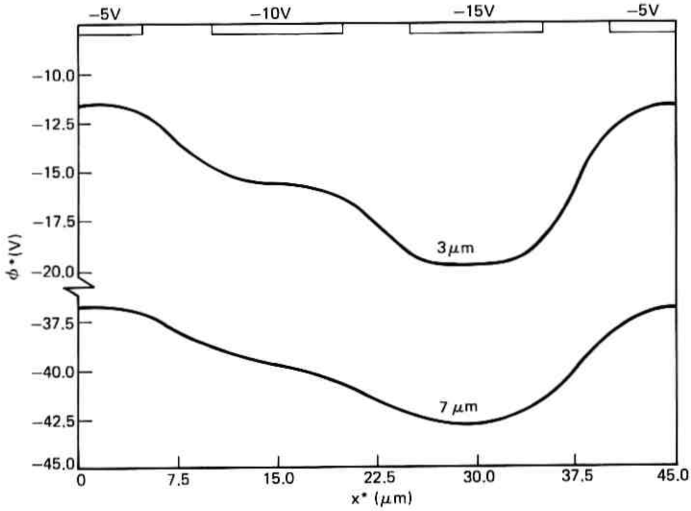


Fig. 14—The channel potential φ^* plotted as a function of x^* along the channel for two buried channel CCD's. The 10-μm plates are at −5 V, −10 V, and −15 V and are separated by 5-μm gaps in which the potential varies linearly between plates. $h_1^* = 0.1$ μm, $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are 3 μm and 7 μm thick, respectively.

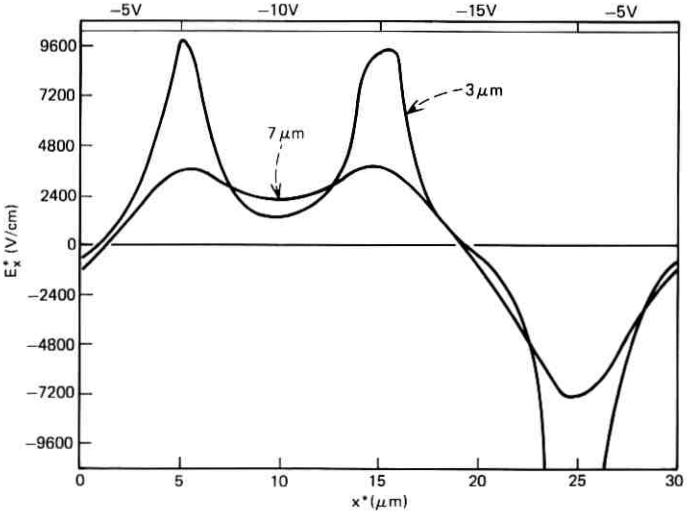


Fig. 15—The channel field $-(\partial\varphi^*/\partial x^*)$ plotted as a function of x^* along the channel for two buried channel CCD's. The 10-μm plates are at −5 V, −10 V, and −15 V, $h_1^* = 0.1$ μm, $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are 3 μm and 7 μm thick, respectively.

and equating to zero the coefficients of $\cos \lambda_n x$ and $\sin \lambda_n x$, $n = 0, 1, 2, \cdots$, we obtain an uncoupled system of second-order, constant-coefficient, ordinary differential equations from which the $A_n(y)$ and $B_n(y)$ can be determined simply. Each $A_n(y)$ and $B_n(y)$ is the sum of two linearly independent solutions and thus each involves two constants of integration which must be determined by making use of the boundary conditions (13), (14), (15), (17), (25), and (26). Since the Fourier series representing the solutions must be equal term by term at the boundaries, this yields a simple set of linear algebraic equations for the unknown constants of integration. These equations can be solved explicitly, yielding the integration constants as linear functions of the coefficients a_n and b_n, and ζ_n and ξ_n, of the Fourier series for $V(x)$ and $Q(x)$ given in (29) and (30). The algebra involved is elementary but involved, and we only record the final answer here.

Let

$$F_n(x) = a_n \cos \lambda_n x + b_n \sin \lambda_n x, \tag{33}$$

$$\Phi_n(x) = \zeta_n \cos \lambda_n x + \xi_n \sin \lambda_n x, \tag{34}$$

so that we can write

$$V(x) = \tfrac{1}{2}a_o + \sum_{n=1}^{\infty} F_n(x), \tag{35}$$

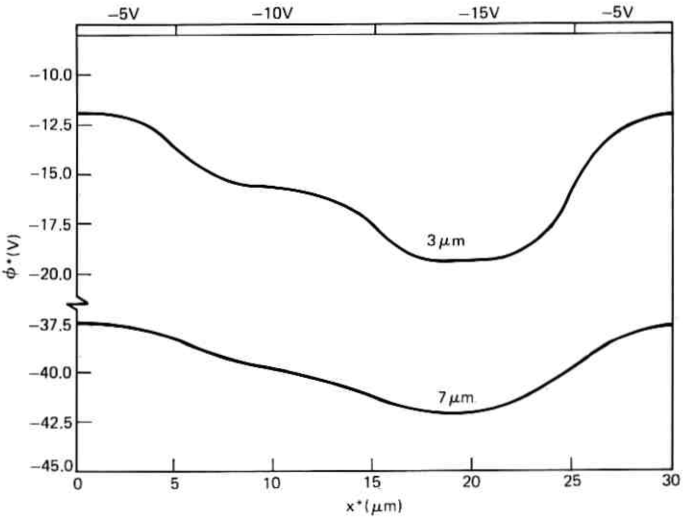


Fig. 16—The channel potential φ^* plotted as a function of x^* along the channel for two buried channel CCD's. The 10-μm plates are at −5 V, −10 V, and −15 V, $h_1^* = 0.1$ μm, $C_S^* = 4.6 \times 10^{15}/\text{cm}^3$, and the p-layers are 3 μm and 7 μm thick, respectively.

$$Q(x) = \tfrac{1}{2}\zeta_o + \sum_{n=1}^{\infty} \Phi_n(x). \tag{36}$$

Furthermore, let

$$E_\pm = 1 \pm 1/\eta, \tag{37}$$

$$\Lambda_n^\pm = 1 \pm (1 + \lambda_n^{-2})^{\frac{1}{2}}, \tag{38}$$

$$\begin{aligned} M_n(y) = {} & \{E_+ \Lambda_n^+ + E_- \Lambda_n^- e^{-2\lambda_n(h_3-h_1)}\} e^{-\lambda_n y} \\ & + \{E_- \Lambda_n^+ e^{-2\lambda_n h_1} + E_+ \Lambda_n^- e^{-2\lambda_n h_3}\} e^{\lambda_n y}, \end{aligned} \tag{39}$$

$$L_n(y) = 2\{\Lambda_n^+ e^{-\lambda_n y} + \Lambda_n^- e^{-\lambda_n(2h_3-y)}\}. \tag{40}$$

Then we can write

$$\begin{aligned} \psi_1(x, y) = {} & (A\bar{a}_o + B) + (C\bar{a}_o + D)y \\ & + \sum_{n=1}^{\infty} \left\{ F_n(x) \frac{M_n(y)}{M_n(0)} + \Phi_n(x) \frac{L_n(h_1)}{M_n(0)} \frac{\sinh \lambda_n y}{\eta \lambda_n} \right\}, \end{aligned} \tag{41}$$

$$\begin{aligned} \psi_2(x, y) = {} & \frac{1}{2} \biggl[\bar{a}_o (1 + h_3 - h_2) - (h_3 - h_2)^2 \\ & - (\bar{a}_o + 2h_2 - 2h_3)(y - h_2) + 2 \int_{h_2}^{y} (y - \xi)\sigma(\xi) d\xi \biggr] \\ & + \sum_{n=1}^{\infty} \left\{ F_n(x) + \Phi_n(x) \frac{\sinh \lambda_n h_1}{\eta \lambda_n} \right\} \frac{L_n(y)}{M_n(0)}, \end{aligned} \tag{42}$$

$$\psi_3(x,y) = \tfrac{1}{2}[\bar{a}_o(1 + h_3 - y) - (y - h_3)^2] + \sum_{n=1}^{\infty} \left\{ F_n(x) + \Phi_n(x) \frac{\sinh \lambda_n h_1}{\eta\lambda_n} \right\} \frac{L_n(y)}{M_n(0)}, \tag{43}$$

$$\psi_4(x,y) = \tfrac{1}{2}\bar{a}_o e^{-(y-h_3)} + 4 \sum_{n=1}^{\infty} \left\{ F_n(x) + \Phi_n(x) \frac{\sinh \lambda_n h_1}{\eta\lambda_n} \right\} \times \frac{\exp[-\sqrt{1 + \lambda_n^2}(y - h_3) - \lambda_n h_3]}{M_n(0)} \tag{44}$$

where

$$A = \frac{1}{2}\left(1 + h_3 - h_1 + \frac{h_1}{\eta}\right), \tag{45}$$

$$B = \int_{h_1}^{h_2} \left(\xi - h_1 + \frac{h_1}{\eta}\right) \sigma(\xi) d\xi - \frac{1}{2}\left[(h_3 - h_2)(h_3 + h_2 - 2h_1) + \frac{h_1}{\eta}(\zeta_o + 2h_3 - 2h_2)\right], \tag{46}$$

$$C = -1/(2\eta), \tag{47}$$

$$D = \left[\zeta_o + 2(h_3 - h_2) - 2\int_{h_1}^{h_2} \sigma(\xi) d\xi\right] \Big/ (2\eta), \tag{48}$$

and

$$\bar{a}_o = (\tfrac{1}{2}a_o - B)/A. \tag{49}$$

Equations for $\partial\psi/\partial x$ and $\partial\psi/\partial y$ can be obtained by differentiating eqs. (41) through (44) term by term. To obtain the equations for a surface CCD, drop equation (42) and set $\sigma = 0$, $h_2 = h_1$, and relabel h_3, ψ_3, and ψ_4 as h_2, ψ_2, and ψ_3 in the remaining equations.

V. FINITE DIFFERENCE SOLUTION OF THE EXACT EQUATIONS

In this section we describe the finite difference (FD) solution[17] of a surface CCD described by equations (10) through (17). We will assume that $Q(x) \equiv 0$, $\sigma(y) \equiv 0$, and $h_1 = h_2 = h$.

The infinite region $0 \leqq x \leqq L$, $h \leqq y < \infty$ is replaced by the finite rectangle $0 \leqq x \leqq L$, $h \leqq y \leqq H$, with $H \gg h$, and the boundary condition (16) is replaced for $0 \leqq x \leqq L$ by

$$\varphi_2(x, H) = 0. \tag{50}$$

This may be done, because the solution tends to zero rapidly as $y \to \infty$. In fact, a one-dimensional analysis[5] shows that φ_2 tends to zero exponentially in y.

A uniform FD net is now placed over region 1 (the oxide layer) and region 2 (the silicon layer) with N points in the x-direction and M_1 and M_2 points in the y-direction in regions 1 and 2, respectively. Let

$$\begin{gathered} h_x = L/(N-1), \qquad h_{1y} = h/(M_1 - 1), \\ h_{2y} = (H-h)/(M_2 - 1), \end{gathered} \tag{51}$$

and then, for $(1 \leqq i \leqq N)$, $(1 \leqq j \leqq M_\alpha)$, and $(\alpha = 1, 2)$, define

$$\varphi_{\alpha,i,j} = \varphi_\alpha((i-1)h_x, (j-1)h_{\alpha y}) \tag{52}$$

where $\varphi_\alpha(x,y)$ is the exact solution of (10) through (17), and define $\hat{\varphi}_{\alpha,i,j}$ as the FD solution which approximates $\varphi_{\alpha,i,j}$.

The FD equations are obtained as follows. The boundary condition (50) is replaced by the N equations

$$\hat{\varphi}_{2,i,M_2} = 0, \qquad (1 \leqq i \leqq N), \tag{53}$$

while the boundary condition (13) is replaced by the N equations

$$\hat{\varphi}_{1,i,1} = V((i-1)h_x) = V_i, \qquad (1 \leqq i \leqq N). \tag{54}$$

If $x_o = (i_o - 1)h_x$ is the edge of a plate, the jump discontinuity in $V(x)$ there is handled by setting

$$V_{i_o} = \tfrac{1}{2}(V_{i_o+1} + V_{i_o-1}). \tag{55}$$

At each interior mesh point $\nabla^2\varphi_{\alpha,i,j}$ is approximated by the standard five-point difference formula[17]

$$\begin{aligned} \nabla_h^2\hat{\varphi}_{\alpha,i,j} = {} & (\hat{\varphi}_{\alpha,i+1,j} + \hat{\varphi}_{\alpha,i-1,j} - 2\hat{\varphi}_{\alpha,i,j})/h_x^2 \\ & + (\hat{\varphi}_{\alpha,i,j+1} + \hat{\varphi}_{\alpha,i,j-1} - 2\hat{\varphi}_{\alpha,i,j})/h_{\alpha y}^2. \end{aligned} \tag{56}$$

Equation (56) can also be used to evaluate $\nabla_h^2\hat{\varphi}_{\alpha,1,j}$ and $\nabla_h^2\hat{\varphi}_{\alpha,N,j}$, $(2 \leqq j \leqq M_\alpha - 1)$ by making use of the periodicity relations

$$\hat{\varphi}_{\alpha,-1,j} = \hat{\varphi}_{\alpha,N-1,j}, \qquad \hat{\varphi}_{\alpha,N+1,j} = \hat{\varphi}_{\alpha,1,j}. \tag{57}$$

Thus, eqs. (10) and (12) are replaced by $(M_1 + M_2 - 4)N$ equations

$$\nabla_h^2\hat{\varphi}_{1,i,j} = 0, \qquad (1 \leqq i \leqq N), \qquad (2 \leqq j \leqq M_1 - 1), \tag{58}$$

$$\nabla_h^2\hat{\varphi}_{2,i,j} = \exp(\hat{\varphi}_{2,i,j}) - 1, \qquad (1 \leqq i \leqq N), \qquad (2 \leqq j \leqq M_2 - 1). \tag{59}$$

There remain the interface conditions (14). The first of these is replaced by the N equations

$$\hat{\varphi}_{1,i,M_1} = \hat{\varphi}_{2,i,1}, \qquad (1 \leqq i \leqq N). \tag{60}$$

To obtain an equivalent set of equations for the second condition, we could replace the derivatives by the first differences from each side.

However, it is well known that this approximation is not very good. This is easily seen from the prototype equation $\nabla^2\varphi = 0$, where, for simplicity, we take $h_x = h_y = h$. Then as is well known,[17]

$$\varphi_{i+1,j} + \varphi_{i-1,j} + \varphi_{i,j+1} + \varphi_{i,j-1} - 4\varphi_{i,j} = h^2\nabla^2\varphi_{ij} + 0(h^4), \quad (61)$$

while

$$\varphi_{i,j+1} - \varphi_{i,j} = h\frac{\partial\varphi}{\partial y}(x_i, y_j) + 0(h^2). \quad (62)$$

Thus, the errors in the FD equations, after scaling the left-hand sides to have coefficients of order 1, are out of balance. The interface condition is $1/h^2$ less accurately modeled than the differential equation. This leads to the following scheme which balances the errors equally. We want $\partial\varphi_1/\partial y$ using only values of φ_1, similarly for $\partial\varphi_2/\partial y$, and the approximation must be good to $0(h^3)$. This may be done using the values of $\varphi_{1,i,j}$, $(M_1 - 3 \leqq j \leqq M_1)$. Simply use the derivative of the cubic interpolation polynomial through these values. It is easily seen that

$$\frac{\partial\varphi_1}{\partial y}(x_i, h) = \{11\varphi_{1,i,M_1} - 18\varphi_{1,i,M_1-1} + 9\varphi_{1,i,M_1-2} - 2\varphi_{1,i,M_1-3}\}/(6h_{1y}) + 0(h_{1y}^3). \quad (63)$$

Similarly,

$$\frac{\partial\varphi_2}{\partial y}(x_i, h) = \{2\varphi_{2,i,4} - 9\varphi_{2,i,3} + 18\varphi_{2,i,2} - 11\varphi_{2,i,1}\}/(6h_{2y}) + 0(h_{2y}^3). \quad (64)$$

Then the second boundary condition (14) is replaced by the N equations

$$\eta\{11\hat{\varphi}_{1,i,M_1} - 18\hat{\varphi}_{1,i,M_1-1} + 9\hat{\varphi}_{1,i,M_1-2} - 2\hat{\varphi}_{1,i,M_1-3}\}/h_{1y} = \{2\hat{\varphi}_{2,i,4} - 9\hat{\varphi}_{2,i,3} + 18\hat{\varphi}_{2,i,2} - 11\hat{\varphi}_{2,i,1}\}/h_{2y}, \quad (1 \leqq i \leqq N). \quad (65)$$

Equations (53), (54), (58), (59), (60), and (65) comprise $(M_1 + M_2)N$ equations in the $(M_1 + M_2)N$ unknowns $\hat{\varphi}_{\alpha,i,j}$, $(\alpha = 1, 2)$, $(1 \leqq i \leqq N)$, and $(1 \leqq j \leqq M_\alpha)$. From the standard FD theory, the solution of this set of transcendental equations differs from the solution of the true boundary value problem by a factor of order $0(h^2)$. We used a nonlinear overrelaxation scheme developed in Refs. 18 through 20 to solve the FD equations for $\hat{\varphi}_2$ and standard overrelaxation methods[21] to solve the $\hat{\varphi}_1$ FD equations.

An initial estimate of the solution, $\hat{\varphi}^{(0)}_{\alpha,i,j}$, was obtained by computing the one-dimensional matching solutions as functions of y along the lines $x = x_i$, $(1 \leqq i \leqq N)$, using the methods of Ref. 5. This provides a solution which is a fairly good estimate under the middle of any plate and a very bad one near the edge of any plate. These one-dimensional solutions also give a good estimate of the greatest depth of the transition region, call it $y_{\max}$. Since $\varphi_2 \rightarrow 0$ exponentially for $y > y_{\max}$, we chose $H = y_{\max} + 20$.

This estimated solution $\hat{\varphi}_{\alpha,i,j}^{(0)}$ is now refined iteratively by the method of successive overrelaxation (SOR). The $(n + 1)$st iterate is obtained from the nth as follows. For all $n = 0, 1, 2, \cdots$, set

$$\hat{\varphi}_{1,i,1}^{(n)} = V_i, \qquad \hat{\varphi}_{2,i,M_2}^{(n)} = 0, \qquad (1 \leqq i \leqq N). \tag{66}$$

For $(2 \leqq i \leqq N - 1)$ and $(2 \leqq j \leqq M_1 - 1)$, let $\tilde{\varphi}_{1,i,j}^{(n+1)}$ be defined by

$$(2h_x^{-2} + 2h_{1y}^{-2})\,\tilde{\varphi}_{1,i,j}^{(n+1)} = (\hat{\varphi}_{1,i+1,j}^{(n)} + \hat{\varphi}_{1,i-1,j}^{(n+1)})h_x^{-2} + (\hat{\varphi}_{1,i,j+1}^{(n)} + \varphi_{1,i,j-1}^{(n+1)})h_{1y}^{-2}. \tag{67}$$

Then set

$$\hat{\varphi}_{1,i,j}^{(n+1)} = \omega_1\,\tilde{\varphi}_{1,i,j}^{(n+1)} + (1 - \omega_1)\,\hat{\varphi}_{1,i,j}^{(n)} \tag{68}$$

where ω_1 is an overrelaxation parameter satisfying $1 \leqq \omega_1 < 2$. For $(2 \leqq i \leqq N - 1)$ and $(2 \leqq j \leqq M_2 - 1)$, let $\tilde{\varphi}_{2,i,j}$ be the solution of

$$(2h_x^{-2} + 2h_{2y}^{-2})\,\tilde{\varphi}_{2,i,j}^{(n+1)} + \exp(\tilde{\varphi}_{2,i,j}^{(n+1)}) = 1 + (\hat{\varphi}_{2,i+1,j}^{(n)} + \hat{\varphi}_{2,i-1,j}^{(n+1)})h_x^{-2} + (\hat{\varphi}_{2,i,j+1}^{(n)} + \hat{\varphi}_{2,i,j-1}^{(n+1)})h_{2y}^{-2}. \tag{69}$$

Equation (69) has the form

$$Aw + e^w = B \tag{70}$$

where A and B are known and $A > 0$. Given any approximate solution w^o of (70), Newton's method[22] yields the sequence

$$w^{k+1} = [A + e^{w^k}]^{-1}[B + (w^k - 1)e^{w^k}], \qquad (k = 0, 1, 2, \cdots), \tag{71}$$

which converges to the solution of (69). The convergence of this scheme is global and quadratic because the function $Aw + e^w$, for $A > 0$, is a monotone increasing, convex function of w. After solving for $\tilde{\varphi}_{2,i,j}^{(n+1)}$, we set

$$\hat{\varphi}_{2,i,j}^{(n+1)} = \omega_2\,\tilde{\varphi}_{2,i,j}^{(n+1)} + (1 - \omega_2)\,\hat{\varphi}_{2,i,j}^{(n)} \tag{72}$$

where $1 \leqq \omega_2 < 2$.

The interface values $\hat{\varphi}_{1,i,M_1}$ and $\hat{\varphi}_{2,i,1}$, $(2 \leqq i \leqq N - 1)$ are relaxed by combining (60) and (65), defining

$$\tilde{\varphi}_{1,i,M_1}^{(n+1)} = \tilde{\varphi}_{2,i,1}^{(n+1)} = (\eta h_{1y}^{-1} + h_{2y}^{-1})^{-1}[\eta h_{1y}^{-1}(18\,\hat{\varphi}_{1,i,M_1-1}^{(n+1)} - 9\,\hat{\varphi}_{1,i,M_1-2}^{(n+1)} + 2\,\hat{\varphi}_{1,i,M_1-1}^{(n+1)}) + h_{2y}^{-1}(18\,\hat{\varphi}_{2,i,2}^{(n)} - 9\,\hat{\varphi}_{2,i,3}^{(n)} + 2\,\hat{\varphi}_{2,i,4}^{(n)})]/(11), \tag{73}$$

and setting

$$\hat{\varphi}^{(n+1)}_{1,i,M_1} = \hat{\varphi}^{(n+1)}_{2,i,1} = \omega\tilde{\varphi}^{(n+1)}_{1,i,M_1} + (1-\omega)\hat{\varphi}^{(n)}_{1,i,M_1} \tag{74}$$

where

$$\omega = \tfrac{1}{2}(\omega_1 + \omega_2). \tag{75}$$

Equations (66) through (75) describe the manner in which the interior nodes of the FD mesh are relaxed. The nodes at $i = 1$ and N involve periodicity and require more detailed study. By using the periodicity relations (57), equations (66) through (75) can be extended to the nodes at $i = 1, N$. We could next do one of two things. First, we could treat $\hat{\varphi}_{\alpha,1,j}$ and $\hat{\varphi}_{\alpha,N,j}$ as separate quantities and relax each of them separately using the periodicity relations (57). Then, each time $\hat{\varphi}_{\alpha,1,j}$ was relaxed, this new value could be substituted for $\hat{\varphi}_{\alpha,N,j}$ to preserve periodicity, and vice versa. This was tried but gave very poor convergence rates. The problem is that in treating $\hat{\varphi}_{\alpha,1,j}$ and $\hat{\varphi}_{\alpha,N,j}$ as separate quantities, the same quantity (that is $\hat{\varphi}_{\alpha,1,j}$ and $\hat{\varphi}_{\alpha,N,j}$) gets relaxed twice rather than once in each SOR sweep. This can be avoided by letting $\hat{\varphi}_{\alpha,N,j} \equiv \hat{\varphi}_{\alpha,1,j}$, $(\alpha = 1, 2)$, $(2 \leqq j \leqq M_\alpha - 1)$ and then relaxing only the quantities $\hat{\varphi}_{\alpha,1,j}$. This produces quite acceptable convergence rates.

The overrelaxation parameters ω_1 and ω_2 were set equal to the optimum values of these parameters for the Laplace equation on regions one and two respectively. These values for ω_1 and ω_2 were estimated as follows. Let $\hat{\boldsymbol{\varphi}}^{(n)}_\alpha$ denote the vector of values of the nth iterate of the solution to $\nabla^2_h\hat{\varphi}_\alpha = 0$ in region α, $(\alpha = 1, 2)$. Define the nth residual vector as $\mathbf{R}^{(n)}_\alpha = \hat{\boldsymbol{\varphi}}^{(n+1)}_\alpha - \hat{\boldsymbol{\varphi}}^{(n)}_\alpha$. Then, starting with any initial guess $\hat{\boldsymbol{\varphi}}^{(0)}_\alpha \neq 0$, standard theory shows[23] that $\lim_{n\to\infty} \|\mathbf{R}^{(n+1)}_\alpha\|/\|\mathbf{R}^{(n)}_\alpha\| = \eta_\alpha$ exists and

$$\omega_{\alpha,\mathrm{opt}} = 2/\{1 + \sqrt{1 - \eta_\alpha}\} \tag{76}$$

where $\|\mathbf{R}^{(n)}_\alpha\|$ denotes the norm of the vector $\mathbf{R}^{(n)}_\alpha$ and is called the residual. In practice, we calculated the residuals for n large enough so that η_α was obtained to the desired accuracy using the L_2 norm.

A further important point is solving the transcendental equation (69), which must be done at each SOR step. In the $(n + 1)$th SOR sweep, the initial estimate for $\tilde{\varphi}^{(n+1)}_{2,i,j}$ in the Newton iteration (71) was $\hat{\varphi}^{(n)}_{2,i,j}$. There is no reason to compute the $\tilde{\varphi}^{(n+1)}_{2,i,j}$ very accurately when $\hat{\varphi}^{(n)}_{2,i,j}$ is far away from its final value. Conversely, the quadratic convergence of Newton's method means that when the error in $\hat{\varphi}^{(n)}_{2,i,j}$ is small, one Newton iteration will produce a very good approximation for $\tilde{\varphi}^{(n+1)}_{2,i,j}$. For this reason, only one Newton iteration was used in solving (69) during each SOR sweep.

Some theory has been developed to show that the SOR scheme we have outlined above converges to the true solution in some mildly nonlinear Dirichlet problems.[18,19] However, to the best of our knowledge, no theoretical analysis exists of the boundary value problem of this paper. Nevertheless, as we will demonstrate by numerical examples in the next section, the scheme works in practice.

We conclude this section with a few remarks on the estimated accuracy of the FD solution. Since little is known of the general theory of a complicated nonlinear boundary value problem such as we are considering, we argue by analogy with the Dirichlet problem for Laplace's equation on a square. Let $\varphi(x, y)$ be the true solution of $\nabla^2\varphi = 0$ in $0 < x,\ y < L$, with φ specified on the boundary, let $h = L/(N-1)$, $\varphi_{i,j} = \varphi((i-1)h,\ (j-1)h)$, and let $\hat{\varphi}_{ij}$ be the solution of the corresponding FD equations. Then it is well known[24] that under reasonable conditions on the boundary values,

$$\|\varphi - \hat{\varphi}\|_\infty = \sup_{1 \leqq i,\ j \leqq N} |\varphi_{ij} - \hat{\varphi}_{ij}| = 0(h^2). \tag{77}$$

This relationship assumes that we know the FD solution exactly. However, we don't know $\hat{\varphi}$ exactly, all we know are the various iterates $\hat{\varphi}^{(n)}$ which have been calculated and the residuals $\|\hat{\varphi}^{(n+1)} - \hat{\varphi}^{(n)}\|_\infty$. Now it is known, though not as widely as it should be, when calculating $\hat{\varphi}$ by the method of SOR that[25]

$$\|\hat{\varphi} - \hat{\varphi}^{(n+1)}\|_\infty = C(\omega)\|\hat{\varphi}^{(n+1)} - \hat{\varphi}^{(n)}\|_\infty, \tag{78}$$

and if ω_b is the optimal choice of ω,

$$C(\omega_b) = 0(N). \tag{79}$$

It can also be shown, if the optimal value of ω is underestimated by ten or fifteen percent, that $C(\omega) = 0(N^2)$. This means that to obtain an approximate solution accurate to $0(h^2) = 0(N^{-2})$ by the method of SOR, we must iterate at least until the residuals are $0(N^{-3})$, and since in the nonlinear problem we can only crudely estimate the optimal ω, we should really iterate until the residuals are $0(N^{-4})$.

As we will show by example in the next section, it is necessary to calculate the potentials with great accuracy if one wishes to obtain the fields from them with any accuracy at all by differencing them. From the previous paragraph, however, we have shown that this is expensive in even a moderately complicated problem, since then the residuals must be made so small. To estimate the cost of increasing the size of the problem or decreasing the mesh size (both equivalent to increasing N), we note that typically in SOR,[26]

$$\|\hat{\varphi}^{(n+1)} - \hat{\varphi}^{(n)}\|_\infty \approx e^{-n|0(N-1)|} \tag{80}$$

for optimal ω, while for nonoptimal ω,

$$\|\hat{\varphi}^{(n+1)} - \hat{\varphi}^{(n)}\|_\infty \approx e^{-n|0(N-2)|}. \tag{81}$$

If we wish to specify that $\|\hat{\varphi} - \hat{\varphi}^{(n+1)}\|_\infty \leqq \epsilon$, then, for optimal ω, it is easy to show from (78) through (80) that the number of iterations must be

$$n = 0(N\ \ell n(N/\epsilon)), \tag{82}$$

while for nonoptimal ω, it follows from (78), (79), and (81) that

$$n = 0(N^2 \ell n(N/\epsilon)). \tag{83}$$

Thus, if we decrease the mesh size by one half, then from (82) or (83), we must double or quadruple the number of iterations to obtain the same accuracy. Since there are now 4 times as many mesh points, the time required to obtain a solution goes up by a factor of 8 or 16, depending on the knowledge of ω.

VI. COMMENTS ON THE ACCURACY OF THE SOLUTIONS

In Section IV we derived the Fourier series solution (FSS) of the linearized problem, and in Section V we outlined the finite difference solution (FDS) of the nonlinear problem. In this section we compare several of these solutions with regard to accuracy and cost. All calculations discussed were performed on a Honeywell 6070 computer, and all programs were written in Fortran IV.

We solved the nonlinear equations (10) through (17) by the method of finite differences for the two-phase surface CCD's discussed in Section III, some of whose properties are presented graphically in Figs. 5 through 10. The three CCD's have plate widths of 45 μm, 5 μm, and 1.5 μm, respectively. The plate voltages are -5 V and -10 V, $Q(x) \equiv 0$, $N_D = 10^{14}/\text{cm}^3$, $\epsilon_1/\epsilon_0 = 4$, $\epsilon_2/\epsilon_0 = 12$, and $\lambda_D = 0.415$ μm. In all cases, a FD net was chosen with $N = 25$, $M_1 = 25$, and $M_2 = 41$.

In the case of the 45-μm-plate device, this corresponds in the dimensionless units to $h_x = 9$, $h_{1y} = 0.02$, $h_{2y} = 1$. After 273 SOR iterations, the residual was $\sim 4 \times 10^{-7}$, the running time was 195 seconds, and 35 K of memory was used. This should ensure that the difference between the true FDS and the iterated solution will never exceed $\sim 25 \times 65 \times 4 \times 10^{-7} = 6.5 \times 10^{-4}$. We have calculated the FSS at the same mesh points, this took 33 seconds to run, and 33 K of memory was used. Let $\varphi_{\mathrm{FD}}(x, y)$ denote the FD solution of the nonlinear problem and $\psi(x, y)$ denote the FS solution of the linearized problem, and

$$e(x, y) = |(\varphi_{\mathrm{FD}}(x ,y) - \psi(x, y))/\varphi_{\mathrm{FD}}(x, y)|. \tag{84}$$

Then we found that along the oxide-semiconductor interface, $e(x, h_1) < 1.14 \times 10^{-2}$, $0 \leqq x \leqq L$, and five Debye lengths below this interface, $e(x, h_1 + 5) < 2.89 \times 10^{-2}$, $0 \leqq x \leqq L$

We have for the 5-μm-plate device $h_x = 1$, $h_{1y} = 0.02$, $h_{2y} = 1$. After 288 SOR iterations, the residual was $\sim 4 \times 10^{-7}$, the running time was 190 seconds, and 35 K of memory was used. This should again ensure that the difference between the true FDS and the iterated solution does not exceed 6.5×10^{-4}. For this case, we found that $e(x, h_1) < 2.8 \times 10^{-3}$, $e(x, h_1 + 5) < 1.4 \times 10^{-3}$, $e(x, h_1 + 10) < 3.14 \times 10^{-3}$, $e(x, h_1 + 15) < 9.81 + 10^{-3}$ for $0 \leqq x \leqq L$. The running time to evaluate the FSS was 10 seconds and 33 K of memory was used.

Finally, for the 1.5-μm-plate device, we have $h_x = 0.3$, $h_{1y} = 0.02$, and $h_{2y} = 1$. After 300 SOR iterations the residual was $\sim 6 \times 10^{-4}$, the running time was 196 seconds, and 35 K of memory was used. Note that in this case the residual is three orders of magnitude greater than in the other two cases. We found that $e(x, h_1) < 1.8 \times 10^{-2}$, the running time to calculate the FSS was 4 seconds, and 33 K of memory was used.

In Ref. 5 it was noted that as long as $|\psi_4(h_3) + 1| < 10$, one could expect the solution of the linearized problem to be a good approximation to the solution of the nonlinear problem, at least in the p-region for buried channel devices or near the oxide-semiconductor interface for surface devices. In the examples considered here, for the 45-μm-plate case, $-5.15 < \psi_4(x, h_3) < 3.00$, for the 5-$\mu$m-plate device, $-1.20 < \psi_4(x, h_3) < -0.96$, and for the 1.5-$\mu$m-plate case, $\psi_4(x, h_3) \equiv -1.077$. This again suggests that the smaller $|\psi_4(x, h_3) + 1|$, the more accurate the approximation.

These examples show that if one only needs a knowledge of the potential in the neighborhood of the oxide-semiconductor interface, the FSS provides a highly accurate approximation to the true solution much more cheaply than can be obtained by FD methods. In fact, to analyze three-phase devices, the cost of obtaining a FDS goes up sharply while the cost of a FSS remains nominal. For example, it took only 15 seconds and 34 K of memory to obtain the solutions presented graphically in Figs. 15 and 16.

In reality, we are as much interested in the fields as we are in the potentials, and it is at this point that the difficulty with using the FD method for solving these problems becomes most acute. In Figs. 6, 8, and 10, the dashed curves are plots of $-(\partial\varphi_{\text{FD}}/\partial x)$ along $y = h_1$, obtained from the FDS just discussed by differencing. In Fig. 6 for example, the two curves differ by nearly an order of magnitude at their peaks. If we take the FSS and difference it to estimate the first derivative, we get a result which, in the neighborhood of the peaks, differs by at most 3 percent from the derivative obtained by differencing the FDS. We can conclude that the fields obtained from the FDS are badly in error. In order to calculate the fields from the FDS with any degree of accuracy, even for these simple examples, we would have to take a mesh so fine that the cost would become prohibitive.

VII. ACKNOWLEDGMENTS

The authors would like to thank G. E. Smith for originally suggesting this research and for many subsequent conversations. They are also indebted to R. H. Krambeck, R. J. Strain, and R. H. Walden for many fruitful conversations on CCD's.

REFERENCES

1. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.
2. Amelio, G. F., Tompsett, M. F., and Smith, G. E., "Experimental Verification of the Charge Coupled Device Concept," B.S.T.J., *49*, No. 4 (April 1970), pp. 593–600.
3. Strain, R. J., and Schryer, N. L., "A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1721–1740.
4. Amelio, G. F., "Computer Modeling of Charge-Coupled Device Characteristics," B.S.T.J., *51*, No. 3 (March 1972), pp. 705–730.
5. McKenna, J., and Schryer, N. L., "On the Accuracy of the Depletion Layer Approximation for Charge Coupled Devices," B.S.T.J., *51*, No. 7 (September 1972), pp. 1471–1485.
6. McKenna, J., and Schryer, N. L., unpublished work.
7. Walden, R. H., Krambeck, R. H., Strain, R. J., McKenna, J., Schryer, N. L., and Smith, G. E., "The Buried Channel Charge Coupled Device," B.S.T.J., *51*, No. 7 (September 1972), pp. 1635–1640.
8. Boyle, W. S., and Smith, G. E., "Charge-Coupled Devices–A New Approach to MIS Device Structures," IEEE Spectrum, *8*, No. 7 (July 1971), pp. 18–27.
9. Engeler, W. C., Tiemann, J. J., and Baertsch, R. D., "Surface Charge Transport in Silicon," Appl. Phys. Lett., *17*, No. 11 (December 1970), pp. 469–472.
10. Berglund, C. N., Powell, R. J., Nicollian, E. H., and Clemens, J. T., "Two-Phase Stepped Oxide CCD Shift Register Using Undercut Isolation," Appl. Phys. Lett., *20*, No. 11 (June 1972), pp. 412–414.
11. Lewis, J. A., McKenna, J., and Wasserstrom, E., "Field of Negative Point, Line or Plane Charges in an n-Type Semiconductor," J. Appl. Phys., *41*, No. 10 (September 1970), pp. 4182–4189.
12. Grove, A. S., *Physics and Technology of Semiconductor Devices*, New York: John Wiley & Sons, 1967, pp. 49–50.
13. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards, 1964, p. 297.

14. Krambeck, R. H., Walden, R. H., and Pickar, K. A., "Implanted-Barrier Two-Phase Charge-Coupled Device," Appl. Phys. Lett., *19*, No. 12 (December 1971), pp. 520–522.
15. Tompsett, M. F., "The Quantitative Effects of Interface States on the Performance of Charge-Coupled Devices," IEEE Trans. Elec. Devices, *ED-12*, No. 1 (January 1973), pp. 45–55.
16. Elson, B. M., "Charge-Coupled Concept Studied for Photo-Sensors," Aviation Week & Space Technology, *96*, No. 21 (May 22, 1972), pp. 73–75.
17. Varga, R. S., *Matrix Iterative Analysis*, Englewood Cliffs, N.J.: Prentice-Hall, 1962, chapter 6.
18. Bers, L., "On Mildly Nonlinear Partial Difference Equations of Elliptic Type," J. Res. Nat. Bur. Standards, *51*, No. 11 (November 1953), pp. 229–236.
19. Ortega, J. M., and Rockoff, M. L., "Nonlinear Difference Equations and Gauss-Seidel Type Iterative Methods," SIAM J. Numer. Anal., *3*, No. 9 (September 1966), pp. 497–513.
20. Wasserstrom, E., and McKenna, J., "The Potential Due to a Charged Metallic Strip on a Semiconductor Surface," B.S.T.J., *49*, No. 5 (May–June 1970), pp. 853–877.
21. Varga, R. S., op. cit., chapter 4.
22. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, New York: McGraw-Hill, 1962, p. 81.
23. Forsythe, G. E., and Wasow, W. R., *Finite Difference Methods for Partial Differential Equations*, New York: John Wiley & Sons, 1960, p. 257.
24. Ibid, section 23.
25. Weinberger, H. F., "A Posteriori Error Bounds in Iterative Matrix Inversion," in *Numerical Solution of Partial Differential Equations* (Proc. Symp. Univ. Maryland, 1965), ed. J. H. Bramble, New York: Academic Press, 1966, pp. 153–163.
26. Varga, R. S., op. cit., p. 204.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

Semilattice Characterization of Nonblocking Networks

By V. E. BENEŠ

(Manuscript received August 30, 1972)

A connecting network is called strictly nonblocking if no call is blocked in any state; it is nonblocking in the wide sense if there exists a rule for routing calls through the network so as to avoid all states in which calls are blocked, and yet still satisfy all demands for connection as they arise, without disturbing calls already present. Characterizations of both senses of nonblocking have been given in previous work, using simple metric and closure topologies defined on the set of states. We give new characterizations based on the natural map $\gamma(\cdot)$ *that carries each state into the assignment it satisfies. This map is a semilattice homomorphism, such that* $\gamma(x) \cap \gamma(y) \geqq \gamma(x \cap y)$. *It turns out that the case of equality in this inequality is very relevant to nonblocking performance. In particular, let a subset* X *of states be said to have the intersection property if for every* x *in* X *and every assignment* a *there exists* y *in* X *such that* y *realizes* a *(i.e.,* $\gamma(y) = a$*) and* $\gamma(x \cap y) = \gamma(x) \cap \gamma(y)$*. Then a network is nonblocking in the wide sense if and only if some subset of its states has the intersection property, and it is strictly nonblocking if and only if the entire set of states has the intersection property.*

I. INTRODUCTION

In a nonblocking network, no call need be lost because of link mismatch or junctor unavailability. Efficient nonblocking networks were invented by Charles Clos and, although they are not in common use at the present time, they are distinct possibilities for practical applications in the future, and they have substantial theoretical interest as outer limits on possible designs.

Two degrees or strengths of the nonblocking property have been distinguished.[1,2] A connecting network is called *strictly* nonblocking if no call is blocked in any state; it is nonblocking *in the wide sense* if there exists a rule for routing calls through the network so as to avoid all states in which calls are blocked, and yet still satisfy all demands for connection as they arise, without disturbing calls already in progress. These properties have been given[1,2] topological characterizations, and examples of each are known, although it must be said that examples of *efficient* wide-sense nonblocking networks are yet to be found.

Our aim in this paper is to give new alternative characterizations of the nonblocking properties in terms of the semilattice structures of the set of network states and of the set of assignments the states realize; a key role is played by the homomorphism $\gamma(\cdot)$ that carries each state into the assignment it realizes.

The property of being nonblocking in the wide sense lies between two other properties: that of being strictly nonblocking (nonblocking in the strict sense) and that of being rearrangeable. In a strictly nonblocking network, no call is blocked in any state; in a rearrangeable network, calls can always be given new routes (rearranged) so as to unblock any blocked call. The three properties (along, doubtless, with others not yet studied) form a spectrum of possible ways of operating switching equipment that exhibits or summarizes the tradeoff obtainable between efficient usage of switches and amount of calculation: the richer the network is in crosspoints, the less one has to do to use it so as to achieve desired load and loss. In a strictly nonblocking network, any path for an idle call will do; there always is one, and no traffic advantage is gained by use of one rather than another. In a wide-sense nonblocking network, the right choice of a path may mean the difference between zero loss and blocking some calls. By calculation, though, one can always find a route that will result in no blocking. In a rearrangeable network, finally, nonblocking behavior is again attainable, but, in general, only at the cost of constantly recalculating new routes for all the desired calls simultaneously, and reswitching them as necessary.

II. PRELIMINARIES

We shall use a model for the combinatorial aspects of a connecting network. This model is called a semilattice,[3] or partially ordered system with intersections, and it can be thought of as arising as follows: a connecting network ν is a quadruple $\nu = (G, I, \Omega, S)$ where G is a graph depicting network structure, I is the set of nodes of G which are inlets, Ω is the set of nodes of G that are outlets, and S is the set of permitted states. Variables w, x, y, and z at the end of the alphabet denote states, while u and v denote a typical inlet and a typical outlet, respectively. A state x can be thought of as a set of disjoint chains on G, each chain joining I to Ω. Not every such set of chains represents a state: sets with wastefully circuitous chains may be excluded from S. It is possible that $I = \Omega$ (one-sided network), that $I \cap \Omega = \phi$ (two-sided network), or that some intermediate condition obtain, depending on the "community of interest" aspects of the network ν.

The set S of states is partially ordered by inclusion $\leqq$, where $x \leqq y$ means that state x can be obtained from state y by removing zero or more calls. If x and y satisfy the same assignment of inlets to outlets, i.e., are such that all and only those inlets $u \in I$ are connected in x to outlets $v \in \Omega$ which are connected to the same v in y (though possibly by different routes), then we say that x and y are equivalent, written $x \sim y$.

We denote by A_x the set of states that are immediately above x in the partial ordering $\leqq$, and by B_x the set of those that are immediately below. Thus

$$A_x = \{\text{states accessible from } x \text{ by adding a call}\}$$
$$B_x = \{\text{states accessible from } x \text{ by a hangup}\}.$$

It can be seen, further, that the set S of states is not merely partially ordered by $\leqq$, but also forms a semilattice, or a partially ordered system with intersections,[3] with $x \cap y$ defined to be the state consisting of those calls and their respective routes which are common to both x and y.

An *assignment* is a specification of what inlets should be connected to what outlets. The set A of assignments can be represented as the set of all fixed-point-free correspondences from subsets of I to Ω. The set A is partially ordered by inclusion, and there is a natural map $\gamma(\cdot)\colon S \rightarrow A$ which takes each state $x \in S$ into the assignment it realizes; the map $\gamma(\cdot)$ is a semilattice homomorphism of S into A, with the properties

$$\begin{aligned}
&x \geqq y \Rightarrow \gamma(x) \geqq \gamma(y),\\
&x \geqq y \Rightarrow \gamma(x - y) = \gamma(x) - \gamma(y),\\
&\gamma(x \cap y) \leqq \gamma(x) \cap \gamma(y),\\
&\gamma(x) = \phi \Rightarrow x = 0 = \text{zero state, with no calls up.}
\end{aligned}$$

Variables a, b are used for members of A.

A *unit* assignment is, naturally, one that assigns exactly one inlet to some one outlet, and it corresponds to having just one call in progress. It is convenient to identify new calls c and unit assignments, and to write $\gamma(x) \cup c$ for the larger assignment consisting of $\gamma(x)$ and the call c together, with the understanding of course that none of the terminals of c is busy in $\gamma(x)$. Not every assignment need be realizable by some state of S. Indeed, it is common for practical networks to realize only a small fraction of the possible assignments.

A simple pseudometric topology on S is defined by the "distance" formula

$$d(x,y) = |\gamma(x)\Delta\gamma(y)|$$

where Δ denotes the symmetric difference, and $|\cdot|$ cardinality, of sets. The distance between states $d(x,y)$ is the number of pairs $(u,v) \in I \times \Omega$ that are either connected in x and not in y, or connected in y and not in x. Clearly, $d(x,y) = 0$ if and only if $x \sim y$, and the d-closure of a set X is just

$$X^d = \{y : y \sim x \text{ for some } x \in X\}.$$

A set X is dense in a set Y in the d-topology iff

$$Y \subseteq X^d.$$

III. INTERSECTION PROPERTY

We shall introduce a property of subsets X of the set S of states, called the *intersection property*, and then show that a network ν is nonblocking in the wide sense if and only if some subset of S has the intersection property. We call it the intersection property because it involves the equality case

$$\gamma(x \cap y) = \gamma(x) \cap \gamma(y) \tag{1}$$

of the semilattice homomorphism inequality

$$\gamma(x \cap y) \leqq \gamma(x) \cap \gamma(y); \tag{2}$$

the latter is *always* true. Our result therefore says roughly that if equality in (2) holds for enough states, then ν is wide-sense nonblocking and this condition is necessary.

A subset $X \subseteq S$ is said to have the *intersection property* if and only if for every $x \in X$ and every $a \in A$, there exists $y \in X$ such that $\gamma(y) = a$ and

$$\gamma(y) \cap \gamma(x) = \gamma(x \cap y).$$

A subset $X \subseteq S$ is *closed below* if $x \in X$ and $y \leqq x$ imply $y \in X$. The *lower closure* of a subset X is the set $\mathbf{X} = \{y : y \in S \text{ and } y \leqq x$ for some

$x \in X\}$; this is just all the states reachable from a member of X by hangups.

Our first result is an important lemma to the effect that the intersection property is preserved by lower closure.

Lemma 1: If X has the intersection property, then so does $\mathbf{X}$.

Proof: Take $x \in \mathbf{X}$, $a \in A$. We are to find $y \in \mathbf{X}$ such that $\gamma(y) = a$ and $\gamma(y \cap x) = \gamma(y) \cap \gamma(x)$. Since $x \in \mathbf{X}$, there is a $z \in X$ such that $x \leqq z$. X has the intersection property, so there is a $w \in X$ such that $\gamma(w) = a$ and

$$\gamma(w \cap z) = \gamma(w) \cap \gamma(z). \tag{3}$$

We show that we can choose y to be w. Obviously $w \in \mathbf{X}$ and $\gamma(w) = a$. Also, intersecting (3) with $\gamma(x)$ we find

$$\gamma(x) \cap \gamma(w \cap z) = \gamma(x) \cap \gamma(w) \cap \gamma(z).$$

Since $x \leqq z$, we have $\gamma(x) \leqq \gamma(z)$, $\gamma(x) \cap \gamma(z) = \gamma(x)$, and so the right-hand side is just $\gamma(w) \cap \gamma(x)$. The left-hand side consists of calls which are in progress in x, and are also in progress in both w and z, on the same routes in each. Since $x \leqq z$, these must use the same routes in x as they do in z and w. Thus the left-hand side comprises exactly those calls which are in progress in each of z, w, and x, on the same routes in each, namely $\gamma(z \cap w \cap x)$. This equals $\gamma(x \cap w)$ because $x \leqq z$. Thus

$$\begin{aligned}\gamma(x \cap w) &= \gamma(z \cap w \cap x)\\ &= \gamma(x) \cap \gamma(z \cap w)\\ &= \gamma(x) \cap \gamma(z) \cap \gamma(w)\\ &= \gamma(x) \cap \gamma(w),\end{aligned}$$

and this proves the Lemma 1. Our next result notes that a subset X having the intersection property must lie entirely in the set N of states in which no call is blocked.

Lemma 2: If X has the intersection property and $x \in X$, then no call idle in x is blocked in x, i.e., $X \subseteq N$.

Proof: Let $x \in X$, c idle in x, $a = \gamma(x) \cup c$. Then there is a $y \in X$ such that $\gamma(y) = a$ and $\gamma(x \cap y) = \gamma(x) \cap \gamma(y)$. Thus the calls in progress in both x and y, and on the same routes in each, are all and only the calls up in x. Hence $x \cap y = x$, or $x \leqq y$, so that $y \in A_x$, and c is not blocked in x. Thus $X \subseteq N$.

IV. WIDE-SENSE NONBLOCKING NETWORKS

We shall need a lemma that identifies the intersection of two states:

Lemma 3: If $z \leqq x$, $z \leqq y$, and $\gamma(z) = \gamma(x) \cap \gamma(y)$, then $z = x \cap y$.

Proof: The hypothesis implies that

$$\gamma(x - z) = \gamma(x) - \gamma(z) = \gamma(x) - [\gamma(x) \cap \gamma(y)]$$
$$\gamma(y - z) = \gamma(y) - \gamma(z) = \gamma(y) - [\gamma(x) \cap \gamma(y)].$$

The right-hand sides are disjoint, so $\gamma(x - z) \cap \gamma(y - z) = \phi$, and the homomorphism inequality for $\gamma(\cdot)$ gives $\gamma[(x - z) \cap (y - z)] = \phi$, whence $(x - z) \cap (y - z) = 0$. Since z is included in each of x and y, we have

$$x = z \cup (x - z), \qquad y = z \cup (y - z)$$
$$x \cap y = z \cup z(y - z) \cup (x - z)z \cup (x - z)(y - z)$$

(here we have used a more convenient notation for intersection on the right-hand side). The last three terms on the right vanish, so $x \cap y = z$.

The following characterization of wide-sense nonblocking was given in an earlier work:[1]

Theorem 1: ν is nonblocking in the wide sense iff *there exists a subset $X \subseteq N$ with $X = \mathbf{X}$, and such that for every $x \in X$, $A_x \cap X$ is d-dense in A_x, i.e., $A_x \subseteq (A_x \cap X)^d$.*

The principal new result is now proved. It is

Theorem 2: ν is nonblocking in the wide sense iff *some subset $X \subseteq S$ has the intersection property.*

Proof (sufficiency): By Lemmas 1 and 2 we can assume that X is closed below, and that $X \subseteq N$. By Theorem 1 it is enough to prove that for every $x \in X$, $A_x \cap X$ is d-dense in A_x, i.e.,

$$A_x \subseteq (A_x \cap X)^d, \quad x \in X. \tag{4}$$

Let $x \in X$ and $z \in A_x$. There exists then $y \in X$ such that $\gamma(y) = \gamma(z)$ and $\gamma(x \cap y) = \gamma(x) \cap \gamma(y)$; thus also

$$\gamma(x \cap y) = \gamma(x) \cap \gamma(z) = \gamma(x),$$

the second equality following from $x \leqq z$. As in Lemma 2, we conclude from $\gamma(x \cap y) = \gamma(x)$ that $y \in A_x$. Then $y \in A_x \cap X$ and $y \sim z$, or

$z \in (A_x \cap X)^d$. Since z was an arbitrary state in A_x, we have shown (4), and so the sufficiency.

Proof (necessity): Since ν is nonblocking in the wide sense, there exists by Theorem 1 a subset X of states which is closed below, is contained in N, and is such that any call new in a state of X can be put up *salva* staying in X. We show that X has the intersection property. Let then $x \in X$ and $a \in A$. Obtain a state $z \leqq x$ by removing from x all the calls that are not part of the assignment $\gamma(x) \cap a$. Next, starting at z, put up the (additional) calls comprising $a - \gamma(z)$ so as to reach a state $y \in X$ with $\gamma(y) = a$. This is possible because any call new in a state of X can be put up so as to keep the system in X. We now claim that

$$\gamma(x \cap y) = \gamma(x) \cap \gamma(y).$$

Since $z \leqq x$, $z \leqq y$, this follows from Lemma 3 as soon as we prove that $\gamma(z) = \gamma(x) \cap \gamma(y)$. To see this, note that $\gamma(z) \leqq \gamma(x)$ and $\gamma(z) \leqq \gamma(y)$, so that $\gamma(z) \leqq \gamma(x) \cap \gamma(y)$. Conversely, by construction, any call up in both x and y is either up in z (never having been disturbed), or else was taken out to reach z and then put back up. However, only calls not up in a were taken down, and only calls up in a were put back. Thus the second alternative is ruled out, and any call up in both x and y is up in z, i.e., $\gamma(x) \cap \gamma(y) = \gamma(z)$. Lemma 3 now implies that $z = x \cap y$, so that

$$\begin{aligned} \gamma(x) \cap \gamma(y) &= \gamma(z) \\ &= \gamma(x \cap y). \end{aligned}$$

Hence X has the intersection property, as claimed.

V. STRICTLY NONBLOCKING NETWORKS

Because of Lemma 2, the intersection property can also be used to characterize the property of being strictly nonblocking, as is shown by the following result:

Theorem 3: ν is strictly nonblocking iff (*the set of states*) S *has the intersection property.*

Proof: Sufficiency is obvious, by Lemma 2. Conversely, if ν is strictly nonblocking, then $\gamma(S) = A$ and

$$A_x \subseteq (A_x \cap S)^d, \quad \text{for every } x \in S.$$

Thus ν is nonblocking in the wide sense; indeed, trivially, S has the property that any call new in a state of S can be put up *salva* staying in S. The necessity argument of Theorem 2 now shows that S has the intersection property.

VI. COMPARISON, EMBEDDING, AND ISOMORPHISM

We next relate the intersection property to a certain partial ordering $\leqq$ of *networks*, introduced in an earlier work,[4] and used there for clarifying some problems of comparison of networks. This partial ordering was defined over the set $N(I,\Omega)$ of all networks $\nu = (G, I, \Omega, S)$ for which the set I of inlets and the set Ω of outlets are fixed, while the graph G and the set S of states may vary in any way consistent with their defining a network in the sense of Ref. 2.

$N(I,\Omega)$ is partially ordered by the following relation $\leqq$: $\nu_1 \leqq \nu_2$ iff $\exists$ domain $D \subseteq S(\nu_1)$ and an onto map μ: $D \rightarrow S(\nu_2)$ such that D is closed below and

(*i*) μ preserves assignments: $\gamma(\mu x) = \gamma(x)$

(*ii*) $x, y \in D, \mu x \geqq \mu y \Rightarrow x \geqq y$.

The relationship $\nu_1 \leqq \nu_2$ means intuitively that one can mimic ν_2 within ν_1. That this is so is not obvious. Indeed, using the notion of isomorphism as a precision of the mimicry in question, it has been proved[4] that $\nu_1 \leqq \nu_2$ if and only if there is an isomorph of ν_2 in ν_1. Roughly, in the definition, μ maps the states of ν_1 doing the mimicking *onto* $S(\nu_2)$; it tells what state mimics what. Condition (*i*) then naturally states that the mimicked state satisfies the same assignment. Condition (*ii*), finally, insists that mimicry preserve inclusion, in the sense that only states x, y with $x \geqq y$ can mimic similarly related states μx, μy.

Remark: In the definition of the partial ordering $\leqq$ for the set $N(I,\Omega)$, the condition $D = \mathbf{D}$, that the domain of the map μ be closed below, may be dropped, because it is implied by the other conditions. To see this, let D, μ be as in the definition of $\leqq$ except omit $D = \mathbf{D}$, and take $x \in D$, $y \leqq x$. We show $y \in D$. Clearly, $\gamma(y) \leqq \gamma(x) = \gamma(\mu x)$, so there is a state $z \in \mu(D)$ with $z \leqq \mu x$ and $\gamma(z) = \gamma(y)$, because $\mu(D)$ is closed below, since μ is onto. Hence there exists $w \in D$ with $z = \mu w$. Thus $\mu w \leqq \mu x$, so by the second property of μ, $w \leqq x$. We now have $y \leqq x$, $w \leqq x$, $\gamma(y) = \gamma(w)$. This implies $y = w$ and so $y \in D$, because a state x can have below it at most one state satisfying a given assignment.

Theorem 4: ν is nonblocking in the wide sense iff $\exists \nu_1$, $\nu \leqq \nu_1$ *and* ν_1 *is strictly nonblocking.*

Proof: If $\nu \leqq \nu_1$, there is a domain $D \subseteq S(\nu)$ and an onto map $\mu: D \rightarrow S(\nu_1)$ such that $\gamma(\mu x) = \gamma(x)$, and $\mu x \geqq \mu y$ implies $x \geqq y$. We show that D has the intersection property. Take $x \in D$ and $a \in A$ and focus on $\mu x \in S(\nu_1)$. Clearly, since ν_1 is strictly nonblocking, there exists a state y of ν_1 with $\gamma(y) = a$ and

$$\gamma(\mu x) \bigcap \gamma(y) = \gamma(\mu x \bigcap y).$$

(It suffices to take down the calls in x not up in a, and then put up the ones in a not up in x.) Since μ is onto, $y = \mu z$ for some $z \in D$, with $a = \gamma(y) = \gamma(z)$. Since $\gamma(\mu x) = \gamma(x)$, we have

$$\gamma(x) \bigcap \gamma(z) = \gamma(\mu x \bigcap \mu z).$$

Since μx and μz are both states of ν_1, so is $\mu x \bigcap \mu z$; there is a state $w \in D$ with $\mu w = \mu x \bigcap \mu z$, since μ is onto. Now note that $\mu x \geqq \mu w$ and $\mu y \geqq \mu w$, so that the second property of μ implies $x \geqq w$ and $y \geqq w$. Together with $\gamma(x) \bigcap \gamma(z) = \gamma(w)$ this implies by Lemma 3 that $w = x \bigcap z$, and so

$$\gamma(x) \bigcap \gamma(z) = \gamma(x \bigcap z).$$

Thus D has the intersection property, and so ν is wide-sense nonblocking, by Theorem 2. Conversely, if ν is nonblocking in the wide sense, there is a subset X of S with the intersection property. Define ν_1 by

$$\nu_1 = (G, I, \Omega, X).$$

Taking $D = X$ and $\mu =$ identity, we conclude $\nu \leqq \nu_1$; Lemma 2 implies that ν_1 is nonblocking, and Theorem 3 is proved.

Our intuitive feeling is that a wide-sense nonblocking network has embedded in it a largest strictly nonblocking network, to whose states the system is restricted by any rule for routing that guarantees no blocking. An appropriate sense of "embedded" is provided by the concept of isomorphism.[3] An isomorphism between two partially ordered systems is a one-to-one correspondence that preserves order in both directions. An isomorph of ν_2 within ν_1 would be a subset $M \subseteq S(\nu_1)$ and a correspondence $i: M \leftrightarrow S(\nu_2)$ such that $x \geqq y$ iff $ix \geqq iy$. In Ref. 4, the existence of an isomorph was related to the partial ordering $\leqq$ for networks, by this result:

Theorem 5: $\nu_1 \leqq \nu_2$ iff $\exists M \subseteq S(\nu_1)$, $\exists$ *correspondence* $i\colon M \leftrightarrow S(\nu_2)$ *such that*

(*i*) $\gamma(ix) = \gamma(x)$

(*ii*) $x \geqq y$ iff $ix \geqq iy$.

Therefore Theorem 4 can be rephrased as

Theorem 6: ν *is nonblocking in the wide sense if and only if there is an isomorph of a nonblocking network embedded in* $S(\nu)$, *and the isomorphism preserves* $\gamma(\cdot)$.

This is a precise form of the intuitive feeling voiced above.

Note added in proof: It should be noticed that Theorems 4 and 6 imply that the quest (mentioned at the top of p. 698) for *efficient* wide-sense nonblocking networks is in a sense vain: there is no "intermediate" amount of switching equipment that will give wide-sense nonblocking behavior but is not so expensive as (it would have to be to give) a strictly nonblocking network; as soon as you have a wide-sense nonblocking network, you have at most to throw away some states to obtain a strictly nonblocking one.

REFERENCES

1. Beneš, V. E., "Algebraic and Topological Properties of Connecting Networks," B.S.T.J., *41*, No. 4 (July 1962), pp. 1249–1274.
2. Beneš, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
3. Birkhoff, G., *Lattice Theory*, Am. Math. Soc. Colloq. Publ. (rev. ed.) *XXV* (1948).
4. Beneš, V. E., "On Comparing Connecting Networks," Journal of Combinatorial Theory, *2*, No. 4 (June 1967), pp. 507–513.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

Efficient Evaluation of Integrals of Analytic Functions by the Trapezoidal Rule

By S. O. RICE

(Manuscript received November 28, 1972)

Definite integrals of analytic functions can often be evaluated efficiently by the trapezoidal rule after a suitable transformation. Here the work of Moran[1] *and Schwartz*[2] *along this line is extended. First the dependence of the error on the spacing is discussed, and then several types of transformations are described and applied to integrals of technical interest.*

I. INTRODUCTION

Quite often the problem of determining the value of a definite integral arises. When the integral cannot be readily evaluated by analysis, we must resort to numerical methods. Here we discuss a method of numerical quadrature which gives promise of being useful in evaluating some types of integrals that are difficult to handle by conventional numerical methods.

In particular, we consider the problem (Moran[1] and Schwartz[2]) of transforming a given integral of an analytic function $f(x)$ into a rapidly converging one (with limits $\pm\infty$) which can be efficiently evaluated by the trapezoidal rule,

$$\int_{-\infty}^{\infty} f(x)dx = h \sum_{n=-\infty}^{\infty} f(nh) - E. \tag{1}$$

The integral and series are assumed to converge. In addition to the trapezoidal error E, a second error is introduced when the series is truncated in the process of computation. It is supposed that both errors are made negligible, E by taking h small, and the truncation error by taking enough terms in the series. The feature which makes the use of (1) attractive is that E often decreases in proportion to $\exp(-C/h)$ as h decreases, C being a constant. Thus if h gives three-figure accuracy, $h/2$ will give six-figure accuracy in many cases.

The transformations, i.e., the changes of variable of integration used to carry the limits of the given integral into $\pm\infty$, are usually constructed by combining functions which are readily computed, such as powers and exponential functions.

Schwartz[2] recommends the following procedure for evaluating the integral of an analytic function $f(u)$: (*i*) change the variable from u to v so as to make the integration with respect to v extend from $-\infty$ to $+\infty$, (*ii*) make the further transformation

$$v = e^{x} - e^{-x}$$

to increase the rate of convergence, and then (*iii*) evaluate the trapezoidal sum, truncating when contributions fall below the desired accuracy, and reduce the spacing h until the answer has the desired accuracy.

Here we present a summary of several variations of Schwartz's procedure. Details are given in a report by the author.[3] The dependence of the error E on the spacing h is first reviewed, and then examples are used to illustrate the evaluation of various types of integrals.

II. THE DEPENDENCE OF E ON h

The trapezoidal error E can be expressed in several ways. For example, it is the remainder in the Euler-Maclaurin sum formula. Again, it can be written as the sum of contour integrals with integrands $f(z)/[\exp(\pm i2\pi z/h) - 1]$ (Ref. 4, p. 145, Problem 7, and Refs. 5, 6, 7, 8, and 9). Here we follow Fettis[5] and use Poisson's summation formula which, when applied to (1), gives

$$E = \left(\sum_{k=-\infty}^{-1} + \sum_{k=1}^{\infty} \right) \int_{-\infty}^{\infty} f(z) \exp(i2\pi zk/h) dz. \tag{2}$$

Let $f(z)$ be analytic throughout a strip in the z-plane containing the real z-axis and assume that suitable convergence conditions are satisfied. Then the paths of integration in the terms for $k > 0$ in (2) can be displaced upwards to make $\text{Im}(z) > 0$. It follows that $|\exp(i2\pi kz/h)| = \exp(-2\pi k \,\text{Im}(z)/h)$ becomes small when h becomes small and z is on the path. Furthermore, as $h \to 0$ the terms for $k > 1$ become negligible in comparison with the term for $k = 1$. A similar argument holds for the $k < 0$ terms, and as $h \to 0$ we have the asymptotic result

$$E \sim R_{+} + R_{-}$$

where R_{+} and R_{-} are the $k = 1$ and $k = -1$ terms, respectively, in (2). For the important case in which $f(z)$ is real on the real z-axis, R_{-}

is equal to the conjugate complex R_+^* of R_+ and E is given asymptotically by

$$E \sim R_+ + R_+^*,$$
$$R_+ = \int_{-\infty}^{\infty} f(z) \exp (i2\pi z/h) dz. \tag{3}$$

When h is small, the trapezoidal error E given by (3) can be viewed as the sum of contributions from singularities of $f(z)$ and saddle points of $f(z) \exp (i2\pi z/h)$. At the saddle points the derivative $d\varphi(z)/dz$ is zero, $\varphi(z)$ being defined to within a multiple of $2\pi i$ by $\exp [\varphi(z)] = f(z) \exp (i2\pi z/h)$.

This picture of E is suggested by the following remarks. The path of integration in (3) can be deformed upwards towards $z = i\infty$ in the complex z-plane until it becomes an optimal path comprised of paths of steepest descent and ascent passing through one or more saddle points. The path runs from $-\infty$ to $+\infty$ and may have detours running out to, and returning from, infinity. It may also have loops around some of the singularities of $f(z)$. When h is small, the factor $\exp (i2\pi z/h)$ decreases rapidly as Im (z) increases, and the only significant contributions to R_+ come from the portions of the path near the singularities and near the saddle points not associated with singularities.

An approximate expression for a typical contribution can be obtained by expanding the contribution about the corresponding singularity or saddle point and taking the leading term. Such expansions are usually asymptotic in h. For estimating orders of magnitude we can use the dominant factors in the leading terms:

$$\text{(Contribution to } R_+ \text{ of a saddle point at } z_0) \approx \exp [\varphi(z_0)],$$
$$\text{(Contribution to } R_+ \text{ of a singularity at } z_1) \approx \exp (i2\pi z_1/h).$$

As h decreases, E may either decrease steadily or may oscillate with decreasing amplitude depending upon how the dominant contributions combine.

If there are no singularities or saddle points in the finite part of the z-plane, the trapezoidal rule may give the exact value of the integral when h is less than some fixed value. This is associated with the sampling theorem for band-limited functions. For example, if m and n are positive integers such that $m - n = 0$ or $2, 4, \cdots$, the integral

$$I = \int_{-\infty}^{\infty} \sin^m x dx/x^n \tag{4}$$

is exactly equal to the trapezoidal sum when $h < 2\pi/m$ (h can equal

$2\pi/m$ if $n \geqq 2$). The proof follows from the fact that all of the terms in the series (2) for E vanish when $h < 2\pi/m$, as can be shown by deforming the paths of integration into infinite semicircles and using Jordan's lemma (Ref. 4, p. 115). As a check, note that for $m = n = 1$ or 2 we can take $h = \pi$. Then $I = \pi$ because there is only one nonzero term in the sum.

The foregoing discussion shows that the structure of E can be determined by computing the saddle points and associated paths of steepest descent for $f(z) \exp(i2\pi z/h)$. This is done in the report[3] for the examples (6) and (25) given below. The path for example (6) is shown in Fig. 1 and discussed in the appendix. However, computations of this sort are laborious. In practice it appears that the dependence of E on h is most easily determined by computing the trapezoidal sum for a sequence of decreasing values of h, bearing in mind the possibility that E may go through zero for some values of h.

Incidentally, arguments similar to those given in this section show that the trapezoidal rule also works well when it is used to evaluate integrals of periodic analytic functions in which the integration extends over a period.

III. CONTRIBUTIONS TO E—EXAMPLE

Goodwin[6] has pointed out that the trapezoidal rule usually performs well for integrals of the type

$$I = \int_{-\infty}^{\infty} g(x) \exp(-bx^2)dx. \tag{5}$$

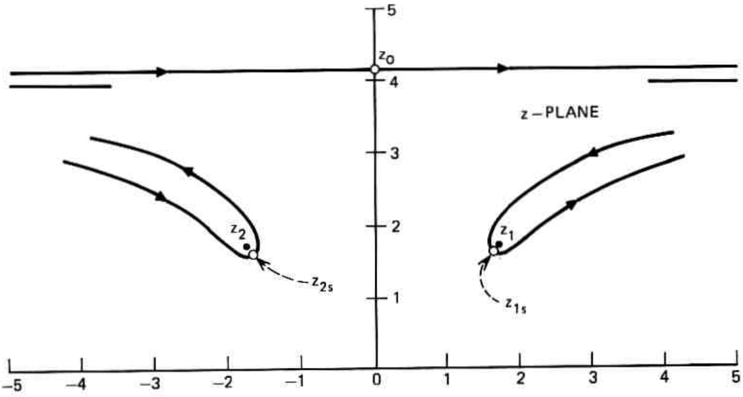


Fig. 1—Steepest descent paths for $\exp[\varphi(z)]$ when $a = 2.4$ and $h = 0.8$ in both (30) and example (6). The points z_0, z_{1s}, z_{2s} are saddle points; z_1, z_2 are branch points.

Computations[3] made with $b = 1$,

$$g(x) = (x^4 + a^4)^{-\frac{1}{2}}, \tag{6}$$

and $a = 2.4$ show that, as h decreases, E first decreases steadily and then near $h = 1$ (where $E/I \approx 10^{-4}$) E starts to oscillate with decreasing amplitude. This behavior can be explained in terms of a contribution to E of approximately

$$2\pi^{\frac{1}{2}} \operatorname{Re} [g(i\pi/h) \exp(-\pi^2/h^2)] \tag{7}$$

from a saddle point near $z = i\pi/h$ (see Fig. 1, eq. (32), and Goodwin[6]) and contributions from the branch points at $z_1 = a \exp(i\pi/4)$ and $z_2 = a \exp(i3\pi/4)$. The expressions for the branch point contributions are somewhat more complicated than (7), but all that need be noted here is that they contain $\exp(i2\pi z_k/h)$, $k = 1, 2$, as a dominant factor. Adding the three contributions and neglecting multipliers such as $g(i\pi/h)$ in (7) shows that E is roughly

$$\exp(-\pi^2/h^2) + (\cos\alpha)\exp(-\beta) \tag{8}$$

where $\beta = 2^{\frac{1}{2}}\pi a/h$ and $\cos\alpha$ oscillates with increasing frequency as h decreases. The steady decrease of E with h, dominated by $\exp(-\pi^2/h^2)$, changes to an oscillating decrease when $\beta = \pi^2/h^2$. Solving for h with $a = 2.4$ gives $h = 0.93$, which agrees with the observed $h \approx 1.0$.

When $g(x)$ in (5) is algebraic and $g(z)$ has no singularities inside the rectangle with corners at $\pm z_0$, $\pm|z_0|$, where $z_0 = i\pi/(bh)$ and $\operatorname{Re}(b) > 0$, the error E tends to be dominated by the saddle point contribution. This contribution is approximately $|\exp(bz_0^2)|$ in the sense that $\exp(-\pi^2/h^2)$ approximates the approximation (7). When the rectangle contains singularities, their contributions dominate. For the example (6), $z_0 = i\pi/h$ and the rectangle becomes a square which expands as h decreases. The behavior of E changes when the sides of the square sweep across the branch points.

IV. INTEGRALS WITH BOTH LIMITS FINITE

Consider the integral

$$I = \int_a^b (u-a)^{\alpha-1}(b-u)^{\beta-1} f(u)du, \qquad \alpha, \beta > 0, \tag{9}$$

where $f(u)$ is analytic and $0(1)$, $f(a)$ and $f(b) \neq 0$, and a and b are finite. The transformation

$$\begin{aligned} u &= (be^v + ae^{-v})/(e^v + e^{-v}), \\ du/dv &= 2(b-a)/(e^v + e^{-v})^2, \end{aligned} \tag{10}$$

carries the limits into $v = \pm\infty$. The associated equations

$$u - a = e^{v}(b - a)/(e^{v} + e^{-v}),$$
$$b - u = e^{-v}(b - a)/(e^{v} + e^{-v}),$$

and du/dv show that the dominant factors in the integrand when $v \to +\infty$ and $v \to -\infty$ are $\exp(-2\beta v)$ and $\exp(2\alpha v)$, respectively. One might expect, and it is confirmed by computation, that in a good transformation the final integral should converge at $-\infty$ at nearly the same rate as it does at $+\infty$. Therefore the further change of variable

$$\begin{aligned} v &= c(\beta^{-1}e^{x} - \alpha^{-1}e^{-x}) \\ dv/dx &= c(\beta^{-1}e^{x} + \alpha^{-1}e^{-x}) \end{aligned} \tag{11}$$

is made to equalize the rates of convergence at $x = \pm\infty$. The transformation (11) makes the integrand behave roughly as $\exp[-2c \exp|x|]$ as $x \to \pm\infty$ when the effect of $f(u)$ is ignored.

The constant c in (11) can be chosen somewhat arbitrarily. It is helpful, but not necessary, to have

$$c \leqq \pi(\beta\alpha)^{\frac{1}{2}}/4. \tag{12}$$

The inequality (12) for c guarantees that the singularities in the complex x-plane due to the vanishing of $e^{v} + e^{-v}$ are at least $\pi/2$ distant from the real x-axis. It says nothing about the singularities and saddle points introduced by $f(u)$.

Thus the integral to be evaluated by the trapezoidal rule is

$$I = \int_{-\infty}^{\infty} (u - a)^{\alpha-1}(b - u)^{\beta-1}f(u)\,\frac{du}{dv}\,\frac{dv}{dx}\,dx \tag{13}$$

which can also be written as

$$I = 2(b - a)^{\alpha+\beta-1}\int_{-\infty}^{\infty} \frac{e^{(\alpha-\beta)v}}{(e^{v} + e^{-v})^{\alpha+\beta}}\,f(u)\,\frac{dv}{dx}\,dx.$$

Here x is the variable of integration and, in writing the program, u, v, du/dv, and dv/dx are given by (10) and (11).

As an example of (9) and (13), consider the beta function

$$\begin{aligned} I &= \int_{0}^{\pi/2} [\sin u]^{\alpha-1}\left[\sin\left(\frac{\pi}{2} - u\right)\right]^{\beta-1} du \\ &= \int_{-\infty}^{\infty} [\sin u]^{\alpha-1}\left[\sin\left(\frac{\pi}{2} - u\right)\right]^{\beta-1} \frac{du}{dv}\,\frac{dv}{dx}\,dx \\ &= \pi c\int_{-\infty}^{\infty} [\sin u]^{\alpha-1}\left[\sin\left(\frac{\pi}{2} - u\right)\right]^{\beta-1}\left[\frac{\beta^{-1}e^{x} + \alpha^{-1}e^{-x}}{(e^{v} + e^{-v})^{2}}\right] dx \end{aligned}$$

where the third line is to be evaluated by the trapezoidal rule. For $\alpha = 0.95$ and $\beta = 0.05$, the value of I is known to be $20.748\ 732 \cdots$ and the inequality (12) for the multiplier c gives $c \leqq 0.171$.

Computations show that the values 0.171, 0.1, and 0.05 for c all give six-figure accuracy with $h = 0.5$ and about 20 terms in the trapezoidal sum (e.g., for $c = 0.1$, $h = 0.5$, and 21 terms, the computed value is 20.748 729). When $c = 1$, about 70 terms (with $h = 0.075$) are required to achieve the same accuracy.

Instead of computing the $\alpha - 1$ power of simply $[\sin u]$, it was found better to compute the $\alpha - 1$ power of $[(e^v + e^{-v})^2 \sin u]$, and similarly for $\sin\left(\frac{\pi}{2} - u\right)\cdot$ In general, it is usually helpful to combine as much as possible of the $1/(e^v + e^{-v})^2$ contained in du/dv with other factors in the integrand in order to avoid underflow and overflow.

When $\alpha = \beta = 1$, the transformations (10) and (11) reduce to the ones used by Schwartz[2] except for the coefficient $c \leqq \pi/4 = 0.785$. To illustrate this case take (Kajfez[10])

$$I = (-\pi/40) \int_{10}^{15} \exp\,(u/4) \sin\,[0.4\pi \exp\,(u/4)]du$$

in which the integrand oscillates through about 6 cycles. Using (10) and (11) with $a = 10$, $b = 15$, $c = 0.785$ shows that the trapezoidal rule with $h = 0.09$ and 60 terms gives $I = -0.0195495$ compared with the true value $-0.0195488 \cdots$. For relative errors less than about 0.01, the trapezoidal rule requires less terms than the spline quadrature methods considered in Ref. 10, but this is offset somewhat by the more complicated terms introduced by (10) and (11).

V. INTEGRALS WITH LIMITS 0, ∞ CONTAINING $u^{\alpha-1}(1 + u)^{-\alpha-\beta}$

The integral

$$I = \int_0^\infty u^{\alpha-1}(1 + u)^{-\alpha-\beta} f(u)du, \qquad \alpha, \beta > 0, \tag{14}$$

where $f(0) \neq 0$ and $f(u)$ is analytic and $0(1)$, can be handled by the transformations

$$u = e^v, \qquad v = c(\beta^{-1}e^x - \alpha^{-1}e^{-x}) \tag{15}$$

where $c \leqq \pi(\alpha\beta)^{\frac{1}{2}}/2$. For $\alpha = 3$, $\beta = 2$ the inequality for c becomes $c \leqq 3.85$, and for $\alpha = 0.2$, $\beta = 0.1$ it becomes $c \leqq 0.222$. Computations were made for these values of α and β with $f(u) = 1$. Values of h and the number of terms N in the trapezoidal sum required for sevenfigure accuracy were found to be as follows:

α	β	c	h	N
3.0	2.0	3.85	0.25	15
		2.00	0.35	15
		5.00	0.10	40
0.2	0.1	0.22	0.45	25
		0.08	0.45	25
		0.45	0.25	35

VI. INTEGRALS WITH LIMITS 0, ∞ CONTAINING $u^{\alpha-1}\exp(-u)$

For the integrals

$$I = \int_0^\infty u^{\alpha-1}e^{-u}f(u)du, \qquad \alpha > 0, \tag{16}$$

where $f(0) \neq 0$ and $f(u)$ is analytic and $0(1)$, we can use

$$u = e^v, \qquad v = x - \alpha^{-1}e^{-x}. \tag{17}$$

Computations for the case $f(u) = 1$ and $\alpha = 1$ gave the following results:

h	N = No. of terms	Trap. values of I
0.4	15	0.9999 9997
0.6	10	0.9999 8711
0.8	7	0.9998 2442

Repeating the computations with $u = c\exp[x - c\exp(-x)]$ and $c = 0.5$, 2.0, and 4.0 showed that the magnitude of the error depends only slightly on c.

VII. INTEGRANDS WHICH CHANGE RAPIDLY NEAR A POINT

When the integrand contains a factor, say $F(t)$ where t is the variable of integration, which changes rapidly near a point it is sometimes helpful to change to a new variable of integration u where $du/dt = F(t)$ and the constant of integration is chosen at our convenience. The success of the transformation depends upon the ease of inverting to get t as an easily computed function of u.

As an example, consider

$$I = \int_{-1}^{1} e^t(t^2 + a^2)^{-\frac{1}{2}}dt \tag{18}$$

where a is small (Smith and Lyness[11]). Taking $du/dt = F(t) = (t^2 + a^2)^{-\frac{1}{2}}$, $u = \text{arcsinh}\,(t/a)$, and $t = a\sinh u$ carries (18) into an integral of the form (9) with $b = -a = A$ and $\alpha = \beta = 1$:

$$I = \int_{-A}^{A} e^t du = \int_{-\infty}^{\infty} e^t \frac{du}{dv}\frac{dv}{dx}\, dx$$
$$= 4Ac \int_{-\infty}^{\infty} e^t(e^x + e^{-x})dx/(e^v + e^{-v})^2.$$

Here $A = \operatorname{arcsinh}\ (1/a)$, $t = a \sinh (A \tanh v)$, $v = 2c \sinh x$, and $c \leqq \pi/4 = 0.785$. Computations with $c = 0.3$, $h = 0.2$, and 40 terms in the trapezoidal sum show that $I = 29.538\,618 \pm 10^{-6}$ when $a = 10^{-6}$.

If the rapidly changing factor has the more general form

$$F(t) = (t^2 + a^2)^{-\theta},$$

the transformation $t = a \sinh u$ can still be used to carry the integral into the form (9), but computation shows that the trapezoidal rule requires more terms as θ moves away from $\frac{1}{2}$. For example, when the exponent $-\frac{1}{2}$ in (18) is replaced by $-\frac{3}{4}$, i.e., $\theta = \frac{3}{4}$, computations with $a = 10^{-6}$, $c = 0.785$, and $h = 0.03$ show that 100 terms give the value 5240.808 for I whereas the true value is $5240.806 \cdots$.

VIII. THE FERMI-DIRAC INTEGRAL

The Fermi-Dirac integral (tabulated by Blakemore[12])

$$I = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} t^{\alpha-1}dt/(1 + e^{t-a}), \qquad \alpha > 0, \tag{19}$$

has an integrand which changes rapidly near $t = a$ when a is large. Section VII suggests taking $du/dt = F(t) = 1/(1 + e^{t-a})$. Choosing the constant of integration to make $u = 0$ at $t = 0$ gives

$$\begin{aligned} u &= \ell n(1 + e^{-a}) - \ell n(e^{-t} + e^{-a}), \\ t &= -\ell n(e^{-u} + e^{-u-a} - e^{-a}), \\ I &= \frac{1}{\Gamma(\alpha)} \int_0^b t^{\alpha-1}du, \end{aligned} \tag{20}$$

where $b = \ell n(1 + e^a)$. When u tends to 0, $t \rightarrow (1 + e^{-a})u$, and hence the integral (20) is of the form (9) with $\beta = 1$.

The transformations (10) and (11) carry I into

$$I = \frac{1}{\Gamma(\alpha)} \int_{-\infty}^{\infty} t^{\alpha-1} \frac{du}{dv}\frac{dv}{dx}\, dx$$
$$= \frac{2bc}{\Gamma(\alpha)} \int_{-\infty}^{\infty} t^{\alpha-1}(e^x + \alpha^{-1}e^{-x})dx/(e^v + e^{-v})^2 \tag{21}$$

where $c \leqq 0.785\alpha^{\frac{1}{2}}$, t is defined in terms of u by (20), and

$$u = be^{v}/(e^{v} + e^{-v}), \qquad v = c(e^{x} - \alpha^{-1}e^{-x}).$$

The integral (21), with $c = 0.5$, was used to compute I for $\alpha = \frac{1}{2}$ and $\frac{3}{2}$ with a between 0 and 20. Difficulty in computing t for small values of u was avoided by using three terms in the expansion of $\ell n(1 - z)$ when $z = \exp(u - a) - \exp(-a)$ was less than 0.001. The following tabulation shows the results for the typical values $\alpha = \frac{1}{2}$ and $a = 10$.

h	N = No. of terms	Trap. values of I
0.2	34	3.5527 792
0.3	22	792
0.4	17	795
0.5	14	742

The above transformation has been used by W. K. Kent in a study of the charge distribution in a charge coupled device and I am indebted to him for helpful discussions regarding his experience with (21).

IX. INTEGRALS WITH LIMITS 0, ∞ AND OSCILLATING INTEGRANDS

Kluyver's Bessel function (random walk) integral for the probability P that the resultant of the sum of m randomly phased unit vectors in a plane be less than r in length is (Bennett[13], Greenwood and Durand[14])

$$P = \int_0^\infty rJ_1(ru)[J_0(u)]^m du. \tag{22}$$

This integral is typical of a class of integrals that are difficult to evaluate by any means. They are characterized by rather slow convergence and an integrand which tends to oscillate at a regular rate as $u \to \infty$. In this section we consider the evaluation of such integrals by the trapezoidal rule when the rate of convergence is not too slow.

Some general remarks can be made concerning integrals that behave like (22). In order that

$$I = \int_0^\infty f(u)du$$

may represent the typical integral of this section, $f(u)$ must tend, as $u \to \infty$, to a form that can be written as a steadily decreasing factor times the sum of a finite number of sinusoidal terms whose periods tend to constant values (as $u \to \infty$). Define h_0 to be the shortest constant period (so that the most rapidly oscillating term varies as $\cos[(2\pi u/h_0) + \beta]$). The interval h_0 is related to the sampling theorem for band-limited functions in that $1/h_0$ plays the role of the "bandwidth" of $f(u)$ at $u = \infty$. Quite often the required accuracy in I can be obtained by using values of h which are close to h_0.

If the integrand is an even function of u, the trapezoidal rule can be applied directly. If the integrand is not even, the integral can be evaluated by setting

$$u = a\ln(1 + e^{x/a}), \qquad du/dx = e^{x/a}/(1 + e^{x/a}) \tag{23}$$

and then using the trapezoidal rule. Computations described below show that the choice $a = 1$ works well for (22). More will be said later about the choice of a. The transformation (23) takes advantage of the fact that the trapezoidal error E is often small, or zero, for regularly oscillating integrands. Although many terms may be needed in some cases, the present method compares favorably with competing ones.

The choice of a in (23) depends upon the behavior of $f(u)$ near $u = 0$. Suppose that $f(u)$ tends to Cu^ν as $u \to 0$. Here C is a constant and $\nu > -1$. After the change of variable (23), the integrand is a function of x which approaches 0 as $\exp[(\nu + 1)x/a]$ when $x \to -\infty$. When $(\nu + 1)h/a$ is too small, successive terms in the negative x portion of the trapezoidal sum are nearly equal and an unduly large number of terms must be taken to achieve a small truncation error on the left. When $(\nu + 1)h/a$ is too large, the successive terms differ by a large amount and the trapezoidal error E tends to be large. The problem is to choose a value of a which balances these two effects and at the same time allows h to be large. The choice $a = (\nu + 1)h_0$ works well for all of the cases that have been tried.

Some insight regarding a good choice of h for (22) can be obtained as follows. Consider the integral, say K, obtained by replacing $J_1(ru)$ by $J_0(ru)$ in (22) and taking the limits of integration to be $u = \pm\infty$. With the help of the asymptotic expression for $J_0(z)$ and the procedure used to deal with (4), it can be shown that when K is evaluated by the trapezoidal rule the error is zero if $h < h_0$ where $h_0 = 2\pi/(r + m)$. Furthermore, when P is transformed by (23) and then evaluated by the trapezoidal rule the error E is relatively small when h is only slightly less than h_0—as might be hoped from the similar behavior of the integrands in P and K as $u \to \infty$. A saddle point analysis of R_+ in (3) leads to the rough estimate

$$|E| \approx 2r^{\frac{1}{2}}(2\pi^2)^{-(m+1)/2} \exp[\pi(r + m - 2\pi h^{-1})] \tag{24}$$

for the error in the trapezoidal sum for P when $a = 1$.

In the example (22) the asymptotic expression for the integrand contains the product $\cos [ru - (3\pi/4)] \cos^m [u - (\pi/4)]$ which can be written as the sum of $m + 1$ sinusoidal terms. The most rapidly oscillating term is proportional to $\cos [(r + m)u - (m + 3)(\pi/4)]$ and the quantity $h_0 = 2\pi/(r + m)$ now appears as the shortest period.

Now we turn to the details of the evaluation of the integral (22) for P. Let r and m have the representative values $r = 4$ and $m = 6$. Then $h_0 = 2\pi/(r + m) = 0.628$. Since $J_0(z) \rightarrow 1$ and $J_1(z) \rightarrow z/2$ as $z \rightarrow 0$, the exponent ν is 1. The suggested value $a = (\nu + 1)h_0$ gives $a = 2h_0 = 1.26$, and since the choice of a is not critical, we take $a = 1$. Putting $a = 1$ in (23), substituting in (22), and using the trapezoidal rule gives the values of P shown in column 3:

h	No. of terms	P	Error: Col. 3	$\lvert E\rvert$ from (24)
0.475	286	0.9375 5485		
0.500	272	5475	$-10. \times 10^{-8}$	3.5×10^{-8}
0.525	259	5437	-4.8×10^{-7}	2.8×10^{-7}
0.550	247	5354	-1.3×10^{-6}	1.3×10^{-6}
0.575	236	5791	$+3.1 \times 10^{-6}$	6.9×10^{-6}
0.600	226	0.9375 9798	$+4.3 \times 10^{-5}$	2.4×10^{-5}
0.625	217	0.9376 9974	$+1.4 \times 10^{-4}$	0.94×10^{-4}

The fourth column gives the error estimated from column 3. The fifth column shows the approximation (24) for $|E|$. The trapezoidal sum was truncated at $x = 124$. Beyond 124 the absolute value of the integrand remains less than 2×10^{-8} and its amplitude decreases as $x^{-7/2}$. Note that the error starts to be appreciable as h approaches the critical value $h_0 = 2\pi/(r + m) = 0.628$.

X. THE INTEGRAL OF $u^k \exp(-u^2 - au^{-1})$ FROM 0 TO ∞

The integral

$$I_k(a) = \int_0^\infty u^k \exp(-u^2 - au^{-1})du, \qquad \mathrm{Re}\,(a) \geqq 0, \qquad (25)$$

is of interest in some physical problems (I wish to thank J. N. Lyness for calling my attention to this example). First let k be 0 and a be positive real. We seek a change of variable from u to x, with new limits $x = \pm\infty$, which will make u^2 tend to $\exp(2x)$ as $x \rightarrow \infty$ and a/u tend to $\exp(-2x)$ as $x \rightarrow -\infty$. This leads to

$$u = ae^x/(a + e^{-x}). \qquad (26)$$

Substituting (26) in (25), taking the special case $k = 0$, $a = 1$, and applying the trapezoidal rule gives the values of $I_o(1)$ shown in column 3:

h	No. of terms	$I_o(1)$	Obs. error	$\lvert E\rvert$ from (27)
0.1	38	0.1500 4597	2×10^{-8}	
0.2	19	0.1500 4597	2×10^{-8}	5×10^{-9}
0.3	12	0.1500 4835	2.40×10^{-6}	7×10^{-6}
0.4	10	0.1501 2711	8.12×10^{-5}	20×10^{-5}

The last column lists a rough approximation obtained by saddle point analysis:

$$|E| \approx \exp\left[-\frac{\pi^2}{2h} + \left(\frac{2\pi}{h}\right)^{\frac{1}{2}}\right]. \tag{27}$$

The observed error shown in column 4 is the trapezoidal sum (column 3) minus the value 0.1500 4595 of $I_o(1)$ computed from

$$\begin{aligned} I_k(a) &= \frac{1}{4\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(-s)\Gamma\left(\frac{k-s+1}{2}\right) a^s ds \\ &= \sum_{n=0}^{k} \frac{(-a)^n}{2(n!)}\Gamma\left(\frac{k-n+1}{2}\right) + \sum_{n=1}^{\infty} \frac{\pi}{2}\frac{(-1)^{n+k}a^{2n+k}}{(2n+k)!\Gamma(n+\frac{1}{2})} \\ &\quad + \sum_{n=0}^{\infty} \frac{[\ell n(a) - \psi(2n+k+2) - \frac{1}{2}\psi(n+1)]}{n!(2n+k+1)!}(-1)^{n+k}a^{2n+k+1} \end{aligned}$$

where $c < \min\ (0,\ k+1)$, $\psi(x) = (d\Gamma(x)/dx)/\Gamma(x)$, and the series holds for $k = -1, 0, 1, 2, 3, \cdots$ except that the first sum is omitted when $k = -1$.

When a is complex, say $a = \rho e^{i\alpha}$ where $|\alpha| \leqq \pi/2$, we tilt the path of integration in (25) by setting $u = ve^{i\theta}$ where $|\theta| \leqq \pi/4$ and $|\alpha - \theta| < \pi/2$. If we choose θ to be $\alpha/3$, the new integral

$$I_k(\rho e^{i\alpha}) = \exp\left[i(k+1)\alpha/3\right]\int_0^{\infty} v^k \exp[-(v^2 + \rho v^{-1})e^{i2\alpha/3}]dv \tag{28}$$

can be evaluated by using the substitution (26) with u and a replaced by v and ρ, and then applying the trapezoidal rule. For the physically important case of $k = 3$ and imaginary a [$I_3(i\rho)$ and $I_k(a)$, $k = 1, 2, 3$, are tabulated in NBS Handbook,[15] Section 27.5], the error can be kept below 1×10^{-6} for $a = i0.001$ by using $h = 0.04$ and 95 terms in the trapezoidal sum. As a increases to $i10.0$, h can increase to 0.08 and the required number of terms decrease to 40.

APPENDIX

Examples of Paths of Steepest Descent for R_+

For the example (6), the integral (3) for R_+ becomes

$$R_+ = \int_{-\infty}^{\infty} \exp\,[\varphi(z)]dz, \tag{29}$$

$$\varphi(z) = -z^2 + i2\pi h^{-1}z - \tfrac{1}{2}\ell n(z^4 + a^4). \tag{30}$$

The saddle point equation $d\varphi(z)/dz \equiv \varphi'(z) = 0$ is of the 5th degree in z. Solving by Newton's rule or otherwise gives 5 saddle points, 3 of which are in the half-plane Im $(z) > 0$. They are shown as small circles in Fig. 1 for the case $a = 2.4$ and $h = 0.8$. One is on the imaginary z-axis at $z_0 = i4.14$, and the other two (z_{1s}, z_{2s}) are at $\pm 1.648 + i1.629$ near the branch points (z_1, z_2) at $a(\pm 1 + i)/2^{\frac{1}{2}} = 1.697(\pm 1 + i)$.

The path of steepest descent through the saddle point z_0 (a path on which Im $[\varphi(z) - \varphi(z_0)] = 0$) was computed by: (*i*) evaluating the phase of the coefficient $[-2\pi/\varphi''(z_0)]^{\frac{1}{2}}$ in the saddle point contribution

$$\int \exp[\varphi(z)]dz \sim [-2\pi/\varphi''(z_0)]^{\frac{1}{2}} \exp\,[\varphi(z_0)] \tag{31}$$

to determine the direction of the path through z_0; (*ii*) selecting two starting points (one for each branch of the path) on opposite sides of, but close to, z_0; and (*iii*) applying

$$z_\ell = z_{\ell-1} + d_{\ell-1}, \qquad d_\ell = -\,|\,\varphi'(z_\ell)\,|\,\Delta/\varphi'(z_\ell)$$

to compute the path step by step, Δ being the step length. The other paths of steepest descent shown in Fig. 1 were computed in the same way. Paths of steepest descent through a saddle point may run down into a "lower" saddle point, i.e., one having a more negative Re $\varphi(z)$, or may end at a point (possibly $z = \infty$) where Re $\varphi(z) = -\infty$.

Figure 1 shows that the path of integration for R_+ can be deformed in a natural way into three portions: the loop around z_2 (which encloses the branch cut from z_2), the path from $-\infty + i\pi h^{-1}$ through z_0 to $+\infty + i\pi h^{-1}$, and lastly the loop around z_1. The path directions are denoted by arrows.

As h increases (from 0.8), z_0 in Fig. 1 moves downward towards the origin. Eventually h reaches a critical value at which the path of steepest descent from z_0 runs directly into z_{1s} and z_{2s}. For still larger values of h, the loops around z_1 and z_2 lie above the path through z_0, and the deformed path of integration for R_+ consists only of the path running from $-\infty + i\pi h^{-1}$ to $+\infty + i\pi h^{-1}$ through z_0. The sudden change in the path as h passes through the critical value is related to Stokes phenomena in the theory of asymptotic expansions.

The approximation (7) for the contribution of z_0 to $E \sim 2\,\mathrm{Re}\,(R_+)$ can be obtained by approximating the saddle point equation $\varphi'(z) = 0$ by $\varphi_A'(z) = 0$ where

$$\varphi_A(z) = -z^2 + i2\pi h^{-1}z \tag{32}$$

is the most important part of the expression (30) for $\varphi(z)$ near $z = z_0$. The equation $\varphi_A'(z) = 0$ gives the approximation $z_A = i\pi/h$ for z_0. Substituting $\varphi_A''(z_A) = -2$ and $\varphi(z_A)$ from (30) in place of $\varphi''(z_0)$ and $\varphi(z_0)$, respectively, in the saddle point contribution (31) to R_+ leads to (7).

Figure 2 shows the paths of steepest descent and the path of integration used to estimate E for the integral (25), $I_o(1)$, when $h = 0.2$. For $k = 0$ and $a = 1$, the $\varphi(z)$ in the integral (29) for R_+ is

$$\varphi(z) = -u^2 - u^{-1} + i2\pi zh^{-1} + \ell n\left[\frac{e^{2z}(e^z + 2)}{(e^z + 1)^2}\right] \tag{33}$$

where u is given by (26) with z in place of x. The deformed path of integration for R_+ shown in Fig. 2 contains the arbitrary bridging segment AB and shows that the saddle points z_1 and z_3 are the main contributors to R_+.

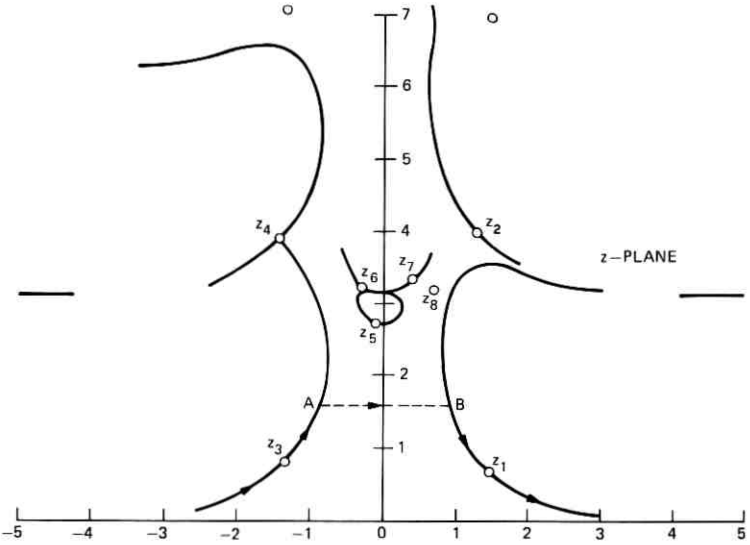


Fig. 2—Steepest descent paths for $\exp[\varphi(z)]$ when $h = 0.2$ in (33). The arrows mark a deformed path of integration for R_+ corresponding to $I_o(1)$ in (25). The points $z_1, z_2, \cdots, z_8$ are saddle points.

Finally, if one wishes to integrate along a path of steepest descent, a combination of the above path computation and the trapezoidal rule with spacing Δ suggests itself. However, greater accuracy can be achieved by either (*i*) first computing the entire path, approximating it by several straight-line segments (sometimes one will do), and using Romberg integration on each segment, or (*ii*) taking the step size Δ relatively large in the path computation, and then using Romberg integration over each linear segment of length Δ (by dividing d_t into lengths of, say, $\Delta/8$).

REFERENCES

1. Moran, P. A. P., "Approximate Relations Between Series and Integrals," MTAC, *12* (1958), pp. 34–37.
2. Schwartz, C., "Numerical Integration of Analytic Functions," J. Comp. Phys., *4* (1969), pp. 19–29.
3. Rice, S. O., "Numerical Evaluation of Integrals of Analytic Functions by the Trapezoidal Rule, "1972. Copies of this report can be obtained from the author.
4. Whittaker, E. T., and Watson, G. N., *A Course of Modern Analysis*, London: Cambridge Univ. Press, 1927.
5. Fettis, H. E., "Numerical Calculation of Certain Definite Integrals by Poisson's Summation Formula," MTAC, *9* (1955), pp. 85–92.
6. Goodwin, E. T., "The Evaluation of Integrals of the Form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$," Proc. Camb. Phil. Soc., *45* (1949), pp. 241–245.
7. McNamee, J., "Error-Bounds for the Evaluation of Integrals by the Euler-Maclaurin Formula and by Gauss-Type Formulas," Math. Comp., *18* (1964), pp. 368–381.
8. Martensen, E., "Zur numerischen Auswertung uneigentlicher Integrale," Z. Ange. Math., *48* (1968), pp. T83–T85.
9. Davis, P. J., and Rabinowitz, P., *Numerical Integration*, Waltham, Mass.: Blaisdell, 1967.
10. Kajfez, D., "Numerical Integration by Deficient Splines," Proc. IEEE, *60* (1972), pp. 1015–1016 (letter).
11. Smith W. E., and Lyness, J. N., "Applications of Hilbert Transform Theory to Numerical Quadrature," Math. Comp., *23* (1969), pp. 231–252.
12. Blakemore, J. S., *Semiconductor Statistics*, New York: Peragmon Press, 1962.
13. Bennett, W. R., "Distribution of the Sum of Randomly Phased Components," Quart. Appl. Math., *5* (1948), pp. 385–393.
14. Greenwood, J. R., and Durand, D., "The Distribution of Length and Components of the Sum of *n* Random Unit Vectors," Ann. Math. Stat., *26* (1955), pp. 233–246.
15. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions*, Nat. Bur. Stand., Appl. Math. Series No. 55, Washington: Government Printing Office, 1964.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

A New Approach to Optimum Pulse Shaping in Sampled Systems Using Time-Domain Filtering

By K. H. MUELLER

(Manuscript received December 5, 1972)

A new approach to time-domain pulse shaping in digital sampled systems is described. The proposed method allows time-limited impulse responses with optimum specified energy distribution in the frequency domain to be generated. Additional constraints to guarantee zero intersymbol interference are easily taken into account. Nyquist-type pulses which have the maximum possible amount of their total energy concentrated below some given frequency are one particularly important application. An example of such an impulse response with only 6 percent excess bandwidth is presented which shows that 99.96 percent of the energy can be concentrated in the desired bandwidth with a pulse 16 baud intervals long that can be generated using a read-only memory (ROM) with only 256 bits of storage. This new class of signals can be used advantageously for waveform generation and processing in digital data systems.

I. INTRODUCTION

The joint optimization of functions in both time and frequency domain is a classical problem in communication theory. Hilberg and Rothe[1] have recently found the lowest possible product of pulse and one-sided spectral widths and have numerically evaluated the impulse and frequency response–which is not Gaussian–that corresponds to this minimum. Landau, Pollak, and Slepian[2–4] in their classical papers have derived the pulse-form of given duration that has a maximum of its energy concentrated below a certain frequency and vice versa; the solutions to this problem are given by the now well-known prolate spheroidal wave functions. Additional comments on this problem have recently been given by Hilberg.[5] A widespread opinion is that pulses with minimum energy at high frequencies should have a rounded form with many continuous derivatives. This is not true; in fact, the optimum pulses based on the prolate spheroidal wave functions are usually not continuous at the limits of their truncation interval. Hilberg[6] has shown that constraints of continuous derivatives tend to increase substantially the total out-of-band energy.

Steep spectral roll-off above the Nyquist frequency and small residual out-of-band energy are desirable properties for signals in data transmission systems to achieve maximum signaling rate over band-limited channels and to avoid fold-over distortion in modulation and demodulation. We have here, however, a very important additional constraint: The generated signal must also have negligible intersymbol interference. One method of deriving shaping filters which simultaneously minimize intersymbol interference and stopband response was proposed by Spaulding.[7] His procedure generates better results than the traditional approach of approximation to the raised cosine roll-off in the frequency domain only.

In this paper we will again carry out optimization in the frequency domain only; but we constrain the intersymbol interference to be exactly zero and we truncate the pulse duration to a chosen number of baud intervals. The impulse response is represented in sampled form. This new class of signals will have particular application in digital modem design.

II. THE SAMPLED APPROACH

A sampled Nyquist-type impulse response with samples a_i is shown in Fig. 1. For convenience, we will assume even symmetry, an integral

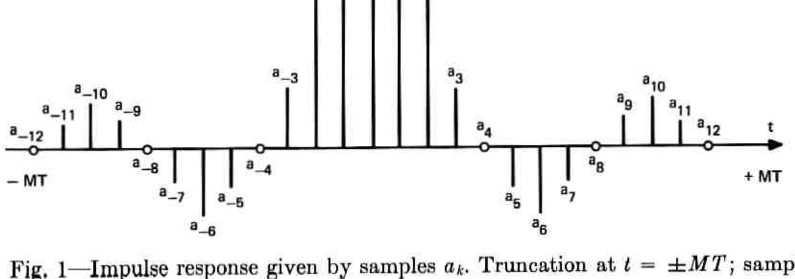


Fig. 1—Impulse response given by samples a_k. Truncation at $t = \pm MT$; sample spacing $\Delta t = T/\mu$.

number μ of samples per baud interval, and coincidence of every μth sample with the zero crossings; thus

$$\left.\begin{aligned} a_{\pm k\mu} &= 0 \qquad \text{for} \quad k = \pm 1, \pm 2, \cdots, \pm M \\ a_i &= a_{-i} \\ a_i &= 0 \qquad \text{for} \quad |i| \geqq \mu M \end{aligned}\right\}. \tag{1}$$

The resulting spectrum is

$$S(\omega) = W(\omega) \sum_{i=-\mu M}^{\mu M} a_i e^{-j\omega i T/\mu} \tag{2}$$

where $W(\omega)$ is an associated weighting function which may take into account the conversion from impulses to a staircase waveform or any other form of interpolation network. Let us define the $(2\mu M + 1)$-sample vector

$$\mathbf{a}^T = (a_{-\mu M}, \cdots, a_0, \cdots, a_{\mu M}) \tag{3}$$

and the transformation vector

$$\mathbf{p}^T = \{p_i\}, \qquad \text{with} \quad p_i = e^{-j\omega i T/\mu}, \tag{4}$$

so that we can write the spectrum in the simple form

$$S(\omega) = W(\omega)\mathbf{a}^T\mathbf{p}. \tag{5}$$

The power density spectrum is given by*

$$|S(\omega)|^2 = |W(\omega)|^2\mathbf{a}^T\mathbf{p}\mathbf{p}^\dagger\mathbf{a}. \tag{6}$$

If we assume that the function $w(t)$ has energy E_w and is nonoverlapping (width $\leqq T/\mu$), the total energy $E(\infty)$ is simply

$$E(\infty) = \mathbf{a}^T\mathbf{a}E_w. \tag{7}$$

The energy below ω_0 is of course

$$E(\omega_0) = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} |W(\omega)|^2\mathbf{a}^T\mathbf{p}\mathbf{p}^\dagger\mathbf{a}\, d\omega. \tag{8}$$

Our goal is again to find **a**, so that

$$\lambda = \frac{E(\omega_0)}{E(\infty)} = \max \tag{9}$$

or, by combining the last three equations,

$$\lambda\mathbf{a}^T\mathbf{a} = \mathbf{a}^T R\mathbf{a} \tag{10}$$

where the elements of the symmetric matrix R are defined by

$$r_{ik} = \frac{1}{\pi E_w}\int_0^{\omega_0} |W(\omega)|^2 \cos\left(\omega T\,\frac{i-k}{\mu}\right) d\omega \tag{11}$$

and **a** has to satisfy the constraints (1). This constraint reduces the

* A dagger is used to indicate the conjugate transpose, $\mathbf{a}^\dagger = \mathbf{a}^{*T}$.

degree of the quadratic form (10), since $a_{\mu k} = \delta_{0k}$ and thus all terms $r_{\mu i,\mu k}$ are immaterial. After elimination of the zero elements and concentration of the remaining elements, the new form

$$\lambda \hat{\mathbf{a}}^T \hat{\mathbf{a}} = \hat{\mathbf{a}}^T \hat{R} \hat{\mathbf{a}} \tag{12}$$

evolves, which is now free of constraints. The desired solution is then simply given by the eigenvector of $\hat{R}$ which corresponds to the largest eigenvalue $\lambda_{\max}$ (since no other choice of $\hat{a}$ will give a larger λ). The original vector **a** can easily be obtained by inserting zeros in the correct positions of $\hat{\mathbf{a}}$.

The elements r_{ik} in (11) will of course depend on the choice of ω_0 and numerical integration generally will be necessary to evaluate them. Because of the Toeplitz and symmetric nature of R, only a small number of terms need really be calculated. Numerical integration and determination of eigenvectors are available as subroutines with most computers, so that no complex programs need be written for the proposed optimization method. We also would like to emphasize that the described method is very flexible. One might, for example, try to minimize the energy contribution within some given frequency range $\omega_1 < \omega < \omega_2$; this may easily be achieved by changing the integration limits in (11).

Two cases of $W(\omega)$ are of practical interest. The first one is the zero-order-hold function which generates a staircase waveform (this is the usual output of D/A converters). In this case, we have*

$$W(\omega) = \frac{T}{\mu} \operatorname{sinc}\left(\frac{\omega T}{2\mu}\right) \tag{13}$$

and therefore

$$r_{ik} = \frac{1}{\mu\pi} \int_0^{\pi(1+\beta)} \operatorname{sinc}^2\left(\frac{x}{2\mu}\right) \cos\left[\frac{x}{\mu}(i-k)\right] dx \tag{14}$$

where we have expressed ω_o in terms of the normalized Nyquist excess bandwidth β.

In the second case we will assume $w(t) = \delta(t)$, so that the spectrum $W(\omega)$ is flat. Due to the periodicity of the resulting spectrum, it is reasonable to consider the energy distribution within one period only. The resulting elements of the matrix R can then be expressed in closed form

$$\left.\begin{aligned} r_{ik} &= \frac{1+\beta}{\mu} \operatorname{sinc} \pi(1+\beta)\left[\frac{i-k}{\mu}\right] \qquad &\text{if} \quad i \neq k \\ r_{ii} &= \frac{1+\beta}{\mu} &\text{if} \quad i = k \end{aligned}\right\}, \tag{15}$$

which further simplifies the optimization procedure.

* We define $\operatorname{sinc}(x) = \sin(x)/x$ for convenience.

III. GENERALIZATION FOR ARBITRARY SPECTRAL BANDS

Equation (10) is a special case of the more general problem of maximizing the energy in one or more specified frequency bands with respect to the energy in some other frequency bands. Taking into account the desired integrating limits and the constraints (1), a quadratic form

$$\lambda \hat{\mathbf{a}}^T Q \hat{\mathbf{a}} = \hat{\mathbf{a}}^T \hat{R} \hat{\mathbf{a}} \tag{16}$$

will then evolve, containing (12) as a special case with $Q = I$. By substituting

$$\mathbf{b} = \sqrt{Q}\hat{\mathbf{a}}, \tag{17}$$

we have now to deal with the new form

$$\lambda \mathbf{b}^T \mathbf{b} = \mathbf{b}^T \sqrt{Q}^{-1} \hat{R} \sqrt{Q}^{-1} \mathbf{b}, \tag{18}$$

which is identical to (12). We are looking for the particular **b** satisfying

$$\sqrt{Q}^{-1} \hat{R} Q^{-1} \mathbf{b} = \lambda \mathbf{b}. \tag{19}$$

By premultiplying both sides with $\sqrt{Q}^{-1}$, we get

$$Q^{-1} \hat{R} \hat{\mathbf{a}} = \lambda \hat{\mathbf{a}}, \tag{20}$$

so that the desired $\hat{\mathbf{a}}$ is simply the eigenvector of $Q^{-1}R$ which corresponds to the largest eigenvalue $\lambda_{\max}$.* The matrix Q is guaranteed to be nonsingular since it is not possible to have zero energy in a finite frequency interval with a time-truncated impulse response.

IV. EXAMPLE

To get some feeling for the capabilities of the described optimization procedure, the samples of a Nyquist-type impulse response were calculated using the following parameters:

Excess bandwidth factor $\beta = 0.06$
Truncation for $|t| > 8T$ $(M = 8)$
$\mu = 4$ samples per baud interval.

The unusually tight roll-off would allow full 4800-baud operation over voice-grade telephone channels with a QAM or a VSB system. If the sample values are coded into 8 bits plus sign, the chosen resolution will bring down the quantization noise to a negligible level of -65 dB. The storage requirement is still only 256 bits, so a rather small bipolar ROM may be used.

* Note that $Q^{-1}\hat{R}$ need not be symmetric, but its eigenvalues are the same as those of the symmetric matrix in (19).

The resulting spectrum is shown in Fig. 2. Attenuation is 6 dB at the Nyquist frequency and 17.6 dB at the 6-percent edge. The sinc(·) weighting caused by the staircase output is not included; it would produce additional attenuation at higher frequencies. The resulting eigenvalue was $\lambda_{\max} = 0.99963$, showing that in fact the residual out-of-band energy is very small.

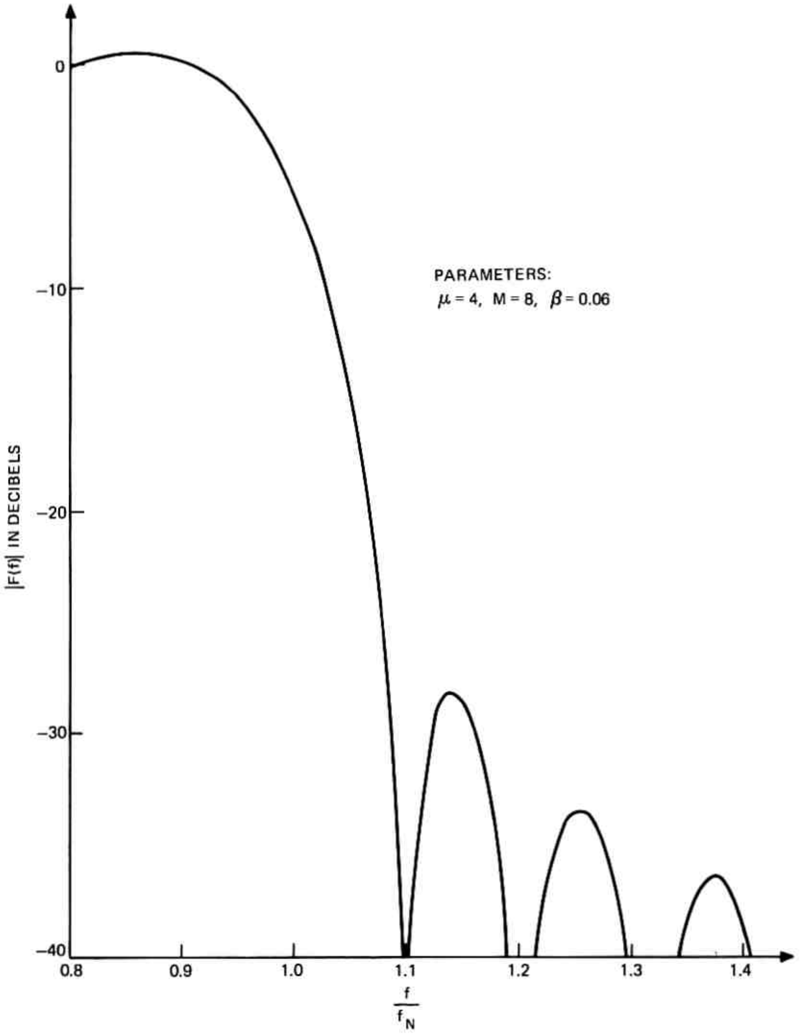


Fig. 2—Spectrum of optimized impulse response with $\lambda = 0.99963$.

V. CONCLUSIONS

A new optimization method for sampled Nyquist-type impulse responses has been proposed. Minimum energy in one frequency band as compared to the energy in any other frequency band is achieved. The computation is straightforward and involves the determination of eigenvectors of a symmetric matrix. It is shown how the constraint for zero intersymbol interference can easily be included. Applications of this method are numerous in digital signal synthesis and processing. Storage can be achieved with high accuracy using ROM's of moderate size. Any desired scaling of time and frequency response is possible with such a system and the well-known disadvantages of traditional filters, namely aging and tuning, are nonexistent.

VI. ACKNOWLEDGMENTS

The author wishes to thank Mrs. B. A. Zimmer for writing the computer program used in Section IV and J. E. Mazo for many stimulating and helpful discussions.

REFERENCES

1. Hilberg, W., and Rothe, P. G., "The General Uncertainty Relation for Real Signals in Communication Theory," Information and Control, *18*, 1971, pp. 103–125.
2. Slepian, D., and Pollak, H. O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—I," B.S.T.J., *40*, No. 1 (January 1961), pp. 43–63.
3. Landau, H. J., and Pollak, H. O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—II," B.S.T.J., *40*, No. 1 (January 1961), pp. 65–84.
4. Landau, H. J., and Pollak, H. O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III: The Dimension of the Space of Essentially Time- and Band-Limited Signals," B.S.T.J., *41*, No. 4 (July 1962), pp. 1295–1336.
5. Hilberg, W., "Zeitlich begernzte Impulse mit einem Maximum an Energie in einem vorgegebenen Frequenzband," Nachrichtentechnische Zeitschrift, Heft 3, 1970, pp. 129–133.
6. Hilberg, W., "Impulse endlicher Dauer mit abgerundeter Impulsform," Nachrichtentechnische Zeitschrift, Heft 6, 1970, pp. 295–301.
7. Spaulding, D. A., "Synthesis of Pulse-Shaping Networks in the Time Domain," B.S.T.J., *48*, No. 7 (September 1969), pp. 2425–2444.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 5, May–June, 1973
Printed in U.S.A.

Utilization of Optical-Frequency Carriers for Low- and Moderate-Bandwidth Channels

By W. M. HUBBARD

(Manuscript received November 10, 1972)

Recent advances in solid-state optical-frequency sources and detectors and low-loss optical fibers make feasible the consideration of optical communication systems for low- and moderate-bandwidth channels (a few kHz to, say, 100 MHz). This paper explores the use of optical-frequency carrier systems for transmission over such channels. Analog intensity modulation, pulse position modulation, delta modulation, and pulse code modulation are considered. This paper is intended to be tutorial in nature.

I. INTRODUCTION

Since the advent of the laser, communication engineers have been intrigued by the promise of fantastic bandwidth capability in optical-frequency communication systems. As a result, attention has been focused on high-capacity, high-bandwidth considerations. Recent advances in component fabrication—for example, light-emitting diodes (LED's), junction lasers, avalanche photodiode detectors, and low-loss optical fibers—have made it feasible to consider the use of optical-frequency carrier systems for moderate- and even low-bandwidth channels.

Fundamental and practical differences between optical-frequency channels and radio-frequency channels* necessitate a reevaluation of concepts acquired from experience with the latter. To this end, in the following sections we consider four potentially attractive forms of modulation of optical-frequency signals and derive results for the required average received signal power in terms of system requirements and parameters. In Section II we consider a system using analog inten-

* We use the term "optical frequency" to mean frequencies roughly in the range 10 to 1000 THz and the term "radio frequency" to indicate frequencies below roughly 3 THz.

sity modulation (IM) of the light source. In some respects such a system is analogous to a baseband system. In Section III we consider pulse position modulation which, because of the nature of optical-frequency sources and detectors, is particularly attractive. In Section IV we consider binary PCM and also delta modulation which we treat as a special case of a binary PCM channel. It is not within the scope of this paper to identify and analyze optimal receivers for these types of modulation. The approach adopted here is rather to analyze the performance of receivers which can be realized and which hold hope of providing an economically attractive approach to transmission of low- and moderate-bandwidth channels.

The four types of systems to be discussed in the following sections are considered in terms of their applicability to the problem of transmitting a comparatively narrow information bandwidth b over a channel with noise (signal) bandwidth $\mathcal{B}$. The features which these systems have in common will be discussed in this section.

Two types of signal sources are considered in this discussion–lasers and light-emitting diodes. For the purposes of this discussion, the difference in these two sources is that the laser possesses a substantial degree of temporal and spatial coherence, while the LED does not. The effect of this incoherence is to give rise to an additional type of noise (later referred to as beat noise) in an LED system. We shall see in the following calculations, however, that this noise is usually negligible in systems of interest. This is because the beat noise is proportional to the ratio $\mathcal{B}/(WJ)$ where W is the spectral width of the LED and J is the number of spatial modes of the signal viewed by the receiver. For typical GaAs LED's, $W \approx 20 \times 10^{12}$ Hz and J is an integer which depends on the details of the channel between transmitter and receiver and which is usually very large.

The receiver in all cases is assumed to begin with an avalanche photodiode with quantum efficiency η and avalanche current gain G. This photodiode is followed by a baseband amplifier which presents a load resistance R to the photodiode and which has a noise figure F_t. Only direct detection receivers are considered in this treatment, since recently developed avalanche photodetectors make heterodyne and homodyne methods look quite unattractive in view of the difficulties encountered in phase-front matching in such systems.*

II. ANALOG INTENSITY MODULATION

The simplest form of modulation is analog intensity modulation. Both light-emitting diodes and double-heterostructure junction lasers

* See appendix for elaboration on this point.

have output power versus bias current characteristics which are sufficiently linear over a reasonable range that they can be modulated directly by modulating their bias currents. Modulation depths of up to about 85 percent can be achieved with suitable light-emitting diodes with very small harmonic distortion. It should be noted that the optical power, not amplitude, is proportional to the drive signal but, since the photodetector is a square-law device, its output current is proportional to the received power. Thus, in many respects, an intensity-modulated optical system can be regarded as equivalent to a baseband system with a transducer (the light-emitting diode or laser) which converts electrons into photons and a subsequent transducer (the photodetector) which converts photons back into electrons.

For a sinusoidally modulated carrier with modulation index m, the mean-square signal current in the photodetector output is given by

$$\langle i_s^2\rangle = \frac{1}{2}\left[\eta \frac{e}{h\nu} Gmp_o\right]^2.$$

When a coherent source is used, the mean-square noise current in the photodetector output is the resultant of the five noise currents described below.

The most important (in most applications) of the noise currents is the quantum noise with its mean-square given by

$$\langle i_Q^2\rangle = 2e\frac{e}{h\nu}\eta p_o G^2 F_d b = N_Q G^2 F_d$$

where in this and the following equations e is the electronic charge, $h\nu$ is the energy per photon, η is the quantum efficiency of the photodiode, p_o is the average received optical power, b is the bandwidth of the information source (which in the case of analog intensity modulation is equal to the bandwidth of the channel), G is the avalanche gain of the photodetector, and F_d is a noise figure associated with the random nature of the avalanche process. F_d is, in general, a function of G which for silicon is well approximated[1,2] by $F_d = \sqrt{G}$. N_Q is the value of the quantum noise in the absence of avalanche gain.

The next most important noise source is the thermal-noise current with mean-square value

$$\langle i_T^2\rangle = \frac{4\mathrm{k}T}{R_{\mathrm{eq}}} bF_t = N_T$$

where kT is Boltzmann's constant times the absolute temperature, R_{eq} is an equivalent load resistance, and F_t is the noise figure of the (baseband) amplifier.

The dark-current noise can be rendered negligible by suitable choice of photodetector. At the present time, this generally dictates that the photodetector be made of silicon. There are actually two kinds of dark-current noise. The first, which will be referred to simply as dark current in the following, consists of electrons (and/or holes) which are thermally liberated in the pn junction and which experience the avalanche gain G. The mean-square value of this current is given by:

$$\langle i_D^2 \rangle = 2eI_d' G^2 F_d b = G^2 F_d N_D$$

when I_d' is the primary detector dark current. The other "dark current," which will henceforth be referred to as leakage current, bypasses the drift region and experiences no avalanche gain. The mean-square value of this current is therefore given by:

$$\langle i_L^2 \rangle = 2eI_L b = N_L$$

where I_L is the leakage current.

Finally, if there is incoherent background radiation with average power p_G incident on the detector, there will be an additional noise current given by:

$$\langle i_G^2 \rangle = 2e \frac{e}{h\nu} \eta p_G G^2 F_d b = G^2 F_d N_G.$$

(This assumes that the background radiation is at about the same wavelength as the signal. This is justified since other wavelengths could be effectively removed by filters.)

Since $\langle i_G^2 \rangle$ and $\langle i_D^2 \rangle$ have the same form, we can simply write

$$I_d = I_d' + \frac{e}{h\nu} \eta p_G$$

and lump both of these terms into an effective dark current. This is done in the following calculations.

When an incoherent source such as a light-emitting diode is used, there is an additional noise term due to the beats between spectral components within the spectral width of the source. This phenomenon gives rise to a noise current with variance

$$\langle i_B^2 \rangle = 2\left(\frac{e}{h\nu} G\eta p_o\right)^2 \frac{b}{JW}\left(1 - \frac{1}{2}\frac{b}{W}\right) = G^2 N_B$$

where W is the spectral width of the source and J is the number of spatial modes of the source which are viewed by the receiver.[3] (In most cases the ratio b/JW renders this term negligible.) The factor $(1 - b/(2W))$ is always very nearly unity in cases of interest.

In each case, $\langle i_X^2 \rangle$ represents the mean-square value of the corresponding noise current after avalanche gain and N_X represents the value it would have in the absence of avalanche gain.

Thus, the signal-to-noise ratio is given by

$$\mathrm{SNR} = \frac{\frac{1}{2}\left(\eta \frac{e}{h\nu} Gmp_o\right)^2}{\langle i_Q^2 \rangle + \langle i_T^2 \rangle + \langle i_D^2 \rangle + \langle i_L^2 \rangle + \langle i_B^2 \rangle}. \tag{1}$$

It is instructive to consider the behavior of (1) for some particular cases. For our examples we use the following numerical values throughout this paper:

$$\lambda = 0.85\ \mu\text{m} \qquad R = 10^3 \text{ ohms}$$
$$\eta = 0.5 \qquad b = 4 \text{ kHz}$$
$$m = 0.85.$$

Figure 1 shows the signal-to-noise ratio (expressed in dB) computed from (1) for $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, and $G = 10$. It must be emphasized that these values are chosen for illustrative purposes only and are not meant to be typical of a real receiver. This choice of parameters allows us to consider the form of the contribution of each noise source to the resulting SNR. The curves labeled Q, T, D, L, and B are the ratio of the mean signal power to the mean quantum-noise power, mean thermal-noise power, mean dark-current-noise power, mean leakage-current-noise power, and mean beat-noise power, respectively. We observe that for this comparatively low value of G the thermal noise is dominant over a large range of incident light power p_o. Then for -30 dBm $< p_o < -15$ dBm the quantum noise dominates the picture. Finally, for larger p_o, the beat noise clamps a ceiling on the SNR. The dark current and leakage current are unimportant.

Figure 2 shows the same curves for the same illustrative parameters as Fig. 1 except that G is now taken to be 100. Two differences between Fig. 1 and Fig. 2 are immediately apparent. First, the increased gain has caused the quantum-limited region to extend to lower values of p_o and, second, the dark-current noise is more important relative to the thermal noise.

In Figs. 1 and 2, F_d has been taken to be given by $G^{\frac{1}{2}}$. Thus, in Fig. 2, for example, $F_d = 10$ and one must be cautious about referring to the behavior as "quantum limited" in the "quantum-limited region." It is "quantum limited" only in the sense that the quantum noise term dominates the other noise terms, but the actual results are 10 dB poorer than could be achieved if the avalanche gain process were noise free. Figure 3 shows the same curves but for the values $I_d = 10^{-10}$ A,

$I_L = 10^{-9}$ A, which are typical of good, but available, silicon photodiodes;[4] $WJ = 10^{17}$ Hz which is typical of GaAs luminescent diodes and multimode optical fibers; and $G = 17$. We see that in this case beat

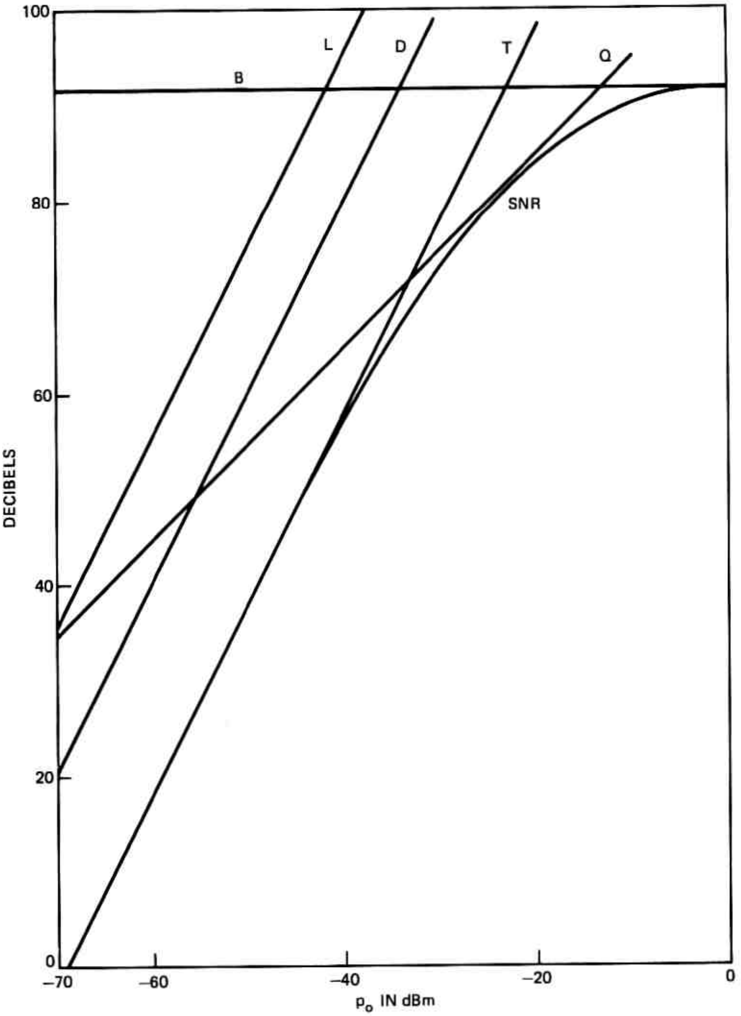


Fig. 1—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the illustrative parameters: $\lambda = 0.85\ \mu$m, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, $b = 4000$ Hz, $G = 10$, for an IM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power.

noise, leakage noise, and dark-current noise are unimportant and the quantum-excess noise controls above $p_o = -30$ dBm with thermal noise controlling below.

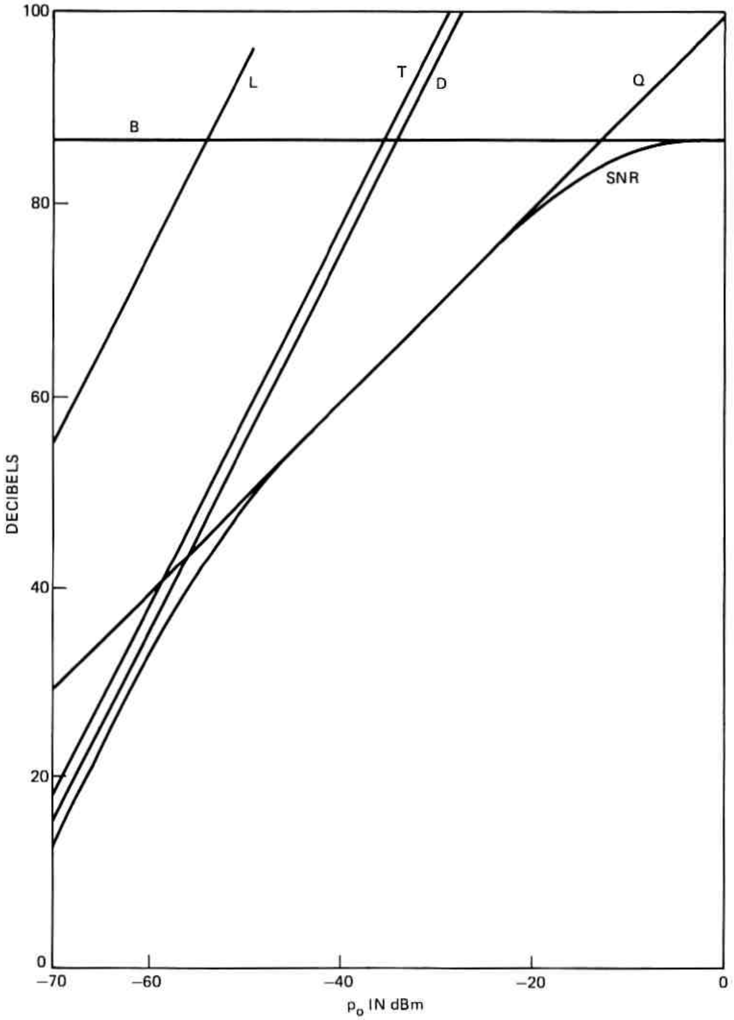


Fig. 2—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the illustrative parameters: $\lambda = 0.85\ \mu\text{m}$, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, $b = 4000$ Hz, $G = 100$, for an IM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power.

It is now well known[1,2,5] that the excess noise figure F_d of an avalanche photodiode increases (in most cases) with increasing gain. In particular, for silicon photodiodes, F_d is well approximated by $G^{\frac{1}{2}}$.

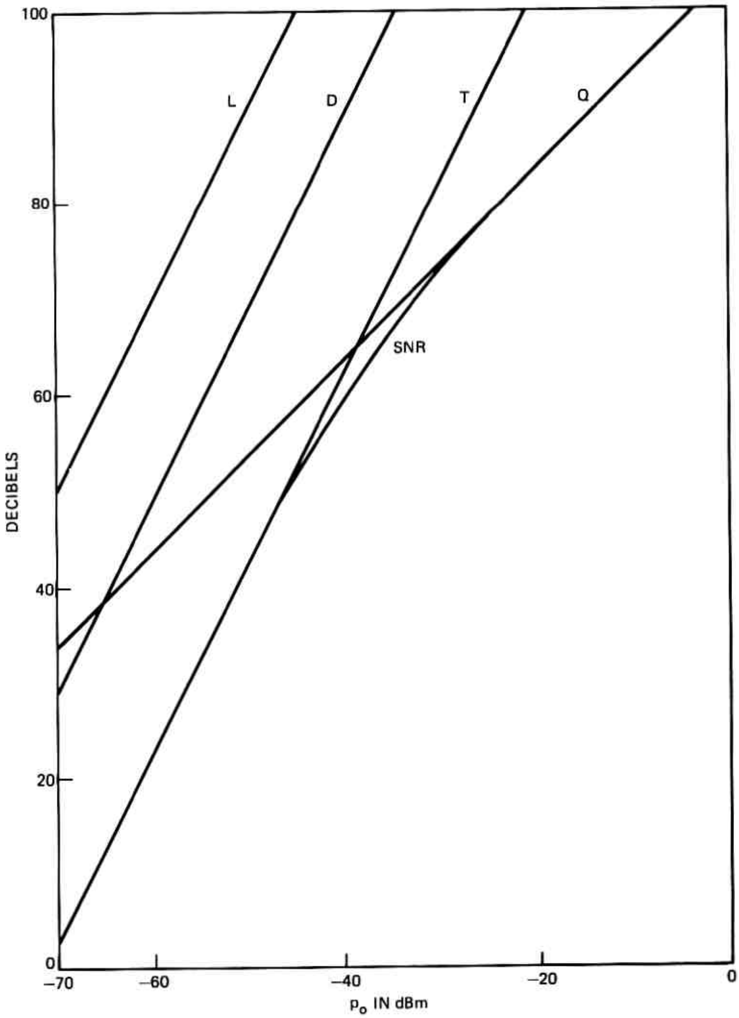


Fig. 3—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the typical parameters: $\lambda = 0.85\ \mu\text{m}$, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-10}$ A, $I_L = 10^{-9}$ A, $WJ = 10^{17}$ Hz, $b = 4000$ Hz, $G = 17$, for an IM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power.

When a form $F_d = G^r$ is assumed, SNR as a function of G has a maximum given by

$$\text{SNR}_{\text{max}} = \frac{\langle i_S^2 \rangle G^{-2}}{\left(\dfrac{2\gamma}{r}\right)^{r/(2+r)} \beta^{2/(2+r)} \left(1 + \dfrac{r}{2}\right) + N_B}$$

for

$$G = G_{\text{opt}} = \left(\frac{2\gamma}{r\beta}\right)^{1/(2+r)}$$

where

$$\gamma = \langle i_T^2 \rangle + \langle i_L^2 \rangle = N_T + N_L,$$

the variance of the gain-independent noise, and

$$\beta = G^{-(1+r)}[\langle i_Q^2 \rangle + \langle i_D^2 \rangle] = N_Q + N_D,$$

the variance of the gain-dependent noise *before* the gain process. For $r = \frac{1}{2}$ (silicon) this reduces to

$$G_{\text{opt}} = \left[\frac{4(N_T + N_L)}{N_Q + N_D}\right]^{2/5} \tag{2}$$

$$\text{SNR}_{\text{max}} = \frac{\dfrac{2}{5}\left[\eta \dfrac{e}{h\nu} m p_o\right]^2}{(4\gamma)^{1/5}\beta^{4/5} + N_B}. \tag{3}$$

It is interesting to note in passing that the condition for G to be optimum is that

$$\beta G_{\text{opt}}^{2+r} = \frac{2}{r}\gamma.$$

But the left-hand side of this equation is the total mean-square current due to gain-dependent noise (excluding beat noise), while the right-hand side is $2/r$ times the total mean-square current due to gain-independent noise. For $r = 0.5$, for example, the gain is optimum when the gain-dependent noise exceeds the gain-independent noise by 6 dB.

This result is illustrated in Fig. 4 for the same parameters used in Figs. 1 and 2 except that in Fig. 4 $G = G_{\text{opt}}$, which is a function of p_o according to (2). First we observe that G_{opt} varies from 132 to 1 over the range of p_o plotted in these figures. These are values which are readily achievable with existing photodiodes. We see that the thermal noise (leakage noise remains negligible) is just a constant 6 dB below the sum of the quantum- and dark-current noises as the condition for optimum gain dictates.

Figure 5 illustrates the optimum gain behavior for the typical device parameters used in Fig. 3. Note that here the optimum gain becomes rather large below about $p_o = -60$ dB and might be difficult or impossible to realize in practice.

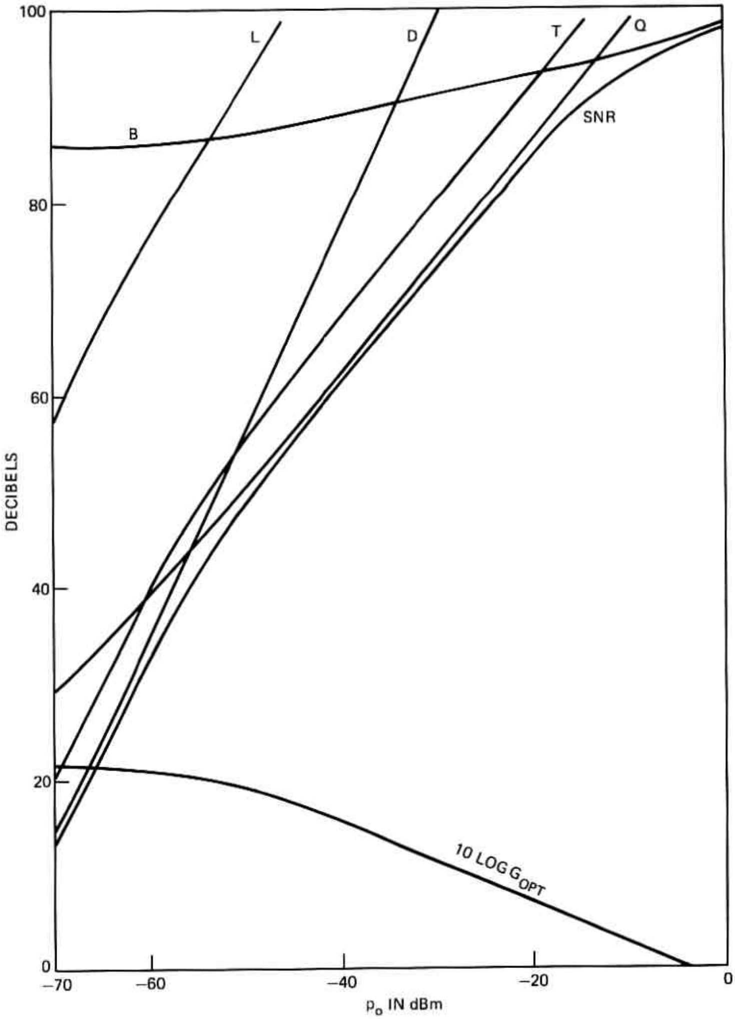


Fig. 4.—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the illustrative parameters: $\lambda = 0.85\ \mu$m, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, $b = 4000$ Hz, $G = G_{\text{opt}}$, for an IM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power.

The value of $R = 10^3$ ohms is used consistently in the numerical examples throughout this paper. The detector c is usually presumed to dictate the maximum value of load resistance R through the relationship

$$R < \frac{1}{4c\mathcal{B}}$$

where $\mathcal{B}$ is the bandwidth of the signal; but in practice it is often beneficial to use a much larger value of R than this and equalize the resultant signal distortion later on in the receiver. It is valid to object to the use of so low a resistance for a 4-kHz channel [the value is more appropriate to the other types of systems (without equalization) to be considered in the following sections]. However, (3) shows that at optimum avalanche gain when, as is usually the case, the leakage current is negligible, SNR $\propto R^{1/5}$, so very little is gained by going to larger load resistors except in the region where G_{opt} is so large that it is difficult to achieve. Here the fact that $G_{\text{opt}} \propto R^{-2/5}$ may be important.

It is useful to present the results of (3) graphically in a form suitable for system design. This can be done in a very general and simple manner, when the leakage current is negligible, by defining the quantities

$$\Psi = 10 \log \left\{ \frac{\text{SNR}_{\max} b}{m^2} \right\}$$

$$x = 10 \log \left\{ \frac{\eta p_o}{h\nu} \right\}$$

where, in the argument of the first logarithm, b is taken to be dimensionless, i.e., it is interpreted as the ratio of bandwidth in Hz to 1 Hz. Ψ vs x is plotted in Fig. 6 for $R/F_t = 10^3$ ohms. Figure 6 can be used as a computational aid as follows. Suppose one needs to design a system with an SNR of 70 dB, a bandwidth of 4 kHz. Suppose further that a modulation index of 0.85 is possible with available devices. Then the value of Ψ which characterizes such a system is 107.4 which, Fig. 6 tells us, can be achieved if $x = 121$. Now $x = 121$ means that $\eta p_o/h\nu = 1.26 \times 10^{12}$; for $\eta = 0.5$ and $\lambda = 0.85$ this gives $p_o = 5.88 \times 10^{-7}$ W which corresponds to -32.5 dBm. If a value of R/F_t which differs from 1000 ohms is desired, Ψ can be modified according to $\Delta\Psi = 2 \log [(R/F_t)/1000]$.

Figure 7 can be used to determine the value of G required to achieve the result computed from Fig. 6. If a value of R/F_t which differs from 1000 ohms was used, recall that $G_{\text{opt}} \propto R^{-2/5}$.

III. PULSE POSITION MODULATION

Considerable improvement in noise immunity can be achieved by properly exploiting the wide available bandwidth of optical systems.

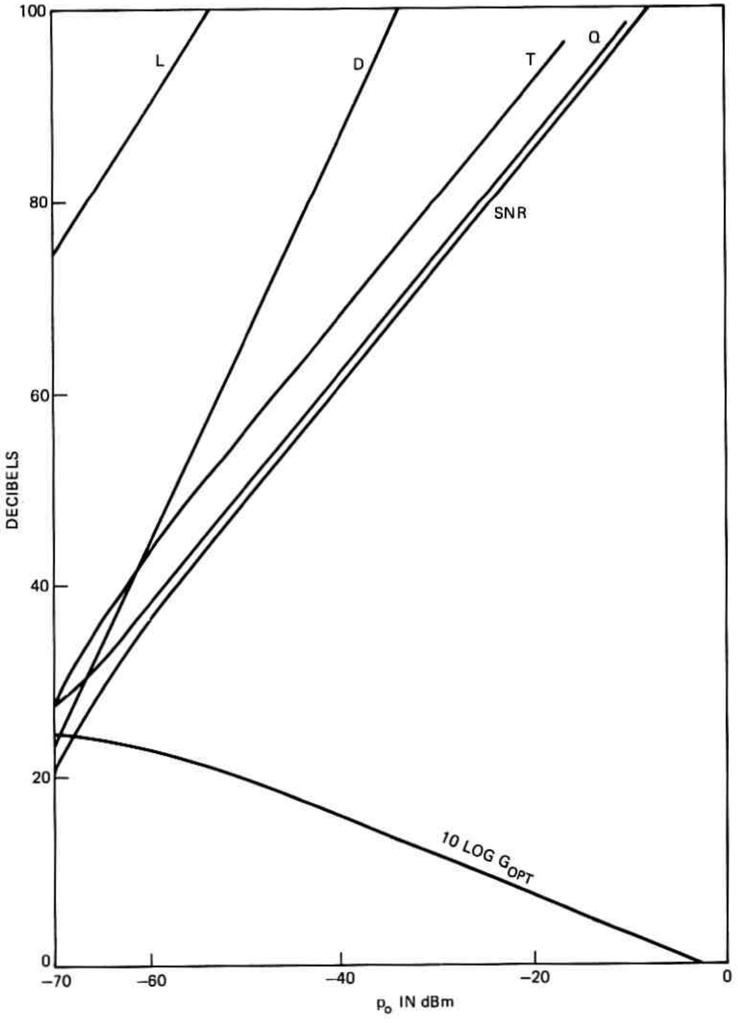


Fig. 5—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the typical parameters: $\lambda = 0.85$ μm, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-10}$ A, $I_L = 10^{-9}$ A, $WJ = 10^{17}$ Hz, $b = 4000$ Hz, $G = G_{\text{opt}}$, for an IM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power.

Pulse position modulation (PPM) offers an attractive method of accomplishing this end. Figure 8 shows a block diagram of the system to be analyzed in this section.

The pulse-position modulation signal is encoded by sampling the message signal periodically at times nT (where n is an integer and T is the sampling interval or time slot duration). The value v_n of the nth sample is transmitted during the nth time slot by sending a short pulse

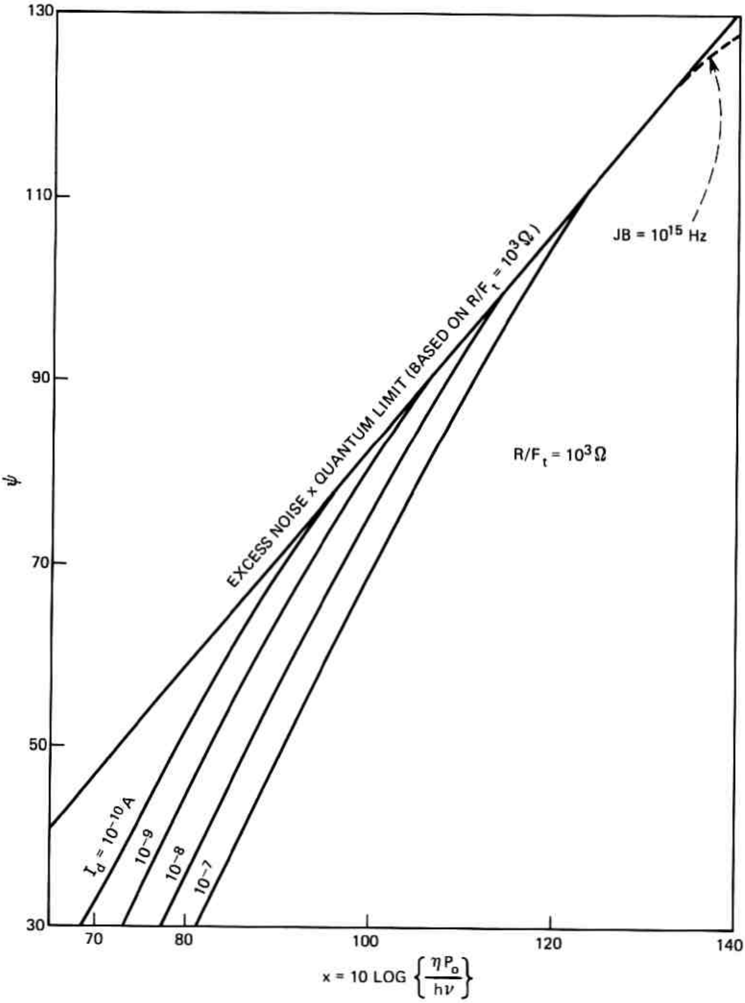


Fig. 6—$\psi = 10 \log \left\{\frac{\text{SNR}_{\max} b}{M^2}\right\}$ vs $x = 10 \log \left\{\frac{\eta P_o}{h\nu}\right\}$ for $R/F_T = 1000$ ohms.

of optical energy at a time which is shifted from the center of the nth time slot by an amount proportional to v_n.

At the receiver, the values of v_n are recovered by measuring the time interval between the center of the time slot and the time at which the amplified output current from the photodetector crosses a threshold. This system is described in some detail in Ref. 6.

Consider a PPM signal consisting of a sequence of light pulses whose

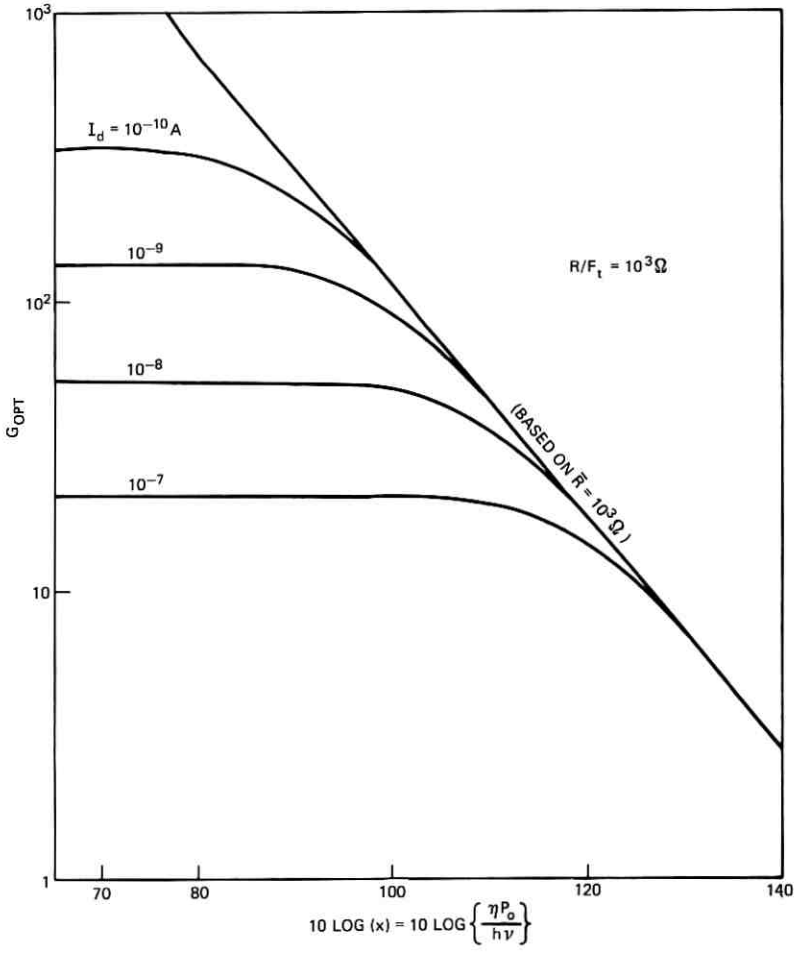


Fig. 7—Optimum gain vs $x = 10 \log \left\{ \frac{\eta P_o}{h\nu} \right\}$ for $R/F_T = 1000$ ohms.

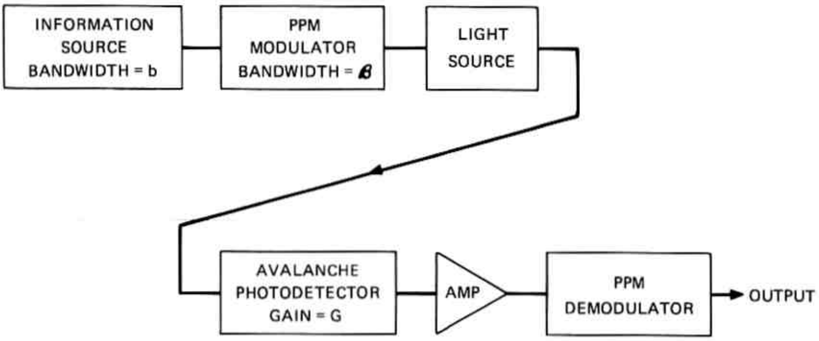


Fig. 8—Block diagram of a PPM channel.

power $P(t)$ is of the form

$$P(t) = \frac{1}{2}\left(1 + \cos\left(\pi \frac{t}{\mathcal{T}}\right)\right) P_m \qquad -\mathcal{T} < t < \mathcal{T} \tag{4}$$

where P_m is the peak power achieved by the pulse and $2\mathcal{T}$ is the total pulse duration. The average power P_o in this signal is related to P_m by

$$P_o = \frac{\mathcal{T}}{T} P_m = \frac{1}{2\kappa} P_m \tag{5}$$

where $\kappa \equiv T/2\mathcal{T}$. The detected current pulse in the receiver (neglecting noise) is given by

$$i(t) = \frac{1}{2}\left(1 + \cos\left(\pi \frac{t}{\mathcal{T}}\right)\right) i_m$$

where $i_m = \eta(e/h\nu)Gp_m = \eta(e/h\nu)G(T/\mathcal{T})p_o$ is the peak current, $p_m = AP_m$, $p_o = AP_o$, A = attenuation between transmitter and receiver.

Noise affects the SNR of a PPM signal in two ways. First, it can perturb the time of the threshold-crossing of the received signal and thereby effectively shift the position of the pulse. This is the predominant effect when the bandwidth expansion is small. Second, the noise can cause the received current to exceed the threshold in the absence of the signal pulse, thereby triggering a "false alarm" in the circuit.

First, we consider the perturbation of the time of threshold-crossing due to the noise. Assume that the threshold current level is one-half of

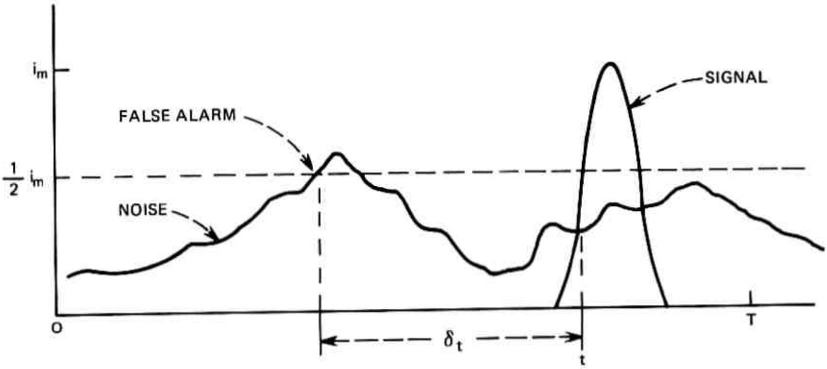


Fig. 9—Illustration of a "false alarm" or threshold violation.

the peak current level,* i.e., $\frac{1}{2}i_m$. The perturbation τ in the position of the pulse due to a noise current i_n is†

$$\tau = i_n \frac{1}{\dfrac{di(t)}{dt}}\Bigg|_{t=\mathcal{T}/2} = \frac{2}{\pi}\frac{i_n}{i_m}\mathcal{T}. \tag{6}$$

The expectation value of τ^2 is then

$$\langle\tau^2\rangle = \frac{4}{\pi^2}\frac{\langle i_n^2\rangle}{i_m^2}\mathcal{T}^2 = \frac{1}{\pi^2}\frac{\langle i_n^2\rangle\mathcal{T}^2}{\left(\eta\dfrac{e}{h\nu}G\kappa p_o\right)^2} \tag{7}$$

where $\langle i_n^2\rangle$ is the expected value of the square of the noise current.

Now consider the contribution to the noise due to "false alarms" generated when the noise current exceeds the threshold value $\frac{1}{2}i_m$. Consider a time slot $(0, T)$ as illustrated in Fig. 9. Let $t \in (0, T)$ be the time at which the signal pulse crosses the threshold. Assume that a false alarm occurs in this time slot. It must occur (with uniform probability density) on the interval $(0, t)$, since the receiver, having sensed a pulse at t, is disabled on the interval (t, T). The mean-square value of the error in t is therefore

$$\langle\delta t^2\rangle = \int_0^t p(\delta t)(\delta t)^2 d(\delta t) = \frac{t^2}{3}.$$

[*] In practice, a slightly different threshold may be optimum due to the details of the pulse shape.

[†] The remainder of this paragraph follows the derivation on pages 256–257 of Ref. 6.

Since we have no statistical information on t, we make an approximation which is clearly very conservative: We assume that t always has its largest possible value, $T - 2\mathcal{T}$. This gives

$$\langle \delta t^2 \rangle = \frac{(T - 2\mathcal{T})^2}{3} = \tfrac{4}{3}\mathcal{T}^2(\kappa - 1)^2. \tag{8}$$

Define Π as the probability of a fase alarm occurring during a time interval $T - 2\mathcal{T}$. The mean-square value of the error in t due to false alarms is then $\Pi \langle \delta t^2 \rangle$.

Let $\pm\theta$ be the limits of the allowable variation of the pulse position. The baseband signal-to-noise ratio *at full load*,* SNR, is determined as follows. The output signal power due to a sinusoidal input signal which swings the pulse position by an amount $\pm\theta$ about its mean is proportional to $\frac{1}{2}\theta^2$, the output noise power due to perturbation of the threshold crossings of the signal is proportional to $\langle \tau^2 \rangle$, and the output noise power due to false alarms is proportional to $\Pi\langle \delta t^2 \rangle$. Thus we can write

$$\text{SNR} = \frac{\frac{1}{2}\theta^2}{\langle \tau^2 \rangle + \Pi\langle \delta t^2 \rangle}. \tag{9}$$

Since T is the duration of time slot, θ is constrained by the requirement

$$2(\theta + \mathcal{T}) \leqq T.$$

Choosing equality in the above expression gives the best possible SNR; substitution of this along with (7) and (8) into (9) gives

$$\text{SNR} = \frac{\dfrac{\pi^2}{2}(\kappa - 1)^2\left(\eta \dfrac{e}{h\nu}\kappa p_o\right)^2 G^2}{\langle i_n^2 \rangle + \dfrac{4\pi^2}{3}(\kappa - 1)\Pi\left(\eta \dfrac{e}{h\nu}\kappa p_o\right)^2 G^2}. \tag{10}$$

The next steps are to evaluate $\langle i_n^2 \rangle$ and Π. We begin by evaluating $\langle i_n^2 \rangle$. The noise currents which make up $\langle i_n^2 \rangle$ for the PPM systems are the same as those which made up $\langle i_n^2 \rangle$ for the analog system except that the noise bandwidth $\mathcal{B}$ is not equal to the signal bandwidth b; and the signal-power-dependent noises are evaluated not at p_o but rather at $p_m/2$ since this is the expected value of the signal when the threshold crossing is to occur.

In the remainder of this section, the reciprocal pulse width, $1/\mathcal{T}$, and the noise bandwidth, $\mathcal{B}$, are assumed to be equal. From (5) one sees that the threshold level $\frac{1}{2}p_m$ is related to the average power by

$$\tfrac{1}{2}p_m = \kappa p_o.$$

* SNR is the ratio of mean signal power when the signal is a sinusoid of maximum allowable amplitude to the mean noise power.

Therefore, the mean-square values of the noise currents given in Section II are appropriate for PPM signals with the substitution of κp_o for p_o and $\mathcal{B}$ for b. Thus,

$$\langle i_n^2 \rangle = \langle i_Q^2 \rangle + \langle i_T^2 \rangle + \langle i_D^2 \rangle + \langle i_L^2 \rangle + \langle i_B^2 \rangle \tag{11}$$

where

$$\langle i_Q^2 \rangle = 2e \frac{e}{h\nu} \eta \kappa p_o G^2 F_d \mathcal{B}, \quad \text{etc.}$$

Now we turn to the problem of evaluating Π, the probability of a threshold violation on the interval $(0, T - 2\mathcal{T})$. We assume for the purposes of this calculation that the noise current,* i_n, during the interval when no signal pulse is present, can be treated as a Gaussian random process strictly bandlimited to the interval $(0, \mathcal{B})$. S. O. Rice[7] computes the probability of such a signal passing a particular value I_1, with positive slope, on the interval Δt to be

$$\frac{1}{\sqrt{3}} \exp - (I_1^2/2\langle \mathrm{i}_n^2 \rangle) \mathcal{B} \Delta t.$$

Thus the probability of the noise alone crossing the threshold $(\frac{1}{2} i_m)$ during the time $T - 2\mathcal{T}$ is

$$\begin{aligned} \Pi &= \frac{1}{\sqrt{3}} \exp \left[(\tfrac{1}{2} i_m)^2 / 2 \langle \mathrm{i}_n^2 \rangle \right] \mathcal{B}(T - 2\mathcal{T}) \\ &= \frac{2}{\sqrt{3}} (\kappa - 1) \exp (-i_m^2 / 8 \langle \mathrm{i}_n^2 \rangle). \end{aligned} \tag{12}$$

Now the noise current i_n, which is important for threshold violation, is not the same as the noise current i_n characterized by (11) because there is no signal during the interval between pulses (when threshold violations can occur) and two of the terms which contribute to i_n, namely, the quantum-noise current i_Q and the beat-noise current i_B, are correspondingly absent. Therefore

$$\langle \mathrm{i}_n^2 \rangle = \langle i_D^2 \rangle + \langle i_T^2 \rangle + \langle i_L^2 \rangle.$$

It will turn out that in many cases of interest $\langle \mathrm{i}_n^2 \rangle \ll \langle i_n^2 \rangle$. This result, which has no classical radio-frequency analog, allows considerably more bandwidth-for-signal-power trade in optical systems than in radio-frequency systems.

It is convenient to write (12) as

$$\Pi = \frac{2}{\sqrt{3}} (\kappa - 1) e^{-\mathrm{XNR}/2} \tag{13}$$

* Note the distinction between i_n and i_n of the preceding paragraphs.

where

$$\text{XNR} \equiv \frac{i_m^2}{4\langle i_n^2\rangle}\cdot \tag{14}$$

We can now substitute (13) into (10) to obtain

$$\text{SNR} = \frac{\dfrac{\pi^2}{2}(\kappa - 1)^2\left(\eta\dfrac{e}{h\nu}\kappa p_o\right)^2 G^2}{\langle i_n^2\rangle + \dfrac{8\pi^2}{3\sqrt{3}}(\kappa - 1)^3 e^{-\text{XNR}/2}}\cdot \tag{15}$$

Equation (15) is plotted in Fig. 10, for the same parameters used in Fig. 2, with $\kappa = 250$. A certain similarity in the *relative* positions of the corresponding curves is evident in the two figures, but two differences are also immediately apparent. First, the curves in Fig. 10 are translated (horizontally) to smaller values of p_o and (vertically) to larger values of SNR; second, a threshold is introduced (by the threshold violation term) below which the SNR degrades extremely rapidly. In fact, this threshold term goes from negligible to dominant over about a 1-dB change in p_o.

As in the case of analog IM, there is an optimum value of G in PPM systems. It can be found by differentiating (15) with respect to G and solving for the value of G, which renders this derivative equal to zero. Unfortunately, the resulting expression for G_{opt} is quite complicated. The fact that the threshold effect sets in so rapidly, however, can be exploited to simplify the determination of G_{opt}. Over the range on which the threshold effect is negligible we neglect it, and, as before, obtain G_{opt} as given by (2) [but with the noise terms redefined as described in connection with (11)]; over the range on which the threshold effect is dominant, one readily obtains

$$G_{\text{opt}} = \left[\frac{4(N_L + N_T)}{N_D}\right]^{2/5}.$$

Figure 11 illustrates (15) for optimum gain for the same set of parameters used in Fig. 4. Figure 12 presents these results for a typical set of parameters.

It is interesting to compare (15) for SNR with the result obtained in Section II for the signal-to-noise ratio in an intensity modulated system with modulation index m. In order to do this, we first observe that, in order to properly sample a signal of bandwidth b, the sampling rate must be (at least) $2b$. This gives the relationships

$$T = \frac{1}{2b}, \qquad \kappa = \frac{T}{2\mathcal{T}} = \frac{\mathcal{B}}{4b}\cdot \tag{16}$$

If we substitute $4b\kappa$ for $\mathcal{B}$ in the noise terms in (15), we find that, above threshold, the expression for SNR in a PPM system is formally identical to that in an analog IM system (1) except that we make the

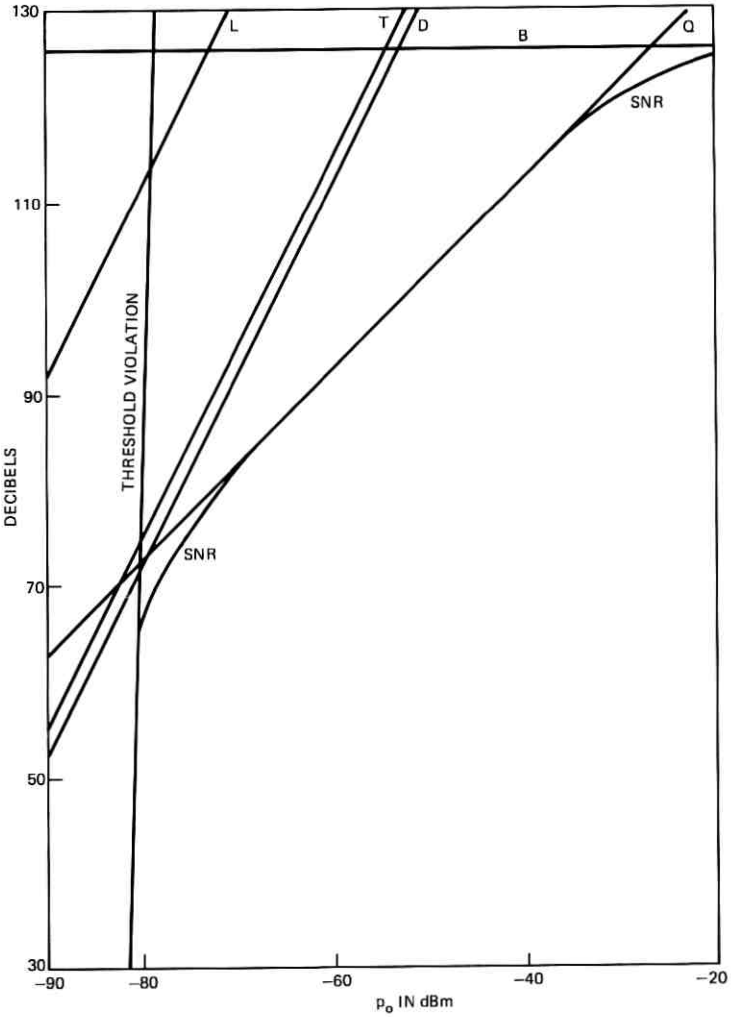


Fig. 10—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the illustrative parameters: $\lambda = 0.85\ \mu$m, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, $b = 4000$ Hz, $G = 100$, for a PPM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power, X = ratio of mean signal power to noise power due to threshold violations.

replacements:

$$p_o \to \kappa p_o, \qquad m \to \frac{\pi}{2}\frac{\kappa - 1}{\sqrt{\kappa}}.$$

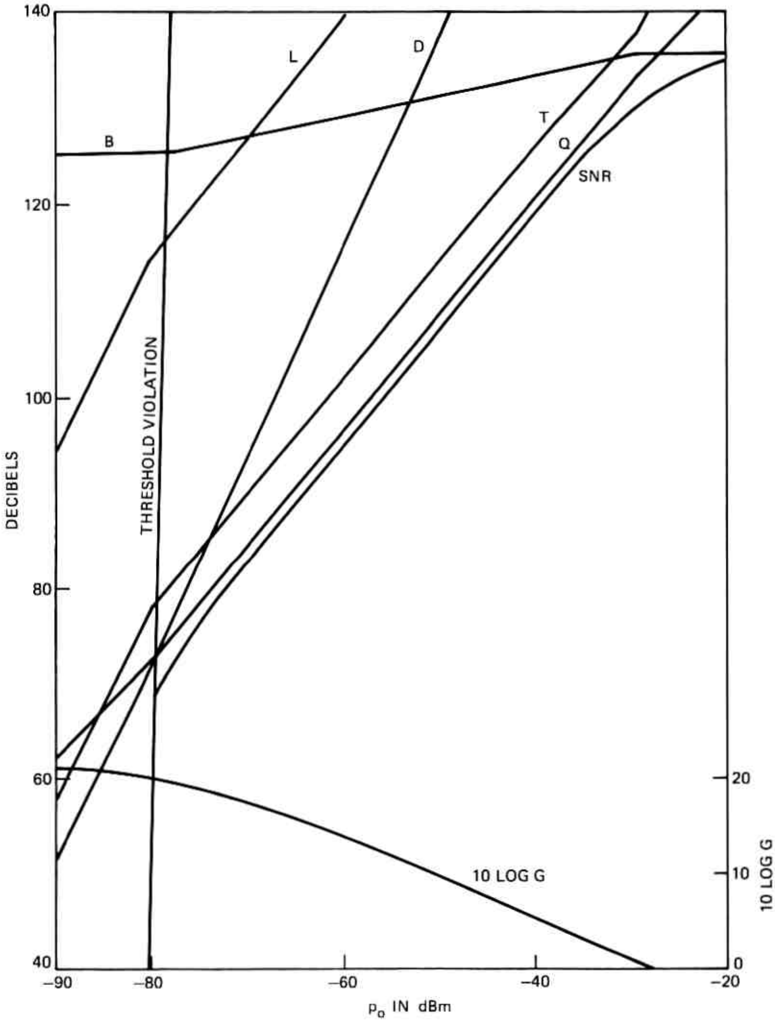


Fig. 11.—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the illustrative parameters: $\lambda = 0.85\ \mu$m, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-9}$ A, $I_L = 10^{-8}$ A, $WJ = 10^{15}$ Hz, $b = 4000$ Hz, $G = G_{\text{opt}}$, for a PPM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power, X = ratio of mean signal power to noise power due to threshold violations.

From this we see that the PPM system yields considerable improvement over the IM system. For PPM, the average signal power is effectively increased by a factor κ and the modulation index (which is less

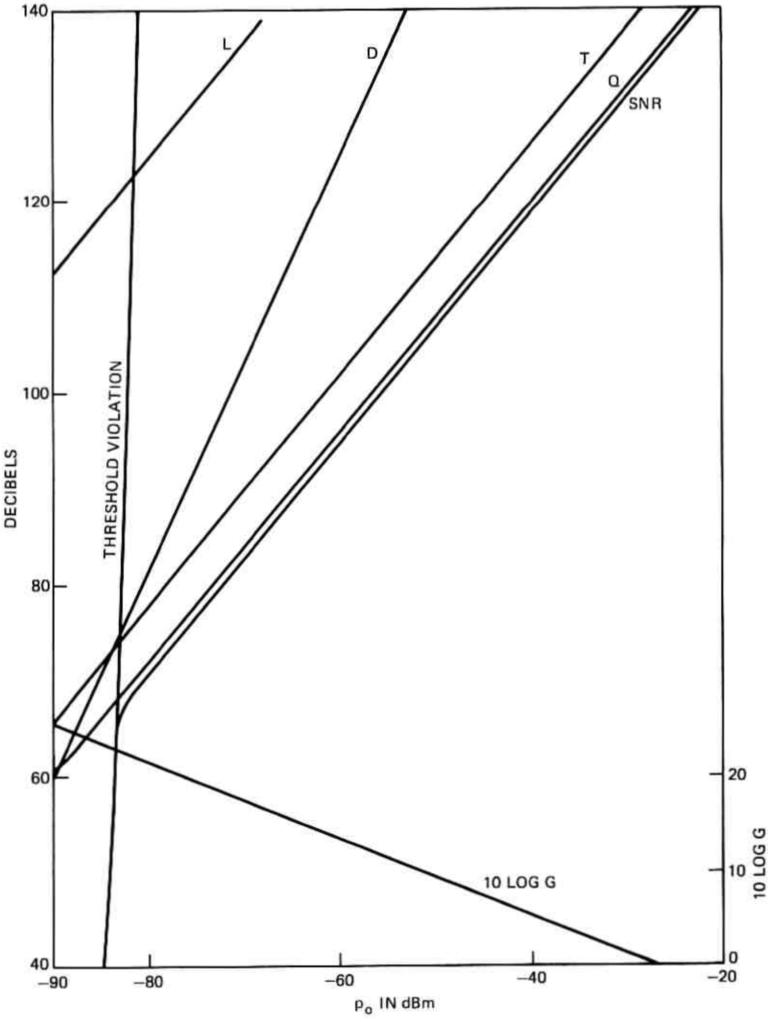


Fig. 12—Signal-to-noise ratio, SNR, and ratio of mean signal power to each component of the mean noise power for the typical parameters: $\lambda = 0.85\ \mu$m, $\eta = 0.5$, $M = 0.85$, $R = 1000$ ohms, $I_d = 10^{-10}$ A, $I_L = 10^{-9}$ A, $WJ = 10^{17}$ Hz, $b = 4000$ Hz, $G = G_{\text{opt}}$, for a PPM channel. B = ratio of mean signal power to beat-noise power, L = ratio of mean signal power to leakage-current-noise power, D = ratio of mean signal power to dark-current-noise power, T = ratio of mean signal power to thermal-noise power, Q = ratio of mean signal power to quantum-noise power, X = ratio of mean signal power to noise power due to threshold violations.

than 1 for an IM system) is replaced by $(\pi/2)(\kappa - 1)/\sqrt{\kappa}$ which can be substantially larger than 1.

In practical applications, bandwidth expansion factors of over a thousand are sometimes possible before threshold violations become important.

3.1 *Power Available*

It is tempting to hypothesize that the average light power obtainable from a given diode is proportional to the average thermal power which can be dissipated in the device without causing catastrophic failure. The thermal power dissipated by the device, P_I [for the signal given by (4)], is

$$\begin{aligned} P_I &= \tfrac{1}{4} I_m^2 \frac{\mathcal{R}}{T} \int_{-\mathcal{T}}^{\mathcal{T}} \left(1 + \cos \pi \frac{t}{\mathcal{T}}\right)^2 dt \\ &= \frac{3}{8\kappa} \mathcal{R} I_m^2 \end{aligned} \tag{17}$$

where $\mathcal{R}$ is the effective resistance of the device. This gives

$$P_o = \mu \sqrt{\frac{2}{3} \frac{P_I}{\kappa \mathcal{R}}} \tag{18}$$

where μ is a constant of proportionality such that $P_m = \mu I_m$. Thus for P_I and $\mathcal{R}$ fixed, P_o (and hence p_o) varies as $\kappa^{-\frac{1}{2}}$.

Unfortunately, the behavior of real LED's and injection lasers is not this simple. First, the heat capacity of some LED's is so small that if a step function change occurs in the diode current, the diode temperature reaches its new steady-state value very quickly. For example, some diodes have such small heat capacity that burnout occurred whenever pulse duration exceeded a few microseconds (independent of duty cycle) if the peak pulse current exceeds the tolerable dc value.[8] Second, even when the pulse is short enough to avoid this problem, the concept of constant power dissipation is not exactly correct. For example, the peak current may be limited by saturation effects.

We observe experimentally for diodes of the type described in Ref. 8 that, for a pulse repetition rate of 8 kHz, the maximum peak pulse power achievable with pulses of 0.1 to 0.25 μs duration is about 5 dB less than that predicted by the constant power dissipation model. Nevertheless, the model is useful as a qualitative guide to diode behavior in pulsed operation.

IV. DIGITAL BINARY PULSE CODE MODULATION

A second method of trading bandwidth for noise immunity is the use of pulse code modulation (PCM). This also has the advantage of being

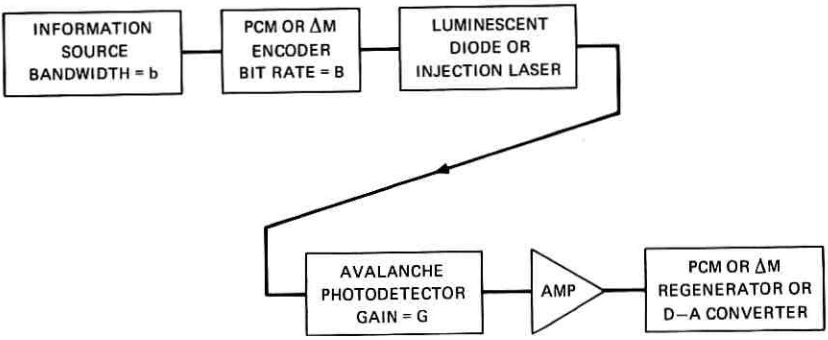

INFORMATION SOURCE BANDWIDTH = b
PCM OR ΔM ENCODER BIT RATE = B
LUMINESCENT DIODE OR INJECTION LASER
AVALANCHE PHOTODETECTOR GAIN = G
AMP
PCM OR ΔM REGENERATOR OR D–A CONVERTER

Fig. 13—Block diagram of a single-channel PCM optical communication system.

readily compatible with digital data transmission. Analysis of a digital PCM channel is somewhat different from that of an analog channel in that the parameter used to characterize a PCM channel is not SNR but rather the error probability P_e.

Memoryless binary optical digital communication systems can be divided into two broad classes–single-channel systems in which the information is coded such that a pulse of energy represents a "1" and no pulse represents a "0"; and twin-channel systems in which energy is transmitted for both information states, but the signal is modulated in such a manner that an appropriate device in the receiver routes the signal energy into one of two channels when a "1" is transmitted and into the other when a "0" is transmitted. It has been shown,[9] however, that unlike the classical radio-frequency case, the twin-channel receiver offers little, if any, advantage in an optical system. In fact, if the transmitter is average-power limited, a single-channel receiver has at least a 1.5-dB advantage in noise immunity over a twin-channel receiver; if the transmitter is peak-power limited, the single-channel receiver suffers, at worst, a 1.5-dB disadvantage. Since the single-channel system is considerably easier to implement than the twin-channel system, and since the twin-channel system offers no significant advantages, we confine our treatment to a single-channel system. Figure 13 is a block diagram of such a system.

It has previously been shown[9,10] that if we assume that the avalanche current gain G is deterministic,* the probability that the receiver

* By this we mean that if m primary electrons are liberated, exactly Gm electrons will be delivered to the load. This artificial constraint will be relaxed in the next section.

mistakes a "0" for a "1" is given by

$$P(1|0) = \tfrac{1}{2} \sum_{n=0}^{\infty} p_o(n) \operatorname{erfc} \left\{ \frac{x_t - nG}{\sqrt{2\langle x_T^2 \rangle}} \right\} \tag{19}$$

and the probability that it mistakes a "1" for a "0" is

$$P(0|1) = \tfrac{1}{2} \sum_{n=0}^{\infty} p_1(n) \operatorname{erfc} \left\{ \frac{nG - x_t}{\sqrt{2\langle x_T^2 \rangle}} \right\} \tag{20}$$

where $\langle x_T^2 \rangle = \langle i_T^2 \rangle/(eB)^2$ is the mean-square thermal noise current expressed as the mean-square average number of electrons flowing during a time slot due to thermal noise and x_t is the decision threshold also expressed in terms of the number of electrons per time slot. This normalization will turn out to be very convenient in that it will allow us to present the results in a form which is independent of bit rate. And where

$p_i(n) = \dfrac{m_i^n}{n!} e^{-m_i}, \qquad i = 0, 1,$

B = bit rate,

m_0 = mean number of primary electrons liberated when a "0" is transmitted,

m_1 = mean number of primary electrons liberated when a "1" is transmitted,

erfc$(\cdot)$ is the complement of the error function.

These equations are derived under the assumption that the thermal noise is Gaussian and that the statistics of the primary electrons liberated in the photodetector due to signal photons, background-illumination photons, and dark current are independent Poisson processes. Then m_0 represents the sum of the means of the background illumination and dark current processes and $m_1 = m_0 + m_s$ where m_s is the mean number of photoelectrons liberated due to the signal. The majority of this section is devoted to the problem of determining the required value of m_s in order to achieve a specified error probability. The required average optical power is, of course, just

$$p_0 = \frac{1}{2} \frac{h\nu m_s}{\eta} B \tag{21}$$

where the factor $\frac{1}{2}$ comes from the assumption that 0's and 1's are equally probable.

The threshold value x_t is, of course, chosen to minimize the total error probability. To a very good approximation, this is achieved when

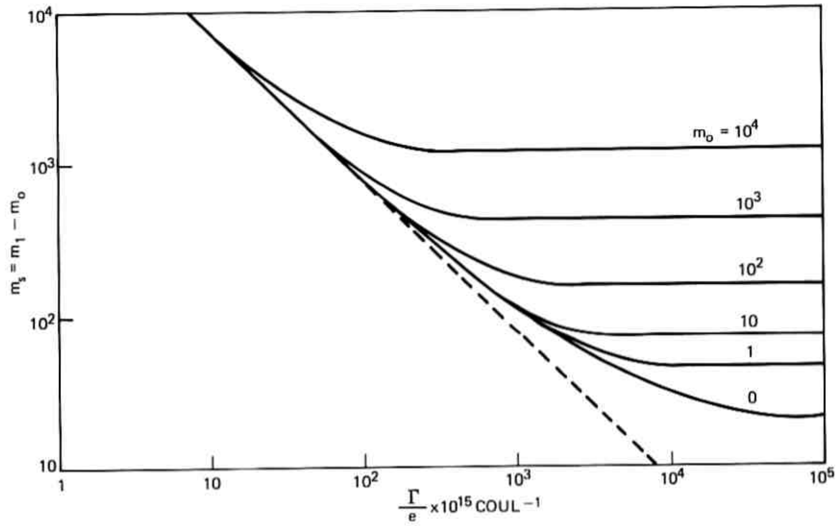


Fig. 14—Number of "signal photoelectrons" for a 10^{-9} error probability from the deterministic gain model vs $\frac{\Gamma}{e} \times 10^{15}$ Coul^{-1}. (Note $\frac{\Gamma}{e} \times 10^{15}$ Coul$^{-1} \approx G$ for reasonable system parameters.)

$P(0|1) = P(1|0)$. Thus, in practice, one can compute x_t from (19) setting $P(1|0)$ equal to the required error probability P_e; and then, knowing x_t, compute the required value of m_1 from (20). In the remainder of this section, it is assumed that the noise bandwidth $\mathcal{B}$ is equal to the bit rate B.

Figure 14 illustrates the results of the calculation described in the preceding paragraph. It is expedient to introduce the dimensionless parameter $\Gamma \equiv G/\langle x_T^2 \rangle^{\frac{1}{2}}$. Examination of this figure reveals that, for small values of Γ, the required signal power is inversely proportional to Γ. However, as Γ is increased, a limiting value, set by the dark current, is soon reached. This sets a maximum value of useful gain (for a given $\langle x_T^2 \rangle$), dependent only on m_0, beyond which no further significant improvement can be achieved. Note that this is true even here for the deterministic gain model with no excess noise factor of the sort to be discussed in Section 4.1.

4.1 *Gain-Dependent Excess Diode Noise Factor*

The gain-dependent excess diode noise factor* F_d played a very important role in the behavior of analog IM and PPM systems. There is

* Recall that F_d results from the random nature of the avalanche gain process.

no reason to believe that it has any less significant role in a digital system. The statistics of the avalanche gain process are very difficult to analyze except in two limiting cases, namely, when the ionization probabilities of holes and electrons in the avalanche region are equal, and when only one carrier contributes to the avalanche process.[1,9,11] Unfortunately, neither case applies to silicon and germanium photodetectors, the two most promising candidates.

Recently S. D. Personick[12] has obtained a rigorous upper bound to the error rate which is applicable to the general case (arbitrary ionization-probability ratio). The result of Personick's calculation is in the form of an integral equation, however, which must be computed numerically. In this section, we derive an approximate relationship between m_s and G which is in excellent agreement* with Personick's result.

This gain-dependent noise figure will have two effects on system behavior: (*i*) it will establish an optimum value of gain in the sense that m_s will have a minimum as a function of G, everything else held constant, and (*ii*) it will cause a larger number of signal photoelectrons to be required, for a given value of G, to satisfy a given error-probability requirement.

The expression we seek is derived as follows. We approximate the Poisson probability density function which describes the primary electron emission by a Gaussian probability density function with the same mean and variance. This approximation, which is discussed in detail in Ref. 13, turns out to be valid for most cases of interest. We also assume, without justification, that the statistics of the output of the photodetector are still Gaussian with variance equal to F_d times the variance of the primary electron distribution† where F_d is the excess detector noise factor used in Sections II and III.

With this assumption, eqs. (19) and (20) can be reduced to:

$$P(1|0) = \tfrac{1}{2}\operatorname{erfc}\left\{\frac{x_t - m_o G}{[2(\langle x_T^2\rangle + m_o F_d G^2)]^{\frac{1}{2}}}\right\} \tag{22}$$

$$P(1|0) = \tfrac{1}{2}\operatorname{erfc}\left\{\frac{m_1 G - x_t}{[2(\langle x_T^2\rangle + m_1 F_d G^2)]^{\frac{1}{2}}}\right\}. \tag{23}$$

We choose x_t in such a way that

$$P(1|0) = P(0|1) = P_e$$

* The results of this approximation are typically less than Personick's upper bound by about 1.0 ± 0.5 dB.

† It is not difficult to show rigorously that this is the correct value of the variance.

where P_e is the error probability. Defining a quantity Q by the relationship

$$P_e = \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{Q}{\sqrt{2}}\right)$$

enables one to write

$$\frac{x_t - m_o G}{(\langle x_T^2\rangle + m_o F_d G^2)^{\frac{1}{2}}} = Q = \frac{m_1 G - x_t}{(\langle x_T^2\rangle + m_1 F_d G^2)^{\frac{1}{2}}}.$$

Eliminating x_t between these equations and setting $F_d = G^{\frac{1}{2}}$ gives

$$m_s = m_1 - m_o = 2Q\left[\frac{\langle x_T^2\rangle}{G^2} + m_o G^{\frac{1}{2}}\right]^{\frac{1}{2}} + Q^2 G^{\frac{1}{2}}. \qquad (24)$$

It is clear from (24) that m_s has a minimum in G.

It was previously stated that $\langle x_T^2\rangle$ is independent of bit rate. This comes about as follows. From its definition and that of $\langle i_T^2\rangle$,

$$\langle x_T^2\rangle = \frac{4\mathrm{k}T}{e^2 B}\frac{F_t}{R},$$

but R is inversely proportional to bit rate. In fact, for a well-designed detector,* one can write

$$R = \frac{1}{4Bc}$$

where c is the capacitance of the photodetector. Then

$$\langle x_T^2\rangle = \frac{16\mathrm{k}TF_t c}{e^2}.$$

A value of $e\langle x_T^2\rangle^{\frac{1}{2}} = 4(\mathrm{k}TF_t c)^{\frac{1}{2}} = 10^{-15}$ Coul implies a value of $F_t c = 15$ pF which is typical of good avalanche photodetectors. This value is used in the example illustrated in Fig. 15. In this figure, we observe that for small values of avalanche gain, where the system performance is thermal-noise limited, the behavior is identical to that of the deterministic gain case. As the gain is increased, however, m_s reaches a broad minimum at $G = G_{\text{opt}}$ and then increases slowly as the gain is further increased.

With proper exclusion of background illumination, typical values of dark and leakage currents are 10^{-10} and 10^{-9} A, respectively. The background count, m_o, is given by

$$m_o = \frac{I_b}{eB} + \frac{I_L}{eBG} = \frac{10^9}{1.6B} + \frac{10^{10}}{1.6BG}.$$

From Fig. 16, we observe that $G_{\text{opt}} \approx 100$ over a wide range of condi-

* Note added in proof: Recent work by S. D. Personick shows that some advantage may be obtained by using larger R and post-detection equalization.

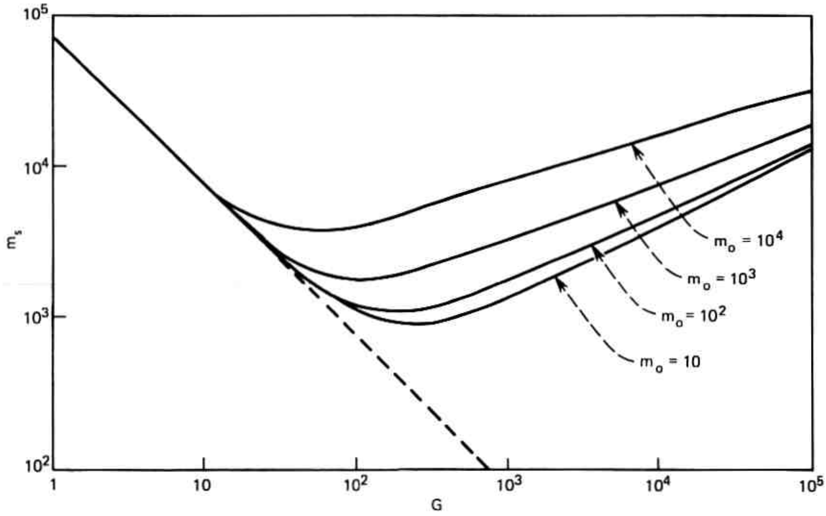


Fig. 15—Number of "signal photoelectrons" for a 10^{-9}-error probability from the gain-dependent noise model ($F_d = G^{\frac{1}{2}}$) vs avalanche gain. $e\langle x_T^2\rangle^{\frac{1}{2}} = 10^{-15}$ Coul.

tions and therefore the dark-current term usually dominates the leakage-current term. Typical values of m_o would therefore be from about 600 at $B = 1$ Mb/s down to about 0.6 at $B = 1$ Gb/s.

Figure 17 shows the value of m_s required for $P_e = 10^{-9}$ ($Q = 6.00$) for $e\langle x_T^2\rangle^{\frac{1}{2}} = 10^{-15}$ Coul [from (24)] for $G = 100$ and for $G = G_{\text{opt}}$. The value of G_{opt} is also shown in Fig. 17. Two important results are apparent from Fig. 17: m_o is not important until it exceeds about 100, and using $G = 100$ instead of $G = G_{\text{opt}}$ costs no more than about 1 dB. This last result is important because G_{opt} is so large over much of the region of interest that it would be difficult to obtain.

We now turn our attention to the use of nonadaptive delta modulation (ΔM) for transmitting analog signals. Delta modulation is a form of digital modulation which allows a trade-off between bandwidth and both terminal cost and signal power.

Noise in ΔM systems has been studied in detail by several authors.[14–17] We present here a sketch of how one might estimate the bit-rate requirements for a ΔM system. Suppose that the frequency band of the information source extends from 0 to b and that the step size of the coder is s. The maximum slope of a sine wave with amplitude A and frequency f is $2\pi A f$, while the maximum slope of the quantized signal with step size s and sampling rate B is sB. The limiting condition for slope overload is then

$$sB = 2\pi A f. \tag{25}$$

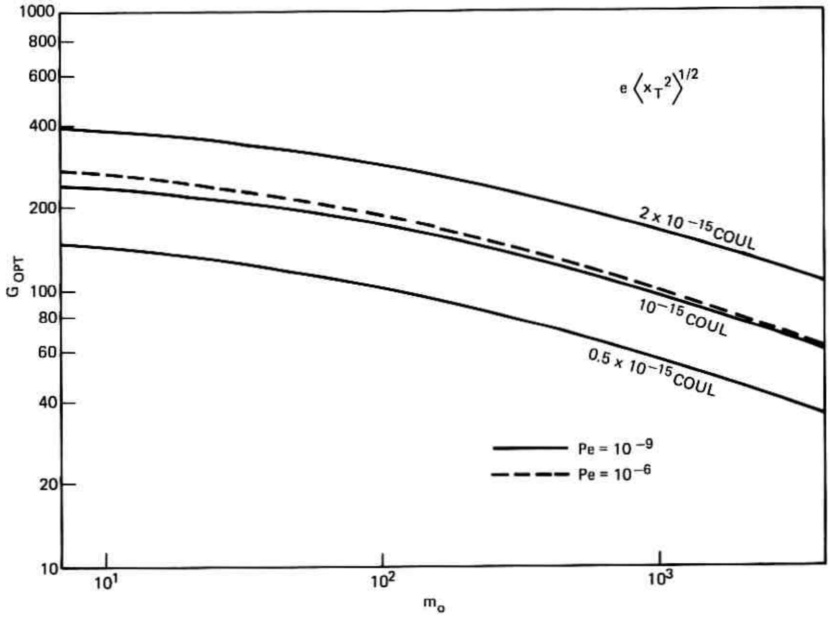


Fig. 16—Optimum gain ($F_d = G^{\frac{1}{2}}$) vs primary background count, m_o.

In order to compute the mean-square error in the quantized signal we assume, following Van de Weg,[14] that there is no correlation between samples and that the difference between the source signal and the quantized signal is uniformly distributed on the interval $(-s, s)$. We obtain:

$$\langle \delta s^2 \rangle = \frac{1}{2s} \int_{-s}^{s} x^2 dx = \tfrac{1}{3} s^2 = \text{mean quantizing noise power.} \quad (26)$$

The spectrum of the noise is quite complicated, but for our immediate purpose it is sufficient to assume that this noise is spread more or less uniformly over a bandwidth B so that the fraction b/B of the quantizing noise power falls into the information band b. (The remainder of the noise can then be eliminated by a low-pass filter of bandwidth b.) Assuming that quantizing noise is the only significant noise source, the signal-to-noise ratio, SNR, is then given for a sinusoidal signal of amplitude A by:

$$\text{SNR} = \frac{\tfrac{1}{2} A^2}{\dfrac{1}{3} s^2 \dfrac{b}{B}} = \frac{3}{2} \frac{B}{b} \left(\frac{A}{s} \right)^2 = \frac{3}{8\pi^2} \frac{B^3}{b f^2} \quad (27)$$

where f is the highest frequency we are required to transmit with the

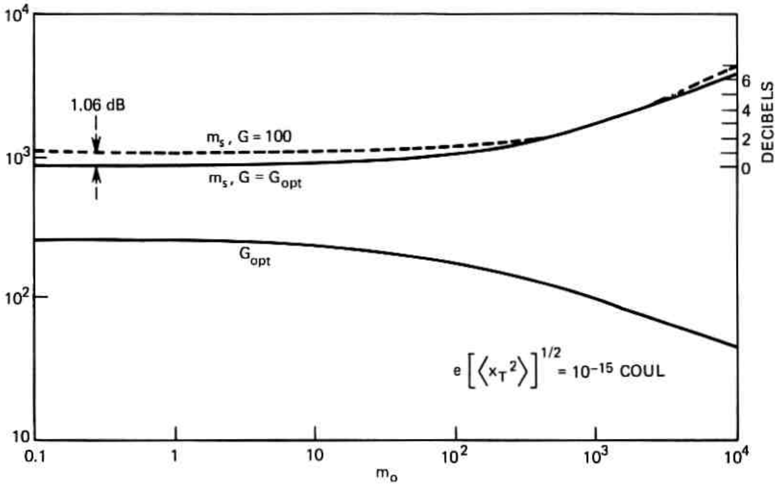


Fig. 17—Number of "signal photoelectrons," m_s, vs background count, m_o, for $P_e = 10^{-9}$.

specified value of SNR.* For a voice channel, we take (as conservative values) $b = 4 \times 10^3$ Hz and $f = 2 \times 10^3$ Hz. Then

$$\mathcal{B} = \left[\frac{8\pi^2}{3} bf^2(\text{SNR})\right]^{\frac{1}{3}} = 7500(\text{SNR})^{\frac{1}{3}} \text{ Hz}. \tag{28}$$

For SNR = 70 dB, this gives $B = 1.6$ MHz. Van de Weg's calculation[14] takes into account the correlation between samples (which we have neglected) and the details of the quantizing noise spectrum. His result (for $B/b \geqq 4$) replaces the factor $8\pi^2/3 = 26.3$ in (28) by the factor 25.0 (and leaves it otherwise unaffected). Thus, his result is essentially identical with the one derived above. Experimental work by Laane and Murphy[17] indicates that a value of $B = 1.5$ MHz is adequate for transmitting a single voice channel by ΔM; we use this value in the following calculation.

For a 1.5-Mb/s rate, $I_D = 10^{-9}$ and $I_L = 10^{-10}$ gives

$$m_o = 417 + 4170/\text{G} \approx 460.$$

From Fig. 17, the required value of m_s is 1450 (which is 2.2 dB poorer

* A ΔM system is limited by slope overload as indicated by (25). The limiting condition is set not by amplitude or by frequency alone, but by their product Af. We define the SNR for the ΔM system in terms of the ratio of full load power in a sinusoidal signal at some frequency f to the mean quantizing noise power. At higher frequencies, the available SNR degrades at a rate of 6 dB per octave. One could, of course, choose f to be the highest frequency in the information source bandwidth, b. For voice signals, however, this turns out to be an unreasonable constraint.[18]

than the case of no dark current and optimum gain, and 1.1 dB poorer than the case of no dark current and $G = 100$). The corresponding value of p_o is (for $\lambda = 0.85\ \mu\text{m}$, $\eta = 0.5$) $0.51 \times 10^{-9}\ \text{W} = -63\ \text{dBm}$. This is based on a tolerable error probability of 10^{-9}, which is probably much better than necessary. However, the error probability varies at the rate of about one order of magnitude for every 0.5 dB change in optical signal power. Therefore it makes little difference what value of error probability is chosen for this calculation.

It has been stated in the preceding sections that dark current and leakage current are generally negligible with good silicon photodiodes. The case of 1.5-Mb/s PCM is just on the border line of being dark-current limited with the numbers used in this example. The values of I_D and I_L used in this example must not, however, be regarded as ultimate performance. Indeed, photodetectors with $I_D \leqq 10^{-11}$ A have been built.[4]

V. CONCLUSIONS

Because light-emitting diodes and diode lasers can be directly modulated, analog intensity modulation is the simplest form of modulation to implement. Considerable improvement in noise immunity can be obtained, however, by judiciously exploiting the wide available bandwidth in optical systems.

Thus, pulse position modulation is particularly attractive because the square-law nature of the detector makes the "bunching" of the optical power beneficial and because the signal-power-dependent nature of the noise makes very large bandwidth-expansion factors feasible. Improvement of over 40 dB (relative to intensity modulation) is theoretically possible with pulse position modulation. Improvement of about 30 dB for a single high-quality 4-kHz voice channel appears to be realizable with existing light-emitting diodes. Use of delta modulation affords a theoretical improvement in noise immunity of about 25 dB relative to analog intensity modulation.

Optical carriers appear attractive for pulse code modulation even at low bit rates. At a bit rate of 6 Mb/s, only about -58 dBm of signal power is required for 10^{-9} error rate.

APPENDIX

Comparison of Practical Direct Detection Receivers With Homodyne Receivers and With Ideal Reception for Binary PCM Channels

It is instructive to compare the performance of a direct detection receiver of the sort described in Section IV with a homodyne receiver operating on a similar signal. Consider a signal in which a "1" is represented by a pulse with peak power P_s and a "0" is represented by the absence of a pulse. Let P_{LO} be the local oscillator power. The detected photocurrent (with no avalanche gain) is then

$$i_s = \eta \frac{e}{h\nu} [2\sqrt{p_m P_{\mathrm{LO}}} + P_{\mathrm{LO}} + p_m] \tag{29}$$

at the peak of the pulse when a "1" is transmitted and is $\eta(e/h\nu)P_{\mathrm{LO}}$ when a "0" is transmitted. In practical operation, $P_{\mathrm{LO}} \gg p_m$ and the last term in (29) is negligible. The quantity $\eta(e/h\nu)P_{\mathrm{LO}}$ is just a dc shift and can be neglected in the following calculation.

The mean-square noise current is given by

$$\langle i_n^2 \rangle = 2e \frac{e}{h\nu} \eta (P_{\mathrm{LO}} + p_m)\mathcal{B} + 2eI_d \mathcal{B} + \frac{4\mathrm{k}TF_t\mathcal{B}}{R}.$$

But once again $P_{\mathrm{LO}} \gg p_m$ and in any reasonable receiver one also has $\eta(e/h\nu)P_{\mathrm{LO}} \gg I_d$, so $\langle i_n^2 \rangle$ becomes

$$\langle i_n^2 \rangle = \left[2e\eta \frac{e}{h\nu} P_{\mathrm{LO}} + \frac{4\mathrm{k}TF_t}{R} \right] \mathcal{B}.$$

For large P_{LO}, this (Poisson) noise can be regarded as Gaussian. The optimum decision threshold will be near $(\frac{1}{2})i_s$. Therefore, the probability of error is well approximated by

$$\begin{aligned} P_e &= \frac{1}{\sqrt{2\pi\langle i_s^2 \rangle}} \int_{\eta(e/h\nu)\sqrt{P_{\mathrm{LO}}p_m}}^{\infty} \exp\left\{ -\frac{x^2}{2\langle i_n^2 \rangle} \right\} dx \\ &= \tfrac{1}{2} \operatorname{erfc} \left\{ \frac{1}{\sqrt{2}} \sqrt{\frac{\eta p_m}{2h\nu\mathcal{B} + \dfrac{4\mathrm{k}TF_t h\nu\mathcal{B}}{eR\eta \dfrac{e}{h\nu} P_{\mathrm{LO}}}}} \right\}. \end{aligned}$$

In order to fully exploit the advantages of homodyne detection, one must require

$$P_{\mathrm{LO}} \gg \frac{2\mathrm{k}TF_t}{e \dfrac{e}{h\nu} \eta R}.$$

When this condition obtains, one has

$$P_e = \tfrac{1}{2} \operatorname{erfc} \left\{ \frac{1}{\sqrt{2}} \sqrt{\frac{\eta p_m}{2h\nu\mathcal{B}}} \right\}.$$

Assuming a square pulse of duration $1/\mathcal{B}$ gives

$$m_s = \frac{\eta p_m}{h\nu \mathcal{B}}$$

and

$$P_e = \tfrac{1}{2}\,\mathrm{erfc}\left\{\frac{1}{\sqrt{2}}\sqrt{\frac{m_s}{2}}\right\}. \tag{30}$$

We model an "ideal receiver" by a device which unerringly distinguishes between the case when no photoelectrons were liberated and the case when one or more were liberated. Since the photoelectrons are Poisson distributed, the probability that none were liberated when the expected number was m_s is just

$$e^{-m_s}$$

and for the ideal receiver this is twice the probability of error. For a $P_e = 10^{-9}$ this gives:

Ideal receiver:	$m_s = 20$
Homodyne receiver:	$m_s = 72.$

Thus the homodyne receiver is 5.5 dB poorer than the ideal. From Section IV we see that a practical direct detection receiver requires $m_s \approx 1000$ which places it 17 dB worse than the ideal receiver and about 12 dB worse than the homodyne receiver. Of course, 12 dB is not insignificant; but homodyne (or heterodyne) detection requires both a coherent source and precise phase-front matching between the signal and the local oscillator. With LED's this is impossible, with existing diode lasers it is at best extremely difficult. Even if adequate phase-front matching could be achieved, phase-lock for the homodyne receiver would be extremely difficult—if at all possible; heterodyne detection would reduce the advantage to 9 dB.

Direct detection without avalanche gain requires $m_s = 7.2 + 10^4$ (for $e\langle x_T^2\rangle^{\frac{1}{2}} = 10^{-15}$ Coul) which is almost 36 dB worse than the ideal receiver. Thus, in the example of Section IV the avalanche gain ($G = 100$) gives almost 19 dB improvement.

These relative performance numbers are based on $P_e = 10^{-9}$ but over the range $10^{-10} < P_e < 10^{-4}$ the relative performance of the ideal detector, the homodyne detector, and the avalanche photodetector varies by less than 1 dB while the performance of direct detector without avalanche gain relative to the ideal detector varies by no more than 2 dB on this range.

REFERENCES

1. McIntyre, R. J., "Multiplication Noise in Uniform Avalanche Diodes," IEEE Trans. Elec. Dev., *ED-13*, No. 1 (January 1966), pp. 164–168.
2. Anderson, L. K., DiDomenico, M., Jr., and Fisher M. B., "High-Speed Photodetectors for Microwave Demodulation of Light," in *Advances in Microwaves*, ed. L. Young, vol. 5, New York: Academic Press, 1970.
3. Davenport, Jr., W. B., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
4. Melchior, H., Fischer, M. B., and Arams, F. R., "Photodectors for Optical Communication Systems," IEEE, *58*, No. 10 (October 1970), pp. 1466–1486.
5. Melchior, H., and Lynch, W. T., "Signal and Noise Response of High-Speed Germanium Avalanche Photodiodes," IEEE Trans. Elec. Dev., *ED-13*, No. 12 (December 1966), pp. 829–838.
6. Schwartz, M., Bennett, W. R., and Stein, S., *Communication Systems and Techniques*, New York: McGraw-Hill.
7. Rice S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23*, No. 3 (July 1944), pp. 282–332.
8. Burrus, C. A., and Dawson, R. W., "Small-Area High-Current Density GaAs Electroluminescent Diodes and a Method of Operation for Improved Degradation Characteristics," Appl. Phys. Lett., *17*, No. 3 (August 1970), pp. 97–99.
9. Hubbard, W. M., "Comparative Performance of Twin-Channel and Single-Channel Optical-Frequency Receivers," IEEE Trans. Commun. COM–20, No. 6, Dec. 1972, pp. 1079–1086.
10. Karp, S., and Gagliardi, R. M., "The Design of a Pulse-Position Modulated Optical Communication System," IEEE Trans. Commun. Tech., *Com-17*, No. 6 (December 1969), pp. 670–676.
11. Personick, S. D., "New Results on Avalanche Multiplication Statistics with Applications to Optical Detection," B.S.T.J., *50*, No. 1 (January 1971), pp. 167–190.
12. Personick, S. D., "Statistics of a General Class of Avalanche Detectors with Application to Optical Communication," B.S.T.J., *50*, No. 10 (December 1971), pp. 3075–3095.
13. Hubbard, W. M., "The Approximation of a Poisson Distribution by a Gaussian Distribution," Proc. IEEE, *58*, No. 9 (September 1970), pp. 1374–1375.
14. Van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep., *8*, 1953, pp. 367–385.
15. O'Neal, J. B., Jr., "Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45*, No. 1 (January 1966), pp. 117–142.
16. Iwersen, J. E., "Calculated Quantizing Noise of Single-Integration Delta-Modulation Coders," B.S.T.J., *48*, No. 7 (September 1969), pp. 2359–2389.
17. Laane, R. R., and Murphy, B. T., "Delta Modulation Codec for Telephone Transmission and Switching Applications," B.S.T.J., *49*, No. 6 (July 1970), pp. 1013–1032.
18. deJager, F., "Delta Modulation, A Method of PCM Transmission Using 1-Unit Code," Philips Res. Rep., *7*, 1952, pp. 442–446.

Contributors to This Issue

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research on traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, and stochastic control. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. In 1971 he taught stochastic processes at SUNY Buffalo, and in 1971–72 he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963) and of *Mathematical Theory of Connecting Networks and Telephone Traffic* (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, Mind Association, Phi Beta Kappa.

Joel Goldman, B.E.E., 1965, The Cooper Union; M.S., 1967, and Ph.D. (E.E.), 1970, Cornell University; Bell Laboratories, 1970—. Mr. Goldman is presently engaged in the analysis of the effects of digital and analog interference on various communications systems. He has also performed research in estimation and prediction theory. Member, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, IEEE, Eta Kappa Nu, Tau Beta Pi.

W. M. Hubbard, B.S., 1957, Georgia Institute of Technology; M.S., 1958, University of Illinois; Ph.D., 1963, Georgia Institute of Technology; Bell Laboratories, 1963–1972. Mr. Hubbard's work has included analyses related to the design of millimeter-wave solid-state repeaters for use in a waveguide transmission system and the construction of prototype high-speed repeaters for this type of system. He was subsequently engaged in optical transmission research with emphasis on repeater techniques. Since January 1973, Mr. Hubbard has been on a temporary assignment at AT&T Engineering, New York City. Member, Sigma Xi, Tau Beta Pi, Phi Kappa Phi, American Physical Society.

James McKenna, B.Sc. (Mathematics), 1951, Massachusetts Institute of Technology; Ph.D. (Mathematics), 1961, Princeton University; Bell Laboratories, 1960—. Mr. McKenna has done research in quantum mechanics, electromagnetic theory, and statistical mechanics. He has recently been engaged in the study of nonlinear partial differential equations that arise in solid state device work, and in the theory of stochastic differential equations.

Kurt H. Mueller, E. E.-Diploma, 1961, and Ph.D., 1967, Swiss Federal Institute of Technology; Bell Laboratories, 1969—. Mr. Mueller has worked on various problems in the fields of high-speed data communication and signal processing. He is presently on leave of absence at the Swiss Federal Institute of Technology. Member, IEEE.

Attilio J. Rainal, University of Alaska; University of Dayton, 1950–52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel University; D. Eng., 1963, Johns Hopkins University; Bell Laboratories, 1964—. Mr. Rainal's early work involved research on noise theory with application to detection, estimation, and radar theory. He has also been engaged in the analysis of FM communication systems. His more recent work includes studies of crosstalk on multilayer boards, voltage breakdown of printed wiring, and electromagnetic compatibility. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

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.

N. L. Schryer, B.S., 1965, M.S., 1966, and Ph.D., 1969, University of Michigan; Bell Laboratories, 1969—. Mr. Schryer has worked on the numerical solution of parabolic and elliptic partial differential equations. He is currently studying problems of this type which arise in semiconductor device theory.

3406 - 7-8 N.Z,