THE BELL SYSTEM
Technical Journal
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 47 March 1968 Number 3

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, B. S. Gilmer, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $7.00 per year; single copies $1.25 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

Volume 47 March 1968 Number 3

A Unified View of Synchronous Data Transmission System Design

By JAMES W. SMITH

(Manuscript received October 5, 1967)

This paper illustrates the basic equivalence of many of the linear data transmission design techniques. It shows the unifying feature of these techniques to be a generalization of Nyquist's original ideas relating time samples and frequency domain constraints. It examines pulse amplitude (with and without constraints on the input data) and pulse shape modulation systems, and shows their relationships. It uses a number of previously-described systems to illustrate the range of possibilities of the very general design approach. This paper presents some new results on noise and channel parameter monitoring and on spectrum shifting by constraining the input data.

I. INTRODUCTION

Over the years, many seemingly different techniques have been proposed for synchronous data transmission. Unfortunately, the literature devoted to these techniques tends to expand the differences between a specific system and all other systems. It is our purpose to show the basic equivalence of the various techniques; hence, to show a unified view of the field. In doing this we examine some well known and some relatively unknown transmission systems in a new light and propose some new techniques.

The unified design view that we take here is basically a generalization of Nyquist's ideas[1] which have recently been expanded upon by

Gibby and Smith.[2] This really is the thread which ties together virtually all of the literature on synchronous data transmission. Section II summarizes the basic ideas of these two papers.

Section III describes the model of the general, linear, data transmission system to be considered. The system uses M channels and is described by means of an input data vector rather than by any statements about the transmitter characteristics. Thus, one type of data vector implies a pulse amplitude modulated system (PAM) while another type of data vector implies pulse shape modulation (PSM) such as frequency shift keying (fsk) or pulse position modulation (PPM.)

In Section IV, we use Nyquist's approach to find the design constraints for distortionless transmission (no intersymbol or interchannel interference). This section shows that the conditions for distortionless transmission depend upon the input data vector description; hence, different design constraints result from PAM or PSM transmission.

Section V illustrates the application of the constraints to some special cases. These include:

(*i*) Linear precoding and decoding for PAM transmission.

(*ii*) The use of band-limited orthogonal signals for multichannel PAM transmission.

(*iii*) Noise monitoring in PAM systems.

(*iv*) Binary PSM transmission (including the specific case of Sunde's FM model with a linear receiver instead of a phase derivative receiver).

(*v*) Parameter monitoring in PSM systems.

(*vi*) Zero stuffing techniques for shifting spectrum location. (The section shows this to have some promise for voice channel transmission without carrier modulation.)

II. THE UNIFYING VIEW

In designing a data system, one usually starts with a desired time response for the total system. Because it is only necessary to examine the output signal at fixed times (for example, at T second intervals where $1/T$ is the rate at which independent symbols are being transmitted), one needs to specify the over-all response at those times (for example, $t = kT$, all k). Since the total response of the transmitting filter, the channel, and the receiving filter is easier to determine in the frequency domain than in the time domain, one must relate the time response constraints to frequency domain constraints.

This briefly is the basic problem attacked by Nyquist. Here is a summary of his results, as expanded by Gibby and Smith.

Any time function $r(t)$ with Fourier transform $R(\omega)$

$$r(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\omega)e^{j\omega t}\,d\omega \tag{1}$$

has sample values at multiples of T seconds which may be written

$$r(qT) = r_q = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\omega)e^{j\omega qT}\,d\omega \tag{2a}$$

$$r_q = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty}\int_{(2n-1)\pi/T}^{(2n+1)\pi/T} R(\omega)e^{j\omega qT}\,d\omega. \tag{2b}$$

Or, changing the variable,

$$r_q = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty}\int_{-\pi/T}^{\pi/T} R(u + 2n\pi/T)e^{juqT}\,du. \tag{2c}$$

Assuming that

$$\sum_n R(u + 2n\pi/T)$$

is a uniformly convergent series, one obtains

$$r_q = \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T}\sum_{n=-\infty}^{\infty} R(u + 2n\pi/T)e^{juqT}\,du. \tag{2d}$$

Noting that r_q is just the q^{th} coefficient of an exponential Fourier series expansion of

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} R(u + 2n\pi/T) \qquad -\pi/T \leqq u \leqq \pi/T,$$

one obtains

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} R(u + 2n\pi/T) = \sum_{q=-\infty}^{\infty} r_q e^{-juqT} \qquad -\pi/T \leqq u \leqq \pi/T \tag{3}$$

which is very closely related to the Poisson sum formula.[3] Throughout this paper the reader should keep in mind the interval $-\pi/T \leqq u \leqq \pi/T$; we will not be repeating it explicitly.

Equation 3 relates a function of the frequency domain characteristic (namely the sum of the values at frequency intervals $2\pi/T$) to the time response constraints r_q which will be chosen for a particular transmission scheme. Fig. 1 illustrates several frequency characteristics which satisfy the time response constraint that $r_q = 0$ for $q \neq 0$. When $r_0 \neq 0$,

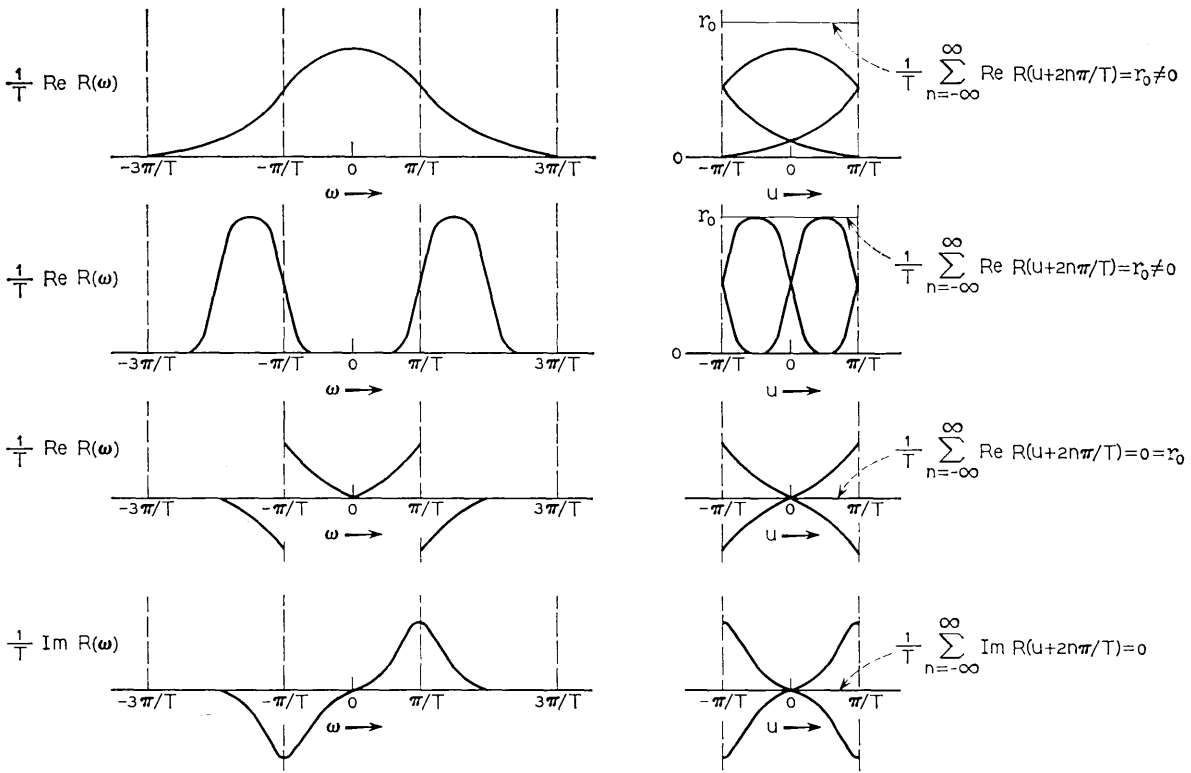


Fig. 1 — Such constraints on $R(\omega)$ that $r(qT) = 0$, $q \neq 0$.

we have the usual design for PAM transmission without intersymbol interference. When $r_0 = 0$, we have a response which is useful when crosstalk between channels is to be eliminated. That is, the output of a signaling path whose time response has $r_q = 0$ for all q contains no information about the input data at the sampling times. For further discussion of equation 3 see Ref. 2.

III. THE GENERALIZED TRANSMISSION MODEL

The optimum theoretical method (in the sense of minimizing error probability) for transmitting data through a Gaussian channel consists of waiting until all data have been accumulated at the transmitter and then sending a single waveform to represent the entire message. The optimum receiver (in the presence of white noise) consists of filters matched to each message waveform. The disadvantage of this form of communication lies in the fact that transmitter and receiver complexity grows exponentially with message length.

Thus, system designers usually restrict system complexity by not waiting for the entire message before transmitting. Short portions of the message can be encoded systematically and transmitted sequentially as they arrive using relatively simple terminal equipment.

Fig. 2 illustrates the general approach to transmission system design considered in this paper. The input data samples, a_{mk}, m = 1, 2, . . . , M are applied at $t = kT$ to the M signal generators $A_m(\omega)$. The sequence $\{a_{mk}\}$ can be considered to be a random sequence of impulses of weight a_{mk} (where a_{mk} is in general multilevel) and spaced T seconds apart. Since there are M signal generators, symbols are be-

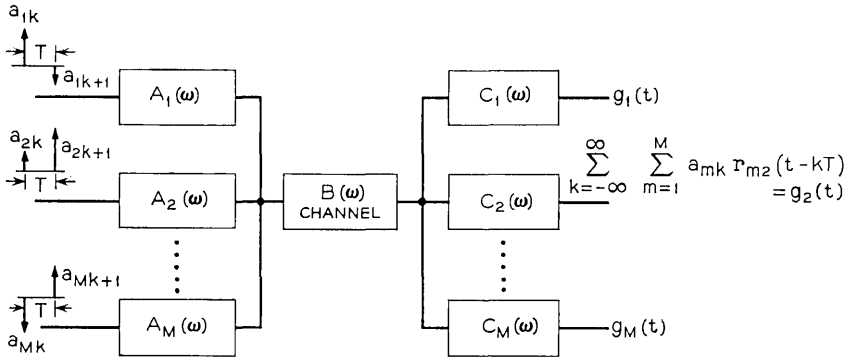


Fig. 2 — General transmission system.

ing sent at the rate M/T symbols per second. The receiver consists of M linear filters, $C_p(\omega)$, $p = 1, 2, \ldots, M$.

The channel input represents the sum of the transmitter outputs and may be written

$$\sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} a_{mk} \mathrm{a}_m(t - kT)$$

where

$$\mathrm{a}_m(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A_m(\omega) e^{j\omega t}\, d\omega.$$

Then, using

$$r_{mp}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A_m(\omega) B(\omega) C_p(\omega) e^{j\omega t}\, d\omega \tag{4}$$

the output of the p^{th} $(p = 1, 2, \ldots, M)$ filter may be written

$$g_p(t) = \sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} a_{mk} r_{mp}(t - kT). \tag{5a}$$

It will be assumed that any decisions will be made on the basis of the output waveform at integral multiples of T seconds. These output samples at the time $t=lT$,

$$g_{pl} = g_p(lT) = \sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} a_{mk} r_{mp}(lT - kT) \tag{5b}$$

may also be written in vector notation as

$$g_{pl} = \sum_{k=-\infty}^{\infty} \vec{r}_p(lT - kT) \cdot \vec{a}_k \tag{5c}$$

where

$$\vec{r}_p(lT - kT) = [r_{1p}(lT - kT), r_{2p}(lT - kT), \cdots, r_{Mp}(lT - kT)] \tag{6}$$

and

$$\vec{a}_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{Mk} \end{bmatrix}. \tag{7}$$

The model described, then, represents a general pulse amplitude modulation system. For example, if the elements of $\vec{a}_k$ are random

multilevel values and the transmitter and receiver are systematically chosen to be

$$A_m(\omega) = A_0(\omega - \omega_m) + A_0(-\omega - \omega_m) \tag{8a}$$

$$C_m(\omega) = C_0(\omega - \omega_m) + C_0(-\omega - \omega_m) \tag{8b}$$

one has a frequency division multiplex PAM system. Likewise, choosing

$$A_m(\omega) = A(\omega) \exp\left[-j\frac{(m-1)\omega T}{M}\right] \tag{9a}$$

$$C_m(\omega) = C(\omega) \exp\left[j\frac{(m-1)\omega T}{M}\right] \tag{9b}$$

leads to a time division multiplex PAM system.

For the model to be as general as possible, it should include the possibility of using multiple waveshapes to convey information. This can be accomplished by restricting the k^{th} data word to be

$$\vec{a}_k = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\\vdots\\0\end{bmatrix} \cdots \text{ or } \begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix}. \tag{10}$$

Thus, the system transmits one of M possible waveforms in each time slot and includes such modulation techniques as FM, PM, pulse position or pulse duration modulation.

IV. DESIGN CRITERIA FOR THE GENERALIZED MODEL

The general design constraints for the two different interpretations of $\vec{a}_k$ are imposed by the requirement of distortionless transmission (that is, no intersymbol or interchannel interference). Each of the two cases leads to a different definition of distortionless transmission and hence to different design constraints.

4.1 *Pulse Amplitude Modulation*

(Elements of $\vec{a}_k$ are random and multilevel.)

The output of the p^{th} receiver filter at $t = lT$ is

$$g_{pl} = \sum_{k=-\infty}^{\infty} \vec{r}_p(lT - kT)\cdot\vec{a}_k \,. \tag{5c}$$

For distortionless transmission it is required that this output depend only upon the input value a_{pl}; that is,

$$g_{pl} = a_{pl}K_p \tag{11}$$

where K_p is a constant that depends on the p^{th} channel $A_p(\omega)B(\omega)C_p(\omega)$. This requirement constrains the time response[4]

$$r_{mp}(lT - kT) = \delta_{mp}\,\delta_{lk}K_p \tag{12}$$

where

$$\begin{aligned} \delta_{mp} &= 0 \qquad m \neq p \\ &= 1 \qquad m = p. \end{aligned}$$

Using equations 3, 4, and 12, the time domain constraint becomes the frequency domain constraint

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u + \frac{2n\pi}{T}\right)B\left(u + \frac{2n\pi}{T}\right)C_p\left(u + \frac{2n\pi}{T}\right) = \delta_{mp}\,K_p\,. \tag{13}$$

This is a generalized Nyquist criterion which applies to all linear PAM systems.

Notice that equation 13 represents M^2 equations which must be satisfied by

$$MT/\pi \times \text{positive frequency range of nonzero } B(\omega)$$

independent variables. Therefore, $B(\omega)$ must have a radian bandwidth of at least $M\pi/T$ for M channel distortionless transmission.

4.2 *Pulse Shape Modulation*

($\vec{a}_k$ given by equation 10.)

The definition of distortionless transmission of the previous part (equation 11) could also be applied here. However, it is possible to use a different definition with quite interesting results because of the constraint upon $\vec{a}_k$. Here, distortionless transmission will require that the output of the p^{th} filter at $t = lT$ be

$$g_{pl} = a_{pl}K_{p1} + K_{p2} \tag{14}$$

where K_{p1} and K_{p2} are constants. That is, the output of the p^{th} receiver takes on one of two values, at $t = lT$, $K_{p1} + K_{p2}$ or K_{p2} depending upon the value of a_{pl}.

This definition of distortionless transmission eases the constraints

upon the various time responses. Examining

$$g_{pl} = \sum_{k=-\infty}^{\infty} \vec{r}_p(lT - kT) \cdot \vec{a}_k \tag{5c}$$

it is seen that all elements of $\vec{r}_p(lT - kT)$ must be identical but not necessarily zero for $k \neq l$; that is,

$$r_{mp}(lT - kT) = r_{qp}(lT - kT) \quad \text{all} \quad m, q. \tag{15}$$

With this condition satisfied, $g_p(lT)$ is independent of $\vec{a}_k$ for $k \neq l$ (that is, the information transmitted at times other than lT). Next it is required that all elements of $\vec{r}_p(0)$ be identical except $r_{pp}(0)$ (that is, $r_{mp}(0) = r_{qp}(0)$ $m, q \neq p$). Thus, $g_p(lT)$ will take on one value if $a_{pl} = 1$ and a different value if any other $a_{ql} = 1$ $q \neq p$. These statements may be summarized by the equation

$$r_{mp}(lT - kT) = F_{p,l-k} + \delta_{mp}\, \delta_{lk} G_p \tag{16}$$

where $F_{p,\,l-k}$ and G_p are constants which depend only on the subscripts and are independent of m.

Using (3), (4), and (16), the time domain constraints become the frequency domain constraints

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A_m\left(u + \frac{2n\pi}{T}\right) B\left(u + \frac{2n\pi}{T}\right) C_p\left(u + \frac{2n\pi}{T}\right)$$

$$= \delta_{mp}\, G_p + \sum_{l,k=-\infty}^{\infty} F_{p,\,l-k} e^{-ju(l-k)T} \tag{17a}$$

$$= \delta_{mp}\, G_p + F_p(u) \tag{17b}$$

recognizing that the last term is really a Fourier series expansion. Notice that there is a good bit of freedom in the design because $F_p(u)$ can be chosen arbitrarily. Alternatively, this means that the time domain response samples, $F_{p,\,l-k}$, can be arbitrarily chosen but, these samples must be the same for the response to each transmitter. Thus, because the input data has been restricted, the definition of distortionless transmission can be relaxed.

V. DESIGN CRITERIA APPLICATIONS

Let us apply the general design criteria derived in the previous section to some special cases to illustrate the principles involved. These examples include PAM, PSM, and systems in which the data vector is partly independent multilevel and partly constrained (that is, where

some of the components of the data vector are unconstrained and the rest are forced to be zero). The examples clearly bring out the relationship between transmitting information with amplitude or waveform variation.

5.1 *Pulse Amplitude Modulation Systems*

5.1.1 *Linear Precoding and Decoding*

Pierce[5] has suggested the use of linear precoding and decoding matrices for data systems to improve performance in the presence of impulse noise. Fig. 3 shows a system using this concept. It differs from normal smear-desmear techniques in that there are M channels instead of just one (that is, the input data are block-encoded).

The customer data, now labeled α_{nk}, $n = 1, \ldots, N$, are applied to the precoder at $t = kT$. The transformed data a_{mk} are then applied to the input of the m^{th} transmitter. In terms of the input data, one has

$$\vec{a}_k = \mathbf{P}\vec{\alpha}_k \tag{18}$$

where $\mathbf{P}$ is the N by M ($M \geqq N$) precoder matrix and the input data is

$$\vec{\alpha}_k = \begin{bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \vdots \\ \alpha_{Nk} \end{bmatrix}. \tag{19}$$

Similarly, the output data may be written

$$\vec{\gamma}_k = \mathbf{D}\vec{g}_k \tag{20}$$

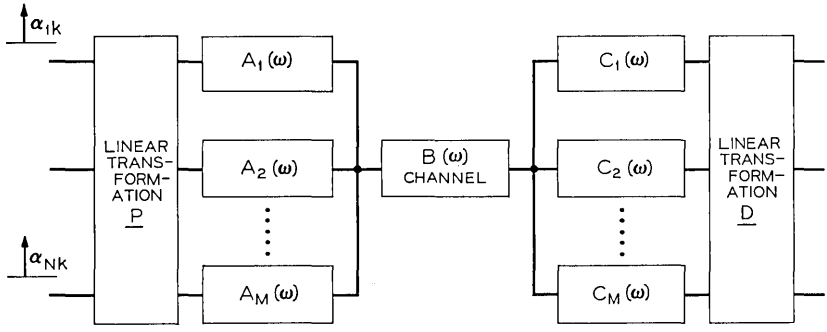


Fig. 3 — Transmission system using linear precoding and decoding transformations.

where **D** is the decoding matrix and

$$\vec{\gamma}_k = \begin{bmatrix} \gamma_{1k} \\ \gamma_{2k} \\ \vdots \\ \gamma_{Nk} \end{bmatrix} \tag{21}$$

and

$$\vec{g}_k = \begin{bmatrix} g_{1k} \\ g_{2k} \\ \vdots \\ g_{Mk} \end{bmatrix}. \tag{22}$$

It may be noted that the transmitted signal may be written

$$\sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} a_{mk} \mathrm{a}_m(t - kT)$$

or in terms of the input signal

$$\sum_{m=1}^{M} \sum_{j=1}^{N} \sum_{k=-\infty}^{\infty} P_{mj} \alpha_{jk} \mathrm{a}_m(t - kT)$$

or

$$\sum_{k=-\infty}^{\infty} \sum_{j=1}^{N} \alpha_{jk} \sum_{m=1}^{M} P_{mj} \mathrm{a}_m(t - kT)$$

or

$$\sum_{k=-\infty}^{\infty} \sum_{j=1}^{N} \alpha_{jk} \mathrm{a}'_j(t - kT)$$

where

$$a'_j(t) = \sum_{m=1}^{M} P_{mj} \mathrm{a}_m(t). \tag{23}$$

Thus, using a linear coder merely corresponds mathematically to using a different set of signal generators with no coder. It might be desirable in some cases to treat the precoder separately,[6] because it could be an easily modified device (that is, one consisting only of gain or delay variables) which could be used to combat noise, change the data rate or shift the spectrum of the signals on the channel.

As an example, consider the two precoding matrices

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

used with time division multiplex transmitters (that is, serial transmission

$$A_m(\omega) = A(\omega) \exp\left[-j\frac{(m-1)\omega T}{M}\right]).$$

Matrix $\mathbf{P}_1$ corresponds to no precoding while matrix $\mathbf{P}_2$ represents interleaving which might be effective in combating burst noise if the input data is redundant (that is, digitally encoded into blocks of length 3, in this case. Notice that interleaving is basically a digital technique for error control in burst noise. This is amply illustrated by the presence of identical values as the single nonzero element in each row and column of the matrix. For analog error control (smearing or spreading the information over several symbols) in burst noise, the elements of $\mathbf{P}$ can be any real values.

The choice of a particular precoding matrix would presumably be based upon some knowledge of the noise characteristics. The decoder can then be designed for distortionless transmission by solving $\mathbf{DP} = \mathbf{I}_{NN}$ if it is assumed that

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u + \frac{2n\pi}{T}\right)B\left(u + \frac{2n\pi}{T}\right)C_p\left(u + \frac{2n\pi}{T}\right) = \delta_{mp}\, K_p\,. \tag{13}$$

Similarly, $\mathbf{D}$ could be obtained by considering the transformed transmitter and receiver and solving

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_i'\left(u + \frac{2n\pi}{T}\right)B\left(u + \frac{2n\pi}{T}\right)C_i'\left(u + \frac{2n\pi}{T}\right) = \delta_{ij}\, K_i'\,. \tag{24}$$

The transformed receiver $C_i'(\omega)$ is just

$$C_i'(\omega) = \sum_{i=1}^{N} D_{ip}C_p(\omega). \tag{25}$$

The two approaches are equivalent.

5.1.2 *A Two-Channel PAM System*

Fig. 4 shows a two-channel PAM system. All of the main features of the design constraints can be easily shown by means of this example. The four equations which must be satisfied are

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C_p\left(u+\frac{2n\pi}{T}\right) = \delta_{mp}K_p$$

$$m, p = 1, 2. \qquad (13)$$

It is apparent that $A_m[u + (2n\pi)/T]$ and $C_p[u + (2n\pi)/T]$ must have nonzero values for at least two values of n (two intervals of π/T bandwidth or two intervals of width $2\pi/T$ when both positive and negative frequencies are considered). Hence, the total bandwidth must be $2\pi/T$ for distortionless transmission. If the bandwidth is greater than $2\pi/T$, an infinite number of designs are possible.

A clearer idea of the implications of equation 13 can be gained by examining the impulse responses in the time domain which are shown in Fig. 5. Notice that $r_{11}(t)$ and $r_{22}(t)$ are the usual pulses required for data transmission. The crosstalk waveforms $r_{12}(t)$ and $r_{21}(t)$ are required to be zero at all $t = kT$ so that the output of either channel at $t = kT$ does not depend upon the input to the other channel. This does not mean, however, that there can be no frequency overlap between the transmitter of one channel and the receiver of the other. It does mean that the characteristics must be chosen so that

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C_p\left(u+\frac{2n\pi}{T}\right) = 0 \qquad m \neq p. \qquad (26)$$

Fig. 1 shows one such characteristic and Fig. 6 gives one possible design for the two-channel system which anticipates the next example.

Notice that if $A_2(\omega)$ were zero (that is, a single channel system), the second receiver could be used for a noise monitor.[7] By taking the

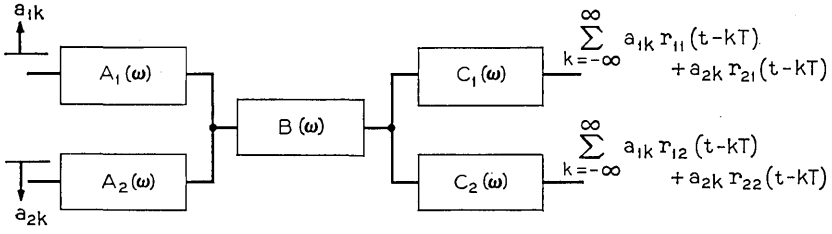


Fig. 4 — Two-channel PAM system.

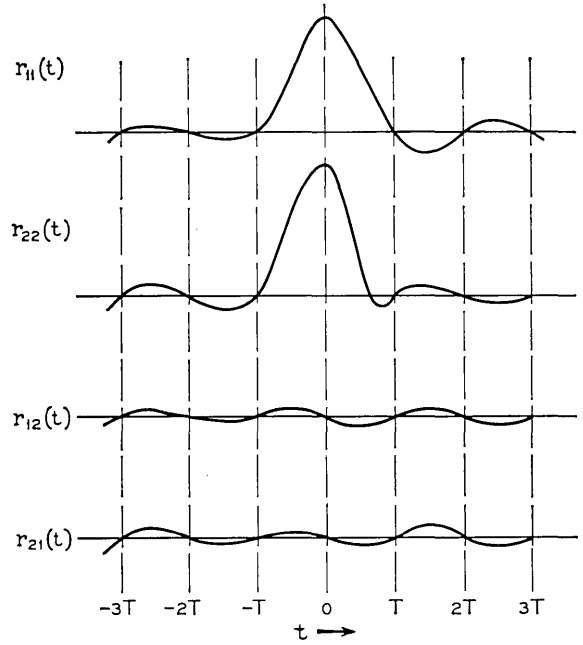


Fig. 5 — Required time domain responses for distortionless transmission.

output of this filter at kT seconds (when the noiseless component is zero), squaring, and averaging, one can get an estimate of the variance of the channel noise.

5.1.3 *Band-Limited Orthogonal Signals for Multichannel Transmission*[8]

Thus far, we have discussed only general design constraints without regard for the specific choices that a designer must make if he has available more than the minimum bandwidth (as he must). In other words, if only the minimum bandwidth were available, the designer would have no choice but to match the M^2 equations with the M^2 variables (a slight choice does arise between serial and parallel formats). However, arbitrary choices can be made when one has more than M^2 variables (bandwidth $> M\pi/T$).

Chang[8] has considered one such possibility; namely, a frequency division multiplex system in which signals at the channel output, $A_m(\omega)B(\omega)$, are orthogonal. In other words, taking $B(\omega) = 1$ for notational simplicity, Chang's signals are chosen to satisfy the time

domain requirement

$$\int_{-\infty}^{\infty} a_m(t)a_p(t - kT)\,dt = \delta_{mp}\,\delta_{k0}\,K_p\,. \tag{27}$$

In the frequency domain this requirement becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} A_m(\omega)A_p^*(\omega)e^{j\omega kT}\,d\omega = \delta_{mp}\,\delta_{k0}\,K_p\,. \tag{28a}$$

Using the technique of equations 1 through 2d, we obtain

$$\frac{1}{2\pi}\int_{-\pi/T}^{\pi/T}\sum_{n=-\infty}^{\infty} A_m\left(u + \frac{2n\pi}{T}\right)A_p^*\left(u + \frac{2n\pi}{T}\right)e^{jukT}\,du = \delta_{mp}\,\delta_{k0}\,K_p \tag{28b}$$

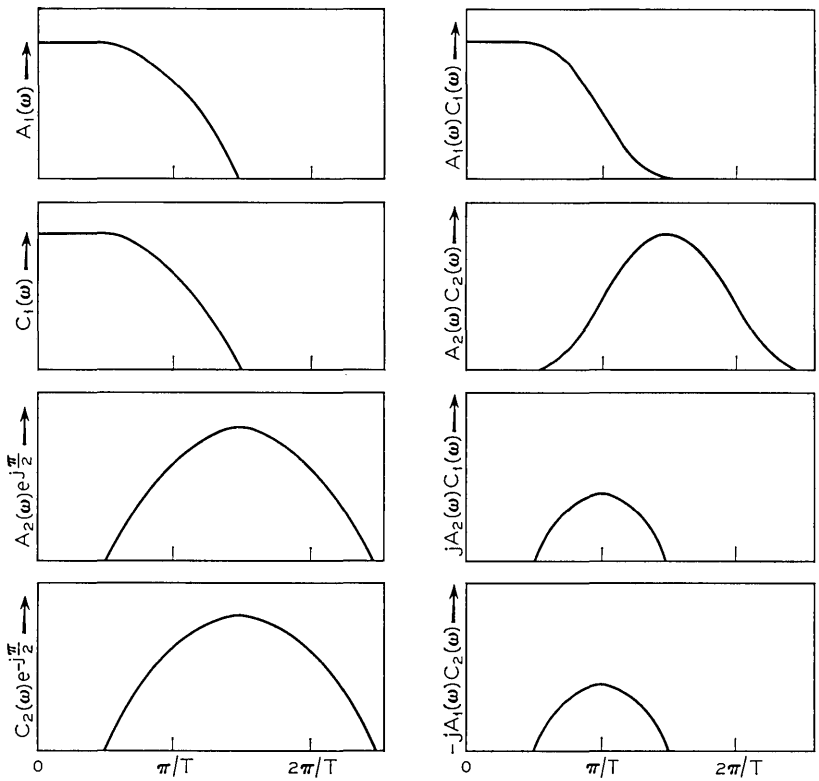


Fig. 6 — A possible two-channel system assuming $B(\omega) = 1$.

or

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u+\frac{2n\pi}{T}\right)A_p^*\left(u+\frac{2n\pi}{T}\right) = \delta_{mp}\,K_p \tag{29a}$$

for the frequency domain representation. This equation is identical to equation 13 if $C_p(\omega) = A_p^*(\omega)$ (which is best in the presence of white noise) assuming $B(\omega) = 1$. [Nonideal $B(\omega)$ can be considered by assuming that $A_m(\omega)$ is the channel output rather than transmitter output]. Thus, it is seen that the requirement of orthogonality is a special case of the general design criteria with the additional constraint that $C_p(\omega) = A_p^*(\omega)$.

In addition to this constraint, Chang chose a frequency division multiplex format with overlapping signal spectra such that

$$|\,A_m(\omega)\,| \neq 0$$

only for

$$\left(m-\frac{3}{2}\right)\frac{\pi}{T} < |\,\omega\,| < \left(m+\frac{1}{2}\right)\frac{\pi}{T}.$$

One can insert these assumptions into equation 29a and arrive at at the design conditions. It is, however, more enlightening to examine the system in the light of the previous discussion of a two-channel system. Fig. 7a shows the spectra of the three transmitters which affect the output of the m^{th} channel under the assumptions outlined above. (For concreteness of the discussion, m is even; odd m would proceed similarly.) No intersymbol interference in the m^{th} channel requires

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u+\frac{2n\pi}{T}\right)A_m^*\left(u+\frac{2n\pi}{T}\right) = K_m\ . \tag{29b}$$

In other words, the characteristic $|A_m(\omega)|^2$ must have vestigial symmetry about $\omega = m\pi/T$ and $(m-1)\pi/T$.

Let us turn now to the crosstalk terms. The equations which must be satisfied for distortionless transmission are

$$\sum_{n=-\infty}^{\infty} A_{m-1}\left(u+\frac{2n\pi}{T}\right)A_m^*\left(u+\frac{2n\pi}{T}\right) = 0 \tag{30a}$$

and

$$\sum_{n=-\infty}^{\infty} A_{m+1}\left(u+\frac{2n\pi}{T}\right)A_m^*\left(u+\frac{2n\pi}{T}\right) = 0. \tag{30b}$$

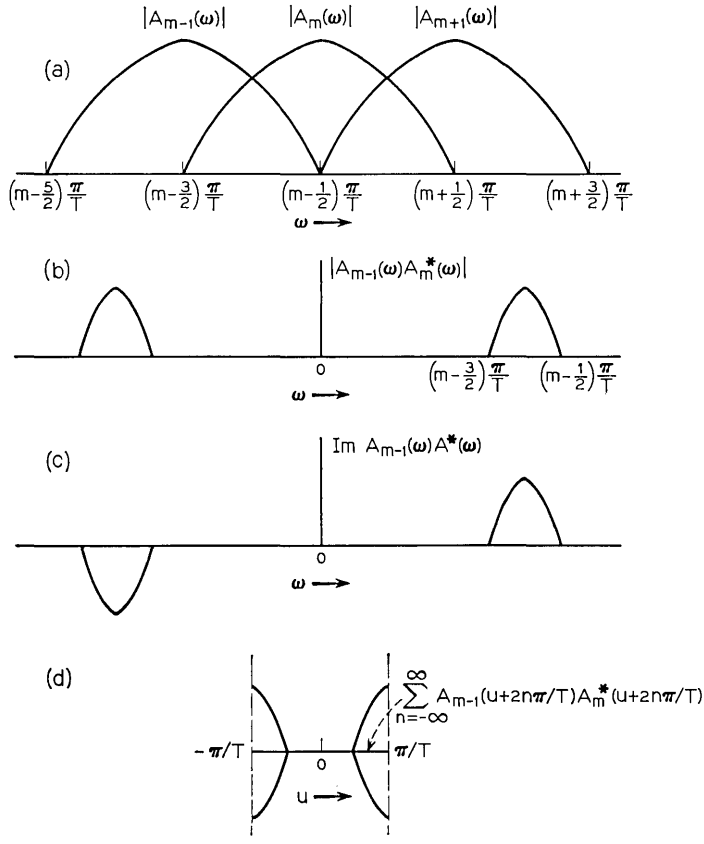


Fig. 7 — (a) Spectra of transmitters which affect m^{th} output. (b) Magnitude of $A_{m-1}(\omega)A_m^*(\omega)$. (c) Required $A_{m-1}(\omega)A_m^*(\omega)$. (d) Demonstration that above $A_{m-1}(\omega)A_m^*(\omega)$ satisfies constraint for periodic zeros.

Fig. 7b shows the magnitude of $A_{m-1}(\omega)A_m^*(\omega)$ which is symmetric about $|\omega| = (m-1)\pi/T$. The only way these components can sum to zero following equation 30a is if they are imaginary as shown in Fig. 7c (with the sum given in Fig. 7d). The same argument applies to the $A_{m+1}(\omega)A_m^*(\omega)$ product and is illustrated in Fig. 8. In other words, $A_m^*(\omega)$ must be ± 90 degrees out of phase with $A_{m+1}(\omega)$ and $A_{m-1}(\omega)$ in the regions of overlap of the functions.

It is seen that the amplitude characteristic design is based upon the condition of no intersymbol interference in each channel and is based upon the usual Nyquist design. The remaining freedom in choosing

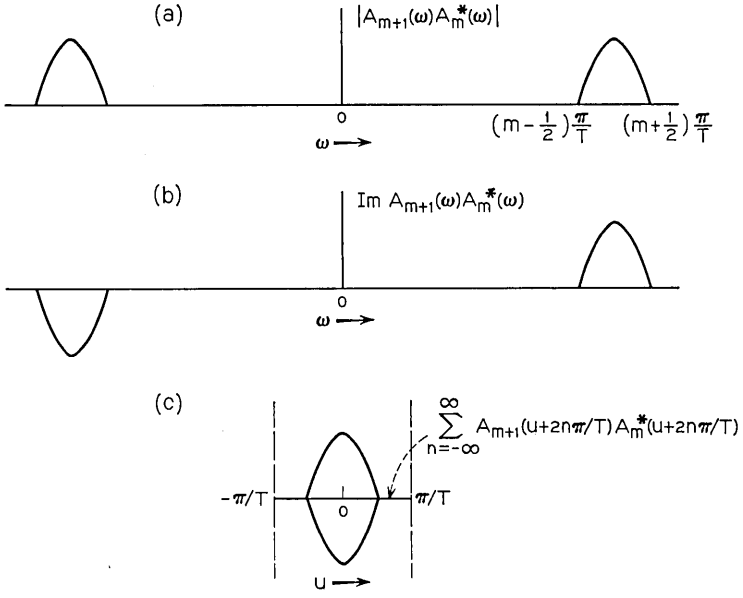


Fig. 8 — (a) Magnitude of, and (b) required $A_{m+1}(\omega)A_m^*(\omega)$. (c) Demonstration that above $A_{m+1}(\omega)A_m^*(\omega)$ satisfies constraint for periodic zeros.

the phase characteristic is then used to eliminate interchannel interference with the requirement being

$$\text{phase of } A_m(\omega) = \text{phase of } A_{m-1}(\omega) \pm 90°. \tag{31}$$

5.1.4 *Noise Monitoring*[7]

The noise monitoring feature mentioned previously can be generalized to the M channel case. The minimum bandwidth of $M\pi/T$ must be exceeded by the practical system. The bandwidth redundancy can be used for noise monitoring by adding another filter $C_{M+1}(\omega)$ at the receiver. This receiver must satisfy the equations

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_m\left(u+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C_{M+1}\left(u+\frac{2n\pi}{T}\right) = 0$$

$$\text{for } 1 \leqq m \leqq M \tag{32}$$

and a nontrivial solution can result because of the bandwidth redundancy. Then, the impulse responses $r_{mM+1}(t)$ go through zero at all $t = kT$ and the noiseless output of the $M+1^{\text{th}}$ filter

$$\sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} a_{mk} r_{mM+1}(t - kT)$$

is zero periodically, independent of the input data.

The filter output at $t = lT$ can be squared and averaged to obtain an estimate of the noise level and hence an estimate of the transmission performance. If the shape of the noise power spectrum is known, one gets a quantitative estimate of the noise power. Timing errors or poor knowledge of $B(\omega)$ can lead to the noiseless output of the $M+1^{\text{th}}$ filter being nonzero at the sample time. The indicated noise variance would be greater than the correct value, thus indicating poorer performance than the noise alone. However, timing or channel characterization errors actually do lead to poor system performance so that the monitor indication is in the right direction. Notice that this monitoring scheme is not tied to any particular choice of transmitter or receiver and is perfectly general.

5.2 *Pulse Shape Modulation Systems*

5.2.1 *A Binary PSM System*

Insight into pulse shape modulation system design constraints can, perhaps, best be gained by examining a binary system such as that shown in Fig. 9a. The equations that must be satisfied are

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A_m\left(u + \frac{2n\pi}{T}\right) B\left(u + \frac{2n\pi}{T}\right) C_p\left(u + \frac{2n\pi}{T}\right) = \delta_{mp} G_p + F_p(u) \tag{17b}$$

for $m, p = 1, 2$. Because it is a binary system, the receiver can be just a single filter

$$C(\omega) = C_2(\omega) - C_1(\omega) \tag{33}$$

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A_1\left(u + \frac{2n\pi}{T}\right) B\left(u + \frac{2n\pi}{T}\right) C\left(u + \frac{2n\pi}{T}\right) = -G + F(u) \tag{34a}$$

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A_2\left(u + \frac{2n\pi}{T}\right) B\left(u + \frac{2n\pi}{T}\right) C\left(u + \frac{2n\pi}{T}\right) = G + F(u) \tag{34b}$$

where it has been assumed without loss of generality that

$$G = G_1 = G_2$$

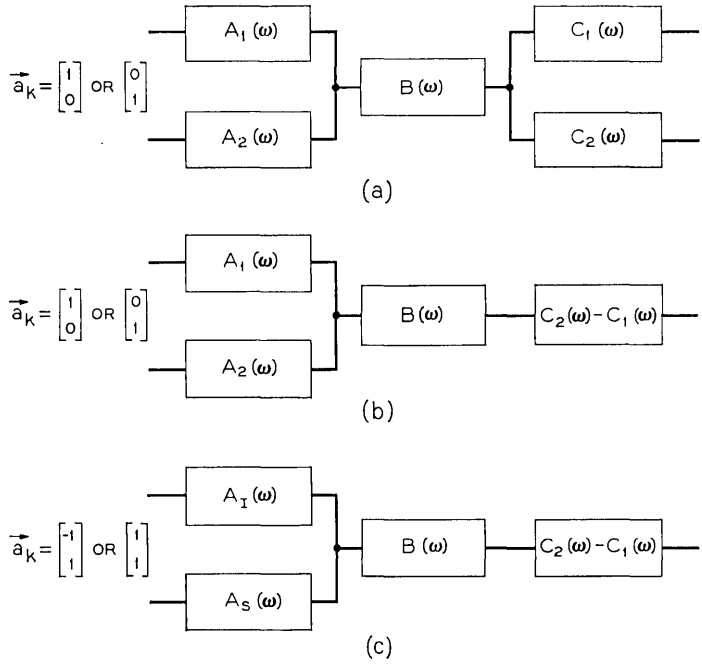


Fig. 9 — (a) Binary PSM system. (b) Modified binary PSM system. (c) Equivalent binary PSM system or PAM system with data constraints.

and where $F(u)$ is an arbitrary function of frequency

$$F(u) = F_2(u) - F_1(u) = \sum_{q=-\infty}^{\infty} f_q e^{-iuqT}. \tag{35}$$

This modified system described above is shown in Fig. 9b.

Fig. 10 shows two possible time domain responses which satisfy equations 34a and b. Notice that the responses differ only at $t = 0$ and are identical to all other $t = kT$. This is the time domain implication of equations 34a and b. This corresponds to the case where two signals are chosen to produce the same intersymbol interference which was discussed briefly by Simon and Kurz.[9]

An alternate way of viewing equations 34a and b is to notice that the two transmitters can each be decomposed into two components. The information component $A_I(\omega)$ of each satisfies

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_I\left(u + \frac{2n\pi}{T}\right)B\left(u + \frac{2n\pi}{T}\right)C\left(u + \frac{2n\pi}{T}\right) = G \tag{36}$$

or the usual Nyquist criterion, and is transmitted with an amplitude of ± 1. The steady component $A_s(\omega)$ satisfies

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_s\left(u+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C\left(u+\frac{2n\pi}{T}\right) = F(u) \qquad (37)$$

and is sent with an amplitude of one regardless of the data stream. The transmitted waveform $A_s(\omega)$ can be anything because $F(u)$ is arbitrary. Fig. 9c shows this system, which is equivalent to the original. The corresponding data vectors are

$$\vec{a}_k = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \qquad (38)$$

The basic equivalence of the PAM and PSM systems is thus made explicit. The difference in the two systems is basically a noninformation bearing signal which represents an inefficient (theoretically) use of power. This point has been brought out by Bennett and Davey[10] in discussing the Sunde[11] model of a synchronous FM system.

5.2.2 *Sunde's FM Model With a Linear Receiver*

In Sunde's[11] model of a synchronous FM system, one of two phase continuous signals

$$a_1(t) = \cos\frac{2\pi q}{T}t + \theta \qquad -\frac{T}{2} < t < \frac{T}{2}$$

$$a_2(t) = -\cos\frac{2\pi(q+1)t}{T} + \theta$$

$$a_1(t) = a_2(t) = 0 \qquad \text{elsewhere}$$

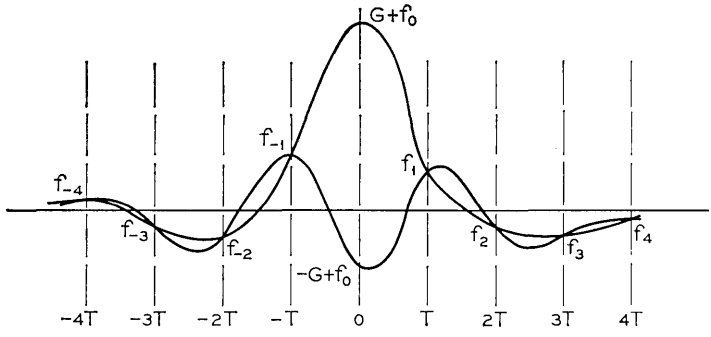


Fig. 10 — Possible time domain responses for distortionless PSM system.

is sent during each interval. The transmitter output may be written

$$\sum_{k=-\infty}^{\infty} a_{1k} \cos\left(\omega_c t - \frac{\pi}{T} t + \theta\right) - (1 - a_{1k}) \cos\left(\omega_c t + \frac{\pi}{T} + \theta\right)$$

or

$$\sin \frac{\pi}{T} t \sin (\omega_c t + \theta) + \sum_{k=-\infty}^{\infty} (2a_{1k} - 1) \cos \frac{\pi}{T} t \cos (\omega_c t + \theta)$$

where

$$\omega_c = \frac{(2q + 1)\pi}{T}.$$

This second form of the output is an explicit example of an information-bearing component (second term where $a_{1k} = 1, 0$) and a steady state component.

To achieve distortionless transmission (with a linear receiver) one must choose any linear filter which satisfies

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A_I\left(u + \frac{2n\pi}{T}\right) B\left(u + \frac{2n\pi}{T}\right) C\left(u + \frac{2n\pi}{T}\right) = G \tag{39}$$

where

$$A_I(\omega) = \int_{-T/2}^{T/2} \cos \frac{\pi}{T} t \cos (\omega_c t + \theta) e^{-i\omega t}\, dt \tag{40a}$$

$$= e^{i\theta} S(\omega - \omega_c) + e^{-i\theta} S(\omega + \omega_c) \tag{40b}$$

where

$$2S(\omega) = \frac{\sin (\omega - \pi/T)T/2}{\omega - \pi/T} + \frac{\sin (\omega + \pi/T)T/2}{\omega + \pi/T}. \tag{40c}$$

If one makes the assumption that $S(\omega+\omega_c)$ is negligible at positive frequencies, then

$$A_I(\omega) = \cos \frac{(\omega - \omega_c)T}{2} \left[\frac{-\pi/T}{(\omega - \omega_c)^2 - \pi^2/T^2}\right] e^{i\theta} \qquad \omega > 0. \tag{40d}$$

By substituting equation 40d into 39 the requirements on $B(\omega)C(\omega)$ can be found. The minimum bandwidth solutions are (neglecting constants)

$$B(\omega)C(\omega) = \frac{(\omega - \omega_c)^2 - \pi^2/T^2}{\cos (\omega - \omega_c)T/2} e^{-i\theta}$$

$$\omega_c + n\pi/T < \omega < \omega_c + (n + 1)\pi/T \quad n = -1, 0 \tag{41}$$

$$= 0 \qquad \text{elsewhere}$$

which are shown in Fig. 11a. Solutions in other regions (other values of n) are possible but require infinite gain.

Sunde's solution for the minimum bandwidth filtering before a phase derivative (nonlinear) detector is given in Fig. 11b for comparison. Notice that the linear receiver requires only half the bandwidth required by the phase derivative detector for distortionless transmission. The response of the linear filter to the steady state term is unimportant because it is deterministic and can be removed.

5.2.3 *Monitoring System Parameters*

If a second filter $C_g(\omega)$ is added to the system of Fig. 9c, one again can monitor some aspect of system performance if the equation

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_I\left(u+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C_g\left(u+\frac{2n\pi}{T}\right)=0 \tag{42}$$

is satisfied. Thus, the output at the sample times will be independent of the input data. However, there will be a constant output value (excluding noise) of

$$\sum_{q=-\infty}^{\infty} f_{gq}$$

where

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A_s\left(n+\frac{2n\pi}{T}\right)B\left(u+\frac{2n\pi}{T}\right)C_g\left(u+\frac{2n\pi}{T}\right)=F_g(u) \tag{43a}$$

$$=\sum_{q=-\infty}^{\infty} f_{gq}e^{-juqT} \tag{43b}$$

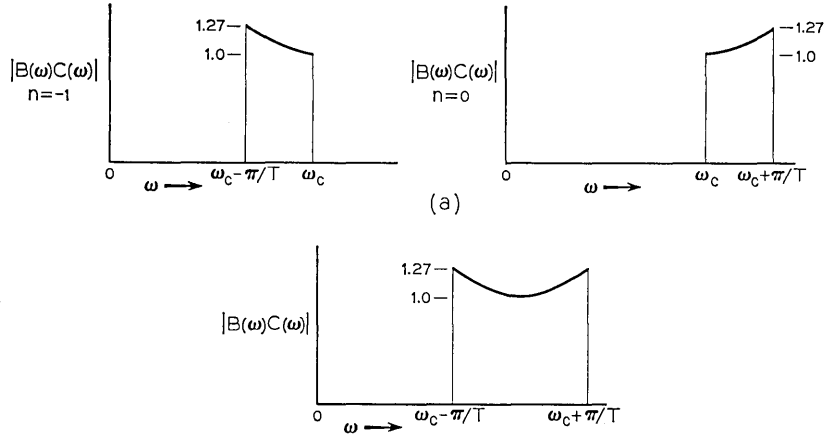


Fig. 11 — Minimum bandwidth filtering for FM system with (a) linear receiver and (b) phase derivative receiver.

because of the steady state transmitter. The total output of this filter is then a (generally) nonzero constant which depends upon the steady state transmitter and channel characteristics and the noise. It does not depend upon the data sequence or the receiver's estimate of the sequence. A change in this constant reflects a change in the transmitter or channel parameters (for example, phase or gain) and can be used to modify the receiver characteristics (such as, phase or threshold level). Thus, the noninformation part of the transmitted signal, in addition to perhaps simplifying implementation, also can be used to provide needed information to the receiver. A simple example is the reference tone for carrier recovery which in fact makes the PAM system into a PSM system.

5.3 *Pulse Amplitude Modulation System With Zero Constraints on Certain Channels*

In section 5.2 we showed that PSM could be considered as PAM with constraints on the input to certain channels. In other words, the equivalent PSM system shown in Fig. 9c contained one channel whose input was constrained to be a one at all times. Now we will discuss a system in which certain channel inputs are constrained to be zero.

Consider a four-channel PAM system and assume it to be serial; that is,

$$A_m(\omega) = A(\omega) \exp\left[-j\frac{(m-1)\omega T}{4}\right] \tag{9a}$$

$$C_m(\omega) = C(\omega) \exp\left[j\frac{(m-1)\omega T}{4}\right]. \tag{9b}$$

If the input data vector is given by

$$\vec{a}_k' = \begin{bmatrix} a_{1k} \\ a_{2k} \\ a_{3k} \\ a_{4k} \end{bmatrix} \tag{44}$$

then a bandwidth of $4\pi/T$ is required. Now, under certain circumstances it might be desirable to reduce this data rate by inserting zeroes for some of the a_{mk} (that is, not transmitting anything at certain times).

Chang[6] has considered this possibility for improving performance in the presence of severe noise. In this case, some a_{mk} can be made zero and the remaining data can be transmitted with increased power (maintaining a fixed average power) to improve the noise margin.

Another purpose of this zero stuffing technique might be to shift (as well as reduce the bandwidth of) the spectrum of the transmitted signals. As a trivial example, the data vector

$$\vec{a}_k = \begin{bmatrix} a_{1k} \\ 0 \\ a_{3k} \\ 0 \end{bmatrix} \tag{45}$$

could be transmitted with a flat spectrum over either the region

$$0 < |\omega| < 2\pi/T \quad \text{or} \quad 2\pi/T < |\omega| < 4\pi/T$$

and zero elsewhere. As a nontrivial example, consider a generalization of a signaling system, invented by Bennett and Feldman,[12] to prevent intersymbol and interchannel interference in multiplex transmission. The original system has been described very briefly by Sunde.[13] Here, the generalized system can be approached by writing the data vector

$$\vec{a}_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ 0 \\ 0 \end{bmatrix} \tag{46}$$

where only the outputs of the first two receivers must be examined. With the assumption (for simplicity) that $B(\omega) = 1$ the constraining equations become

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A\left(u + \frac{2n\pi}{T}\right) C\left(u + \frac{2n\pi}{T}\right) = r_0 \tag{47}$$

and

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} A\left(u + \frac{2n\pi}{T}\right) C\left(u + \frac{2n\pi}{T}\right) \exp \pm j\left(u + \frac{2n\pi}{T}\right) \frac{T}{4} = 0. \tag{48a}$$

Recognizing that exp $(\pm ju\ T/4)$ is a nonzero term which can be removed, 48a becomes

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} A\left(u+\frac{2n\pi}{T}\right)C\left(u+\frac{2n\pi}{T}\right)\exp \pm j\frac{n\pi}{2}=0. \qquad (48b)$$

Fig. 12a illustrates the type of characteristic $A(\omega)C(\omega)$ which satisfies the constraints. (There are others of larger bandwidth which also will satisfy the constraints.) If the characteristic is limited to $|\omega| < 4\pi/T$ and zero elsewhere, it has symmetry about $\omega = 2\pi/T$ and vestigial symmetry about $|\omega| = \pi/T$ and $3\pi/T$. It can easily be verified that the equations are satisfied when one uses the value of the multiplying factor exp $\pm jn\pi/2$ which is shown in each region.

Notice that a response which is flat from $\pi/T < |\omega| < 3\pi/T$ and zero elsewhere satisfies the equations and represents the minimum bandwidth approach to this scheme. The time response one obtains at the receiver using this technique is illustrated in Fig. 12b. It is constrained to be zero at all $t = kT$ except $k = 0$ and $\pm(4q-2)$ for all q and can be used for transmission, as explained, without distortion.

The advantage of this particular scheme is that it represents a baseband technique for shifting the transmission spectrum without modulation merely by inserting zeros into the data stream. It appears particularly attractive for placement within a voice channel (for example, 200Hz-3KHz) as Fig. 13 shows. Here no modulation has been required, the energy is concentrated in the center of the band and

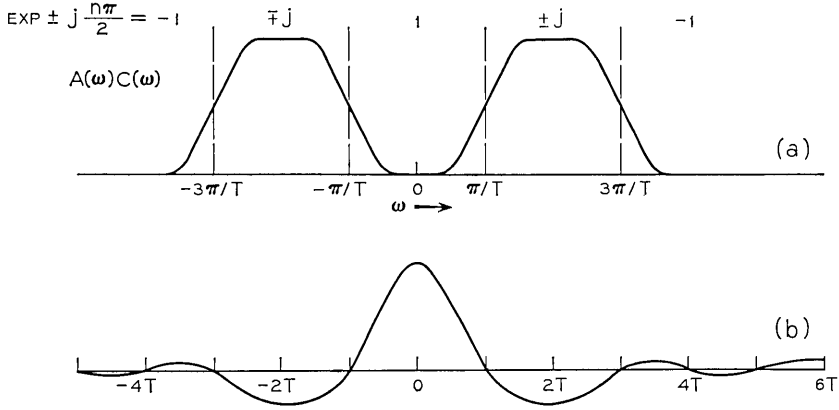


Fig. 12 — (a) Frequency characteristic, and (b) time response, for distortionless transmission with zero stuffing.

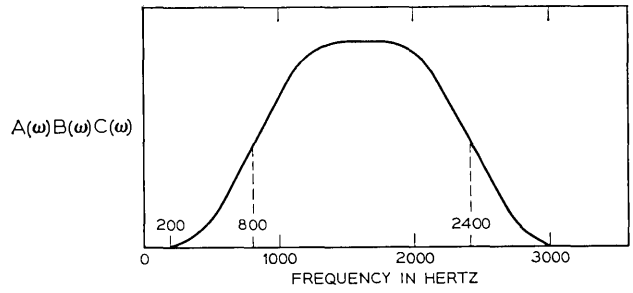


Fig. 13 — Zero stuffing spectrum for voiceband transmission.

one could obtain symbol rates of 3200 symbols per second with easily realized filtering. The primary disadvantage would be increased sensitivity to timing errors.

VI. CONCLUDING COMMENTS

The thesis of this paper has been that all linear data system designs are based on the modified Poisson sum formula

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} R\left(u+\frac{2n\pi}{T}\right)=\sum_{q=-\infty}^{\infty} r_q e^{-iuqT} \tag{3}$$

which relates the time domain samples to the frequency domain constraints. Various types of systems which a designer may choose require a variety of constraints on the time samples r_q. These values of r_q, which depend upon the type of system chosen, then specify the frequency domain requirements.

Section 5 gave a sampling of the range of systems which can be designed using equation 3. That section certainly does not exhaust the possibilities and we hope that it does not limit the reader's imagination. Most of the examples (as well as most real systems) assume systematic choice of transmitted signals; usually related by integral time or frequency shift. There may, however, be potential gain in considering nonsystematic transmitters and receivers. This may easily be done using equation 3. The last case examined, that of spectrum shifting by adding zeros, is just such a nonsystematic function when viewed from a serial transmission point of view.

REFERENCES

1. Nyquist, H., "Certain Topics in Telegraph Transmission Theory," AIEE Transactions, *47*, (April 1928), pp. 617–644.

2. Gibby, R. A., and Smith, J. W., "Some Extensions of Nyquist's Telegraph Transmission Theory," B.S.T.J., *44,* No. 7, (September 1965), pp. 1487–1510.
3. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics,* New York: McGraw-Hill Book Co., Inc., 1953.
4. Shnidman, D. A., "A Generalized Nyquist Criterion and an Optimum Linear Receiver for a Pulse Modulation System," B.S.T.J., *46,* No. 8 (November 1967), pp. 1773–1796.
5. Pierce, W. H., "Linear Real Coding," 1966 IEEE International Convention Record, Part VII, pp. 44–53.
6. Chang, R. W., "Precoding for Multiple-Speed Data Transmission," B.S.T.J., *46,* No. 7 (September 1967), pp. 1633–1649.
7. Smith, J. W., unpublished work.
8. Chang, R. W., "Synthesis of Band-Limited Orthogonal Signals for Multichannel Data Transmission," B.S.T.J., *45,* No. 10 (December 1966), pp. 1775–1796.
9. Simon, M. K., and Kurz, L., "A Method of Signal Design for Gaussian Noise and Intersymbol Interference Immunity Using Maximum Likelihood Detection," Symposium on Signal Transmission and Processing, New York: Columbia University, May 14, 1965, pp. 62–68.
10. Bennett, W. R., and Davey, J. R., *Data Transmission,* New York: McGraw-Hill Book Co., Inc., 1965.
11. Sunde, E. D., "Ideal Binary Pulse Transmission by AM and FM," B.S.T.J., *38,* No. 6 (November 1959), pp. 1357–1425.
12. Bennett, W. R., and Feldman, C. B. H., "Prevention of Interpulse Interference in Pulse Multiplex Transmission," U. S. Patent 2,719,189, applied for May 1, 1951; issued September 27, 1955.
13. Sunde, E. D., "Theoretical Fundamentals of Pulse Transmission I," B.S.T.J., *33,* No. 3 (May 1954), pp. 721–788.

Click Comparison of Digital and Matched Filter Receivers

By H. L. SCHNEIDER

(Manuscript received May 25, 1967)

We applied the click theory of errors to determine the performance of a digital FM receiver. The receiver had binary orthogonal FSK modulation in a channel that had a single random-phase echo at the symbol duration. We use practical bandwidth assumptions to show that this error performance is identical to that calculated for a matched filter receiver. Numerical results show, for example, that an increase in signal-to-noise ratio of 10 dB is required to maintain a 10^{-4} error rate when an echo of half the signal power is added.

I. INTRODUCTION

The concept of clicks in an FM receiver was originally used by S. O. Rice[1] and J. Cohn[2] to explain the effect of noise on analog signal demodulation near threshold. Recently, several theoretical investigations of digital FSK signal demodulation have applied the concept of clicks in analyses of low pass filter processing of the discriminator output. Klapper,[3] and Mazo and Salz[4] modelled the low pass filter with an integrate-and-dump function, while Schilling, Hoffman, and Nelson considered a gaussian low pass filter.[5] In all cases, additive gaussian noise was assumed to be the sole source of interference in the signal channel.

In this paper, we consider intersymbol interference that is induced by delay dispersion in the signal channel. Analysis is limited to a practical single-echo channel* and binary orthogonal modulation. Although the analysis seems to be tractable for only special cases, we gain insight into the error mechanism of digital FM reception.

The relation between clicks and errors is viewed as follows. Since the fundamental description of clicks concerns a random angular encircle-

* The single-echo channel was used by Bennett, Curtis, and Rice[6] in their study of analog angle modulated transmission systems.

ment of the origin by the received signal plus noise vector, it is convenient to choose the integrate and dump model for the post-discriminator filter. Then the filter output is a measure of the angular change of the received vector over a symbol duration. Signal and click angular changes are readily compared. The particular signal angular modulation considered here is $\pm\pi$ radians; in this case an error occurs if and only if a click occurs, to a good approximation.† Intersymbol interference is considered as a perturbation of the signal modulation. This distortion affects the instantaneous signal-to-noise ratio and the instantaneous frequency which, as shown by Rice, are the controlling parameters for the click (error) probabilities.

After the calculations described above are used to compute the error rate for the digital FM receiver, another computation for error rate is made using a noncoherent orthogonal matched filter receiver. The error performances of the two receivers are the same for this binary signal having angular modulation $\pm\pi$ radians.

In the following sections, the modulation and the channel parameters are first defined. An expression is derived for the distorted output of the channel. The derivations of the FM receiver performance and the matched filter receiver performance are explained, then the significance of the work is discussed. Two appendices give detailed derivations of the receivers' performances.

II. FSK MODULATION IN THE SINGLE ECHO CHANNEL

Since the two receivers are applied in turn to the same channel as shown in Figure 1, we shall first express the output of this single echo channel. The input waveform is either $s_1(t)$ or $s_2(t)$.

$$\begin{aligned} s_1(t) &= \text{Re}\ \{e^{j2\pi(f_c+f_d)t}\} \qquad 0 \leqq t \leqq T \\ s_2(t) &= \text{Re}\ \{e^{j2\pi(f_c-f_d)t}\} \end{aligned} \tag{1}$$

where $s_1(t) = s_2(t) = 0$ otherwise

f_c is a center frequency
f_d is the frequency deviation
T is the symbol duration.

(We shall consider only the deviation: $2f_dT = 1$.) These input waveforms are applied in some arbitrary sequence to the channel; the

† Mazo and Salz[4] have considered the approximations involved in some detail, and their work relates different angular modulations.

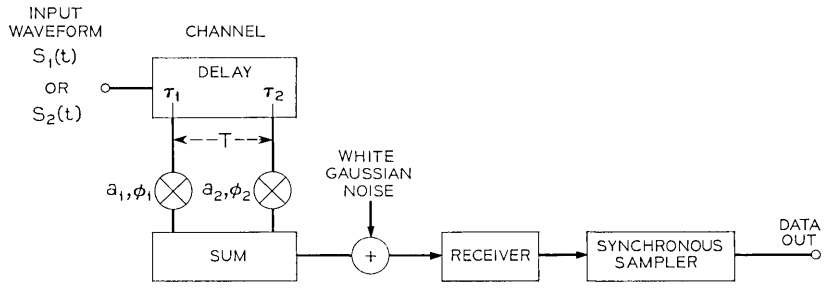


Fig. 1 — Data transmission system model.

channel input voltage can be described as

$$e_{in}(t) = \sum_{n=-\infty}^{\infty} s_{i_n}(t - nT) \qquad i_n = 1, 2. \tag{2}$$

The output of the channel (receiver input) with noise added is

$$e(t) = \text{Re}\ \{a_1 e^{i\varphi_1} e_{in}(t - \tau_1) + a_2 e^{i\varphi_2} e_{in}(t - \tau_2)\} + e_N(t) \tag{3}$$

where

$a_i e^{i\varphi_i}$ $(i = 1, 2)$ are the complex tap gains and we shall consider that $a_1 > a_2$

τ_i are the tap delays and we shall consider the case $\tau_2 - \tau_1 = T$

$e_N(t)$ is additive white Gaussian noise.

In the ensuing work we shall refer to the first term in the braces of equation (3) as the signal, and to the second term as the echo. Since the echo is displaced by one symbol duration, the receiver input in simply a superposition of signal and echo s_i, s_j $(i, j = 1, 2)$. Because each combination is assumed to be equally probable, and because corresponding conditional error probabilities are equal, it is sufficient to evaluate the cases s_1, s_1 and s_1, s_2. Thus the receiver input is either

$$e(t) = \text{Re}\ \{[a_1 e^{i(2\pi f_d t + \varphi_1)} + a_2 e^{i(2\pi f_d t + \varphi_2)}] e^{i2\pi f_c t}\} + e_N(t)$$
$$0 \leqq t \leqq T \tag{4}$$

or

$$e(t) = \{[a_1 e^{i(2\pi f_d t + \varphi_1)} + a_2 e^{i(-2\pi f_d t + \varphi_2)}] e^{i2\pi f_c t}\} + e_N(t). \tag{5}$$

III. FM RECEIVER PERFORMANCE

The FM receiver model used here includes a predetection filter, limiter, discriminator, and a postdetection integrate and dump circuit,

as shown in Figure 2. The predetection filter is used to reduce noise and is supposed to have negligible effect on the modulation. Then the receiver output is proportional to the angular change of the input modulation over a symbol duration.

We proceed by first rewriting the input voltage, represented by equation (4) or equation (5), in a form that shows explicitly the amplitude and angular variations of the noise-free envelope in the form

$$e(t) = \text{Re}\ \{A(t)e^{j\varphi(t)}e^{j2\pi f_c t}\}.$$

The envelope which describes the signal-echo pair s_1, s_1, corresponding to equation (4), is

$$\begin{aligned} A &= [a_1^2 + a_2^2 + 2a_1a_2 \cos(\varphi_1 - \varphi_2)]^{\frac{1}{2}} \\ \varphi(t) &= 2\pi f_d t + \varphi_0 \end{aligned} \tag{6}$$

where φ_0 is a constant. The envelope which describes the signal-echo pair s_1, s_2, corresponding to equation (5), is

$$\begin{aligned} A(t) &= [a_1^2 + a_2^2 + 2a_1a_2 \cos(4\pi f_d t + \varphi_1 - \varphi_2)]^{\frac{1}{2}} \\ \varphi(t) &= 2\pi f_d t + \varphi_1 - \tan^{-1}\left[\frac{a_2 \sin(4\pi f_d t + \varphi_1 - \varphi_2)}{a_1 + a_2 \cos(4\pi f_d t + \varphi_1 - \varphi_2)}\right]. \end{aligned} \tag{7}$$

(In these equations, $A(t)$ and $\varphi(t)$ have been obtained by straightforward trigonometric relations from equations (4) and (5).)

Thus in the absence of noise for signal and echo pairs s_1, s_1 or s_1, s_2, the receiver output is proportional to

$$\Delta\varphi = \varphi(T) - \varphi(0) = 2\pi f_d T = +\pi \qquad (a_2 < a_1),$$

Similarly, complementary signal-echo pair s_2, s_2 or s_2, s_1 would give an output $\Delta\varphi = -\pi$.

The noise perturbation is considered an additive error angle $\theta(t)$, illustrated in Fig. 3. Now the FM receiver output is proportional to $\psi(t)$

$$\psi(t) = \varphi(t) + \theta(t)$$

Fig. 2 — Digital FM receiver.

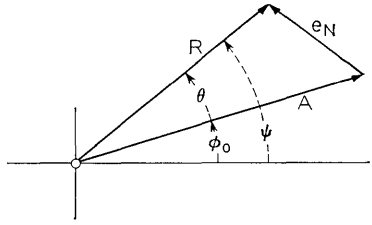


Fig. 3 — Angular perturbation caused by noise.

and the angular change over a symbol duration is

$$\Delta\psi = \Delta\varphi + \Delta\theta.$$

The decision threshold is placed at $\Delta\psi = 0$, midway between $\pm\pi$. When the transmitted signal has an angular modulation $\Delta\varphi = +\pi$, an error is made if $\Delta\theta < -\pi$.

Fig. 4 illustrates possible loci of the signal plus noise envelope R. For signal alone, the locus is simply a semicircle. With echo and noise added, no error is made provided $\Delta\psi > 0$. We observe that the locus encircles the origin in a counterclockwise direction. But when a negative click occurs, the locus encircles the origin in a clockwise direction, $\Delta\psi < 0$, and an error is made.

The probability of error is obtained from the probability of a negative click during a symbol interval. Rice[4] defines $H_-\, dt$ as the proba-

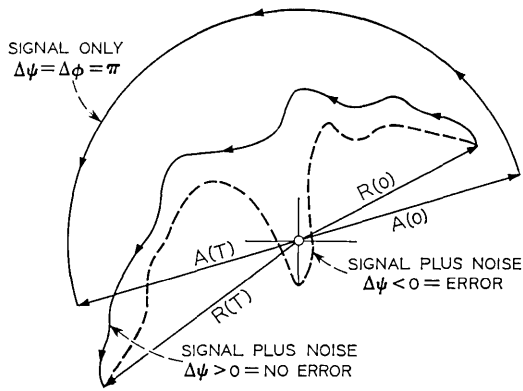


Fig. 4 — Possible loci of angular change.

bility of a negative click: the angle $\theta(t)$ decreases* through an odd multiple of π between t and $t + dt$. $H_- \, dt$ is a function of signal-to-noise ratio and the time derivative of the angular modulation, which are time- and phase-dependent according to equations (6) or (7). The desired probability of error is obtained by integrating H_- over a symbol duration and averaging over the random channel phase angle:

$$P = \frac{1}{2\pi} \int_0^{2\pi} dx \int_0^T H_-(\rho, \dot{\varphi}) \, dt \tag{8}$$

where

$\rho = \dfrac{A^2(t)}{2\overline{e_N^2}}$ is the instantaneous signal-to-noise ratio

$\overline{e_N^2}$ is the noise power passed by the predetection filter

$\dot{\varphi}$ is the time derivative of the modulation angle $\varphi(t)$

$x = \varphi_1 - \varphi_2$ is the relative echo phase, assumed uniformly distributed over $(0, 2\pi)$.

As Appendix A shows, the error probability obtained when $A(t)$ and $\varphi(t)$ from equation (6) are substituted in equation (8) is

$$P_{e1} = \frac{1}{2} I_0\left(\frac{a_1 a_2}{\overline{e_N^2}}\right) \exp\left(-\frac{a_1^2 + a_2^2}{2\overline{e_N^2}}\right). \tag{9}$$

The error probability corresponding to equation (7) is

$$P_{e2} = Q\left[\frac{a_2}{(\overline{e_N^2})^{\frac{1}{2}}}, \frac{a_1}{(\overline{e_N^2})^{\frac{1}{2}}}\right] - \frac{1}{2} I_0\left(\frac{a_1 a_2}{\overline{e_N^2}}\right) \exp\left(-\frac{a_1^2 + a_2^2}{2\overline{e_N^2}}\right), \tag{10}$$

where $Q[\cdot,\cdot]$ is the Marcum Q function. The average of P_{e1} and P_{e2} is simply

$$P_e = \tfrac{1}{2}(P_{e1} + P_{e2}) = \frac{1}{2} Q\left[\frac{a_2}{(\overline{e_N^2})^{\frac{1}{2}}}, \frac{a_1}{(\overline{e_N^2})^{\frac{1}{2}}}\right]. \tag{11}$$

The noise power $\overline{e_N^2}$ depends on the predetection filter bandwidth, which can be estimated using Carson's rule with the Nyquist criterion for video bandwidth. These assumptions give a bandwidth B

$$B = \frac{1}{T}(1 + 2f_d T) = \frac{2}{T} Hz \tag{12}$$

*Decrease means in a direction opposite the time derivative of the modulation $\varphi(t)$. It is possible that $\theta(t)$ can also increase by π and thus cancel the decrease; the probability of this occurrence is asymptotically negligible for low error rates.

and thus a noise power

$$\overline{e_N^2} = BN \text{ watts} \tag{13}$$

where N is the noise density in watts/Hz. Substitution of equations (12) and (13) into equation (11) gives

$$P_e = \frac{1}{2} Q\left[a\left(\frac{E}{N}\right)^{\frac{1}{2}}, \left(\frac{E}{N}\right)^{\frac{1}{2}}\right] \tag{14}$$

where

$$a = \frac{a_2}{a_1} \text{ is the echo/signal voltage ratio}$$

$$E = \tfrac{1}{2}a_1^2 T \text{ is the signal energy/bit.}$$

IV. MATCHED FILTER RECEIVER PERFORMANCE

We are concerned here with the incoherent matched filter receiver shown in Fig. 5. The mark and space filters are matched (except for phase) to the waveforms $s_1(t)$ and $s_2(t)$ defined by equation (1). As Fig. 5 indicates, the combined operations of filtering, square law rectifying, and time sampling produce R_1^2 and R_2^2 which are the squared envelopes of the filter outputs at the end of the symbol interval. Assuming mark is transmitted, the probability of error is

$$P = \text{Prob}\,\{R_2^2 > R_1^2\} = \int_0^\infty dR_1 \int_{R_1}^\infty p(R_1, R_2)\, dR_2 \tag{15}$$

where $p(R_1, R_2)$ is the joint density function of R_1 and R_2.

As shown in Appendix B, the error probability corresponding to equation (4), with the signal-echo pair s_1, s_1 is

$$P_{e1} = \frac{1}{2} I_0\left(\frac{aE}{N}\right) \exp\left(-\frac{E + a^2E}{2N}\right). \tag{16}$$

The error probability corresponding to equation (5), with the signal-echo pair s_1, s_2 is

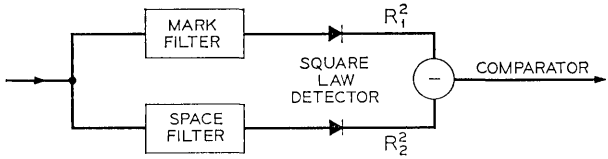


Fig. 5 — Matched filter receiver.

$$P_{e2} = Q\left[a\left(\frac{E}{N}\right)^{\frac{1}{2}}, \left(\frac{E}{N}\right)^{\frac{1}{2}}\right] - \frac{1}{2} I_0\left(\frac{aE}{N}\right) \exp\left(-\frac{E + a^2E}{2N}\right). \quad (17)$$

The average of P_{e1} and P_{e2} is

$$P_e = \tfrac{1}{2}(P_{e1} + P_{e2}) = \frac{1}{2} Q\left[a\left(\frac{E}{N}\right)^{\frac{1}{2}}, \left(\frac{E}{N}\right)^{\frac{1}{2}}\right]. \quad (18)$$

This is identical with the error performance of the FM receiver, specified in equation (14).

V. DISCUSSION OF RESULTS

The concept of clicks has made possible a unique comparison between digital FM and matched filter receivers. When a suitable predetection filter is chosen for the FM receiver and the assumption made that this filter does not significantly process the signal, then the error performance of the two receivers is described identically.

We have gained particular insight into the error mechanism of the digital FM receiver under conditions of intersymbol interference. The analysis shows how the rate of occurrence of the noise clicks is critically dependent on this distortion of the signal waveform. This is in direct contrast to the usual AM systems where intersymbol interference manifests itself by a gradual degradation caused by "eye" closing.

Numerical results, illustrated in Fig. 6, show that the receivers'

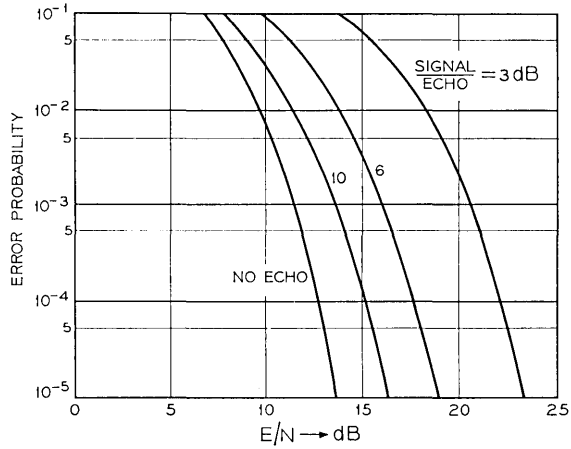


Fig. 6 — Error performance in single echo channel.

performance can be summarized in terms of the increase in signal-to-noise ratio required to maintain a desired error rate when an echo is superposed. For example, 10 dB higher signal-to-noise ratio is required to maintain a 10^{-4} error rate when an echo having half the signal power is added. It is easy to show that the asymptotic deterioration in performance with echo behaves as 20 $\log_{10}$ $(1-a)$ dB.

VI. ACKNOWLEDGMENT

The author wishes to thank J. E. Mazo and J. Salz for discussions of this work.

APPENDIX A

Click Probabilities

This appendix concerns the application of an average click probablity to FM receiver performance. The mathematical details are given here which relate equations (6), (7), and (8) to equations (9) and (10).

It has been shown in Refs. 1 and 2 that the probability of a click in time dt can be approximated as*

$$H_- \, dt \cong \frac{\dot{\varphi}}{2\pi} e^{-\rho} \, dt \tag{19}$$

where $\dot{\varphi}$ is the time derivative of the envelope angular variation and

$$\rho = \frac{A^2}{2\overline{e_N^2}}$$

is the signal-to-noise ratio. Substitution in equation (8) gives the error probability

$$P = \int_0^{2\pi} \frac{dx}{2\pi} \int_0^T \frac{dt}{2\pi} \dot{\varphi}(x, t) \exp\left[-\frac{A^2(x, t)}{2\overline{e_N^2}} \right] \tag{20}$$

where $x = \varphi_1 - \varphi_2$. Substitution in equation (20) of $A(x, t)$ and the time derivative of $\varphi(x, t)$, $\dot{\varphi}(x, t) = 2\pi f_d$, from equation (6) gives for

* This validity of this approximation depends on sufficiently large ρ for $\dot{\varphi} \neq 0$. Klapper[3] has discussed this in some detail. Although the approximation is not as good with intersymbol interference, it appears adequate.

the signal-echo pair s_1 , s_1

$$P_{e1} = \frac{1}{2\pi}\int_0^{2\pi} dx \int_0^T f_d \exp\left[-\frac{1}{2e_N^2}(a_1^2 + a_2^2 + 2a_1a_2\cos x)\right] dt$$
$$= \frac{1}{2} I_0\left(\frac{a_1a_2}{e_N^2}\right) \exp\left(-\frac{a_1^2 + a_2^2}{2e_N^2}\right) \tag{21}$$

where we have used the integral definition of $I_0(\cdot)$ and noted $f_dT = \frac{1}{2}$.

We now take $A(x, t)$ and $\varphi(x, t)$ from equation (7) and notice that the time derivative of the latter can be expressed as

$$\dot{\varphi}(t) = 4\pi f_d\left[\frac{a_1^2 + a_1a_2\cos(4\pi f_d t + \varphi_1 - \varphi_2)}{a_1^2 + a_2^2 + 2a_1a_2\cos(4\pi f_d t + \varphi_1 - \varphi_2)}\right] - 2\pi f_d . \tag{22}$$

When the first term of equation (22) and the expression for $A(x, t)$ are substituted into equation (20) the integral can be recognized as a Q function representation given by Helstrom:[7]

$$Q(\alpha, \beta) = \frac{1}{2\pi}\exp\left(-\frac{\alpha^2 + \beta^2}{2}\right)\int_0^{2\pi} \frac{\left(1 - \frac{\alpha}{\beta}\cos u\right)e^{\alpha\beta\cos u}}{1 + \left(\frac{\alpha}{\beta}\right)^2 - 2\left(\frac{\alpha}{\beta}\right)\cos u} du$$

where $\alpha < \beta$ and we have replaced $4\pi f_d t + \varphi_1 - \phi_2 + \pi$ by the variable u.

We make the identifications

$$\alpha \triangleq \frac{a_2}{(e_N^2)^{\frac{1}{2}}}, \qquad \beta \triangleq \frac{a_1}{(e_N^2)^{\frac{1}{2}}}.$$

Substitution of the second term of equation (22) gives an integral identical to the right-hand side of equation (21). Thus, for the signal-echo pair s_1, s_2

$$P_{e2} = Q\left[\frac{a_2}{(e_N^2)^{\frac{1}{2}}}, \frac{a_1}{(e_N^2)^{\frac{1}{2}}}\right] - \frac{1}{2} I_0\left(\frac{a_1a_2}{e_N^2}\right)\exp\left(-\frac{a_1^2 + a_2^2}{2e_N^2}\right).$$

APPENDIX B

Matched Filter Receiver

This appendix concerns the application of the single echo channel waveforms to filters which are matched to the waveforms in the absence of the echo. The mathematical form is similar to the form

illustrated by Helstrom,[7] and we give only a brief summary here to show the effect of the echo.

The error probability depends on a comparison of the sampled outputs, R_1 and R_2, of the two matched filters. Helstrom shows

$$R_i^2 = X_i^2 + Y_i^2 , \qquad i = 1, 2 \tag{23}$$

where

$$X_i = \int_0^T e(t) \cos (2\pi f_i t)\, dt \tag{24}$$

$$Y_i = \int_0^T e(t) \sin (2\pi f_i t)\, dt \tag{25}$$

where $e(t)$ is signal plus echo plus noise defined by equation (4) or (5).

f_i is one of the signal frequencies
$f_1 = f_c + f_d$
$f_2 = f_c - f_d$.

Substitution of $e(t)$ as given in equation (4) for the signal-echo pair s_1, s_1 gives

$$\begin{aligned} X_1 &= \frac{a_1 T}{2} \cos \varphi_1 + \frac{a_2 T}{2} \cos \varphi_2 + I_{N1} \\ Y_1 &= -\frac{a_1 T}{2} \sin \varphi_1 - \frac{a_2 T}{2} \sin \varphi_2 + I_{N2} , \\ X_2 &= I_{N3} \\ Y_2 &= I_{N4} \end{aligned} \tag{26}$$

where $I_{N1}, I_{N2}, I_{N3}, I_{N4}$ are zero mean independent Gaussian variables having equal variances $\sigma^2 = NT/4$ for noise density N(watts/Hz).

From these terms, we find that the joint distribution of R_1 and R_2 is $p(R_1, R_2)$

$$p(R_1 , R_2) = \frac{R_1 R_2}{\sigma^4} I_0\left(\frac{R_1 C_1}{\sigma^2}\right) \exp\left(-\frac{R_1^2 + R_2^2 + C_1^2}{2\sigma^2}\right) \tag{27}$$

where

$$C_1^2 = \left(\frac{a_1 T}{2}\right)^2 + \left(\frac{a_2 T}{2}\right)^2 + 2\left(\frac{a_1 T}{2}\right)\left(\frac{a_2 T}{2}\right) \cos (\varphi_1 - \varphi_2).$$

The error probability is

$$\begin{aligned} P &= \int_0^\infty dR_1 \int_{R_1}^\infty dR_2\, p(R_1 , R_2) \\ &= \frac{1}{2} \exp\left(-\frac{C_1^2}{4\sigma^2}\right). \end{aligned} \tag{28}$$

Averaging this value over $x = \varphi_1 - \varphi_2$ yields

$$P_{e1} = \frac{1}{2\pi} \int_0^{2\pi} P(x)\, dx = \frac{1}{2} I_0\left(\frac{aE}{N}\right) \exp\left[-\frac{E + a^2E}{2N}\right]. \tag{29}$$

Similarly, substitution of $e(t)$ as given by equation (5) into equations (24) and (25) gives, for the signal-echo pair s_1, s_2

$$\begin{aligned} X_1 &= \frac{a_1T}{2} \cos \varphi_1 + I_{N1} \\ Y_1 &= -\frac{a_1T}{2} \sin \varphi_1 + I_{N2} \\ X_2 &= \frac{a_2T}{2} \cos \varphi_2 + I_{N3} \\ Y_2 &= -\frac{a_2T}{2} \sin \varphi_2 + I_{N4} . \end{aligned} \tag{30}$$

Now the joint distribution of R_1 and R_2 is found to be

$$p(R_1 , R_2) = \frac{R_1R_2}{\sigma^4} I_0\left(\frac{R_1C_1}{\sigma^2}\right) I_0\left(\frac{R_2C_2}{\sigma^2}\right) \exp\left(-\frac{R_1^2 + R_2^2 + C_1^2 + C_2^2}{2\sigma^2}\right) \tag{31}$$

where

$$\begin{aligned} C_1 &= \frac{a_1T}{2} \\ C_2 &= \frac{a_2T}{2}. \end{aligned}$$

The error probability in this case is found via the following steps.

$$\begin{aligned} P_{e2} &= \int_0^\infty dR_1 \int_{R_1}^\infty dR_2\, p(R_1 , R_2) \\ &= \int_0^\infty dR_1\, \frac{R_1}{\sigma^2} I_0\left(\frac{R_1C_1}{\sigma^2}\right) Q\left(\frac{C_2}{\sigma} , \frac{R_1}{\sigma}\right) \exp\left(-\frac{R_1^2 + C_1^2}{2\sigma^2}\right) \end{aligned} \tag{32}$$

where we have substituted $p(R_1, R_2)$ from equation (31) and used the Q function definition:

$$Q(\alpha, \beta) = \int_{\beta}^{\infty} tI_0(\alpha t) \exp\left(-\frac{\alpha^2 + t^2}{2}\right) dt.$$

We see, by this manipulation, that equation (32) is integrable; for example, as shown by Stein.[8] Thus

$$P_{e2} = Q\left(\frac{C_2}{\sigma\sqrt{2}}, \frac{C_1}{\sigma\sqrt{2}}\right) - \frac{1}{2} I_0\left(\frac{C_1 C_2}{2\sigma^2}\right) \exp\left(-\frac{C_1^2 + C_2^2}{4\sigma^2}\right). \quad (33)$$

Appropriate substitutions of the terms from equations (31) and (14) give equation (17).

REFERENCES

1. Rice, S. O., "Noise in FM Receivers," in *Proc. Symposium Time Series Analysis,* ed., M. Rosenblatt, New York: John Wiley & Sons, 1963.
2. Cohn, J., "A New Approach to the Analysis of FM Threshold Extension," Proc. Nat. Elec. Conf., *12* (1956), pp. 221–236.
3. Klapper, J., "Demodulator Threshold Performance and Error Rates in Angle Modulated Digital Signals," RCA Review, *27* (June 1966), pp. 226–244.
4. Mazo, J. E. and Salz, J., "Theory of Error Rates for Digital FM," B.S.T.J., *45* (November 1966), pp. 1511–1535.
5. Schilling, D. C., Hoffman, E., and Nelson, E. A., "Error Rates for Digital Signals Demodulated by an FM Discriminator," IEEE Trans. Commun. Technology, *COM-15* (August 1967), pp. 507–517.
6. Bennett, W. R., Curtis, H. E., and Rice, S. O., "Interchannel Interference in FM and PM Systems," B.S.T.J., *34* (May 1955), pp. 601–636.
7. Helstrom, C. W., "The Resolution of Signals in White, Gaussian Noise," Proc. IRE, *43* (September 1955), pp. 1111–1118.
8. Stein, S., "Unified Analysis of Certain Coherent and Noncoherent Binary Communication Systems," IEEE Trans. Inform. Theory, *IT-10* (January 1964), pp. 43–51.

Eliminating Broadband Distortion in Transistor Amplifiers

By LEE C. THOMAS

(Manuscript received July 6, 1967)

This paper presents the results of a study directed toward understanding the basic distortion mechanisms in transistors. (i) We develop an analytic model for the transistor which describes small signal linear performance and nonlinear effects. The linear model is matched to the measured h-parameters of the device over a wide range of frequency and bias current. We superimpose three distinct nonlinear effects on this linear skeleton model, all approximated to third order terms. (ii) We show experimental confirmation that, for some bias-load conditions, the second order distortion can be minimized and we show that it is possible to simultaneously minimize both second- and third-order distortion under the same bias-load condition. This result also is confirmed experimentally. (iii) We derive and discuss in detail an analytic expression for the optimum load. Based on this expression, we present detailed procedures for finding this optimum condition for any transistor, and give experimental corroboration. (iv) We give a qualitative description of the interaction among these three nonlinear effects based on an analog computer simulation of the model. This description makes it easier to visualize the distortion cancellation phenomena derived in this paper, and indicates a technique for extending the effect to a broad band of frequencies. We conclude that proper use of the distortion cancellation effect can greatly improve intermodulation performance in existing transistors.

I. INTRODUCTION

System studies have indicated that very broad band (greater than 20 mHz) AM coaxial cable systems will be modulation-limited. Intensive investigations to understand and characterize the inherent modulation properties of devices and repeater circuits have been called for. We made one such study directed toward understanding the basic distortion mechanisms in transistors.

The history of transistor distortion literature can be characterized as an erosion process in which highly restricted parts of the total problem are attacked leaving fresh complexities exposed for future work. In early work by Akgun and Strutt, the analysis is restricted to nonlinearities in the emitter resistance assuming an ac short at the input and output.[1] Observed nulls in second and third order distortion do not correlate with the theory, which does, however, include frequency effects. Using many of the same assumptions, Mallinckrodt and Gardner extended this earlier work to account for a third order null at low frequencies when the nonlinear emitter resistance is dominant.[2]

More recently Riva, Beneteau, and Dalla Volta considered all important sources of distortion by breaking the problem into three distinct operating regions with expressions for minimizing second order distortion in each.[3] They do not treat of third order minimization, and they use a dc model. Reynolds analyzes third order minimization at particular nonzero frequencies for dominance of the emitter resistance nonlinearity.[4]

There are two reasons for the specialized nature of these efforts. First, transistors, as contrasted with vacuum tubes, have at least three dominant nonlinearities. It would be difficult to consider all of these in a general expression for second and third order distortion. Second, frequency effects can be important in many applications. In general, the analysis of nonlinear effects as a function of frequency requires the use of extremely powerful and, as a result, cumbersome analytic techniques. In the special case of an exponential input v-i relation it is possible to avoid a general analysis, which explains why analyses which include frequency effects have been limited to emitter nonlinearities. Even in this exponential case, however, the third order null predicted by Reynolds is a narrowband effect, applicable only at a particular frequency.

This paper extends these earlier efforts in four important respects.

(*i*) We conclude that the distortion measured at the terminals results from algebraic cancellation between distortion components produced by nonlinear effects within the transistor. This conclusion originated from empirical observations made on an analog computer simulation of a transistor. An analytic argument reinforces this conclusion by comparing plots of algebraic cancellation to measured distortion curves. Also, we give experimental support of the cancellation phenomenon.

(*ii*) We present a low-frequency analysis of a complete extrinsic model including three nonlinearities: emitter resistance, nonlinear current gain, and avalanche multiplication, all approximated by a third order polynominal. We avoid considerable complexity by directing the analysis strictly to the question of minimizing distortion and by not developing a general distortion expression. This analysis is independent of any assumptions concerning distortion cancellation, but yields the same results.

(*iii*) From this analysis we show that it is possible to *simultaneously* null both second and third order distortion under the same bias-load condition. The analytic technique we use to determine a null is linearization of the input-output relation up to and including third order, thus implying a minimum in harmonic distortion, intermodulation, or any other specialized figure of merit. The existence of this simultaneous null is verified in the laboratory.

(*iv*) Extension of the cancellation effect to a broad band of frequencies can be accomplished by external reactive compensation. This compensation maintains a 180° phase shift between the collector-base voltage and the real component of the emitter current, a relation that exists automatically at low frequencies where the rigorous analysis is performed. This phase shift is the fundamental requirement for total cancellation, based on the qualitative insight mentioned in item *i*.

PRINCIPAL SYMBOLS

A	Parameter in the $\beta(I_c)$ relation.
$\alpha(I_e)$	Current dependence of the dependent current source.
α_1 , α_2 , α_3	Taylor series coefficients in the expansion of $\alpha(I_e)$ around I_{eo} .
α_{max}	Maximum value of α with respect to I_c .
β	Common emitter ac gain.
β_{max}	Maximum value of β with respect to I_c .
I_c	Total collector current.
I_{cp}	Collector current where β_{max} occurs.
i_c	Small signal collector current.
I_e	Total emitter current.
I_{eo}	Emitter current bias level.
i_e	Small signal emitter current.
I_o	Collector current bias level.
I_{out}	Current in the load resistor.

$M(V_{cb})$	Voltage dependence of the dependent current source.
M_1 , M_2 , M_3	Taylor series coefficients in the expansion of $M(V_{cb})$ around V_o .
r_e	Emitter resistance.
r_1 , r_2 , r_3	Taylor series coefficients relating v_e to i_e .
R_L	Load resistance.
$R_{(2)\ opt}$	Load which minimizes second order distortion when third order distortion is negligible.
R_s	Source resistance.
V_{cb}	Total collector-to-base voltage.
v_{cb}	Small signal collector-to-base voltage.
v_e	Small signal voltage across r_e .
V_o	Collector-to-base bias level.
V_{out}	Voltage across the load resistor.

II. A QUALITATIVE MODEL FOR THE DISTORTION MECHANISM

Let us describe the qualitative insight (i) to get a broad look at the cancellation phenomenon before rigorous analysis obscures a simple concept.

An analog computer simulation of the model of Fig. 1 allows us to examine the interaction of the three nonlinearities by examining their effects one at a time. Thus, for example, we may allow only $\alpha(I_e)$ to be nonlinear and observe the second harmonic distortion components of the output voltage. If we then make α constant and allow $M(V_{cb})$ to vary, we observe that the resulting waveform is 180° out of phase with the first waveform as shown in Fig. 2. This is plausible since V_{cb} and I_e are inherently 180° out of phase at low frequencies. Thus any cancellation that we obtain between current dependent nonlinearities

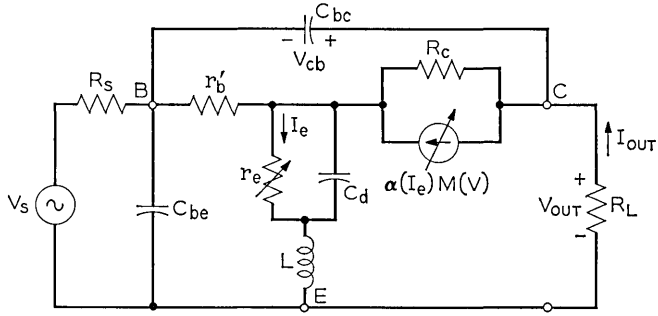


Fig. 1 — High frequency nonlinear model.

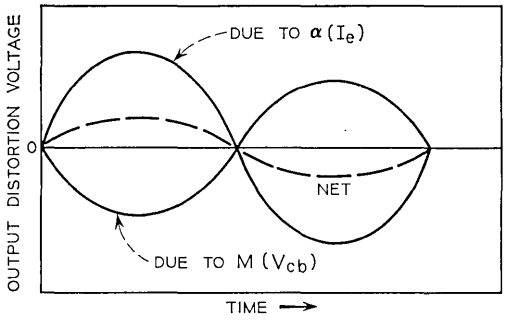


Fig. 2 — Cancellation of distortion components.

and voltage dependent nonlinearities will *not* require *phasing* I_e and V_{cb} properly, but will result from properly adjusting the relative magnitudes of V_{cb} and I_e.

Since $I_e \approx I_{out}$, the most direct way to adjust the magnitude of $V_{cb} \approx V_{out}$ relative to I_e is to change the load resistance. Hence the strong dependence of distortion on R_L as shown in Fig. 3 for a fixed bias level of $V_{cb} = V_o$ and $I_{out} = I_o$. To obtain cancellation in second and third order distortion at the same time, not only the relative magnitudes are important but the absolute level must be correct. This cancellation model explains the sharpness of the null: since the net distortion is a small difference between large distortion components, a small percentage change in the ratio of the larger components will yield a large percentage change in the difference. Experimentally, as a null is passed the output distortion waveform changes phase by 180° as we would expect from one component's becoming dominant over the other.

It is important to notice that this cancellation effect is not some artificial phenomenon that we are forcing to occur. According to the model presented here, some degree of cancellation always occurs in any transistor at any level of distortion. We give a more quantitative argument supporting this exact cancellation model for visualizing the transistor distortion mechanism in Appendix C.

It has been the author's experience that a disturbingly large percentage of published technical material is exclusively concerned with presenting conclusions. In most cases, these conclusions were arrived at by the rigorous manipulation of symbols long after the original insight which prompted the investigation. The purpose of this section is to describe the insights first in the belief that the reader will have

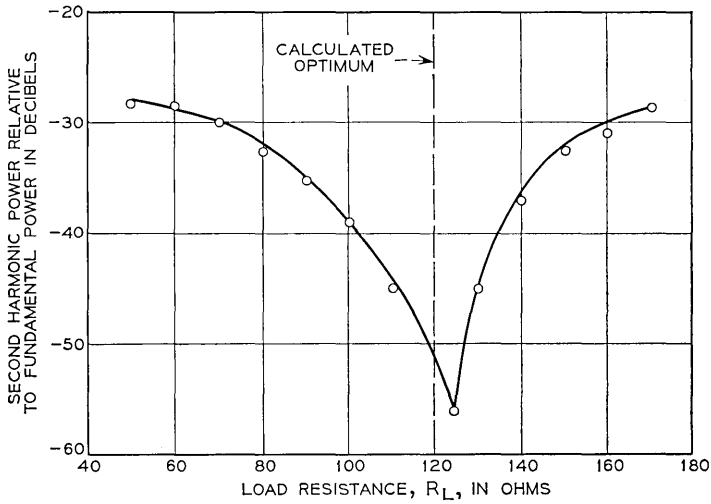


Fig. 3 — Experimental null in second harmonic distortion as a function of R_L, using a Western Electric 20J transistor with $V_o = 30$ volts, $I_o = 100$ milliamperes, and $R_s = 500$ ohms.

at least one less handicap if he is allowed to see the simple ideas on which the rather interesting conclusions of this paper are based. These ideas are:

(*i*) The nonlinearities of the transistor (including some, such as the base spreading resistance and the diffusion capacitance, which are not considered in this paper) are dependent on the emitter current, I_e, and the collector-base voltage, V_{cb}. At low frequencies I_e and V_{cb} are 180° out of phase.

(*ii*) As a result of this phase difference, distortion components resulting from these independent variables will subtract at low frequencies.

(*iii*) On an analog computer simulation, we observe the ability to extend this subtraction effect to the extent of total cancellation by manipulating external circuit parameters. Thus it should be possible to analyze a low frequency model by imposing the condition of zero distortion and solve for the required circuit parameters. We would expect the load resistance to be an important parameter in this analysis since it determines the ratio of V_{cb} to I_e.

(*iv*) Considering the low frequency phase difference between I_e and V_{cb} as the most important factor in achieving total cancellation, we suggest a technique for extending the low frequency results to a broad

band of frequencies. This extension is achieved by the simple expedient of compensating the load to achieve a constant real part, R_L, and still maintain the proper phase between V_{cb} and I_e as frequency increases.

The following sections develop the rigorous analysis (most of which is relegated to Appendix A) and examine in some detail the analytic conditions for a null and the implications of these conditions in the area of circuit and device design.

III. TRANSISTOR MODEL

The model in Fig. 1 has been matched closely to the h parameters of the Western Electric type 46A transistor over a wide range of frequency (5 to 100 mHz) and bias current (50 to 150 mA). Figs. 4 and 5 show a typical match, obtained from a general purpose optimization program. Three distinct nonlinear effects were then superimposed on this small signal linear skeleton model. The current dependence of the dependent current source is changed from αI_e to the expansion around the emitter current bias point, I_{eo},

$$\alpha(I_e) = I_o + \alpha_1(I_e - I_{eo}) + \tfrac{1}{2}\alpha_2(I_e - I_{eo})^2 + \tfrac{1}{6}\alpha_3(I_e - I_{eo})^3 + \cdots \quad (1)$$

where I_o is the quiescent collector current. The voltage dependence of the dependent current source is changed from the constant, M, to the expansion around the collector-to-base bias voltage, V_o,

$$M(V_{cb}) = 1 + M_1(V_{cb} - V_o) + \tfrac{1}{2}M_2(V_{cb} - V_o)^2 + \tfrac{1}{6}M_3(V_{cb} - V_o)^3 + \cdots \quad (2a)$$

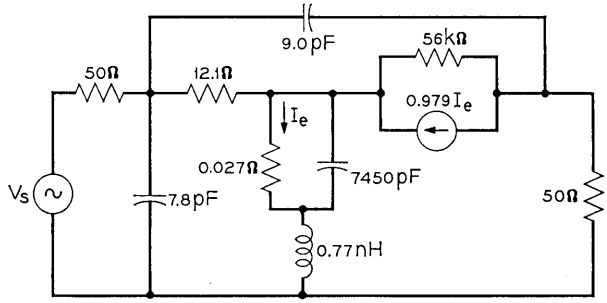


Fig. 4 — Linear model for $I_e = 150$ mA, $V_{cb} = 10$V. With the indicated element values, this model matches the measured h-parameters shown in Fig. 5. The quality of the match at this bias point ($I_e = 150$ mA, $V_{cb} = 10$V) is typical.

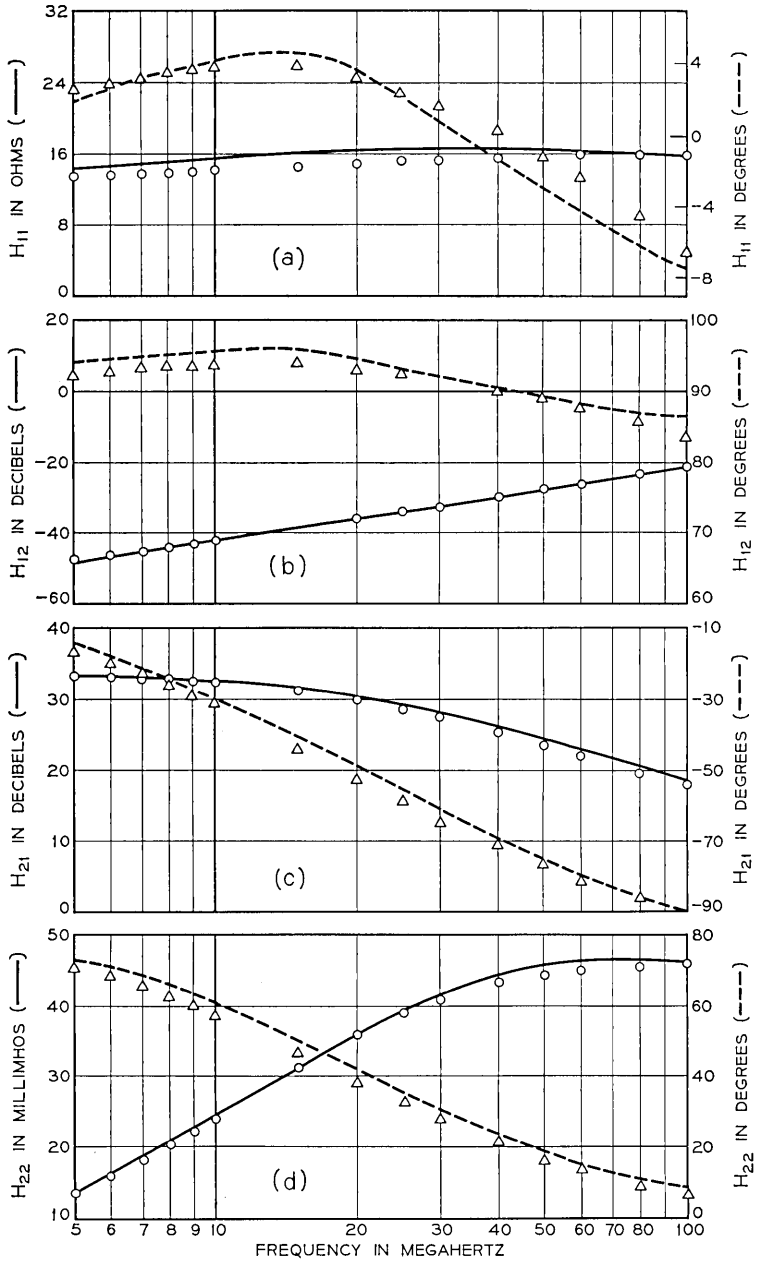


Fig. 5 — Magnitude and phase, as a function of frequency match to (a) h_{11}, (b) h_{12}, (c) h_{21}, and (d) h_{22}. Solid lines were measured; points were plotted from the model.

So that the total dependent current source relationship is

$$I_c = \alpha(I_e)M(V_{cb}). \tag{2b}$$

And finally the emitter resistance, r_e, is replaced by the expansion around I_{eo}

$$v_e = r_1(I_e - I_{eo}) + \tfrac{1}{2}r_2(I_e - I_{eo})^2 + \tfrac{1}{6}r_3(I_e - I_{eo})^3 + \cdots \tag{3}$$

where the coefficients in equations 1, 2, and 3 are the corresponding derivatives of the Taylor series expansion. Define the small signal quantities as

$$i_c = I_c - I_o \tag{4}$$

$$i_e = I_e - I_{eo} \tag{5}$$

$$v_{cb} = V_{cb} - V_o\,. \tag{6}$$

Using these relations, equations 1, 2, and 3 become

$$\alpha(I_e) = I_o + \alpha_1 i_e + \tfrac{1}{2}\alpha_2 i_e^2 + \tfrac{1}{6}\alpha_3 i_e^3 + \cdots \tag{7}$$

$$M(V_{cb}) = 1 + M_1 v_{cb} + \tfrac{1}{2}M_2 v_{cb}^2 + \tfrac{1}{6}M_3 v_{cb}^3 + \cdots \tag{8}$$

$$v_e = r_1 i_e + \tfrac{1}{2}r_2 i_e^2 + \tfrac{1}{6}r_3 i_e^3 + \cdots \tag{9}$$

Substituting equations 7 and 8 into 2b, and retaining third order terms

$$I_c = (I_o + \alpha_1 i_e + \tfrac{1}{2}\alpha_2 i_e^2 + \tfrac{1}{6}\alpha_3 i_e^3) \cdot (1 + M_1 v_{cb} + \tfrac{1}{2}M_2 v_{cb}^2 + \tfrac{1}{6}M_3 v_{cb}^3) \tag{10}$$

$$I_c - I_o = i_c = \alpha_1 i_e + I_o M_1 v_{cb} + \tfrac{1}{2}\alpha_2 i_e^2 + \tfrac{1}{2}I_o M_2 v_{cb}^2 + \alpha_1 M_1 i_e v_{cb} + \tfrac{1}{2}\alpha_1 M_2 v_{cb}^2 i_e + \tfrac{1}{2}\alpha_2 M_1 v_{cb} i_e^2 + \tfrac{1}{6}\alpha_3 i_e^3 + \tfrac{1}{6}I_o M_3 v_{cb}^3 \tag{11}$$

At this point we have developed a model for the transistor, indicating the nature and form of the particular nonlinearities considered in both the analog computer simulation of the complete, frequency-dependent model of Fig. 1 and the analysis of the dc model of Fig. 6.

IV. THE ANALYSIS

4.1 *Optimization Equations*

An analog computer simulation of the complete, frequency-dependent model just discussed suggests that a simpler model is sufficient to describe the distortion characteristics of the transistor at low

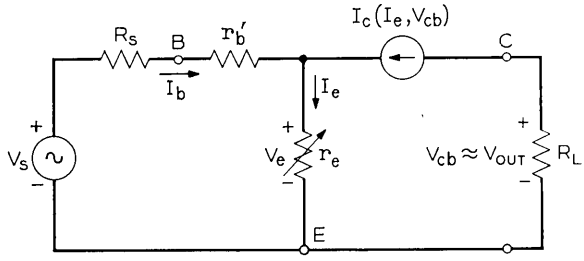


Fig. 6 — Low frequency nonlinear model.

frequencies. Fig. 6 shows this simplified dc model. The following analysis of this model is detailed in Appendix A.

(*i*) The incremental output voltage, v_{out}, is related to the input voltage, V_S, retaining third order terms as in equation 11.

(*ii*) This input-output relation is constrained to be linear, thus forcing both second and third order distortion to zero.

(*iii*) This constraint requires certain coefficients in the nonlinear v_{out} (V_S) relation to be zero. These coefficients are, of course, functions of the linear and nonlinear parameters of the system. Thus, when these functions are made zero, v_{out} is a linear function of V_S (to third order), and the derivation of the optimization equations is complete. These equations are:

$$-R_L^2(I_oM_2) + R_L(2\alpha_1M_1) + \delta - \alpha_2 = 0 \tag{12}$$

$$R_L^3(I_oM_3) - R_L^2(3\alpha_1M_2) + \xi - \alpha_3 = 0 \tag{13}$$

where

$$\delta = r_2/(R_s + r_b') < 0, \quad \text{since} \quad r_2 < 0, \quad \text{(equation 19)}, \tag{14}$$

and

$$\xi = r_3/(R_s + r_b') > 0, \quad \text{since} \quad r_3 > 0 \quad \text{(equation 20)}. \tag{15}$$

For the simpler case where the amplitude of third order distortion is sufficiently low so that third order terms are negligible, equation 13 is satisfied identically and only equation 12 remains, which is easily solved to yield

$$R_{(2)\,\text{opt}} = \alpha_1M_1/I_oM_2 + [(\alpha_1M_1/I_oM_2)^2 - (\alpha_2 - \delta)/I_oM_2]^{\frac{1}{2}}. \tag{16}$$

Thus $R_{(2)\text{opt}}$ is the value of load resistance which causes second order distortion to be zero for the case where third order terms are

negligible. Notice that the analytic technique used to determine a distortion null here is linearization of the input-output relation, and thus implies a minimum in harmonic distortion, intermodulation distortion, or any other specialized figure of merit. Of course, $R_{(2)\text{opt}}$ is a function of bias current and voltage because of the dependence of M_1 and M_2 on voltage and r_2, α_1 and α_2 on current. The implications of equations 12, 13, and 16 become more clear when the dependence of these parameters on bias is considered.

4.2 *Relating Parameters to More Directly Measurable Quantities*

It is revealing to express the parameters of equations 12 and 13 in terms of the bias variables and other directly measurable parameters of the transistor.

Assuming the standard exponential i-v relation at the emitter-base junction we can immediately derive from

$$I_e = I_s[\exp(\lambda q V_e/kT) - 1] \tag{17}$$

the following relations:

$$r_1 = kT/\lambda q I_o = r_o/I_o , \tag{18}$$

$$r_2 = -kT/\lambda q I_o^2 = -r_o/I_o^2 , \tag{19}$$

$$r_3 = 2kT/\lambda q I_o^3 = 2r_o/I_o^3 . \tag{20}$$

Similarly, if we assume that the avalanche effect in the common-emitter mode is described by an equation of the same form as Miller's[5]

$$M(V_{cb}) = [1 - (V_{cb}/V_A)^n]^{-1} \tag{21}$$

where V_A is the common-emitter breakdown voltage as shown in Fig. 7. Then, at $V_{cb} = V_o$:

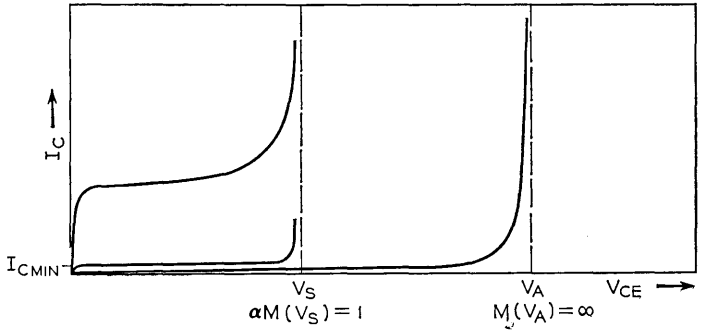


Fig. 7 — Avalanche characteristics.

$$M_1 = n(V_o/V_A)^n/V_o\,, \tag{22}$$

$$M_2 = n(n-1)(V_o/V_A)^n/V_o^2\,, \tag{23}$$

$$M_3 = n(n-1)(n-2)(V_o/V_A)^n/V_o^3\,. \tag{24}$$

The avalanche voltage, V_A, can be determined on a curve tracer oscilloscope by leaving the emitter open-circuited in a grounded base configuration and sweeping the collector-base voltage. The sustaining voltage, V_S, shown in Fig. 7, is obtaining with the transistor in the common-emitter mode and at least enough base current flowing to produce $I_{C\min}$ at the output. At V_S the avalanche factor $M(V_{cb})$ has increased above unity sufficiently so that $\alpha(I_C)M(V_S) = 1$. As a result the common-emitter current gain (β) at this voltage is infinite. Choosing the smallest α at which this occurs ($\alpha_{\min}$) allows us to determine the exponent, n, in equation 21:

$$\alpha_{\min}M(V_S) = 1 = \alpha_{\min}[1 - (V_S/V_A)^n]^{-1}. \tag{25}$$

Therefore

$$n \approx \log \beta_{\min}/\log (V_A/V_S), \tag{26}$$

where $\beta_{\min}$ corresponds to $\alpha_{\min}$ and may be determined from equation 27 using $I_o = I_{C\min}$. Notice that equation 21 constitutes an empirical relationship in this study and is not intended to be rigorously tied to any one of the various avalanche mechanisms. It is apparent, too, that the measurements determining equations 25 and 26 will be influenced by other voltage-dependent mechanisms (for example, the Early effect); hence they are not strictly related to the avalanche multiplication effect alone. Equation 21 has the virtue of mathematical tractability; equation 25 allows the parameters of 21 to be determined conveniently; and, finally, the excellent experimental agreement with the theory described in Section V provides adequate justification of the original assumptions. In any case, the derivation of equations 12 and 13 is based on a general power series expansion for $M(V_{cb})$ around V_o; hence it remains valid for any M_1, M_2, and M_3.

Finally we require α_2 and α_3. We show in Appendix B that β can be empirically related to collector bias by

$$\beta \cong \beta_{\max}/[1 + A \ln^2 (I_o/I_{cp})] \tag{27}$$

where $\beta_{\max}$ is the maximum β which occurs at $I_o = I_{cp}$, as shown in Fig. 8, and A is a parameter of the equation. Determination of α_2,

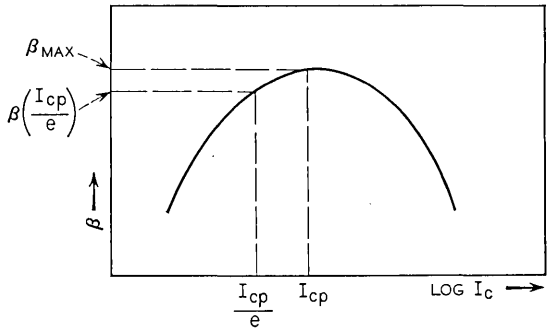


Fig. 8 — Current gain nonlinearity as a function of the bias current, I_c.

α_3, and A is derived in Appendix B. They can be expressed as

$$\alpha_2 \cong -\left(\frac{4A\alpha_{\max}^2}{\beta_{\max} I_o}\right) \ln\left(\frac{I_o}{I_{cp}}\right) \tag{28}$$

$$\alpha_3 \cong \left(\frac{6A\alpha_{\max}^2}{\beta_{\max} I_o^2}\right) \ln\left(\frac{I_o}{eI_{cp}}\right) \tag{29}$$

$$A = \left[\beta_{\max} - \beta\left(\frac{I_{cp}}{e}\right)\right]/\beta_{\max} \tag{30}$$

so that A may be determined by finding I_{cp} and measuring β at that current and at $1/e$ times that current. Thus equations 18 through 30 give the functional relations for the various parameters in equations 12 and 13 and indicate the method of measuring the more fundamental parameters such as n and A. In the next section we use these relations in existence conditions for a simultaneous null, in order to guide an experimental search for this condition.

4.3 *Existence Conditions for Realizability*

While the simultaneous solution of equations 12 and 13 has not been accomplished in closed form, it is possible to derive the conditions under which a solution exists. Expressed in terms of the bias variables, such conditions can then be used as a guide in an experimental search for simultaneous nulling of second and third order distortion.

Basically we require R_L to be real and positive. For the second order equation, solved in equation 16, this simply requires that

$$(\alpha_1 M_1/I_o M_2)^2 \geqq (\alpha_2 - \delta)/I_o M_2 . \tag{31}$$

The condition for the existence of a positive, real solution to a cubic of the form

$$x^3 - px^2 + r = 0 \tag{32}$$

where

$$p = 3\alpha_1 M_2 / I_o M_3 \tag{33}$$

$$r = (\xi - \alpha_3)/I_o M_3 \tag{34}$$

$$x = R_L \tag{35}$$

is easily derived. Basically require

$$x^3 = px^2 - r. \tag{36}$$

Now, from equations 34 and 29, $r > 0$ for $I_o < eI_{cp}$. Thus, at $x = 0$, the parabola on the right side of equation 36 will be below the cubic on the left. There will be a positive intersection only if the equation is satisfied before the cubic term begins increasing more rapidly (larger slope) than the parabola. The slopes are equal at

$$x_1 = \tfrac{2}{3}p. \tag{37}$$

Therefore require

$$x_1^3 \leqq px_1^2 - r \tag{38}$$

or

$$\tfrac{4}{27}p^3 \geqq r. \tag{39}$$

Expressing this existence condition in terms of the problem variables and rearranging terms gives

$$(\alpha_1 M_2)^3 \geqq \tfrac{1}{4}(\xi - \alpha_3)(I_o M_3)^2. \tag{40}$$

Substituting in equations 31 and 40 with 18 through 29 and arranging terms we obtain

$$(V_o/V_A)^n > \text{the Greater of } [Q_1, Q_2] \tag{41}$$

where

$$Q_1 = \left(1 - \frac{1}{n}\right)\left[\frac{r_o}{I_o}(R_s + r_b') - \left(\frac{4A\alpha_{\max}^2}{\beta_{\max}}\right)\ln\left(\frac{I_o}{I_{cp}}\right)\right] \tag{42}$$

$$Q_2 = \left[\frac{r_o}{I_o}(R_s + r_b') - \left(\frac{3A\alpha_{\max}^2}{\beta_{\max}}\right)\ln\left(\frac{I_o}{eI_{cp}}\right)\right]\frac{(n-2)^2}{2n(n-1)}. \tag{43}$$

For most ranges of parameters and bias variables $Q_1 > Q_2$, thus, we will examine the condition

$$\left(\frac{V_o}{V_A}\right)^n > \left(1 - \frac{1}{n}\right)\left[\frac{r_o}{I_o}(R_s + r_b') - \left(\frac{4A\alpha_{\max}^2}{\beta_{\max}}\right)\ln\left(\frac{I_o}{I_{cp}}\right)\right] \quad (44)$$

in greater detail. This distinction between Q_1 and Q_2 is not critical, however, because they are similar in form. Thus, many of the qualitative considerations to be developed in the next section are the same for Q_1 or Q_2.

4.4 *Searching for a Simultaneous Null*

A careful examination of the existence condition (44) is useful in guiding an experimental search for a simultaneous null. Starting at the left side of the inequality, it is obvious that the bias voltage, V_o, must be as large as possible relative to V_A. Since, in any case, $V_o < V_A$, the exponent, n, should be as small as possible. The value of n, according to Rogers,[5] depends on whether the collector or base has the higher resistivity, and whether the high resistivity side is n or p type.

Where the collector has the higher resistivity, the lowest values of n are for npn silicon, and for pnp germanium. A second, less important, advantage of small n is that the multiplier on the right side of the inequality is reduced. The first term in the brackets tends to be the major contributor to the right side of the inequality and is therefore the term which is most desirable to reduce. This term, which represents input distortion resulting from a nonlinear emitter resistance, can be reduced by increasing the bias, I_o, and by increasing R_s to approximate a current source drive, thereby reducing input distortion.

The second term in the brackets will favorably reduce the right side of the inequality only if the logarithm is positive. This will be true if the bias current, I_o, is greater than I_{cp}, which is consistent with the earlier requirement for a larger I_o. Finally, the multiplier λ, in equation 18 should be small in order to reduce r_o.

Thus, it appears that the most likely candidate for a simultaneous null is a silicon power transistor to allow large values of I_o and V_o. The structure should be either pnp or npn, depending on which type gives the smaller n.

V. EXPERIMENTAL PROCEDURE AND RESULTS

Let us illustrate the application of these existence conditions in an experimental determination of $R_{(2)\text{opt}}$ as well as a simultaneous null in second and third order distortion.

As a fundamental check on the theoretical results, we decided to determine the accuracy of equation 16 with 19, 22, 23, and 28 substituted for the Taylor series coefficients. Also, it was desirable to verify the existence of a simultaneous null using the existence conditions of the previous section. Because of the low frequencies involved (input frequency of 1 kHz), the simplest approach was to simulate the measurement apparatus on the analog computer, using the same oscillator and bandpass filters already available on the original simulation.[6] The transistor used was the Western Electric 20J, and npn power transistor.

Using this equipment, the parameters of the $\beta(I_c)$ characteristic curve of the transistor were measured:

$$\beta_{\max} = 78$$

$$I_{op} = 15mA$$

$$\beta(I_{cp}/e) = 73.$$

From a curve tracer oscilloscope, the avalanche parameters were determined:

$$V_A = 60V$$

$$V_S = 35V$$

$$\beta_{\min} = 45.$$

These measurements yield the information to compute

$$n = 7$$

$$A = .064$$

from equations 26 and 30. From the manufacturer's data, $r_o = 50$ mV and $r_b' = 50\ \Omega$. The output power was maintained at one watt.

These parameters give all the information required by equation 16 to compute the function $R_{(2)\ \mathrm{opt}}(I_o)$ for various values of V_o. The curves in Figs. 9 and 10 show this computation compared to the plotted points which were measured. The agreement here is quite adequate. The quality of the match is further emphasized by comparing the computed values of $R_{(2)\ \mathrm{opt}}$ indicated in Fig. 3 and Fig. 11 to the measured nulls. The computed value shown in Fig. 11 is based on a solution to equation 16 only.

A typical simultaneous null obtained in the laboratory is shown in Fig. 11. This data indicates the high voltages (to emphasize avalanche

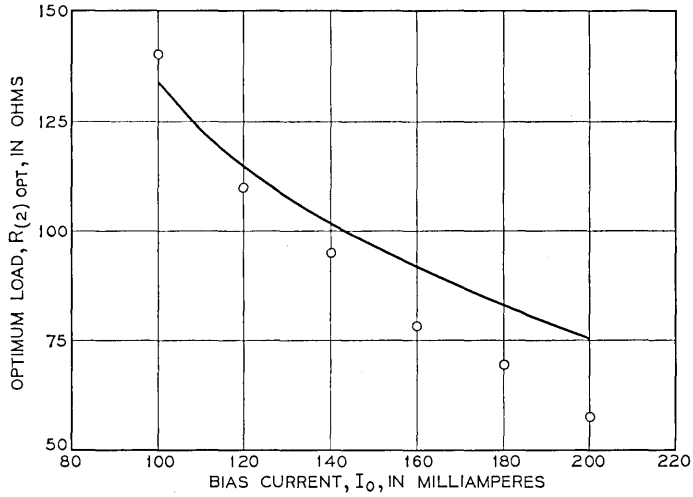


Fig. 9 — Measured (plotted points) and computed values (curve) of $R_{(2)opt}$ as a function of bias current, I_o, using a Western Electric 20J transistor with R_s = 500 ohms and V_o = 29 volts.

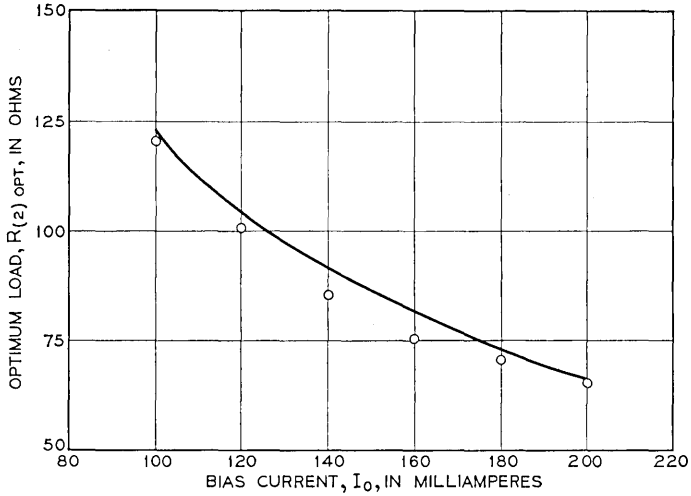


Fig. 10 — Measured (plotted points) and computed values (curve) of $R_{(2)opt}$ as a function of bias current, I_o, using a Western Electric 20J transistor with R_s = 500 ohms and V_o = 25 volts.

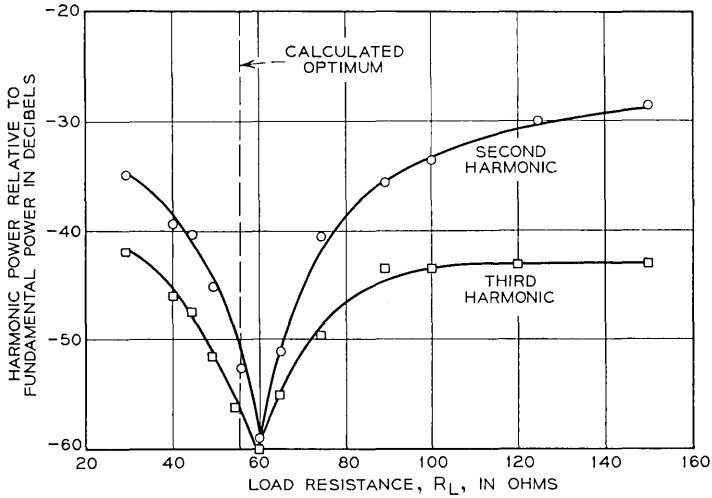


Fig. 11 — Simultaneous null in second and third harmonics as a function of load, R_L, with the same transistor and R_s as in Figs. 9 and 10. Here $V_o = 29$ volts and $I_o = 240$ mA.

distortion) and currents (to minimize input distortion) required for a simultaneous null. Conditions for a simultaneous null exist on an (R_L, I_o, V_o, R_s) surface, giving some redundant control to achieve desired power and impedance levels as well as minimum second and third harmonic distortion. It is apparent that a transistor manufactured with a lower value of n would allow a broader range of control over the bias voltage and current level required. Measurements on different units of the WE20J show a maximum spread of ± 10 per cent in measured values of the optimum load for a simultaneous null.

Experimentally, as R_L is varied, the second harmonic displayed on the oscilloscope decreases in amplitude, goes to zero, and begins to increase in amplitude. As it goes through a null, the second harmonic changes sign, giving additional weight to the qualitative distortion model discussed in Section II.

VI. EXTENSION OF CANCELLATION TO A BAND OF FREQUENCIES

Up to this point, our discussion has been limited to low frequency effects. Now let us consider why the above results do not apply at high frequencies and look at a straightforward approach to extend the validity of all previous results to a broad band of frequencies.

If we accept the qualitative picture given in Section II, it is obvious that we should not expect to maintain exact cancellation as frequency increases, since the phase of V_{cb} relative to I_e will change. A small change in phase will have an effect similar to changing the relative magnitudes of the current and voltage-dependent distortion components: the amplitude of their difference (the net distortion) will change by a large percentage near a null. In fact, at higher frequencies (on the order of $f_T/100$), the null of Fig. 2 vanishes altogether. It is apparent, then, that a solution to this problem is to apply external reactive compensation in such a way as to keep V_{cb} and I_c 180° out of phase as frequency increases. In the model shown by Fig. 6 if we consider a capacitor, C_D, in parallel with r_e, it is straightforward to derive the relation,

$$-V_{cb}/I_e \cong \text{Re}\,\{Z_L\} + j\left[\frac{\omega r_b'}{\omega_T} + \text{Im}\,\{Z_L\}\right], \tag{45}$$

where

$$\omega_T = \frac{1}{C_D r_e}. \tag{46}$$

Ideally, we would desire $Z_L = R_L - j\omega r_b'/\omega_T$ but this would require a negative inductor. A simple first order approximation to this function would be to parallel R_L with a capacitor, C. Then

$$Z_L = \frac{R_L}{1 + \omega^2 R_L^2 C^2} - j\omega \frac{R_L^2 C}{1 + \omega^2 R_L^2 C^2}. \tag{47}$$

From equation 47 choose

$$C_{\text{opt}} = r_b'/R_L^2 \omega_T . \tag{48}$$

Now

$$-\frac{V}{I_e} = \frac{R_L}{1 + \omega^2 R_L^2 C_{\text{opt}}^2} + j\frac{\omega^3}{\omega_T}\left[\frac{r_b' R_L^2 C_{\text{opt}}^2}{1 + \omega^2 R_L^2 C_{\text{opt}}^2}\right]. \tag{49}$$

For small angles the phase is given by

$$\varphi(\omega) = \omega^3 (r_b'/R_L \omega_T)^3. \tag{50}$$

Thus the phase is reduced below the uncompensated case up to the frequency

$$\omega_{\max} = \frac{R_L \omega_T}{r_b'}.$$

At which point the cubic dependence of equation 50 intersects the linear phase of the uncompensated transistor.

Obviously additional compensating elements can be used to cause higher derivatives of $\varphi(\omega)$ to be zero. A complication may arise if C_{opt} is less than the parasitic C_{CE} of the transistor. In this case we extend the required low-pass structure of the compensating network to include an inductor in series with R_L. In this case the required inductance is given by

$$L_{\text{opt}} = R_L^2 C_{CE} - r_b'/\omega_T$$

which is greater than zero for $C_{CE} > r_b'/R_L^2\omega_T = C_{\text{opt}}$.

VII. CONCLUSION

Our conclusions are based on simulation of the transistor on an analog computer, analysis, and experimental results. The rigorous analysis predicts the existence of a simultaneous null in second and third harmonic distortion under the same bias-load conditions. This null has been observed in the laboratory. In addition, experiments on the simulation provide qualitative insight into the nature of the distortion mechanism.

We conclude that this mechanism consists of the algebraic subtraction, at low frequencies, of distortion components from various sources within the transistor such as the nonlinear emitter resistance, current gain, and avalanche multiplication effect. This interaction between distortion components yields a net distortion which is the difference between the contributing components, and can be made zero by a proper choice of the bias and load.

With this mechanism in mind we developed a technique for extending the cancellation phenomenon to a broad band of frequencies. This technique consists of external reactive compensation which maintains 180° phase shift between the distortion components, a condition which exists inherently at low frequencies.

We have obtained experimental confirmation of the theoretical dependence of the optimum load for second order distortion on bias variables. The theory predicting a simultaneous null in second and third order distortion has been confirmed. We have also obtained experimental support for the distortion cancellation phenomenon. We discussed methods to aid future measurement efforts in implementing this distortion reduction phenomenon. These methods are based on interpretation of the theoretical expressions developed in the paper

which reveal the necessity for high levels of bias voltage and current to obtain a simultaneous null.

This study opens several fruitful areas for future work, both in device and circuit areas. Primarily, the phenomenon described uses circuit techniques to minimize distortion (optimizing the bias-load point). Additional effort in the circuit aspects of minimizing distortion should be directed toward desensitizing the null condition to variations in the bias-load point. For example, if the bias current is forced to change with R_L as shown in Fig. 9, optimum conditions could be maintained over a range of changes in the load.

In the realm of device design, effort should be directed toward adjusting device parameters to allow nulling in useful regions of the bias-load space. For example, a softer avalanche characteristic (lower value of n) would allow the use of lower bias voltages.

ACKNOWLEDGMENT

In a long term study of this kind, a large number of people share in the effort. There are certain individuals whose critical contributions are gratefully acknowledged.

Primarily, the author thanks S. Narayanan, who has analyzed the distortion problem using the Volterra Series representation which is amenable to digital simulation.[7]

J. S. T. Huang did much of the ground work on which the simulation is based and has been an invaluable consultant since that time.

Miss R. Mayorkas, M. J. Magelnicki, and R. Ukeiley took measurements for the linear and nonlinear characterization and Miss D. M. Bohling and Mrs. J. L. Murray helped with digital computer calculations. N. J. Chaplin contributed valued data and comments on initial results.

REFERENCES

1. Akgun, M., and Strutt, M. J. O., "Cross Modulation and Nonlinear Distortion in RF Transistor Amplifiers," IRE Trans. Elec. Devices, *6* (October 1959), pp. 457–469.
2. Mallinckrodt, A. J., and Gardner, F. M., "Distortion in Transistor Amplifiers," IEEE Trans. Elec. Devices, *10* (July 1963), pp. 288–289.
3. Riva, G. M., Beneteau, P. J., and Dalla Volta, E., "Amplitude Distortion in Transistor Amplifiers," Proc. IEE, *111* (March 1964), pp. 481–490.
4. Reynolds, J., "Nonlinear Distortions and their cancellation in Transistors," IEEE Trans. Elec. Devices, *16,* No. 11 (November 1965), pp. 595–599.
5. Phillips, A. B., *Transistor Engineering,* New York: McGraw-Hill, 1962, pp. 136–137.
6. Thomas, Lee C., "An Application of the Analog Computer to Electronic Circuit Simulation," IEEE Trans. Elec. Comp., *16,* No. 4 (August 1967).

7. Narayanan, S., "Transistor Distortion Analysis Using Volterra Series Representation," B.S.T.J., *46*, No. 5 (May–June 1967) pp. 991–1024.

APPENDIX A

Derivation of the Optimization Equations

In Fig. 6 the following relations hold

$$i_e = i_b + i_c \tag{51}$$

$$i_b = (V_s - v_e)/(R_s + r_b') \tag{52}$$

$$v_{\text{out}} = -R_L i_c \approx v_{cb} . \tag{53}$$

Let

$$V_s/(R_s + r_b') = I \tag{54}$$

$$r_1/(R_s + r_b') = \gamma \tag{55}$$

$$r_2/(R_s + r_b') = \delta \tag{56}$$

$$r_3/(R_s + r_b') = \xi. \tag{57}$$

Substituting (9) and (54) through (57) in (52)

$$i_b = I - \gamma i_e - \tfrac{1}{2}\delta i_e^2 - \tfrac{1}{6}\xi i_e^3 . \tag{58}$$

Combining (51) and (58)

$$i_c = i_e(1 + \gamma) + \tfrac{1}{2}\delta i_e^2 + \tfrac{1}{6}\xi i_e^3 - I. \tag{59}$$

Now i_c is given by equation 11. Therefore

$$\begin{aligned} i_e(1 + \gamma) + \tfrac{1}{2}\,\delta i_e^2 + \tfrac{1}{6}\xi i_e^3 - I \\ = \alpha_1 i_e + I_o M_1 v_{cb} + \tfrac{1}{2}\alpha_2 i_e^2 + \tfrac{1}{2} I_o M_2 v_{cb}^2 + \alpha_1 M_1 i_e v_{cb} \\ + \tfrac{1}{2}\alpha_1 M_2 v_{cb}^2 i_e + \tfrac{1}{2}\alpha_2 M_1 v_{cb} i_e^2 + \tfrac{1}{6}\alpha_3 i_e^3 + \tfrac{1}{6} I_o M_3 v_{cb}^3 . \end{aligned} \tag{60}$$

Substituting (51) and (53) into (60) and gathering terms:

$$\begin{aligned} & i_c^3[\tfrac{1}{6} I_o M_3 R_L^3 + \tfrac{1}{2}\alpha_2 M_1 R_L - \tfrac{1}{2}\alpha_1 M_2 R_L^2 - \tfrac{1}{6}\alpha_3 + \tfrac{1}{6}\xi] \\ & + i_c^2[\tfrac{1}{2}\,\delta - \tfrac{1}{2}\alpha_2 - \tfrac{1}{2} I_o M_2 R_L^2 + \alpha_1 M_1 R_L \\ & + i_b(\tfrac{1}{2}\xi - \tfrac{1}{2}\alpha_1 M_2 R_L^2 - \tfrac{1}{2}\alpha_2 M_1 R_L - \tfrac{1}{2}\alpha_3)] \\ & + i_c[1 + \gamma - \alpha_1 + I_o M_1 R_L + i_b(\delta - \alpha_2 + \alpha_1 M_1 R_L) \\ & + i_b^2(\tfrac{1}{2}\xi - \tfrac{1}{2}\alpha_3 + \tfrac{1}{2}\alpha_2 M_1 R_L)] + i_b[1 + \gamma - \alpha_1] \\ & + i_b^2[\tfrac{1}{2}\,\delta - \tfrac{1}{2}\alpha_2] + i_b^3[-\tfrac{1}{6}\xi - \tfrac{1}{6}\alpha_3] - I = 0 \end{aligned} \tag{61}$$

The only approximation that we have made up to this point is that $v_{out} \approx v_{cb}$, the collector-to-base voltage, assuming that v_s is small. Now we would like to express the variables of (61) in terms of the independent driving voltage, V_s, and the output current, i_c, which is linearly related to the output voltage. To accomplish this, we start with (58) and make the approximation

$$i_b \approx I - \gamma i_c \tag{62}$$

where we have ignored the high order terms in (58) and used the linear relation $i_c \approx i_e$.

Notice that (62) certainly does not imply that we have fixed a linear relationship between i_b, I, and i_c. We are simply using this new approximate variable in the highly nonlinear (61) for convenience. The approximation is justified by the fact that second order and higher terms ignored in (62) would appear as fourth order and higher terms in (61).

Substituting (62) into (61) we have

$$\begin{aligned}
&i_c^3\{\tfrac{1}{6}I_oM_3R_L^3 - \tfrac{1}{6}\alpha_3(1-\gamma)^3 \\
&\quad + \tfrac{1}{6}\xi(1-\gamma)^3 + \tfrac{1}{2}\alpha_2M_1R_L(1-\gamma)^2 - \tfrac{1}{2}\alpha_1M_2R_L^2(1-\gamma)\} \\
&\quad + i_c^2\{\tfrac{1}{2}\,\delta(1-\gamma)^2 + \alpha_1M_1R_L(1-\gamma) - \tfrac{1}{2}I_oM_2R_L^2 - \tfrac{1}{2}\alpha_2(1-\gamma)^2 \\
&\quad + I[\alpha_2M_1R_L(1-\gamma) + \tfrac{1}{2}\xi(1-\gamma)^2 - \tfrac{1}{2}\alpha_3(1-\gamma)^2 - \tfrac{1}{2}\alpha_1M_2R_L^2]\} \\
&\quad + i_c\{(1-\gamma^2) - \alpha(1-\gamma) + I_oM_1R_L \\
&\quad + I[\delta(1-\gamma) - \alpha_2(1-\gamma) + \alpha_1M_1R_L] \\
&\quad + I^2[\tfrac{1}{2}\xi(1-\gamma) - \tfrac{1}{2}\alpha_3(1-\gamma) + \tfrac{1}{2}\alpha_2M_1R_L]\} \\
&\quad + I[\gamma - \alpha_1] + I^2[\tfrac{1}{2}\,\delta - \tfrac{1}{2}\alpha_2] + I^3[\tfrac{1}{6}\xi - \tfrac{1}{6}\alpha_3] = 0.
\end{aligned} \tag{63}$$

At the 100 mA bias levels where we are assumed to be operating, $r_1 \leqq 0.5\ \Omega$. Also $r_b' \approx 10\text{–}20\ \Omega$ and R_s can only increase the $R_s + r_b'$ sum in (55). Hence $\gamma \ll 1$ and will be ignored in (63). Thus we have effectively substituted I for i_b in (61) to obtain (63). This substitution is *not* justified by requiring the assumption $I \gg \gamma i_c$ in (62) (that is, a current source drive); but is justified on the grounds that the substitution of (62) into (63) did not generate new terms in (63) for $\gamma \ll 1$. Equation (63) is of the form

$$ai_c^3 + i_c^2(b + cI) + i_c(d + eI + fI^2) + gI + hI^2 + jI^3 = 0. \tag{64}$$

Now to force linearity we would like to require

$$v_{out} = kV_s, \qquad \text{where } k \text{ is a constant.} \tag{65}$$

But from (53) and (54), (65) can be expressed in terms of the variables of (64) as

$$i_c = -\frac{k(R_s + r_b')}{R_L} I = BI. \tag{66}$$

Substituting (66) into (64) and gathering terms

$$I^3[aB^3 + cB^2 + fB + j] + I^2[bB^2 + eB + h] + I[dB + g] = 0. \tag{67}$$

Now I is an independent variable so that this equation can hold only if each coefficient is simultaneously zero. In the linear term

$$dB + g = 0$$
$$B = -\frac{g}{d}. \tag{68}$$

Ignoring terms in γ and noticing that $I_o M_1 R_L \ll 1$ in (63)

$$B \approx \frac{\alpha_1}{1 - \alpha_1 + I_o M_1 R_L}. \tag{69}$$

The constant B should be easy to identify. For small M_1 (low levels of V_o), $B = \beta_1$. However, at the higher values of I_o and V_o, $I_o M_1 R_L$ can be on the order of $(1 - \alpha_1)$. Thus, roughly speaking

$$B \geqq \tfrac{1}{2}\beta \gg 1. \tag{70}$$

Substituting (69) into (67) our final coefficients to be equated to zero in (67) become

$$aB^3 + cB^2 + fB + j = 0 \tag{71}$$
$$bB^2 + eB + h = 0. \tag{72}$$

Substituting for a, c, f, j in (71) by comparison between (64) and (63); ignoring terms in γ:

$$B^3[\tfrac{1}{6}I_o M_3 R_L^3 - \tfrac{1}{6}\alpha_3 + \tfrac{1}{6}\xi + \tfrac{1}{2}\alpha_2 M_1 R_L - \tfrac{1}{2}\alpha_1 M_2 R_L^2] + B^2[\alpha_2 M_1 R_L + \tfrac{1}{2}\xi - \tfrac{1}{2}\alpha_3 - \tfrac{1}{2}\alpha_1 M_2 R_L^2] + B[\tfrac{1}{2}\xi - \tfrac{1}{2}\alpha_3 + \tfrac{1}{2}\alpha_2 M_1 R_L] + [\tfrac{1}{6}\xi - \tfrac{1}{6}\alpha_3] = 0. \tag{73}$$

Gathering terms in R_L:

$$R_L^3[I_o M_3][\tfrac{1}{6}B^3] + R_L^2[-\alpha_1 M_2][B^3 + B^2] + R_L[\alpha_2 M_1][\tfrac{1}{2}B^3 + B^2 + B] + \xi[\tfrac{1}{6}B^3 + \tfrac{1}{2}B + \tfrac{1}{6}] - \alpha_3[\tfrac{1}{6}B^3 + \tfrac{1}{2}B^2 + \tfrac{1}{2}B + \tfrac{1}{6}] = 0 \tag{74}$$

Using (70) we can ignore lower powers of B, and $\alpha_2 M_1$, being the product of second order terms, is very small compared to the other coefficients in (74). Thus (74) becomes

$$(I_o M_3) R_L^3 - R_L^2(3\alpha_1 M_2) + \xi - \alpha_3 = 0. \tag{75}$$

Now substituting for b, e, and h in (72) by comparison between (63) and (64); ignoring terms in γ:

$$B^2[\tfrac{1}{2}\,\delta - \tfrac{1}{2}\alpha_2 + \alpha_1 M_1 R_L - \tfrac{1}{2} I_o M_2 R_L^2] + B[\delta - \alpha_2 + \alpha_1 M_1 R_L] + [\tfrac{1}{2}\,\delta - \tfrac{1}{2}\alpha_2] = 0.$$

Gathering terms in R_L:

$$R_L^2[-I_o M_2][\tfrac{1}{2}B^2] + R_L[\alpha_1 M_1][B^2 + B] + \delta[\tfrac{1}{2}B^2 + \tfrac{1}{2}B] - \alpha_2[\tfrac{1}{2}B^2 + \tfrac{1}{2}B] = 0. \tag{76}$$

Using (71), (76) becomes

$$-(I_o M_2) R_L^2 + (2\alpha_1 M_1) R_L + \delta - \alpha_2 = 0. \tag{77}$$

Equations (75) and (77) are the relations that must be satisfied to satisfy (67), which in turn results from the requirement of a linear input-output relation, (65).

APPENDIX B

Relating Current Gain Nonlinearities to the Bias Current

Riva[3] has shown that the small signal gain of a transistor can be closely matched to an expression of the form

$$\beta = h_{fe\max}[a \log_{10}^2 (I_c/I_{c\max}) + 2a \log_{10} e \log_{10} (I_c/I_{c\max}) + 1]^{-1}. \tag{78}$$

Where

$h_{fe\max}$ = maximum dc current gain
$I_{c\max}$ = collector current bias where $h_{fe\max}$ occurs
a = a constant characteristic of the transistor.

Differentiating the denominator of (78) reveals that the maximum ac current gain ($\beta_{\max}$) occurs for $I_c = I_{c\max}/e$. Call this current I_{cp}. Then

$$\beta = h_{fe\max}[a \log_{10}^2 (I_c/I_{cp}) - a \log_{10}^2 e + 1]^{-1}. \tag{79}$$

At the peak in the $\beta(I_c)$ curve, $I_c = I_{cp}$, and

$$\beta_{\max} = h_{fe\max}/(1 - a \log_{10}^2 e). \tag{80}$$

Substituting for h_{femax} in (79)

$$\beta = \beta_{\max}[1 + a(1 - a \log_{10}^2 e)^{-1} \log_{10}^2 (I_c/I_{cp})]^{-1}. \tag{81}$$

Then, for

$$A = a(1 - a \log_{10}^2 e)^{-1} \log_{10}^2 c \tag{82}$$

$$\beta = \beta_{\max}/[1 + A \ln^2 (I_c/I_{cp})] \tag{83}$$

$$\alpha = \frac{\beta}{1+\beta} = \frac{\alpha_{\max}}{1 + \dfrac{A\alpha_{\max}}{\beta_{\max}} \ln^2 \left(\dfrac{I_c}{I_{cp}}\right)}. \tag{84}$$

Where

$$\Delta I_c = \alpha \, \Delta I_e = \left(\alpha_1 + \frac{d\alpha}{dI_e} \Delta I_e + \frac{1}{2} \frac{d^2\alpha}{dI_e^2} \Delta I_e^2\right) \Delta I_e$$

$$\Delta I_c = \alpha_1 \, \Delta I_e + \frac{d\alpha}{dI_e} \Delta I_e^2 + \frac{1}{2} \frac{d^2\alpha}{dI_e^2} \Delta I_e^3$$

$$= \alpha_1 \, \Delta I_e + \tfrac{1}{2}\alpha_2 \, \Delta I_e^2 + \tfrac{1}{6}\alpha_3 \, \Delta I_e^3 .$$

Thus

$$\alpha_2 = 2 \frac{d\alpha}{dI_e} \tag{85}$$

and

$$\alpha_3 = 3 \frac{d^2\alpha}{dI_e^2} = 1.5 \frac{d\alpha_2}{dI_e}. \tag{86}$$

Now, taking $I_e \approx I_c$, from (84)

$$\alpha_2 = -\frac{4A\alpha_{\max}^2}{I_o \beta_{\max}} \frac{\ln (I_o/I_{cp})}{\left[1 + \dfrac{A\alpha_{\max}}{\beta_{\max}} \ln^2 \left(\dfrac{I_o}{I_{cp}}\right)\right]^2} \tag{87}$$

at $I_c = I_o$. In essentially all cases

$$0.04 < \frac{I_c}{I_{cp}} < 25$$

$$\beta_{\max} > 30$$

$$A\alpha_{\max} < 0.15.$$

Thus, to within 10 per cent in the most extreme case

$$\alpha_2 \cong -(4A\alpha_{\max}^2/\beta_{\max} I_o) \ln (I_o/I_{cp}). \tag{88}$$

Then, from (86)

$$\alpha_3 \cong (6A\alpha_{\max}^2/\beta_{\max} I_o^2) \ln (I_o/eI_{cp}). \tag{89}$$

Now to solve for A, notice that, from (78)

$$\beta\left(\frac{I_{c\,\max}}{e^2}\right) = h_{\text{femax}}[a \log_{10}^2 e^2 - 2a \log_{10} e \log_{10} e^2 + 1]^{-1} \quad (90)$$

$$= h_{\text{femax}}$$

But from (82) and (90)

$$h_{\text{femax}} = \beta_{\max}/(1 + A). \quad (91)$$

Therefore,

$$A = (\beta_{\max} - h_{\text{femax}})/h_{\text{femax}} \quad (92)$$

where h_{femax} may be measured at

$$I_c = I_{c\,\max}/e^2 = I_{cp}/e. \quad (93)$$

APPENDIX C

The Qualitative Distortion Model

The purpose of this appendix is to support the qualitative picture of algebraic distortion cancellation given in the text. The development here is not intended to be rigorous, but rather to strengthen the reader's ability to share the author's insight into the cancellation mechanism. We have argued that the net distortion current, D, is the algebraic difference between positive and negative distortion current components, A and B, dependent on output voltage and current, respectively. Express this relation as

$$D = A - B. \quad (94)$$

But, for A and B monotonic in voltage and current, the ratio A/B is a measure of the load. Define this measure as

$$R = \frac{A}{B}. \quad (95)$$

Now

$$D = B(R - 1). \quad (96)$$

On a dB basis

$$20 \log \frac{D}{B} = 20 \log |R - 1|. \quad (97)$$

Fig. 12 is a plot of 20 log $|R\text{-}1|$ as a function of R. Compare this plot with that of Fig. 2, which was measured in the laboratory. The simi-

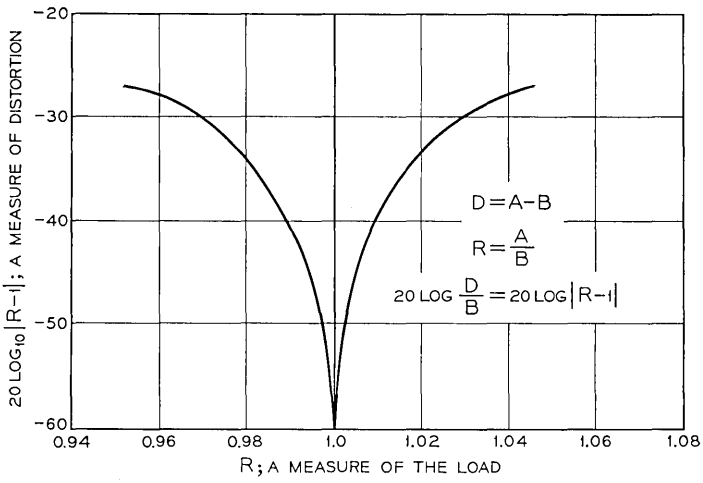


Fig. 12 — Decibel measure of the small difference between large numbers.

larity between the nature of these two minima adds additional weight to the idea that exact algebraic cancellation is involved in producing the net distortion frequencies. Thus any dependence of distortion on frequency should be compensated at distortion frequencies and not at input frequencies, since it is at the distortion frequency that cancellation takes place.

Identification and Synthesis of Linear Sequential Machines

By P. J. MARINO

(Manuscript received July 14, 1967)

This paper considers the identification and synthesis of linear sequential machines from their state transition tables. Necessary and sufficient conditions for linearity are derived which form the basis of identification tests. A sufficient condition leads to a method for coding the system's state vectors in a fashion consistent with linearity but which does not entail trial and error. The coding process is analytic in nature and allows the coding of state vectors independently of the coding or linearity of the output table. Both the Moore and Mealy models are considered in deriving coding procedures for the input and output vectors.

I. INTRODUCTION

This paper develops a method for identifying and synthesizing linear sequential machines using their state transition table representation. The basic objective is to construct a procedure which can be efficiently implemented by a digital computer. Towards that end, we develop simple and easily used preliminary tests which reject nonlinear systems to precede the time consuming synthesis, or state coding, process. The method for the coding of states is completely analytic, with the result that trial and error processes are not required.

Consider the symbolic state transition table, Table I.

The input vectors, u, have m components ($2 \leqq M \leqq 2^m$),* and the next state vectors, s_{ix}, and present state vectors, s_i, have n unspecified components ($N \leqq 2^n$). The vector components are defined over a modular field, and here this field is taken as $GF(2)$. Most of the results obtained below can be easily extended to other prime fields.

In terms of the state transition table, a linear sequential machine

* This paper considers tables which have at least two distinct columns of next states (nonautonomous systems).

is defined as a system which changes state according to the equation

$$s_{ix} = As_i + Bu_x . \tag{1}$$

A and B are $n \times n$ and $n \times m$ matrices, respectively.* A linear sequential machine is called fully linear when its symbolic output vector, z_{ix}, obeys

$$z_{ix} = Cs_i + Du_x \tag{2}$$

where C and D are matrices of proper size.

When D is the null matrix, the last two equations represent a Moore model of the linear system. Otherwise, the equations describe a Mealy model. Cohn and Even have given a method for model conversion in linear systems.[1]

TABLE I

	u_1	u_2	$\cdots$	u_x	$\cdots$	u_M
s_1	s_{11}	s_{12}	$\cdots$	s_{1x}	$\cdots$	s_{1M}
s_2	s_{21}	s_{22}	$\cdots$	s_{2x}	$\cdots$	s_{2M}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
s_i	s_{i1}	s_{i2}	$\cdots$	s_{ix}	$\cdots$	s_{iM}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
s_N	s_{N1}	s_{N2}	$\cdots$	s_{Nx}	$\cdots$	s_{NM}

In recent years linear sequential machines have been studied extensively. The motivation for this activity stems from two sources. Not only do linear sequential machines exhibit interesting mathematical and theoretical properties, but they have found a wide range of practical applications; for example, memory addressing circuits, computing over finite fields, counting and timing circuits, error correcting codes, encoding and decoding circuits, and generating pseudo-random and minimum-time test sequences.

As persistent research led to greater understanding, several investigators developed synthesis procedures for linear sequential machines. Davis and Brzozowski[2] have reported a method for the synthesis of nonsingular systems (systems in which, under each input, every pre-

* The addition and multiplication operations are modulo 2. Also, the entries in all matrices are from GF(2).

sent state goes into a unique next state). Their technique is based upon an iterative search over partitions of the system states.

In a mathematically elegant treatment, Cohn and Even[1] have derived a synthesis procedure which is free of trial-and-error processes. Coded output vectors are used to generate the state vector codes. Not only is it necessary that the system have a linear output, but the more severe restriction that the output vectors have been given, *a priori*, a linear coding is also required.

Yau and Wang[3] have disclosed a synthesis technique which does not require a linear and coded output. The construction of the A matrix by examination of a transition graph, which describes the state transitions owing to a given input, leads to the coding of the state vectors. The method requires the system to have 2^n states. When $N < 2^n$, a sufficient number of "don't care" states are introduced to complete the state transition table; however, no suitable procedure is given for the specification or coding of the "don't care" states. The lack of complete freedom from trial-and-error routines is another disadvantage of the method.

In this paper, necessary and sufficient conditions for linearity of the state transition table are derived which lead to the development of the procedure for coding the state vectors. The method accommodates linear systems in general (both singular and nonsingular). The synthesis procedure is analytic and, therefore, no trial-and-error routines are necessary. Also, the state vectors are coded independently of the output table so that the coding process is able to treat systems that have linear or nonlinear, coded or uncoded, output vectors. Both the Moore and Mealy models are considered in deriving coding procedures for the input and output vectors.

II. NECESSARY CONDITIONS FOR LINEARITY

Forming the sum of two next states, say s_{ix} and s_{iy}, under the same present state, s_i, yields

$$s_{ix} + s_{iy} = B(u_x + u_y),$$

since $A(s_i + s_i) = A0 = 0$, mod (2). Since the sum is independent of the present state, it follows that

$$s_{1x} + s_{1y} = s_{2x} + s_{2y} = \cdots = s_{ix} + s_{iy} = \cdots = s_{Nx} + s_{Ny}$$

for each x and y. Let the equality of these sums, for a particular x and y, be denoted by the term state sum, and call the individual sums

of pairs of next states component sums. For example, the state sum, $s_{11} + s_{12} = s_{21} + s_{22} = s_{31} + s_{32}$, consists of the component sums $s_{11} + s_{12}$, $s_{21} + s_{22}$ and $s_{31} + s_{32}$.

As a direct consequence of the state sum: if a present state has two identical next states, under two different inputs, then the columns which correspond to the inputs in question are identical, or the table represents a nonlinear machine.*

The state sums of a linear system must be consistent over all pairs of inputs. For example, assume that a state sum contains component sums (written in terms of present state symbols) $s_1 + s_2$ and $s_1 + s_3$. The state sum is consistent only if $s_2 = s_3$; however, if the output table does not allow the reduction of the state transition table by merging s_2 and s_3, then the state sum is inconsistent and the system is nonlinear.

In order to check state sums over the entire table only $M - 1$ state sums are required. Taking the input u_1 as a reference, the state sums

$$s_{11} + s_{1y} = \cdots = s_{i1} + s_{iy} = \cdots = s_{N1} + s_{Ny}$$

for $y = 2, 3, \ldots, M$ cover the table. Since, if $s_{j1} + s_{jy} = s_{i1} + s_{iy}$ for all y of interest, then for any x

$$\begin{aligned} s_{jx} + s_{jy} &= s_{jx} + s_{j1} + s_{jy} + s_{j1} \\ &= s_{ix} + s_{i1} + s_{iy} + s_{i1} \\ &= s_{ix} + s_{iy} \, . \end{aligned}$$

Therefore, it is not necessary to form state sums for all possible pairs of inputs. For example, consider Table II.

TABLE II

	u_1	u_2	u_3	u_4
s_1	s_2	s_1	s_3	s_4
s_2	s_4	s_3	s_1	s_2
s_3	s_3	s_4	s_2	s_1
s_4	s_1	s_2	s_4	s_3

From inputs u_1 and u_2, $s_1 + s_2 = s_3 + s_4$ (redundant components sums have been deleted) is consistent in the first two columns. From u_1 and

* A similar result has been obtained by Davis and Brozozowski[2] using a different approach.

u_3, $s_2 + s_3 = s_1 + s_4$ and from u_1 and u_4, $s_2 + s_4 = s_1 + s_3$. The last two state sums are rearrangements of the first and therefore, the state sums are consistent over the entire table.

Another symmetry feature appears in singular linear machines (those characterized by a singular A matrix*). If A is singular, then for some present states the rows of next states are identical.[2] This follows since a singular A has rank $r < n$, and therefore, the null space of A has dimension $n - r$, so that $As_i + Bu_x = s_{ix}$ has more than one solution for s_i given s_{ix} and u_2.

The reduced state transition table of a linear sequential machine has additional interesting properties. In what follows only reduced state transition tables are considered unless stated otherwise.

As a preliminary, consider

Theorem 1: If A is nonsingular, then the reduced table of a linear system has an even number of states.

Proof: If A is nonsingular, under each input, each state will appear once, and only once as a next state. Thus, the next state columns are permutations of the present state column. As a consequence, each of the state sums involves all of the system's states. If the number of states, N, were odd, then the same state must appear in two distinct component sums of the same state sum. That is, the state sum contains an equality $s_i + s_j = s_e + s_i$ which implies $s_j = s_e$. But this contradicts the statement that the state table is reduced.

Next, a starting result which connects the number of system states to the number of distinct inputs† is described by

Theorem 2: For a reduced, nonsingular, linear sequential machine which has N states, the number of distinct inputs cannot exceed 2, if $N/2$ is odd, or

$$2 + N \sum_{i=2}^{t} 2^{1-i}$$

where t is the smallest integer for which $N/2^t$ is odd.

Proof: Consider the state sums associated with the first two distinct inputs, u_1 and u_2.

$$s_{11} + s_{12} = s_{21} + s_{22} = \cdots = s_{i1} + s_{i2} = \cdots = s_{N1} + s_{N2} . \qquad (3)$$

* Common terms from linear and abstract algebra which appear in this paper are treated in several texts; for example, see Birkhoff and MacLane.[4]

† u_x is said to be distinct from u_y if and only if the columns of next states under u_x and u_y are distinct.

Taking u_1 with u_x, a third distinct input, the following state sum is obtained:

$$s_{11} + s_{1x} = s_{21} + s_{2x} = \cdots = s_{i1} + s_{ix} = \cdots = s_{N1} + s_{Nx} . \quad (4)$$

Writing segments of equations 3 and 4 in terms of present state symbols as $s_i + s_j = s_c + s_k = s_r + s_u = \ldots$, and $s_i + s_c = s_{j1} + s_{jx} = \ldots$, respectively, leads to the conclusion that state sum consistency requires that equation 4 contain a sum $s_j + s_k$. For if it did not, then locating the terms of equation 4, which contains s_k, say $s_k + s_l$, leads to $s_i + s_c + s_k + s_l = 0$. From equation 3 $s_i + s_j + s_k + s_c = 0$. For compatibility, $s_j = s_l$; so that equation 4 must contain $s_k + s_j$ or contradict the reduction of the table. It then follows that the state sums over distinct pairs of inputs must be mutually derivable via component sums.

Let the component sums obtained from the first state sum from u_1 and u_2 be denoted as follows:

$$S_1 = s_{11} + s_{12} , \cdots , S_i = s_{i1} + s_{i2} , \cdots , S_N = s_{N1} + s_{N2} .$$

Since there are $N/2$ distinct sums, let the symbols $\bar{S}_1 , \bar{S}_2 , \cdots , \bar{S}_{N/2}$ denote the distinct component sums. Then equation 3 can be represented by $\bar{S}_1 = \bar{S}_2 = \cdots = \bar{S}_{N/2} .$

In view of the foregoing, a necessary condition for linearity is that all other state sums must be derived from the sums $\bar{S}_1 , \cdots , \bar{S}_{N/2} .$ In generating new state sums the $\bar{S}_i$s are paired and the component sums which are consistent with equation 3 are formed by transposing terms in the resulting equation. For example, pairing $\bar{S}_i$ and $\bar{S}_j$ can yield either of the two equations which do not appear in 3:

$$s_{i1} + s_{j1} = s_{j2} + s_{i2} ,$$

or

$$s_{i1} + s_{j2} = s_{j1} + s_{i2} .$$

Then it is clear that each pairing of the $\bar{S}$s yields two possible state sums. Therefore, the number of unique pairings of the $\bar{S}$s, where each pairing occurs only once (this insures that no component sum will appear in two distinct state sums), is equal to half the maximum number of distinct inputs in excess of the first two.

Separating the $\bar{S}_i$ according to subscript parity gives:

$$\bar{S}_1 \quad \bar{S}_3 \cdots \bar{S}_{N/2-1}$$
$$\bar{S}_2 \quad \bar{S}_4 \cdots \bar{S}_{N/2} .$$

If $N/2$ is odd, then one $\bar{S}$ cannot be paired. Therefore, one sum will occur in more than one state sum. Since this is inconsistent with a reduced linear table, a system for which $N/2$ is odd can have only two distinct inputs. If $N/2$ is even, then the number of unique odd to even subscript pairings is $N/4$. The odd-to-odd and even-to-even pairings can be enumerated by considering a single row. It is advantageous to transform the subscripts as follows:

$$S'_1 = \bar{S}_2 , \qquad S'_2 = \bar{S}_4 , \cdots , S'_i = \bar{S}_{2i} .$$

Then, separating the new symbols by subscript parity gives:

$$S'_1 \quad S'_3 \cdots S'_{N/4-1}$$

$$S'_2 \quad S'_4 \cdots S'_{N/4} .$$

When $N/4$ is odd, no pairing is possible. If $N/4$ is even, the odd-even subscript pairings number $N/8$. Clearly, the odd-to-even pairings of the S' can be treated by reapplying the same transformation to the subscripts. Therefore, the number of allowed pairings is

$$N \sum_{i=2}^{t} 2^{-i},$$

where t is the smallest integer for which $N/2^t$ is odd, if $N/2$ is even. Since each pairing provides for the generation of two distinct columns, in addition to the first two columns, the number of distinct inputs is 2, if $N/2$ is odd or not greater than

$$2 + N \sum_{i=2}^{t} 2^{1-i}, \qquad \text{otherwise.}$$

This completes the proof.*

Theorem 1 leads to a very simple test for the identification of nonlinear tables. The number of states in the table is used to determine the maximum number of distinct inputs. Then, the table is rejected as nonlinear if the number of its distinct inputs, or, equivalently, the number of distinct columns, exceeds the maximum. Table III illustrates the restriction which linearity imposes upon the form of the state transition table.

There are similar, but weaker restrictions associated with the state transition tables of singular linear machines. Consider a system which has N states such that each next state column contains $d(<N)$ distinct states. (The singularity of the A matrix requires that some rows of

* A smaller upper bound can be obtained when $N \neq 2^n$. See the Appendix.

TABLE III

N	Maximum Number of Distinct Columns or Inputs
4	4
6	2
8	8
10	2
12	8
14	2
16	16
70	2
72	56
74	2
76	40
526	2
528	464
530	2
2086	2
2088	1568
2090	2
2092	1048
2094	2
2096	1836

state transition table to be identical.) If the singular matrix A has rank r, then the maximum number of present states which yields the same next state (that is, the maximum number of times a row can be repeated) cannot exceed 2^{n-r}. Since there are $N(\leqq 2^n)$ states, $N2^{r-n} \leqq d \leqq 2^r$. $r \neq n$ implies that each column cannot contain all of the system's states so that the reasoning of the last theorem cannot be applied. In order to gain some insight into how linearity limits the form of the state transition table, consider the case where $N = 2^n$. That is, the set of state vectors form a complete set of n-dimensional vectors with components over $GF(2)$. Let S denote the set of present states, $(s_1, s_2, \cdots, s_N)$, S_{1x} is defined as the set of distinct next states, $(s_{1x}, s_{2x}, \cdots, s_{dx})$ under the input u_x, and it is assumed that $u_1 = 0$. Consider the following

Theorem 3: *A linear system which is associated with a singular A matrix of rank r, a null input vector, and which has all states appearing as*

next states must have at least 2^{n-r} distinct input vectors, and S_{ix} and S_{iy} are either identical or disjoint for all x and y.

Proof: First, it will be shown that the S_{1x} are cosets of the group $\{S, +\}$. Since S is a vector space, 0, $s_i + s_j$ are members of S for any s_i , s_j which belong to S. If $u_1 = 0$, then the vectors of S_{11} are a subspace of S because:

(*i*) $A(0) = 0 \in S_{11}$ and
(*ii*) As_i , $As_j \in S_{11}$ implies $A(s_i + s_j) \in S_{11}$

(This follows since $s_i + s_j = s_k \in S$ and $As_k \in S_{11}$.)
The nonull input, u_x , generates cosets of the group $\{S, +\}$, because Bu_x is an n-component vector which must belong to S and therefore, $S_{11} + Bu_x = S_{1x}$.

It is well known that cosets are either disjoint or identical. Since $0 \in S_{11}$, $Bu_x \notin S_{11}$ implies that S_{11} and S_{1x} are disjoint. Therefore, no member of S_{11} can be used as Bu_x if the table is to have a column which contains states not found in column 1. If S_{11} and S_{1x} are to be disjoint, then $Bu_x \in (S - S_{11})$; that is, Bu_x can be selected from a set of $2^n - d$ vectors. Continuing, the next distinct coset is associated with an input, u_y , such that Bu_y has not appeared in any of the preceding cosets. (If it has, then $0 \in S_{11} + Bu_y$.) Then, Bu_y is among $2^n - 2d$ vectors. The last unique cosets is generated from a set of d vectors, or $2^n - kd = d$. So that there are $k + 1 = 2^n/d$ unique cosets of next states in the table. If each present state is to appear as a next state, the table which contains the minimum number of distinct columns must be comprised of one column from each unique coset. Since A has rank r, $d = 2^r$; consequently, there must be 2^{n-r} distinct inputs.

Also, since each unique coset can form 2^r distinct columns, the number of distinct inputs is not greater than 2^n, as expected.

This section has derived several properties that must be exhibited by the state transition table of a linear sequential machine. The consistent state sum requirement will play a central role in the code assignment problem.

III. SYNTHESIS: THE ASSIGNMENT OF LINEAR STATE CODES

To each symbolic state, s, it is necessary to assign a p-dimensional* vector, v, with components over $GF(2)$. That is, $s_{ix} \rightarrow v_{ix}$.

The vector assignment must preserve linearity;

$$v_{ix} = Av_i + Bu_x .$$

* $p \geqq n$

Linear systems must give rise to consistent state sums; therefore, the vectors must obey the same sums. Since state sums are only necessary conditions for linearity, a nonlinear state transition function may exhibit consistent state sums.

For any sequential machine the general state transition equation can be written as

$$v_{ix} = Av_i + Bu_x + f_{ix}\,, \tag{5}$$

where f_{ix} is a p-dimensional vector which is a nonlinear function of the present state and input vectors. When the symbolic states obey the state sums, (equation 3), the vectors must be assigned such that

$$v_{1x} + v_{1y} = v_{2x} + v_{2y} = \cdots = v_{ix} + v_{iy} = \cdots = v_{Nx} + v_{Ny}\,, \tag{6}$$

for each x and y. Therefore, from equation 5 it is clear that the nonlinear function obeys the same restriction,

$$f_{1x} + f_{1y} = f_{2x} + f_{2y} = \cdots = f_{ix} + f_{iy} = \cdots = f_{Nx} + f_{Ny}\,. \tag{7}$$

The selection of the A and B matrices exerts some control over the nonlinear function. When $u_1 = 0$,

$$A[v_1 \mid v_2 \mid \cdots \mid v_p] + [f_{11} \mid f_{21} \mid \cdots \mid f_{p1}] = [v_{11} \mid v_{21} \cdots \mid v_{p1}]$$

where

$$[v_1 \mid v_2 \mid \cdots \mid v_p]$$

is a $p \times p$ matrix whose columns are p linearly independent vectors.* Then,

$$f_{11} = f_{21} = \cdots = f_{p1} = 0 \tag{8}$$

can be achieved by

$$A = [v_{11} \mid v_{21} \mid \cdots \mid v_{p1}][v_1 \mid v_2 \mid \cdots \mid v_p]^{-1}$$

Similarly,

$$A[v_1 \mid v_1 \mid \cdots \mid v_1] + B[u_2 \mid u_3 \mid \cdots \mid u_{m+1}] + [f_{12} \mid f_{13} \mid \cdots \mid f_{1,m+1}] = [v_{12} \mid v_{13} \mid \cdots \mid v_{1,m+1}],$$

where $u_2, \ldots, u_{m+1}$ are m linearly independent input vectors, yields

$$f_{12} = f_{13} = \cdots = f_{1,m+1} = 0 \tag{9}$$

* In cases where the system is singular the matrix of next states (the matrix on the right side of the last equality) must be selected so that A has the required rank. The rank of A can be determined directly from the repetition of rows in the transition table.

when

$$B = [v_{12} + v_{11} \mid v_{13} + v_{11} \mid \cdots \mid v_{1,m+1} + v_{11}][u_2 \mid u_3 \mid \cdots \mid u_{m+1}]^{-1}.$$

Additional constraints on the nonlinear function become clear when equation 7 is examined in light of equations 8 and 9. For $x = 1$, a sample equality in equation 7 is

$$f_{i1} + f_{iy} = f_{j1} + f_{jy}.$$

When $i, j \leqq p$, $f_{i1} = f_{j1} = 0$ (by equation 8); so that $f_{iy} = f_{jy}$. Equation 9 indicates that $f_{iy} = 0$ for $y \leqq m + 1, j = 1$. Therefore, f_{iy} vanishes when $i \leqq p$ and $y \leqq m + 1$. Equation 7 implies $f_{ix} = f_{iy} = 0$ and $f_{jx} = f_{jy}$ for $x, y \leqq m + 1, i \leqq p < j$. Finally, $f_{iy} = f_{ix} + f_{jy}$ when $x \leqq m + 1 < y$, and $j \leqq p < i$. Table IV below summarizes the restrictions on the nonlinear function for systems which exhibit consistent state sums.

TABLE IV

	u_1	u_2	$\cdots$	u_{m+1}	u_{m+2}	$\cdots$	u_M
v_1	0	0	$\cdots$	0	$f_{1,m+2}$	$\cdots$	f_{1M}
v_2	0	0	$\cdots$	0	constant	$\cdots$	constant
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\downarrow$		$\downarrow$
v_p	0	0	$\cdots$	0	$f_{1,m+2}$	$\cdots$	f_{1M}
v_{p+1}	$f_{p+1,1}$	constant $\rightarrow$		$f_{p+1,1}$	$f_{p+1,1} + f_{1,m+2}$	$\cdots$	$f_{p+1,1} + f_{1M}$
$\vdots$	$\vdots$			$\vdots$	$\vdots$		$\vdots$
v_N	$f_{N,1}$	constant $\rightarrow$		f_{N1}	$f_{N1} + f_{1,m+2}$	$\cdots$	$f_{N1} + f_{1M}$

A particular code assignment can be verified by comparing state transitions along segments of one row and one column with the transitions predicted by the linear equation. If f_{j1} and f_{1x} are found to vanish for $p < j \leqq N$ and $m + 1 < x \leqq M$, respectively, then the code assignment is acceptable. The necessary state transition checks number $M + N - m - p - 1$ (compared with $MN - p - m - 1$ checks if state sum consistency is not verified before a code assignment is attempted). The implication is that sufficiently large values of m and p will force the nonlinear function to vanish over the entire table. While it is undesirable to increase m and p (since this requires more memory

elements) it is certainly possible to do so in principle.* However, in computing the A matrix it was convenient to construct a nonsingular matrix of p linearly independent vectors. The state sums (equation 6) which contain no more than $N/2$ component sums imply that at most $N/2 + 1$ states can be assigned vectors independently.

Therefore, not more than $N/2 + 1$ of the coded state vectors are linearly independent with the consequence that it is impossible to form a nonsingular matrix of coded state vectors when $p > N/2 + 1$. Where p exceeds this limit it is possible, in some cases, to express some elements of the A matrix in terms of the remaining elements. This method for finding A is far less attractive than forming a nonsingular matrix of coded state vectors; accordingly, the bound $p \leqq N/2 + 1$ will be enforced. The development which follows demonstrates that the limitation on p does not obscure a system's linearity. Similar considerations lead to $m \leqq M$.

Turning to the state coding problem, the state sum will play an important role in the generation of equations which lead to the linear coding of states. First, attention will be concentrated on nonsingular systems, then a later section will treat singular systems.

3.1 *Nonsingular Systems*

Consider the sum of two component sums

$$v_{ix} + v_{iy} + v_{jx} + v_{jy} = 0;$$

it is true that

$$\begin{aligned} A(v_{ix} + v_{iy} + v_{jx} + v_{jy}) \\ &= v_{ix1} + v_{iy1} + v_{jx1} + v_{jy1} + f_{ix1} + f_{iy1} + f_{jx1} + f_{jy1} \\ &= 0. \end{aligned}$$

(The additional subscript indicates that v_{ix} is the present state which goes into v_{ix1} under the null input, u_1.) Taking v_{ix}, v_{iy}, and v_{jx} among the p independent vectors implies $f_{ix1} = f_{iy1} = f_{jx1} = 0$. Furthermore, assigning vectors such that

$$v_{ix1} + v_{iy1} + v_{jx1} + v_{jy1} = 0 \tag{10}$$

forces f_{jy1} to vanish. The sum (10) must be consistent with the state sums (that is, no state sum can contain an equality of component sums which contradicts equation 10).

* When the input vectors are given as coded it is possible to increase their dimension by a translation.

By similar treatment of all component sums of a state sum, the nonlinear function can be made to vanish in the first column (and therefore, over the first $m + 1$ columns) provided that the attendant increase in the value of p (owing to the designation of independent vectors) does not cause it to exceed its bound.

If all component sums are treated, it is possible to generate an equation of the type 10 in which all four of the vectors have been previously designated linearly independent. Clearly, the equation contradicts the independence of one of the vectors; then, any one of the vectors must be deleted from the set of linearly-independent vectors. The nonlinear function corresponding to the deleted vector can be made to vanish by satisfying the type 10 equation which is obtained when the A matrix operates on the generated equation in question. Consider the following process for treating a single state sum.

(*i*) Select a component sum as a reference sum. Add another component sum to the reference sum.

(*ii*) Operate on the resulting sum with the A matrix to obtain an equation of the type in equation 10. Designate linearly-independent vectors as required and mark the vectors for which the associated nonlinear function has been forced to vanish.

(*iii*) Verify that the equation obtained in step *ii* is consistent with the state sum and the other equations obtained in *ii*. If all vectors in the equation generated in *ii* are linearly-independent, delete one of the vectors from the set of linearly-independent vectors and repeat step *ii* using the generated equation as the sum upon which A operates.

(*iv*) If one of the type 10 equations has three linearly-independent vectors, use it as the sum in repeating step *ii*. Otherwise, add another component sum to the reference sum and repeat step *ii*. Repeat *ii* and *iii* until an inconsistent equation is generated (the system is nonlinear), or until all vectors have been used (the system is linear).

The first time the process passes through step *ii*, three linearly-independent vectors are required; in subsequent passes at most one additional independent vector is needed. After the first pass through the process $N/2 - 2$ unused component sums remain. Therefore, not more than $N/2 + 1$ independent vectors are required for the process, precisely the upper bound on p. If the state vectors were coded using this process, the system would have an undesirably large number of memory elements. Therefore, this process is not an efficient design procedure. However, completion of the process implies linearity of the first $m + 1$ columns with the result that the A and B matrices can be

determined. Linearity of the table over the remaining columns is dependent upon the coding of the input vectors.

As previously indicated, linearity over all of the columns can be attained, in the worst case, by increasing the dimension of the input vectors. In this case, $m = M - 1$ would yield a linear system. If it is assumed that the input vectors are uncoded, or can be recoded, then it follows that completion of the process is a sufficient indication that the system is linear. (The problem of coding input vectors is treated in a later section.) The process will be referred to as the maximum memory process.

In order to illustrate the maximum memory process consider the reduced table, Table V.

TABLE V

	0	1
s_1	s_1	s_6
s_2	s_7	s_8
s_3	s_5	s_3
s_4	s_2	s_4
s_5	s_8	s_7
s_6	s_4	s_2
s_7	s_6	s_1
s_8	s_3	s_5

In terms of the coded vectors the state sum is

$$v_1 + v_6 = v_7 + v_8 = v_3 + v_5 = v_2 + v_4 \tag{11}$$

(which is consistent over the table). Using $v_1 + v_6$ as the reference sum,

$$A(v_1 + v_6 + v_7 + v_8)$$
$$= f_{1,0} + f_{6,0} + f_{7,0} + f_{8,0} + v_1 + v_4 + v_6 + v_3 = 0.$$

Designating v_1, v_6, and v_7 as linearly-independent vectors and satisfying the equation

$$v_1 + v_4 + v_6 + v_3 = 0 \tag{12}$$

yields

$$f_{1,0} = f_{6,0} = f_{7,0} = f_{8,0} = 0.$$

However, vectors v_1, v_6, and v_4 appear in two component sums of equation 11, the first and last; this implies

$$v_1 + v_6 + v_2 + v_4 = 0.$$

Adding this to equation 12 leads to $v_2 = v_3$. This contradicts the reduction of the table; therefore, equation 12 is inconsistent with the state sum. The system is nonlinear.

As pointed out, the maximum memory process is not a suitable vehicle for the economical design of linear sequential machines. The process extracts a limited amount of information from the state transition table. This can be improved by considering all of the unique state sums (that is, from the $M - 1$ state sums by pairing the first input with every other input), and using such equations as 10, which the process generates, to better advantage.

Consider the following procedure:

(*i*) Form the $M - 1$ unique state sums.

(*ii*) Select a reference sum from one of the state sums. Add another component sum (from the same state sum) to the reference sum.

(*iii*) Operate on the sum with the A matrix. Designate linearly-independent vectors as required. Obtain an equation like equation 10 and verify that it is consistent with the state sums. Mark the vectors in all state sums and equations of this type which have been guaranteed a linear state transition under the null input by this step. If all vectors in the equation obtained have been designated linearly independent, delete any one of these vectors from the set of linearly-independent vectors and repeat this step using the equation as the sum upon which A operates.

(*iv*) In the state sums where at least one component sum has had both vectors marked in step *iii*, search for a component sum which has one vector marked or, search over the type 10 equations for one which has three terms marked. If such a component sum or equation is found, use it in repeating step *iii*. (Since three of the vectors make a linear transition, the nonlinear function which is associated with the fourth vector can be made to vanish in step *iii* without designating another linearly-independent vector.) Otherwise,

(*v*) If the sum of the type 10 equation is unique and has two vectors marked, then use it in repeating step *iii*. Otherwise,

(*vi*) Form a new sum for use in step *iii* by adding the reference sum to another component sum (from the same state sum). Repeat step *iii*.

The process is repeated until all of the vectors have been marked in step *iii* (the system is linear) or until an inconsistent equation is generated in step *iii* (the system is nonlinear). The coding process for the state vectors is initiated by assigning arbitrary, but linearly independent, vectors to the state vectors so designated by passes through step *iii*. The remaining state vectors are coded using the type 10 equations which were generated in step *iii* in conjunction with the state sums.

The application of this synthesis procedure is more straightforward than its description would indicate. This is best illustrated by an example. Consider Table VI.

TABLE VI

	$\binom{0}{0}$	$\binom{1}{0}$	$\binom{0}{1}$
s_1	s_6	s_4	s_2
s_2	s_1	s_3	s_8
s_3	s_2	s_7	s_6
s_4	s_3	s_1	s_5
s_5	s_4	s_6	s_7
s_6	s_5	s_8	s_3
s_7	s_8	s_5	s_1
s_8	s_7	s_2	s_4
s_9	s_{11}	s_{12}	s_9
s_{10}	s_9	s_{10}	s_{11}
s_{11}	s_{10}	s_9	s_{12}
s_{12}	s_{12}	s_{11}	s_{10}

In terms of the coded vectors, the state sums are: from inputs $\binom{0}{0}$ and $\binom{1}{0}$,

$$v_6 + v_4 = v_1 + v_3 = v_2 + v_7 = v_5 + v_8 = v_{11} + v_{12} = v_9 + v_{10} , \quad (13)$$

from inputs $\begin{pmatrix}0\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$,

$$v_6 + v_2 = v_1 + v_8 = v_3 + v_5 = v_4 + v_7 = v_{11} + v_9 = v_{10} + v_{12} \ . \quad (14)$$

Equations 13 and 14 are mutually consistent, and both are consistent over the table.

The observation that $Av_{12} = v_{12}$ leads to the assignment of the null vector* to v_{12} (therefore, $f_{12,0} = 0$). Then in step *ii* of the synthesis process, take $v_{11} + v_{12}$ of equation 13 as the reference sum, and add it to $v_5 + v_8$ of equation 13. In step *iii* form

$$A(v_{12} + v_{11} + v_5 + v_8)$$
$$= f_{12,0} + f_{11,0} + f_{5,0} + f_{8,0} + v_{12} + v_{10} + v_4 + v_7 \ .$$

Since $f_{12,0} = 0$, taking v_{11} and v_5 among the linearly-independent vectors and the satisfying equation

$$v_{12} + v_{10} + v_4 + v_7 = 0 \quad (15)$$

leads to $f_{11,0} = f_{5,0} = f_{8,0} = 0$. Also, the set of vectors for which the nonlinear function vanishes, denoted by L, is

$$L = \{v_{12} \ , v_{11} \ , v_8 \ , v_5\}.$$

Equation 15 is the sum of two component sums of equation 14. Therefore, equation 15 is automatically satisfied.

At this point, component sums do not satisfy the conditions of step *iv*. In step *v* equation 15 is not unique. In step *vi* taking $v_2 + v_7$ with the reference sum yields

$$A(v_{12} + v_{11} + v_2 + v_7)$$
$$= f_{12,0} + f_{11,0} + f_{2,0} + f_{7,0} + v_{12} + v_{10} + v_1 + v_8 \ .$$

Since $f_{12,0} = f_{11,0} = 0$, designate v_2 as a linearly-independent vector. Then,

$$f_{2,0} = f_{7,0} = 0$$

if

$$v_{12} + v_{10} + v_1 + v_8 = 0.$$

The last equation is a rearrangement of terms in state sum (14) and therefore it is automatically satisfied.

* The assignment of the null vector is somewhat arbitrary. It has been shown (Yau and Wang[3]) that the null vector can be assigned to any state which is mapped into itself under the null input.

Updating the set L,

$$L = \{v_{12}, v_{11}, v_8, v_7, v_5, v_2\}.$$

The conditions of steps *iv* and *v* are not satisfied; in step *vi*, adding $v_1 + v_3$ and the reference sum yields

$$A(v_{12} + v_{11} + v_1 + v_3) = f_{12,0} + f_{11,0} + f_{1,0} + f_{3,0} + v_{12} + v_{10} + v_6 + v_2 .$$

Since $f_{12,0}$ and $f_{11,0}$ have been shown to vanish, including v_1 among the linearly-independent vectors leads to

$$f_{1,0} = f_{3,0} = 0$$

if

$$v_{12} + v_{10} + v_6 + v_2 = 0.$$

The last equation is the sum of two component sums of equation 14. The set L becomes

$$L = \{v_{12}, v_{11}, v_8, v_7, v_5, v_3, v_2, v_1\}.$$

In equation 14, $v_1 + v_8$, and $v_3 + v_5$ have both vectors marked in step *iii* while $v_6 + v_2$, $v_4 + v_7$, $v_{11} + v_9$ and $v_{10} + v_{12}$ each have one vector marked.

In step *iv*, forming

$$A(v_1 + v_8 + v_6 + v_2) = f_{1,0} + f_{8,0} + f_{6,0} + f_{2,0} + v_6 + v_7 + v_5 + v_1 .$$

Since $f_{1,0}$, $f_{2,0}$, and $f_{8,0}$ have been shown to vanish, $f_{6,0} = 0$ if

$$v_6 + v_7 + v_5 + v_1 = 0. \tag{16}$$

Equation 16 is unique and consistent with the state sums.

$$L = \{v_{12}, v_{11}, v_8, v_7, v_6, v_5, v_3, v_2, v_1\}.$$

Proceeding more quickly, A operating on $v_1 + v_8 + v_4 + v_7$ (from step *iv*) yields $f_{4,0} = 0$ and

$$v_6 + v_7 + v_3 + v_8 = 0. \tag{17}$$

The last equation is unique and consistent with the state sums. Update the set L; let A operate on $v_1 + v_8 + v_{11} + v_9$ (from step *iv*) to obtain $f_{9,0} = 0$ and

$$v_6 + v_7 + v_{10} + v_{11} = 0 \tag{18}$$

Equation 18 is unique and consistent and also has three vectors in the set L. Update the set L.

From step *iv*, $A(v_6 + v_7 + v_{10} + v_{11})$ yields $f_{10,0} = 0$ and $v_5 + v_8 + v_9 + v_{10} = 0$ (which is the sum of two component sums of equation 14). Update the set L. Step *iv* gives $v_{11} + v_{12} + v_9 + v_{10}$ which, when operated on by A, yields

$$f_{10,0} = 0 \qquad \text{and} \qquad v_{10} + v_{12} + v_{11} + v_9 = 0$$

(which is the sum of two component sums). L contains all of the state vectors; therefore, the system can be coded.

There are four linearly-independent vectors ($p = n = 4$). Make the following assignment of the linearly-independent vectors:

$$v_{11} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \qquad v_5 = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \qquad v_2 = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \qquad v_1 = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

Since $v_{12} = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}$, all of the component sums in equation 13 equal

$$v_{11} + v_{12} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}.$$

Then, from equation 13 it follows that

$$v_8 = v_5 + \begin{bmatrix}1\\0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\1\\0\\0\end{bmatrix},$$

$$v_7 = v_2 + \begin{bmatrix}1\\0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\0\\1\\0\end{bmatrix},$$

and

$$v_3 = v_1 + \begin{bmatrix}1\\0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\0\\0\\1\end{bmatrix}.$$

Similarly, the component sums in equation 14 are each equal to $v_9 +$

$v_{11} = \begin{bmatrix}1\\1\\0\\1\end{bmatrix}$ with the result

$$v_6 = v_2 + \begin{bmatrix}1\\1\\0\\1\end{bmatrix} = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

$$v_4 = v_7 + \begin{bmatrix}1\\1\\0\\1\end{bmatrix} = \begin{bmatrix}0\\1\\1\\1\end{bmatrix}$$

$$v_9 = v_{11} + \begin{bmatrix}1\\1\\0\\1\end{bmatrix} = \begin{bmatrix}0\\1\\0\\1\end{bmatrix}$$

$$v_{10} = v_{12} + \begin{bmatrix}1\\1\\0\\1\end{bmatrix} = \begin{bmatrix}1\\1\\0\\1\end{bmatrix}.$$

To calculate A consider

$$A[v_{11} \mid v_5 \mid v_2 \mid v_1] = [v_{10} \mid v_4 \mid v_1 \mid v_6]$$

$$A\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} = A = \begin{bmatrix}1&0&0&1\\1&1&0&1\\0&1&0&1\\1&1&1&1\end{bmatrix}.$$

The matrix B satisfies

$$B\begin{bmatrix}1&0\\0&1\end{bmatrix} = [v_4 + v_6 \mid v_2 + v_6],$$

with the result

$$B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since equation 16 and 18 are consistent with the state sums and since they were not used in the coding process, they are redundant. It is easy to verify that the equations are satisfied by the code assignment.

The process produced a code assignment for which the coded state vectors had the minimum number of components. This will not be true for all tables. Considering the way the process extracts information from the table leads to the conclusion that the attainment of the minimum component coding vector depends upon the connectivity of the sequential machine. In a case where the machine is not strongly connected there is a possibility the process will require $p > n$.* (The last example involved a machine which was not strongly connected.) In order to illustrate this consider Table VII.

TABLE VII

	0	1		0	1
s_1	s_6	s_4	s_9	s_{11}	s_{12}
s_2	s_1	s_3	s_{10}	s_9	s_{10}
s_3	s_2	s_7	s_{11}	s_{10}	s_9
s_4	s_3	s_1	s_{12}	s_{12}	s_{11}
s_5	s_4	s_6	s_{13}	s_{14}	s_{15}
s_6	s_5	s_8	s_{14}	s_{15}	s_{14}
s_7	s_8	s_5	s_{15}	s_{13}	s_{16}
s_8	s_7	s_2	s_{16}	s_{16}	s_{13}

The coded vectors obey the state sum:

$$v_4 + v_6 = v_1 + v_3 = v_2 + v_7 = v_5 + v_8$$
$$= v_{11} + v_{12} = v_9 + v_{10} = v_{13} + v_{16} = v_{14} + v_{15} . \qquad (19)$$

* A sequential machine is said to be strongly connected if it is possible to reach any state of the system starting in any initial state.

Taking $v_{16} = 0$ leads to the seection of $v_{13} + v_{16}$ as the referencc sum since it is an advantage to exploit the fact that $f_{16,0} = 0$. For brevity, the results obtained from step *iii* of the synthesis process are shown in Table VIII.

TABLE VIII

The Sum A Operates Upon	The Equation Obtained	Additions to The Set L	Linearly Independent
$v_5 + v_{13} + v_8 + v_{16}$	$v_{16} + v_7 + v_{14} + v_4 = 0$	v_{16} , v_{13} , v_8 , v_5	v_5 , v_{13}
$v_{16} + v_7 + v_{14} + v_4$	$v_{16} + v_8 + v_3 + v_{15} = 0$	v_{14} , v_7 , v_4	v_7 , v_{14}
$v_{16} + v_{13} + v_6 + v_4$	$v_{16} + v_{14} + v_5 + v_3 = 0$	v_6	
$v_{16} + v_{13} + v_{14} + v_{15}$	$v_{16} + v_{14} + v_{13} + v_{15} = 0$	v_{15}	
$v_{16} + v_{13} + v_7 + v_2$	$v_{16} + v_{14} + v_1 + v_8 = 0$	v_2	
$v_{16} + v_{14} + v_1 + v_8$	$v_{16} + v_{15} + v_6 + v_7 = 0$	v_1	
$v_{16} + v_{13} + v_1 + v_3$	$v_{16} + v_{14} + v_6 + v_2 = 0$	v_3	

At this point it is observed that v_9, v_{10}, v_{11}, and v_{12} are not in L, and more importantly, it is not possible to involve these vectors in a relationship by application of step *iii* without introducing another linearly-independent vector. By continuing the process it can be demonstrated that the system is linear.

$$A(v_{16} + v_{13} + v_{11} + v_{12}) \quad \text{yields} \quad f_{11,0} = f_{12,0}$$

when

$$v_{16} + v_{14} + v_{10} + v_{12} = 0. \tag{20}$$

Then, $A(v_{16} + v_{14} + v_{10} + v_{12})$ leads to $f_{10,0} = f_{12,0}$

when

$$v_{16} + v_{15} + v_{12} + v_9 = 0. \tag{21}$$

A, operating on the last equation, gives $f_{9,0} = f_{12,0}$

when

$$v_{16} + v_{13} + v_{12} + v_{11} = 0 \qquad \text{(automatically satisfied).} \tag{22}$$

The system is linear since designating v_{12} as a linearly-independent vector leads to

$$f_{12,0} = f_{9,0} = f_{10,0} = f_{11,0} = 0.$$

It is also of interest that the set L which was obtained by the process can be coded using the four linearly-independent vectors. This will be true in general.

In order to insure realization of the transition table with the least number of memory elements, it is important to develop a means for keeping p at its minimum value, n. First, a general method for the reduction of the number of vector components will be developed. Then, the method will be applied to the problem at hand.

Let the set L_n denote the largest set L generated by the synthesis process using n linearly independent vectors; let $\bar{L}_n$ denote the set of vectors which require additional linearly independent vectors in order to become members of L. The n-component vectors y are members of L_n, and the n-component vectors $\bar{y}$ are in $\bar{L}_n$.

The members of L_n can be coded. Taking the first n vectors of L_n as linearly independent (since order is unimportant), the calculation of A, after the coding, leads to

$$A[y_1 \mid y_2 \mid \cdots \mid y_n] = [y_{10} \mid y_{20} \mid \cdots \mid y_{n0}].$$

This can be simplified by coding the linearly-independent vectors such that $[y_1 \mid y_2 \mid \cdots \mid y_n] = I_n$ (the $n \times n$ identity matrix). Then,

$$A = [y_{10} \mid y_{20} \mid \cdots \mid y_{n0}].$$

Suppose the set L_{n+1} is tentatively formed by designating another linearly-independent vector. It is clear that the vectors in L_{n+1} (and $\bar{L}_{n+1}$) have $n + 1$ components. In order to preserve the coding of L_n take $\begin{pmatrix} y \\ 0 \end{pmatrix} \varepsilon L_{n+1}$ where $y \varepsilon L_n$. That is, the vectors which have been coded over n linearly-independent vectors are increased by one component (which is taken as zero. Members of $\bar{L}_n$ which become members of L_{n+1} will be denoted as $\begin{pmatrix} \bar{y} \\ 1 \end{pmatrix}$. Let $\begin{pmatrix} \bar{y}_{n+1} \\ 1 \end{pmatrix}$ denote the $(n+1)^{\text{th}}$ linearly-independent vector where $\bar{y}_{n+1}$ is any coded vector not in L_n. The $(n + 1) \times (n + 1)$ matrix $\bar{A}$ is given by

$$\bar{A}\left[\begin{array}{c|c|c|c|c} y_1 & y_2 & \cdots & y_n & \bar{y}_{n+1} \\ 0 & 0 & & 0 & 1 \end{array}\right] = \left[\begin{array}{c|c|c|c|c} y_{10} & y_{20} & \cdots & y_{n0} & \bar{y}_{n+1,0} \\ 0 & 0 & & 0 & 1 \end{array}\right]^*.$$

* If $\bar{y}_{n+1}$ and $\bar{y}_{n+1,0}$ are in different sets (L_n or $\bar{L}_n$), then the known A matrix can be used to determine the coding of the vector which is in $\bar{L}_n$.

From the previous coding of L_n, it follows that the left side of the last equation can be written as

$$\bar{A}\left[\begin{array}{c|c} I_n & \bar{y}_{n+1} \\ \hline 0 & 1 \end{array}\right].$$

Therefore, it can be verified that

$$\bar{A} = \left[\begin{array}{c|c} A & A\bar{y}_{n+1} + \bar{y}_{n+1,0} \\ \hline 0 & 1 \end{array}\right]. \tag{23}$$

Observe that if $\bar{A}$ has the form

$$\bar{A} = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & 1 \end{array}\right], \tag{24}$$

then $\bar{A}$ operating on any vector which has the form $\begin{pmatrix} \bar{y} \\ 1 \end{pmatrix}$ will yield a vector of the same form. Similarly, all vectors $\begin{pmatrix} y \\ 0 \end{pmatrix}$ are mapped into another vector where the last component is zero. Since all of the n component vectors y (of L_n) and all n component $\bar{y}$ vectors have different codes, then the last component of the vectors in L_{n+1} can be deleted. Also, the matrix A (in equation 23) is the required matrix.

In view of the foregoing, a code transformation, acting on the coded $\bar{y}$, must be found such that $\bar{A}$ has the form of equation 24. It is well known (for example, Cohn and Even[1]) that the code transformation $y' = Ry$, where R is a nonsingular matrix, cannot alter the linearity of a system. From the state transition equation 1 it is easy to show that this type of code transformation produces a new matrix of the form $R\bar{A}R^{-1}$.

It is required to find an R such that

$$R\bar{A}R^{-1} = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & 1 \end{array}\right]. \tag{25}$$

Comparison of equations 23 and 24 indicates that R must have the form

$$R = \left[\begin{array}{c|c} I_n & T \\ \hline 0 & 1 \end{array}\right]. \tag{26}$$

Using equation 26 in equation 25 leads to the following restriction on the vector T:

$$A\bar{y}_{n+1} + \bar{y}_{n+1,0} = (A + I_n)T. \tag{27}$$

Also, applying the coding transformation yields

$$R\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

and

$$R\begin{pmatrix} \bar{y} \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{y} + T \\ 1 \end{pmatrix}.$$

$\bar{y} + T$ cannot be a member of L_n if the $(n + 1)^{\text{th}}$ component is to be deleted.

Therefore, the process of reducing the vector components by one is:

(*i*) Determine the vectors T which satisfy (27).
(*ii*) Of the vectors obtained in *i*, select one which preserves the identity of $\bar{L}_n$.

Returning to the example, it can be verified that the following is an acceptable code for all of the states except v_9, v_{10}, v_{11}, and v_{12}:

$$v_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad v_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad v_5 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$v_6 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad v_7 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad v_8 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad v_{13} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$v_{14} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad v_{15} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad v_{16} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The corresponding A matrix is

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

The set $\bar{L}_4$ is $\{v_9, v_{10}, v_{11}, v_{12}\}$. Set v_9 equal to any vector not in L_4, for example

$$v_9 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

From the table $v_{9,0} = v_{11}$. To determine v_{11}, add equations 21 and 22. The result is

$$v_{11} = v_9 + v_{15} + v_{13} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Then,

$$Av_9 + v_{9,0} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$

and

$$A + I_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

Substituting in equation 27 yields

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} T = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

An acceptable solution is

$$T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

since the new v_9, $\begin{bmatrix}1\\0\\1\\0\end{bmatrix} + \begin{bmatrix}0\\0\\0\\1\end{bmatrix} = \begin{bmatrix}1\\0\\1\\1\end{bmatrix}$, is not a member of L_4. From equations 20 through 22,

$$v_9 = \begin{bmatrix}1\\0\\1\\1\end{bmatrix}, \quad v_{10} = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}, \quad v_{11} = \begin{bmatrix}1\\0\\1\\0\end{bmatrix}, \quad \text{and} \quad v_{12} = \begin{bmatrix}1\\1\\1\\0\end{bmatrix}.$$

It can be verified that this is an acceptable coding.

3.2 *Singular Systems*

The synthesis procedure which was developed in the last section can be readily extended to treat the coding of singular systems.

Consider a state transition table that has d ($d < N$) distinct rows. For linear systems, tables of this type have columns of next states comprised of sets of d distinct entries. Therefore, the state sums (with respect to the first column) have the form

$$v_{11} + v_{1x} = \cdots = v_{i1} + v_{ix} = \cdots = v_{d1} + v_{dx}, \quad \text{for } x = 2, 3, \cdots, M. \tag{28}$$

In general a given table can have columns that are disjoint sets of states. Also, when $2d < N$, these columns can generate state sums such that the only state vectors two state sums may have in common are from the first column. State sums of this type will be called distinct state sums.

An attempt to apply the synthesis algorithm of the last section to a table which has distinct state sums can lead to undesirable results. Since the equations determined in step *iii* of the algorithm can contain only the d distinct vectors of the first column, and since all state vectors must be guaranteed a linear transition to column one, it is clear that in step *iv* at least one linearly-independent vector must be designated for each distinct state sum. As a consequence, a system coding that requires a large number of components may result. In order to avoid this situation it is useful to introduce the concept of independent inputs.

The set of independent inputs is the set of input vectors that spans all of the input vectors. That is, if $u_2, \ldots, u_{m+1}$ are the independent inputs (the set of linearly-independent input vectors), then any other

vector, say u_y, can be written as

$$u_y = \sum_{x=2}^{m+1} a_{yx} u_x ,$$

where the a_{yx} are constants which can be arbitrarily assigned when the inputs are uncoded, or they can be determined by inspection of u_y when the inputs are coded.

Recall that $Bu_y = v_{iy} + v_{i1}$, for any i; allowing B to operate on the expansion for u_y yields

$$v_{iy} + v_{i1} = \sum_{x=2}^{m+1} a_{yx}(v_{j(x)x} + v_{j(x)1}), \tag{29}$$

where each $j(x)$ can assume any row index.

The state vector v_{iy} can be forced to make a linear transition under the null input if all vectors in column one and all of the $v_{j(x)}x$ have been guaranteed a linear transition in step *iii*, and if (29) holds. To demonstrate this, use the A matrix to operate on the sum of $v_{iy} + v_{i1}$ and another component sum; then, compare the result to that obtained by repeating the operation after $v_{iy} + v_{i1}$ has been replaced by its expansion. The equations of the type (10) which are generated are required to meet the conditions of step *iii*. Since each distinct state sum not associated with an independent input can be treated in this way, these state sums meet the conditions of step *iv* without the designation of a linearly-independent state vector. Also, since all vectors in column one make linear transitions, all other vectors in the state sum can be treated via step *iv*.

After the independent inputs have been identified (for coded inputs), or designated (for uncoded inputs) the synthesis procedure of the last section* can be applied over the state sums associated with independent inputs. The remaining vectors could be treated by employing equations of the type (29) and entering the synthesis process at step *iv*.

A modified synthesis procedure of this nature would require consideration of the equations of the type (29) in the coding process. The modifications are compatible with nonsingular systems.

The remainder of this section develops an alternative synthesis procedure that is better suited to capitalize on the redundancies found in transition tables of singular systems.

* It is necessary to provide for the selection of another distinct state sum after all component sums of the state sum under consideration have been exhausted. This can be accomplished by selecting the reference sum in step *vi* from another state sum.

When, for example, columns 1 and 2 are disjoint sets of states, (28) contains $2d$ distinct vectors. It is asserted that the $2d$ vectors of this distinct state sum can be coded over the $r + 1$ components (where r is the rank of the A matrix) by a modified form of synthesis procedure of the last section and that the remaining vectors can be coded by a simple relationship. Since only one distinct state sum is to be considered, the information concerning the designation and deletion of linearly-independent vectors (in step *iii*) which is implicit in the other state sums must be incorporated. This is readily accomplished by generating equations of the type (10) over the remaining state sums and equations like (29). Then, the unique equations, which must be compatible with state sums, are used to augment the state sum; that is, these equations, as well as the state sum, are used as state sums in coding over $r + 1$ components. (Notice that the equations in question contain only vectors from column one.)

Partition the vectors such that the upper partition contains $r + 1$ components. That is,

$$v_{ix} = \begin{bmatrix} v'_{ix} \\ \hline v''_{ix} \end{bmatrix}.$$

Then, using the augmenting equations,

$$v'_{11} + v'_{12} = \cdots = v'_{i1} + v'_{i2} = \cdots = v'_{d1} + v'_{d2}$$

and the state sums formed by columns which may appear in the table that are permutations of the first or second columns in the synthesis process, will yield a coding of these vectors and an $(r+1) \times (r+1)$ A matrix. Let A' denote this matrix of rank r. The relations

$$N2^{r-n} \leqq d \leqq 2^r$$

(from Section II) and

$$2^{n-1} < N \leqq 2^n$$

imply

$$2^r < 2d \leqq 2^{r+1} \tag{30}$$

which verifies that $r + 1$ components are required to code the vectors in the first two columns.

Before continuing, it is convenient to separate the present states into sets such that all members of a particular set give rise to identical

rows of next states. Let V_j denote the set of present states which are associated with the j^{th} row. If one or more members of the set has been coded over $r + 1$ components, then select one such $r + 1$ component vector as the characteristic vector of the set. Denote it by the symbol cv_j'. If no vector in V_j has been coded over $r + 1$ components, the characteristic vector can be determined from

$$A'cv_j' = v_{j1}' .$$

This last equation has at least one solution for cv_j' since the columns of A' must span all of the v' vectors of the first column. Also, the cv_j' so determined is not a characteristic vector of any other set. Then, taking the system A matrix $(n \times n)$ as

$$A = \left[\begin{array}{c|c} A' & 0 \\ \hline 0 & 0 \end{array}\right]$$

will make the linearity of the transitions in the first column dependent only upon the coding over v'. For the vectors which have been coded over v', set the remaining components, the v'', equal to the $n-r-1$ component null vector. (This is necessary, if $u_1 = 0$.) That is,

$$v_{i1}'' = v_{i2}'' = 0 \qquad \text{for} \qquad i = 1, 2, \cdots, d.$$

Thus, all vectors in the first two columns are completely coded.

The remaining vectors can be coded by the following process. Consider the column of uncoded vectors under an independent input u_y. From the state sum of u_1 and u_y, it follows that

$$v_{1y} = v_{11} + v_{1y} + v_{j1} , \quad \text{for} \quad j = 2, 3, \cdots, d. \tag{31}$$

Since v_{11} and v_{j1} are known, once v_{1y} is coded, the coding of all other vectors in the column is determined by equation 31. v_{1y}' can be set equal to the characteristic vector of the set V in which v_{1y} (as present state) belongs. This will insure that the present state v_{1y} makes a linear transition under input u_1 . v_{1y}'' can be set equal to any $n - r - 1$ component vector that has not been previously used as a v'' vector. When u_y is not an independent input, v_{1y} can be obtained from equation 29 after the independent inputs have been treated; then equation 31 gives the the coding of the remaining vectors in column y.

To illustrate the process consider Table IX.

TABLE IX

	$\begin{bmatrix}0\\0\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\1\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$
s_1	s_1	s_3	s_5	s_7	s_9
s_2	s_2	s_4	s_6	s_8	s_{10}
s_3	s_1	s_3	s_5	s_7	s_9
s_4	s_2	s_4	s_6	s_8	s_{10}
s_5	s_1	s_3	s_5	s_7	s_9
s_6	s_2	s_4	s_6	s_8	s_{10}
s_7	s_1	s_3	s_5	s_7	s_9
s_8	s_2	s_4	s_6	s_8	s_{10}
s_9	s_1	s_3	s_5	s_7	s_9
s_{10}	s_2	s_4	s_6	s_8	s_{10}

The state sums are

$$v_1 + v_3 = v_2 + v_4 ,$$
$$v_1 + v_5 = v_2 + v_6 ,$$
$$v_1 + v_7 = v_2 + v_8 ,$$
$$v_1 + v_9 = v_2 + v_{10} .$$

The state sums are consistent over the table. Take the second, third, and fourth inputs as independent. All state sums and equations like (29) generate augmenting equations identical to the null vector. To determine r, note $d = 2$; therefore, from equation 30, $r = 1$. The vectors in the first state sum are coded over two components using the synthesis process. The result is

$$v_1' = \begin{pmatrix}0\\0\end{pmatrix}, \quad v_2' = \begin{pmatrix}1\\0\end{pmatrix}, \quad v_3' = \begin{pmatrix}0\\1\end{pmatrix}, \quad v_4' = \begin{pmatrix}1\\1\end{pmatrix},$$

and

$$A' = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}.$$

The sets V are:

$$V_1 = \{v_1 , v_3 , v_5 , v_7 , v_9\},$$
$$V_2 = \{v_2 , v_4 , v_6 , v_8 , v_{10}\}.$$

The characteristic vectors are taken as v_1' and v_2', respectively; that is,

$$cv_1' = \begin{pmatrix}0\\0\end{pmatrix} \quad \text{and} \quad cv_2' = \begin{pmatrix}1\\0\end{pmatrix}.$$

The vectors in the first two columns are fully coded by setting the last two components of each vector to zero:

$$v_1 = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \quad v_2 = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad v_3 = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad v_4 = \begin{bmatrix}1\\1\\0\\0\end{bmatrix},$$

and

$$A = \begin{bmatrix}1 & 0 & 0 & 0\\0 & 0 & 0 & 0\\0 & 0 & 0 & 0\\0 & 0 & 0 & 0\end{bmatrix}.$$

From the second state sum,

$$v_5 = v_1 + v_2 + v_6 .$$

Take

$$v_6'' = \begin{pmatrix}1\\0\end{pmatrix}.$$

Since $v_6 \in V_2$,

$$v_6' = cv_2' = \begin{pmatrix}1\\0\end{pmatrix}$$

therefore,

$$v_6 = \begin{bmatrix}1\\0\\1\\0\end{bmatrix}$$

and

$$v_5 = \begin{bmatrix}1\\0\\0\\0\end{bmatrix} + \begin{bmatrix}1\\0\\1\\0\end{bmatrix} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}.$$

Similarly, taking

$$v_8'' = \begin{pmatrix}1\\1\end{pmatrix}$$

and observing that

$$v_8' = cv_2' \quad \text{and} \quad v_7 = v_1 + v_2 + v_8$$

lead to

$$v_8 = \begin{bmatrix}1\\0\\1\\1\end{bmatrix} \quad \text{and} \quad v_7 = \begin{bmatrix}0\\0\\1\\1\end{bmatrix},$$

respectively. Since the last input is the sum of the third and fourth inputs, $v_1 + v_9 = v_1 + v_5 + v_1 + v_7$. Then,

$$v_9 = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \quad \text{and} \quad v_{10} = \begin{bmatrix}1\\0\\0\\1\end{bmatrix}.$$

Finally,

$$B = \begin{bmatrix}0 & 0 & 0\\1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}.$$

IV. CODING THE OUTPUT VECTORS

4.1 *The Mealy Model*

Let z_{ix} denote the symbolic output vector for input u_x and present state s_i. Let the output table contain L distinct vectors; then, z_{ix} is an l-component vector ($L \leqq 2^l$) over $GF(2)$. A k-dimensional vector ($l \leqq k$), w_{ix}, is to be assigned to each symbolic output vector.

If the output is linear,

$$w_{ix} = Cv_i + Du_x . \tag{32}$$

This equation implies that all of the equations obtained in terms of the coded state vectors (state sums and equations derived in the coding process) can be converted directly into relationships involving the coded output vectors. For example, if $v_1+v_2 = v_3+v_4$, then from the output table, for any u_y,

$$w_{1y} + w_{2y} = w_{3y} + w_{4y} .$$

Let the equations obtained in this way be referred to as state derived equations.

From equation 32 it follows (by the same reasoning leading up to equation 6) that

$$w_{1x} + w_{1y} = \cdots = w_{ix} + w_{iy} = \cdots = w_{Nx} + w_{Ny} . \tag{33}$$

Such equalities will be denoted by the term output sum.

Clearly, output sums, and state derived equations must be consistent if the output is linear.

It is known that $N = 2^n$ for a reduced table of a fully linear system.* Therefore, if $N < 2^n$ for a reduced table, the output is nonlinear. On the other hand, where $N = 2^n$, the set of coded state vectors forms a complete set of n component vectors. Then, when $l > n$ and there is a null input, say u_1, it is easy to show that the output vectors in column 1 are a subspace of the space of all output vectors. This suggests that the output vectors in column 1 can be coded over n components setting the remaining $l - n$ components to zero (that is, code over the subspace only).

The upper submatrix of the partitioned C matrix,

$$\left[\frac{C_n}{C_{l-n}}\right] (C_n \text{ is an } n \times n \text{ matrix})$$

can be determined by considering the outputs associated with the independent coded state vectors. Then,

$$C_n[v_1 \mid \cdots \mid v_n] = [w'_{11} \mid \cdots \mid w'_{n1}]$$

where the w' are n component vectors. Recalling that $[v_1 \mid \ldots \mid v_n] = I_n$ (for convenience in the state coding process) leads to

* Cohn and Even[1]

$$C_n = [w'_{11} \mid \cdots \mid w'_{n1}].$$

Since the subspace property of the first column must be preserved, it is true that C_{l-n} is the $n \times l - n$ null matrix. Furthermore, C_n and the code assignment over n components can be determined by designating n independent output vectors in conjunction with the state derived equations and outputs sums.

For all columns which are premutations of the first column, the corresponding Du_x must have zero as components $l - n$ through l (again, to preserve the subspace property). More generally, the entries in the first n components of Du_x generate permutations of the first column, while the entries in the remaining $l - n$ components force the output vector out of the n dimension subspace.*

In order to code the output vectors which are not members of the first column, notice that an output sum which involves such a column and the first column has N distinct sums and $2N$ vectors. N of these vectors have been coded (the members of the first column) and one of these vectors appears in each sum of vectors in the output sum. Therefore, if a linearly-independent vector is designated, by assigning a nonzero entry in $l - n$ components, then all of the remaining $N - 1$ vector codes are determined.

For example, if, in equation 33, $x = 1$, then w_{ix}, for $i = 1, 2, \ldots, N$ are vectors in the subspace (and can be coded), then setting w_{1y} equal to a linearly-independent vector or any vector not in the subspace gives the code assignment for *all* other vectors in column y since $w_{iy} = w_{1y} + w_{1y} + w_{ix}$ for any i. The procedure is basically the same as in coding state vectors of a singular system.

The case where $L \leqq N$ can be treated as an obvious special case of the above.

4.2 *The Moore Model*

In the Moore model of a sequential machine the output is a function of the system's state alone. That is, $D = 0$. It is then obvious from equation 32 that the number of distinct output vectors cannot exceed the number of state vectors. Therefore, the method for coding over the n-dimensional subspace introduced in Section 4.1 can be applied directly, using the state-derived equations.

---

* This property is the identity or disjointness of sets of vectors which can be rigorously proven by an argument parallel to the proof of theorem 3.

V. CODING THE INPUT VECTORS

5.1 *The Mealy Model*

In the Mealy model of a sequential machine it is possible to have inputs which do not generate distinct columns in the state transition table and yet give rise to distinct output columns. Such cases require an interaction in the determination of the B and D matrices.

Let there be M different inputs, and $K(K \leqq 2^k)$ distinct columns in the state transition table. Take one input as the null input, say $u_1 = 0$. Considering one component sum from each state sum, it follows that

$$Bu_x = v_{11} + v_{1x}$$

for each u_x that generates a distinct column. For convenience number such inputs u_1 through u_K. Similar to the approach of Section 4.1, code the input vectors over the first k components. That is, from

$$B = [B_k \mid B_{m-k}] \qquad \text{(where } B_k \text{ has } k \text{ columns)}$$

select

$$B_k = [v_{11} + v_{12} \mid \cdots \mid v_{11} + v_{1,k+1}][u_2' \mid \cdots \mid u_{k+1}']^{-1}, \tag{34}$$

where the u' are linearly independent vectors formed by the first k components of the input vectors, and where $v_{11} + v_{12}, \ldots, v_{11} + v_{1,k+1}$ are the k vectors that span the set of the K distinct sums of the form $v_{11} + v_{1x}$. Since it is only the first k components of the input vectors that influence the state transitions, it must be true that B_{m-k} is the $n \times (m - k)$ null matrix. The remaining k component input vectors can be obtained by solving the set of linear equations* (in matrix form)

$$B_k u_y' = v_{11} + v_{1y}.$$

At this point k components of all input vectors have been coded.

Considering the output table, the D matrix can be determined from

$$D[u_a \mid \cdots \mid u_r] = [w_{11} + w_{1a} \mid w_{11} + w_{1b} \mid \cdots \mid w_{11} + w_{1r}] \tag{35}$$

where the columns of the matrix on the right side span all sums of the form $w_{11} + w_{1y}$, and where $u_a, \ldots, u_r$ are m inputs which are assigned as linearity independent vectors.

* These linear equations must be consistent since the columns of B_k span all vectors of the form $v_{11} + v_{1y}$. (See equation 34.)

Consider the example in Table X.

TABLE X

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
s_1	s_1/z_1	s_1/z_2	s_1/z_5	s_1/z_6	s_2/z_7	s_2/z_8	s_2/z_9	s_2/z_{10}
s_2	s_3/z_2	s_3/z_1	s_3/z_6	s_3/z_5	s_4/z_8	s_4/z_7	s_4/z_{10}	s_4/z_9
s_3	s_4/z_3	s_4/z_4	s_4/z_{11}	s_4/z_{12}	s_3/z_{13}	s_3/z_{14}	s_3/z_{15}	s_3/z_{16}
s_4	s_2/z_4	s_2/z_3	s_2/z_{12}	s_2/z_{11}	s_1/z_{14}	s_1/z_{13}	s_1/z_{16}	s_1/z_{15}

In order to code the system states, consider u_1 and u_8 which generate the only distinct columns. Take $u_1 = 0$. It can be verified that an acceptable coding is

$$v_1 = \begin{pmatrix}0\\0\end{pmatrix}, \qquad v_2 = \begin{pmatrix}0\\1\end{pmatrix}, \qquad v_3 = \begin{pmatrix}1\\0\end{pmatrix}, \qquad v_4 = \begin{pmatrix}1\\1\end{pmatrix};$$

giving

$$A = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}.$$

For the output table, $L = 16$, $l = 4$. Calculating the upper partition of the C matrix, C_2,

$$C_2[v_3 \mid v_2] = [w_3' \mid w_2'],$$

or

$$C_2 = [w_3' \mid w_2'].$$

Take

$$w_3' = \begin{pmatrix}1\\0\end{pmatrix}, \qquad w_2' = \begin{pmatrix}0\\1\end{pmatrix};$$

then, $C_2 = I_2$ or

$$C = \begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\\ 0 & 0\end{bmatrix}.$$

From the state sum, $w_4' = \begin{pmatrix}1\\1\end{pmatrix}$, and since $v_1 = 0$, it follows that $w_1' = \begin{pmatrix}0\\0\end{pmatrix}$. Clearly, w_1, w_2, w_3, and w_4 are given by including two additional zero components. In order to code column 3 consider the output sum

$$w_1 + w_5 = w_2 + w_6 = w_3 + w_{11} = w_4 + w_{12}.$$

Taking $w_5 = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}$ leads to

$$w_6 = w_1 + w_5 + w_2 = \begin{bmatrix}0\\0\\0\\0\end{bmatrix} + \begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}0\\1\\0\\0\end{bmatrix} = \begin{bmatrix}0\\1\\1\\0\end{bmatrix}.$$

Similarly,

$$w_{11} = \begin{bmatrix}1\\0\\1\\0\end{bmatrix}, \qquad w_{12} = \begin{bmatrix}1\\1\\1\\0\end{bmatrix}.$$

By taking

$$w_7 = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \quad \text{and} \quad w_9 = \begin{bmatrix}0\\0\\1\\1\end{bmatrix}$$

the remaining output vectors can be coded.

Coding the first component of the input leads to

$$B_1 = \begin{bmatrix}1\\0\end{bmatrix}.$$

With the results

$$B = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\end{bmatrix},$$

$$u_1' = u_2' = u_3' = u_4' = 0, \tag{36}$$

and

$$u_5' = u_6' = u_7' = u_8' = 1. \tag{37}$$

The output vectors w_2, w_5, and w_7 span all sums of the form $w_1 + w_x$ (notice $w_1 = 0$) $x = 2, 5, 6, 7, 8, 9, 10$. Then,

$$D[u_2 \mid u_3 \mid u_5] = [w_2 \mid w_5 \mid w_7] = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Considering equations 36 and 37, make a linearly-independent assignment of u_2, u_3, and u_5. Say

$$u_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad u_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad u_5 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix};$$

which leads to

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

To determine u_4, for example, set up the linear equations

$$Du_4 = w_6 \qquad (\text{recall } u_4' = 0)$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} u_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix},$$

or

$$u_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Similarly,

$$u_6 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad u_7 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad u_8 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

5.2 *The Moore Model*

For the Moore model all columns of the state transition table are distinct; therefore, the method for coding over k components of u, introduced in the last section, can be applied directly.

VI. LINEARITY AND INCOMPLETELY SPECIFIED SYSTEMS

This section considers the problem of specifying "don't care" entries in the state transition table in a way consistent with linearity; the results can be extended to the output table.

It is obvious that if an incompletely specified table is to have a linear realization, then the entries in the table must obey the same relationships which were developed in the preceding sections. For example, the table must exhibit consistent state sums, and allow completion of the maximum memory process. The resulting restrictions on the unspecified entries may be used to deduce their appropriate assignments.

Consider Table XI in which the "don't care" next states are denoted by the symbol t.

TABLE XI

	0	1
s_1	t_1	s_2
s_2	s_4	s_6
s_3	s_8	t_4
s_4	s_6	s_4
s_5	s_3	s_1
s_6	s_1	s_3
s_7	t_2	s_8
s_8	t_3	s_5

The state sum is

$$v_2 + t_1 = v_4 + v_6 = t_4 + v_8 = v_1 + v_3 = t_2 + v_8 = t_3 + v_5 . \quad (38)$$

From the third and fifth terms, it follows that $t_4 = t_2$. Forming

$$A(v_1 + v_3 + v_4 + v_6)$$

leads to

$$v_6 + v_1 + v_8 + t_1 = 0. \tag{39}$$

Also,

$$A(v_4 + v_6 + v_2 + t_1)$$

yields

$$v_6 + v_1 + v_4 + \bar{t}_1 = 0,$$

where $\bar{t}_1 = At_1$. Locating v_6, v_1, and v_4 in the state sum leads to the conclusion $\bar{t}_1 = v_3$. The present state which gives s_3 as its next state under zero input is s_5. So that, $t_1 = v_5$. The first and last terms of the state sum imply that $t_3 = v_2$. From $A(v_4 + v_6 + v_8 + t_2)$ obtain

$$v_6 + v_1 + t_3 + \bar{t}_2 = v_6 + v_1 + v_2 + \bar{t}_2 = 0.$$

Adding the last equation to equation 39 gives

$$v_8 + t_1 + v_2 + \bar{t}_2 = v_8 + v_5 + v_2 + \bar{t}_2 = 0.$$

However, equation 38 indicates $t_2 + v_8 = v_2 + v_5$; so that $\bar{t}_2 = t_2$. From the table, $t_2 = v_7$. Table XII shows the fully specified table.

TABLE XII

	0	1
s_1	s_5	s_2
s_2	s_4	s_6
s_3	s_8	s_7
s_4	s_6	s_4
s_5	s_3	s_1
s_6	s_1	s_3
s_7	s_7	s_8
s_8	s_2	s_5

APPENDIX

An Addition to Theorem 2

For N not equal to 2^n there is a set of $2^n - N$ vectors that cannot be used as state vectors. However, when the system is linear operat-

ing on any unused vector with the A matrix, and adding Bu_z must produce a vector which is also an unused vector. If this were not the case, there would exist a state vector which would give rise to one of the unused vectors as a next state for some input. It follows that a transition table can be constructed containing only the $2^n - N$ unused vectors. Let this system be called a virtual system. From the foregoing, it is clear that the states of the virtual system must obey the restriction of Theorem 2; that is, the number of distinct inputs cannot exceed

$$2 + (2^n - N) \sum_{i=2}^{t} 2^{1-i},$$

where t is as before. Since the virtual and original systems must remain disjoint, the original system must observe the bound. This is a smaller upper bound than obtained in Theorem 2.

REFERENCES

1. Cohn, M. and Even, S., "Identification and Minimization of Linear Machines," IEEE Trans. Elec. Computers, *EC-14,* No. 3 (June 1965), pp. 367–376.
2. Davis, W. A. and Brzozowski, J. A., "On the Linearity of Sequential Machines," IEEE Trans. Elec. Computers, *EC-15,* No. 1 (February 1966), pp. 21–29.
3. Yau, S. S. and Wang, K. C., "Linearity of Sequential Machines," IEEE Trans. Elec. Computers, *EC-15,* No. 3 (June 1966), pp. 337–354.
4. Birhoff, G. and MacLane, S., *A Survey of Modern Algebra,* 3rd ed., New York: Macmillan, 1965.

Laser Machining of Thin Films and Integrated Circuits

By M. I. COHEN, B. A. UNGER, and J. F. MILKOSKY

(Manuscript received September 5, 1967)

The feasibility of using the YAG laser as a tool to thermally machine integrated circuits has been studied. Results suggest that defining the resistor geometry, trimming the resistors to value, fabricating gap capacitors, and defining interconnecting circuitry might be performed by such a laser.

Pattern generation by laser machining has been demonstrated on various thin and electroplated films. Vaporized lines (gaps) are readily attainable as fine as 0.25 mil in thin films and 0.4 mil in plated films. Much thinner lines may be obtained under particularly well-controlled conditions. These films may be removed with minimum effect to the substrate surface. The heat-affected region of the substrate may be confined to less than one film thickness. Better laser output control and shorter pulse widths will diminish this thickness.

Gap capacitors have been made on sapphire substrates with capacitance approaching 20 pf in 0.04 square inches, and experiments suggest improvements.

Tantalum films may be shaped to resistor geometries and trimmed to tolerance by removing metal or by oxidizing the resistive film with the laser. Resistors usually can be trimmed to tolerances of less than ±0.1 per cent.

With further development it might be possible to combine these laser machining processes into a single-step, automated fabricating procedure for certain types of integrated circuits. We review some of the technical aspects of this and discuss using Q-spoiled YAG lasers to directly machine masks for photoetching.

I. INTRODUCTION

The steps to fabricate thin film passive elements and interconnecting circuitry on hard substrates are well documented.[1, 2] These procedures, defining resistor geometry and interconnecting circuitry, resistor trim-

ming to value, and capacitor fabrication, are the essential steps for producing precise integrated circuits. It is our intention to show that a laser can be used to vaporize, in a controlled manner, thin film structures, and that it therefore might be capable of supplementing or performing these four processes; or it might be used to supplement photolithography by directly machining thin film masks for making circuits by photochemical methods.

We studied methods for performing these steps with a continuous neodymium doped yttrium-aluminum-garnet (YAℓG:Nd or YAG) laser. Laboratory work has demonstrated the feasibility of fabricating circuit building blocks and eliminating many steps for film structures which are common to tantalum integrated circuits; that is, circuits in which the resistive films are compounds of tantalum. Similar techniques will be applicable to other types of circuits.

Tantalum thin film resistors have been shaped and trimmed to value with a laser beam. A controlled laser has removed various combinations of thin films from substrates without adverse effects to the substrates. Tantalum films have been thermally oxidized with the YAG laser as the heat source. Controlled parasitic or gap capacitors have been made with specific capacitance up to 4.5×10^{-4} pf per square mil by laser machining narrow lines across thin films.

The YAG laser has demonstrated considerable potential as a thin film machining tool. The simplicity of operating and controlling the device, and the characteristics of its output beam render it particularly well suited to this application.

II. GENERAL CONSIDERATIONS

2.1 *Characteristics of Lasers as Fabricating Tools*

The utility of a laser as a tool for fabricating thin film circuits results primarily from the spectral purity and degree of collimation of the laser light. These characteristics allow the beam to be focused to a very fine and intense spot. The high heat flux which occurs when the light is absorbed by the target material, and the sharp definition and localized nature of the working region allow heating, melting, or vaporizing minute amounts of material, with minimum effect to adjacent material or components.

Among other useful characteristics of light as a working tool is its small absorption depth in metallic materials. This property renders the laser particularly applicable to operating on thin materials such

as films without damaging the substrate or material beneath the film. In addition, energy may be transferred to the workpiece through an optically transparent material or atmosphere, and without physical contact with the workpiece. Encapsulated or otherwise inaccessible parts may be machined, and working regions may be kept free of contamination which might result from contact with a tool. References 3 and 4 give more detailed discussion of the suitability of lasers for fabrication.

Many types of lasers have been shown to be applicable in processes related to thin-film circuit fabrication. Pulsed lasers have been used to vaporize slots and lines in metallic targets[5] and have been included in an experimental automated procedure for trimming thin-film resistors.[6] Pulse-pumped He-Ne gas lasers have been used to scribe lines on metals.[7] Reported linewidths are 12.5 microns. Ionized argon lasers operating at one-half watt cw output have been used to scribe lines as fine as 10 microns in iron-nickel films deposited on glass.[8]

The neodymium-doped yttrium-aluminum garnet (YAG) laser discussed by Geusic and others,[9–11] is particularly well suited to thin-film fabrication because of its good combination of such characteristics as the adequate intensity, stability, and optical quality of the output beam, and its simple and compact design. Such YAG lasers may be operated continuously at about 1 watt output, or may be repetitively Q-switched by rotating the rear reflector. The Q-switched output is a continuous train of pulses with peak power exceeding 1 kw, and pulse duration about 200 ns when the repetition rate is 400 cps. Both types of operation may be obtained by using as a pump source an inexpensive lamp powered directly from line current. In both cases the laser may be adjusted to oscillate in a sufficiently low order mode that the output beam may be focused conveniently to the fine spot needed for precise thin-film machining.

2.2 *Machining Thin Films*

The processes that we studied, except thermal oxidation, use the laser's ability to vaporize material. It is desirable, therefore, to discuss briefly a few of the parameters and phenomena of material removal by laser. Cohen and Epperson give more detail.[3]

We notice first that it is the optical power density in the focused spot rather than the laser power output, itself, that determines the suitability of a laser for removing material. Greater power densities often may be obtained from a laser that oscillates in a low order

mode, than from a higher-powered multimoded device. For the useful case of a laser oscillating in the fundamental Gaussian mode, the minimum focal spot radius, w_f, may be determined from the relationship. (See Ref. 3.)

$$\frac{1}{w_f^2} = \frac{1}{w_0^2}\left(1 - \frac{(z_0 + d)}{f}\right) + \frac{1}{(f\theta)^2} \tag{1}$$

where, f is the focal length of a lens with sufficiently great aperture to admit the entire beam, w_0 and z_0 are parameters which depend upon laser cavity configuration, d represents the distance of the focusing lens from the cavity, and θ is the far-field beam divergence angle. In many metal-working applications, including the one we are discussing, the lens is sufficiently far removed from the laser output reflector that equation 1 becomes

$$\lim_{[f/(z_0+d)]\to 0} w_f = \frac{\lambda f}{\pi w} \tag{2}$$

where w is the radius of the beam as it enters the lens. Equation 2 predicts, therefore, that spot sizes of the order of wavelength, λ, may be obtained with lenses having small f numbers (that is, small ratios of focal length to aperture).

The size of the affected zone in the target material will depend on the thermal properties of that material as well as the optical spot size and the intensity distribution across the spot. Edge definition of the affected zone depends primarily on thermal properties of the target and the duration of exposure. Metals with high thermal diffusivity and a large difference between melting and vaporizing temperatures, such as gold and copper, tend to develop a lip formed by molten metal around the region from which material was removed. The lip may be minimized by using a very short exposure such as that which might result from operating the laser in the Q-switched mode.

The use of Q-switched output also is desirable to minimize thermal damage to the substrate. Damage results both from heat conducted from the film into the substrate, and from direct impingement of the focused laser light on the substrate after the film has been removed. Using laser pulses of high peak power and short duration substantially decreases both effects.[3]

Many metals reflect an appreciable portion of the incident laser light. Such metals might need much higher laser output levels than nonreflecting materials. In most laser micrometalworking processes,

the surface remains solid and reflecting for only a small portion of the laser pulse duration. The reflectance of the surface may decrease abruptly when it melts or reacts with its atmosphere, and subsequent absorption will occur with greater efficiency. The initial laser output, however, must be sufficient to break down the surface.

2.3 *Experimental Apparatus*

Figs. 1 and 2 show the apparatus used in our experiments. The YAG laser was Q-switched by rotating the rear reflector at 400 hertz. The output pulse parameters at the rated voltage of the pump lamp (120 V) were about 1 kW peak power and 200ns duration at the half power point. Output, at 1.06 micron wavelength, was monitored by means of a photomultiplier tube behind the cavity, and the photomultplier frequently was used in conjunction with a thermopile which measured the mean power of the output beam, so that the peak power could be calculated. The laser output was attenuated by neutral density filters. Laser mode pattern was observed by means of an image converter tube, and the laser cavity was adjusted periodically so that most of the output was contained in the fundamental mode.

An x-y-z micropositioner was used to focus and move the workpiece. For applications such as line scribing, in which the work is moved across the beam, one of the micrometer barrels is rotated by means of an hydraulic drive coupled to a gear and belt mechanism. A wide range of continuously-variable speed was available. The maximum

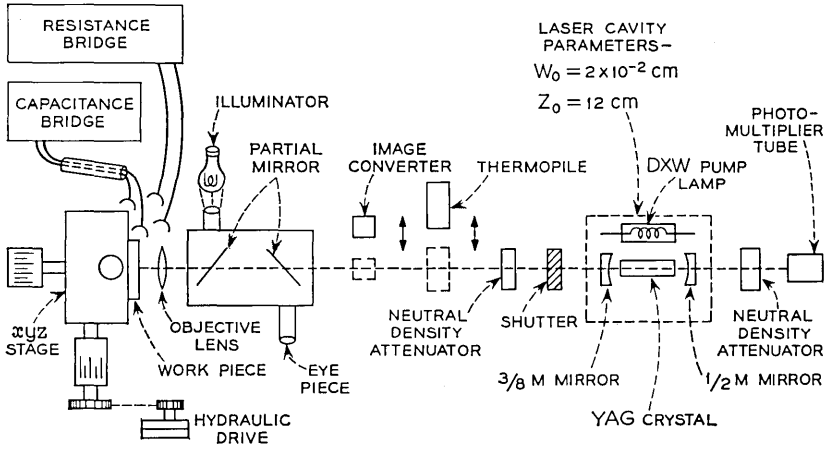


Fig. 1 — Schematic diagram of laser machining apparatus.

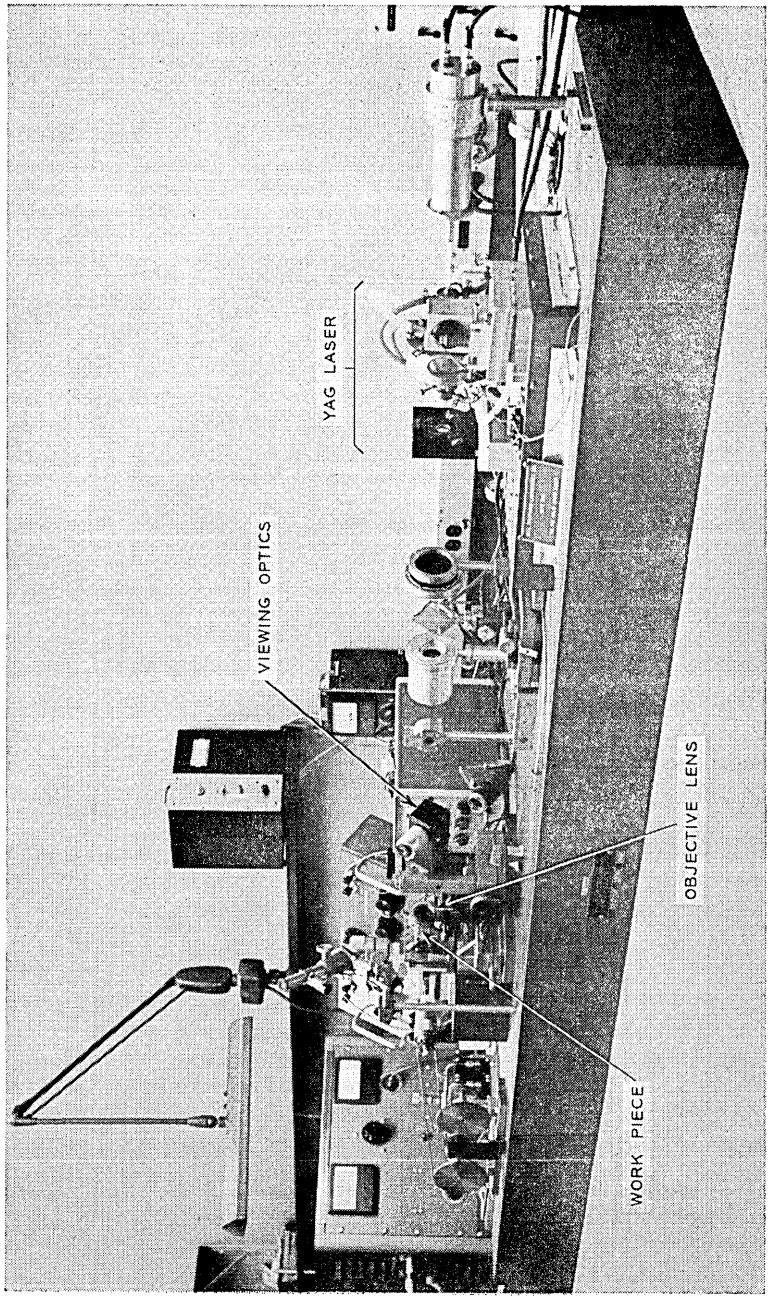


Fig. 2 — Laser machining apparatus.

permissible sweep speed, however, was dependent upon the diameter of the focal spot and the stability of the laser output. With excessive sweep speed, successive laser pulses did not overlap, causing scalloping at the edge of the line. Speeds of 0.1 inch per second were permissible with 0.0002-inch gaps, and correspondingly lower speeds were required with smaller gaps. Working atmosphere generally was air, although limited tests suggested that somewhat better edge definition is possible in oxygen-rich environment.

The viewing and focusing device consisted of two partial mirrors and an eyepiece mounted on a traverse mechanism, and a separately-mounted microscope objective. Objectives with 23 mm focal length (10X) were used for most of the applications studies associated with direct machining of tantalum integrated circuits, and lenses with focal length as short as 4 mm (approximately 70X) were studied for thin-film mask-making application.

The 23 mm objective was a particularly good lens for general-purpose work. Its depth of focus was sufficiently great (approximately ±0.001 inch) that fine lines of appreciable length (greater than 0.5 inch) usually could be scribed with good uniformity on nonuniform surfaces such as that of a glazed alumina substrate without elaborate alignment of the workpiece. The spot diameter, $2\ w_f$, calculated from equation 1 and the parameters appearing in Fig. 1, was about 8 microns (about 0.0003 inch) for this lens. Increasing the strength of the objective to about 40X will decrease the calculated spot size to about 2 microns, but the severe loss of depth of focus restricts use of such lenses to very flat and well-aligned targets. Such lenses with short working distances also necessitate careful attention to laser output intensity in order to prevent lens damage due to the laser plume of vaporized material.

The apparatus described has been used to vaporize spots and scribe lines in a wide variety of thin-film structures. Figs. 3 through 5 show some of the general characteristics of such lines.

Figure 3 indicates the effect of laser beam intensity and stage sweep speed on width and definition of lines vaporized in a gold thin film (approximately 3000 Å) on sapphire substrate and a nichrome film (2000 Å) on quartz. An 8 mm lens was used to scribe the gold film and a 4 mm lens was used for the nichrome. The finest lines were about 0.00025 inch wide for the gold and 0.0001 inch for the nichrome. Such gaps generally contain no metallic debris, as evidenced by the very low electrical conductance that is measured across them.[3]

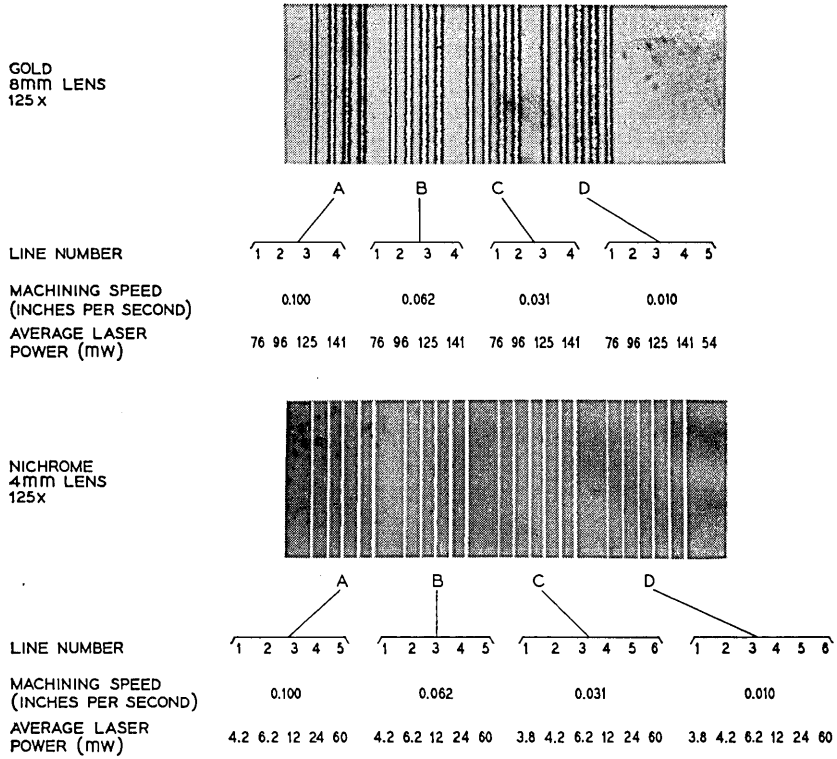


Fig. 3 — Effect of machining parameters on lines machined in 3000 Å gold film on sapphire and 2000 Å nichrome film on quartz.

Lines as fine as 0.00004 inch (1.0 micron) have been scribed in tantalum nitride, titanium, and nichrome films.

It has been shown that for thin film samples similar to those shown in Fig. 3, the extent of damage to the sapphire substrate may be minimized by proper control of power density and sweep speed.[3] Fig. 4, representing a typical cross section through such a sample, indicates that the extent of the affected zone may be limited to a depth approximately equal to the film thickness. Recent results have shown that some metal films may be removed with no substrate damage observable by optical means.

Fig. 4 also suggests the presence of a small gold lip bounding the laser-machined gap. The size and nonuniformity of the lip, and the depth of the thermally affected zone in the substrate increase as the film thickness is increased. Fig. 5 shows typical depth of penetration

into the sapphire when the gold thin film is plated with approximately 0.3 mil of copper and then 0.01 of gold. Although gaps as narrow as 0.00020 inch have been machined in such films with a 23 mm lens, gaps smaller than the film thickness require precise control of laser and stage parameters. The lateral extent of the heat-affected zone resulting from thermal conduction in YAG laser-machined thick films often is the same as the film thickness.

Results similar to these have been obtained when the substrate material is glazed alumina, quartz, or silicon rather than sapphire, although the body of data for these materials is not nearly so large as it is for sapphire. Damage to the surface, apparently caused by melting, is confined to a depth less than the film thickness. Cracking of the glaze in the vicinity of the working zone has not been observed when laser parameters have been adjusted properly.

III. DIRECT MACHINING OF TANTALUM INTEGRATED CIRCUITS

3.1 *Gap Capacitors*

3.1.1 *Characteristics*

Interelectrode effects normally are present in miniaturized or high density integrated circuits. However, these parasitic or stray capaci-

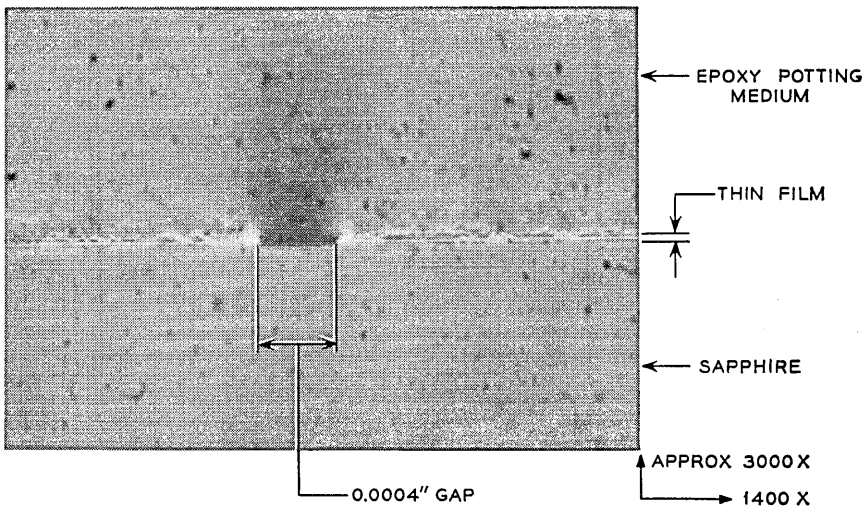


Fig. 4 — Cross section through gap machined in 3000 Å gold film showing the affected zone in the sapphire substrate.

tances usually are minimized by judicious spacing or isolating the components. These parasitics, gap capacitors as we call them in this article, can be produced with values high enough for use as discrete circuit elements by varying the spacing and length of the coupling electrodes, and the dielectric constant of the surrounding media.

Kaiser and Castro have calculated the interelectrode effects between two thin film conductors deposited on a substrate.[12] The calculations were directed at predicting parasitic capacitances between parallel conductors on substrates of various dielectric constants. The analysis was based on a model of two parallel conductors a distance d apart, of equal width l, on the same side of a substrate with dielectric constant k, of finite substrate thickness t, and of infinite extent. It has been assumed that the dielectric constant of the environment is negligible compared with that of the substrate. Fig. 6 presents, for various d/t and l/t ratios, calculations based upon this analysis.

Fig. 6 shows experimental data for some gap capacitors fabricated in the laboratory by means of a Q-switched YAG laser. Gap width was varied from 0.0005 to 0.025 inch for thin chrome-gold films on 0.025 inch thick barium titanate substrates (dielectric constant approximately 500), and 0.0003 to 0.015 inch for tantalum-chrome-gold thin film composite on 0.03 inch sapphire. In the case of the high dielectric substrate, agreement with numerical calculations is best for the larger gaps. Behavior for the fine gaps probably is affected by the granular and nonuniform nature of the substrate, and further study

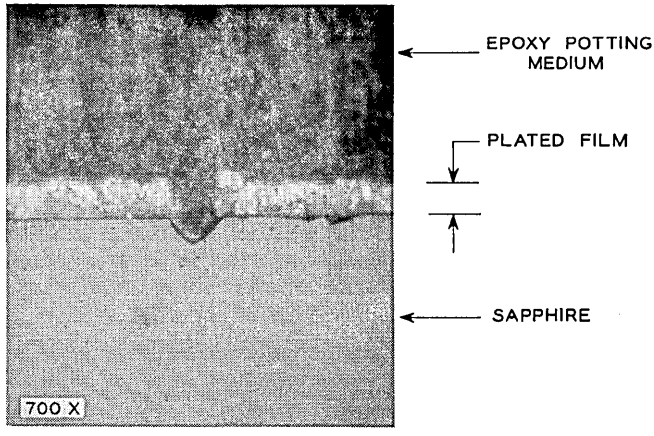


Fig. 5 — Cross section through gap machined in plated thin film showing the affected zone in the sapphire substrate.

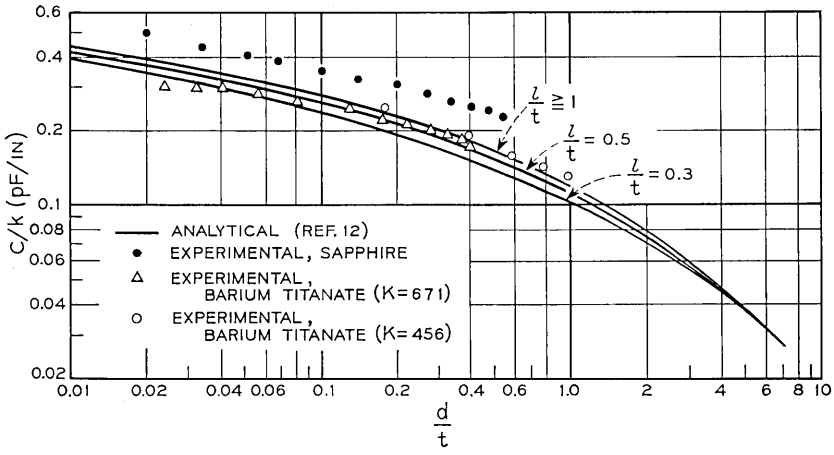


Fig. 6 — Capacitance of gap capacitors as a function of gap width. Comparison of analytical and experimental results.

is necessary. The shape of the curve for sapphire is similar to that predicted by the analytical model. The displacement between the curves may be explained by considering the fringing capacitance through the environment.

3.1.2 *Laser Fabrication*

A Q-switched YAG laser was used to machine fine lines in thin film conductors to form the gap capacitors. The effects of gap width, line length and geometry have been investigated. Most of the data were obtained with gap capacitors on single crystal polished sapphire substrates. Limited data also have been obtained for other substrate materials.

Fig. 7 demonstrates the linear relationship that has been observed between capacitance and gap length. There is no observed effect of bending the cut into a serpentine configuration until, as will be shown later, the parallel legs become sufficiently closely spaced. These results suggest that there is no significant contribution due to the field concentration at the corners, since such an effect would cause a deviation from linearity.

Most of the data in Fig. 7 were taken on tantalum-chrome-gold thin film composites on sapphire. Some data also are included for gaps machined in 0.3 mil thick copper-plated conductors. Gap width in

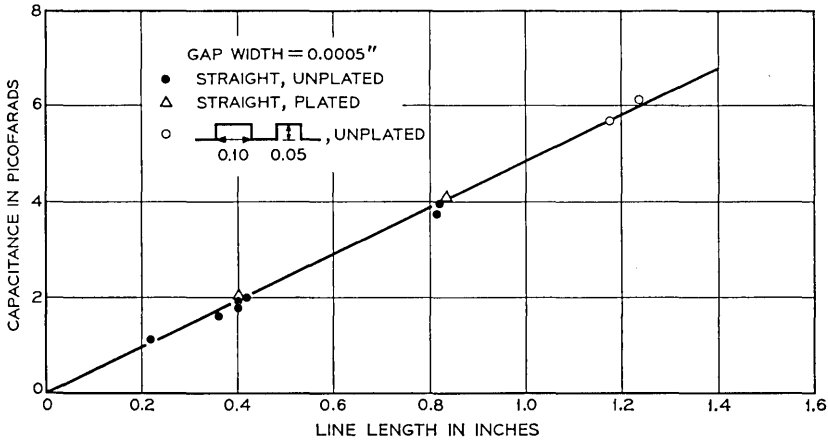


Fig. 7 — Capacitance of gap capacitors as a function of gap length.

both cases was 0.5 mil. The plated films appear to produce capacitance values similar to those for unplated thin films.

The response of gap capacitors to frequency was determined by measuring a gap capacitor at frequencies to 4 GHz. A plated 0.5 mil wide gap capacitor cut in a plated 50 ohm transmission line on an alumina substrate (Fig. 8) was measured with the following results:

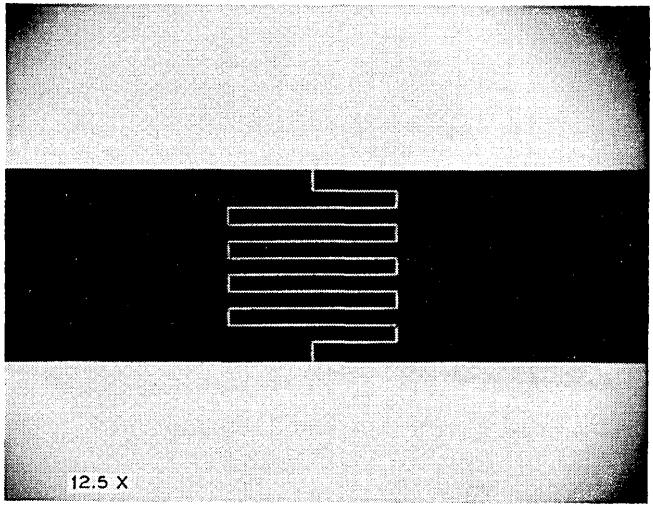


Fig. 8 — Gap capacitor laser machined in plated 50 ohm transmission line on glazed alumina substrate.

At frequencies of 465 kHz, 1 MHz, 3 GHz, and 4 GHz, the measured capacitance was 2.49, 2.48, 2.65, and 2.49 pf, respectively. The response of a laser machined gap capacitor appears relatively flat up to 4 GHz.

A further study was made of the effect of gap width on capacitance value. Fig. 9 shows the results of a tantalum-chrome-gold composite film on a polished sapphire substrate, and of a chrome and gold composite on an unglazed barium titanate substrate. Calculated points from Fig. 6 are indicated to show the agreement with the analytical model. The data indicates that the gap capacitors are insensitive to changes in gap dimensions when the gap is a few mils wide, but very sensitive when the gap is made less than a mil across. These results suggest that in order to achieve the highest possible specific capacitance it is desirable to achieve the smallest possible gap width. On a normalized plot, such as Fig. 9, the elements are also insensitive to substrate dielectric constant and the dielectric properties of the test environment. Gap capacitance therefore can be presented on a normalized basis with some constant as a dielectric scaling factor.

We show in Fig. 7 that capacitance varies linearly with length for straight-line gaps. To achieve a high capacitance per given area,

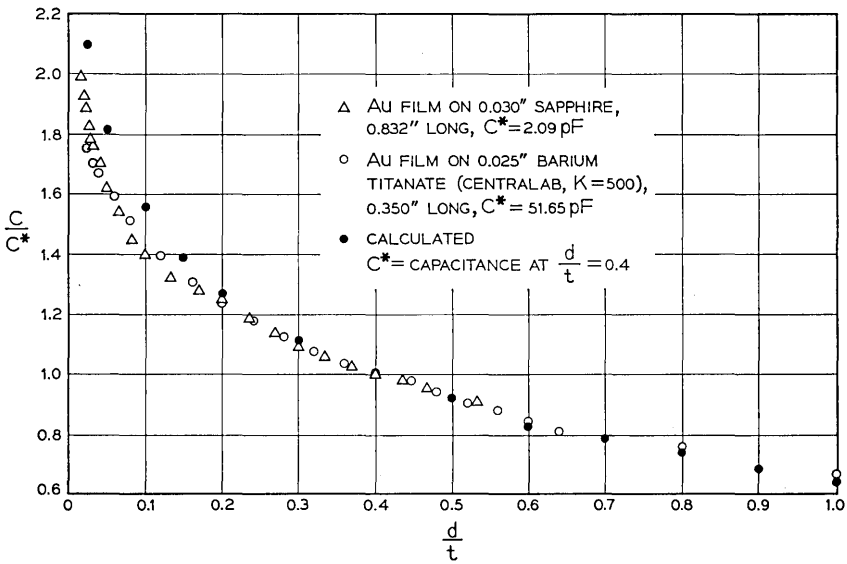


Fig. 9 — Capacitance of gap capacitors as a function of gap width.

however, the gap may be fabricated in a serpentine geometry. Fig. 10 shows a portion of a serpentine capacitor with 0.5 mil gap and 5 mil spacing. Fig. 11 shows the results of capacitance as a function of total line length for serpentine gap capacitors of equal heights on a conductor 0.225 inch wide. The curve is linear down to about a 50 mil spacing between the parallel legs. As this dimension decreases, the relationship between capacitance and line length becomes nonlinear. These results suggest the interaction of fields remote from the gap edges that significantly affects the value of the element.

Specific capacitance is one means of comparing different thin film capacitors. A specific capacitance can be given for gap capacitors provided the substrate dielectric constant and gap width are stated. Laser machined gaps 0.5 mil wide have been made in films on 200 mils-square sapphire substrates ($K \sim 10$) in a serpentine fashion with values up to 18 pf. This corresponds to a specific capacitance of 4.5×10^{-4} pf per square mil. It is estimated that this value can be increased to about 1×10^{-2} pf per square mil by adjusting the serpentine geometry and decreasing the gap width to 0.2 mil. Such gap widths have been machined in thin films with the YAG laser when the optical alignment and beam control have been precise. With careful control of techniques and equipment, repeatable gap widths of a few microns or less are practical.

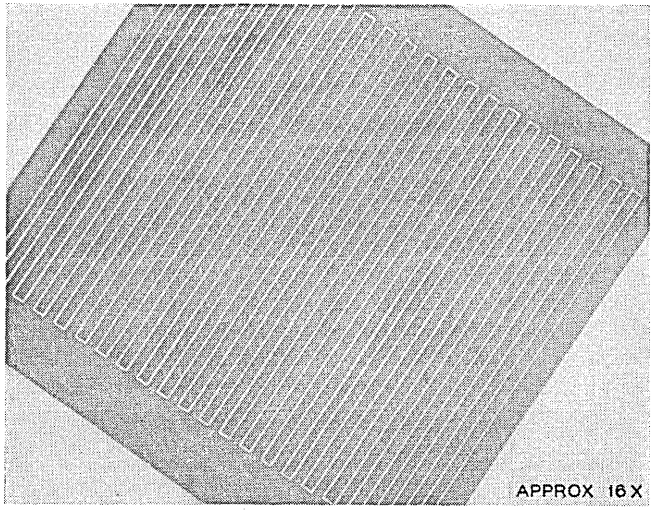


Fig. 10 — Serpentine gap capacitor on sapphire substrate. Gap width 0.5 mil, separation between legs 5.0 mils.

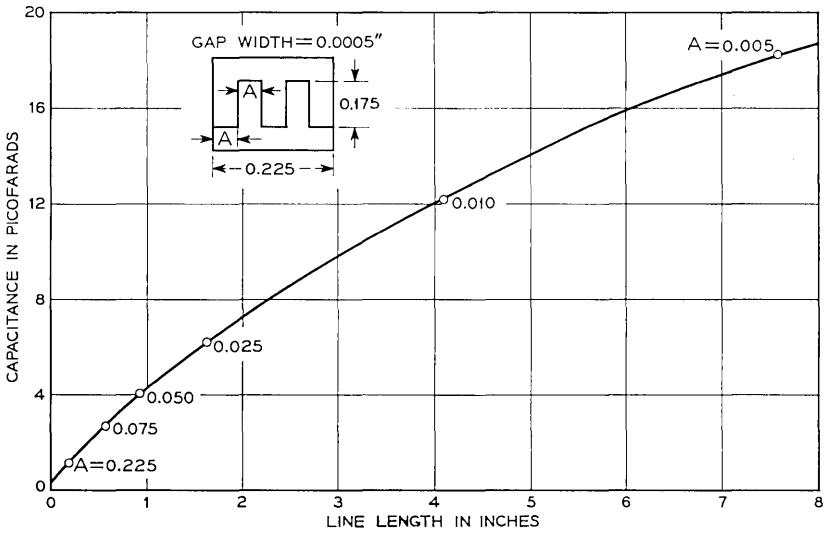


Fig. 11 — Capacitance of serpentine gap capacitors as a function of total length of gap.

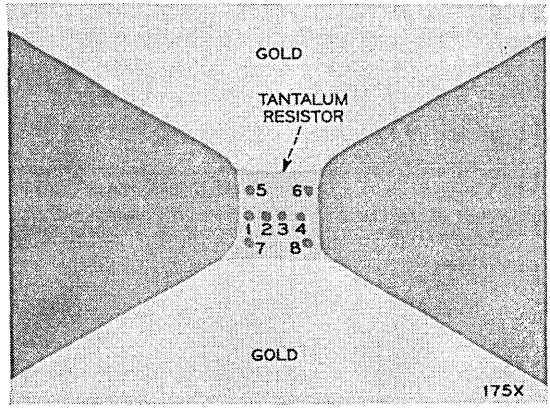
Leakage currents less than 10^{-9} amperes at 40 V are normal for these units. These values indicate that there is very little debris remaining in the gap.

3.2 *Trimming Thin Film Resistors*

Two techniques for increasing the resistance of tantalum nitride resistors with the YAG laser have been explored. Initial studies have established the feasibility of trimming to value by removing material and by thermal oxidation.

Trimming by removing material may be accomplished either by changing the dimensions of the resistor or by vaporizing small spots in the interior of the resistor. Fig. 12 shows a resistor whose value has been increased 19 per cent by means of a series of internal spots, each vaporized with a 1/50-second exposure to the Q-switched beam. Included in Fig. 12 are the results of resistance measurements taken on this resistor.

Changes in resistance from less than 0.01 to over 0.1 per cent per spot have been demonstrated by varying the size and location of the spot on the resistor. Fig. 13 demonstrates typical resistance changes for a resistor, one of whose edges has been removed progressively in



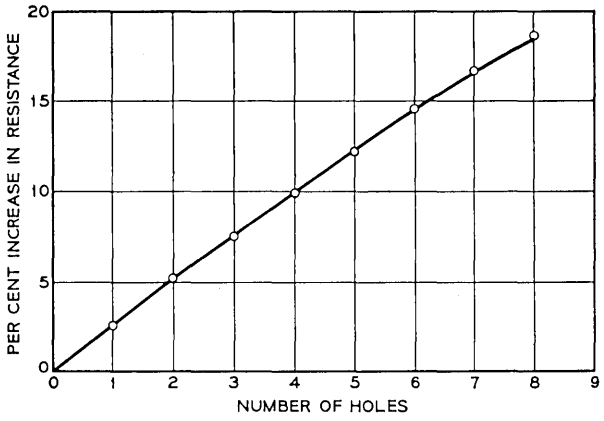


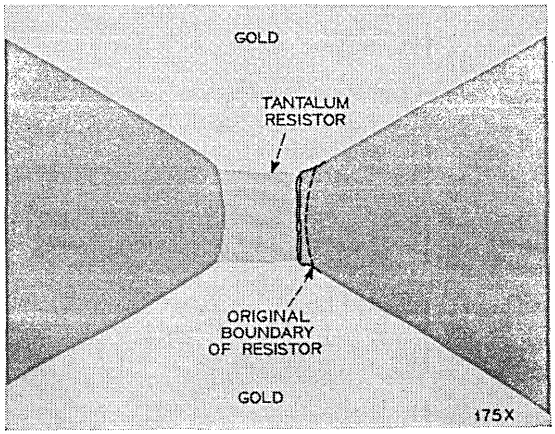
Fig. 12 — Resistor trimming by spot material removal.

0.0001-inch steps. With proper instrumentation and control, either the spot or edge laser vaporization method could provide a rapid means of trimming resistors to better than 0.1 per cent. We notice that resistors trimmed by material removal change less than 0.1 per cent in resistance after such resistors are exposed to the laboratory environment for several months. These results suggest the temporal stability of laser trimmed resistors. A definitive measure of stability, however, will necessitate a power aging test under controlled conditions.

Our studies of thermal oxidation were exploratory because a proper

quantitative investigation will require a laser with sufficiently high output to achieve the required power density over a spot large enough to completely expose the entire resistor. In the present study, a spot about 0.001 inch across was used and the resistor surface (0.005 inch square) was swept past the beam. A uniform brown color resulted, accompanied by a change in resistance from 16.96 to 17.98 ohms. The same color and increase in resistance may be obtained by wet anodization to a potential of about 20 V.

A much greater increase in resistance (for example, 16.29 to 38.26 ohms) was obtained by further exposure. The surface, however, ap-



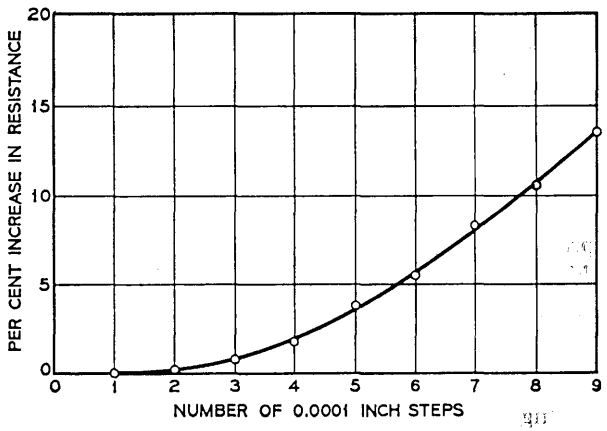


Fig. 13 — Resistor trimming by edge material removal.

pears nonuniform in color, indicating the varying degrees of oxidation. As treatment progressed, the process appeared to lose stability in the sense that it required greater attention to laser output and stage sweep speed to prevent the resistor from burning through locally. The indication is, however, that both uniformity and stability will improve when larger laser spots are used.

We noticed that oxidation of tantalum areas adjacent to gold conductors is possible without damage to the gold. The high reflectivity of the gold raises the damage threshold of gold well above the power required for converting the tantalum to oxide.

3.3 *Miscellaneous Applications*

The YAG laser may be used to remove metal films in order to define interconnecting circuitry. Such pattern generation requires no considerations other than those already discussed. With the present laser the small spot size necessitates many passes by means of an index-and-repeat method, in order to remove large areas of film. Such a method would be more economical if the present YAG laser were Q-switched at great speeds so as to allow correspondingly greater sweep speeds. An alternative would be to use a higher powered laser to define roughly the circuit dimensions, and then to trim precisely with the present laser.

The initial results suggest the possibility of fabricating, in one step, a complete integrated circuit with all the passive elements. Such a process would start with a metallized substrate and would use a programmable laser and work stage. Complete laser fabrication of hybrid circuits will require a process in which a metal film is removed selectively, exposing a different film. For example such a process may be necessary in order to remove the conductor and expose the resistor film.

In the present tantalum-chrome-gold technology such a selective removal of the gold presents substantial difficulties because the reflectivity of the gold is much greater than that of the tantalum nitride. It is quite probable, however, that some combination of films will be found for which the upper film can be removed from the resistor without damaging it.

IV. MACHINING THIN-FILM MASKS

The machining of masks involves considerations similar to those for pattern generation. Experience has indicated that a variety of

sputtered and deposited films (1000—3000 Å) on quartz, glass, and sapphire may be machined conveniently. Patterns machined in titanium and nichrome films, however, appear to have superior edge definition and uniformity of transmitted illumination compared with patterns produced in other films. Mechanisms governing the behavior of the metal film during laser machining have not yet been defined, but appear to depend upon the surface tension and bonding characteristics of the film as well as the thermal properties of both film and substrate.

V. SUMMARY

The feasibility of machining resistive and capacitive components directly on thin film metallized substrates with a laser has been demonstrated. Tantalum films can be shaped into resistor geometries and trimmed to tolerance by removing metal. These films also can be oxidized to value using the laser beam as the heat source. Resistors can be made with tolerances in value of less than ± 0.1 per cent.

Gap capacitors have been made on sapphire substrates with capacitance measuring up to 20 pf in an area of 0.04 square inch with 0.5 mil gap spacing. Limited studies have indicated that these elements are stable with time, have leakage currents at 40 V of less than 10^{-9} amperes and do not change significantly with frequency up to 4 GHz. Variation of capacitance with gap width has been studied and experimental results show good agreement with numerical results based on an analytical model for determining parasitic capacitance.

Pattern generation by laser machining has been demonstrated on various thin films as well as on electroplated films. Vaporized lines as fine as 0.25 mil are readily attainable in thin films, as are 0.4 mil lines in plated films. Much narrower lines may be obtained under particularly well-controlled conditions. Uniform lines as fine as 1 micron have been scribed in thin films on sufficiently flat substrates. These films have been removed with minimum effect to the substrate surface.

The present work has been accomplished with a Q-switched YAG laser operating with a repetitive output, at 400 hertz, of 1 kw peak power and 200 ns pulse duration. Advantages of such a laser as a machining tool include the optical quality of the output, and simplicity and economy in both design and operation. Maximum machining speed presently approaches 0.1 inch per second for 0.2 mil gaps, and a further increase of one to two orders of magnitude may be expected as the

repetition rate of the laser is increased. Continued development of the YAG laser, particularly with regard to decreasing the width of the pulse, is expected to provide for increased definition of the working zone, improved selectivity in removing films of various metals, and a further decrease in the already small damage to the substrate.

Our study has been exploratory and has served only to establish the feasibility of machining thin film circuits with existing lasers. Further attention to the details of the specific machining processes as well as to the combination of these processes into an automated procedure is necessary in order to evaluate their practicality. It is necessary, also, to better define the mechanisms governing laser material removal processes in order to realize fully the potential applicability of laser machining of thin film circuits.

ACKNOWLEDGMENTS

The YAG lasers that we used in this study were provided and maintained by J. E. Geusic and members of the Solid State Optical Device Department at Bell Laboratories including M. L. Hensel, R. G. Smith, W. W. Benson, and M. F. Galvin. We appreciate the interest, thoughtfulness, and kind cooperation of these gentlemen. We are grateful to have had the skillful assistance of S. P. Hourin and T. G. Melone in the laboratory and in preparing our photographs.

Miss E. B. Murphy performed the numerical calculations in Section III on a digital computer. We appreciate the interest of J. W. West, C. Maggs and L. Rongved, all of Bell Laboratories.

REFERENCES

1. McLean, D. A., Schwartz, N., and Tidd, E. D., "Tantalum Film Technology," Proc. IEEE, *52* (Dec. 1964), pp. 1450–1462.
2. Integrated Circuits, Design Principles (R. M. Warner, Jr., and J. N. Fordemwalt, ed.), McGraw-Hill (1965).
3. Cohen, M I. and Epperson, J. P., "Applications of Lasers to Microelectronic Fabrication," Advances in Electronics and Electron Physics, ed., A. B. El-Kereh, Academic Press (to be published).
4. Bahun, C. J. and Engquist, R. D., "Metallurgical Applications of Lasers," Proc. Nat. Elec. Conf., *19* (1963), pp. 607–619.
5. Williams, D. L., "The Laser as a Drilling Tool," Eng. Proc., New Ind. Technologies, Pennsylvania State University (June 27–July 2, 1965), p. 44.
6. Lins, S. J. and Morrison, R. D., "Laser-Induced Resistivity Changes in Film Resistors," Wescon '66 Technical Papers, Western Electronic Show and Convention (Aug. 23–26, 1966), Pt. 2, pp. 1–9.
7. Boot, H. A. D., Clunie, D. M., and Thorn, R. S. A., "Micromachining With a Pulsed Gas Laser," Nature, *198*, No. 4882 (1963), pp. 773–774.
8. "Laser Machining of Thin Metal Films," Laboratoire Central de Telecommunicatios, Electrical Communications, *41*, No. 2 (1966).

9. Geusic, J. E., Marcos, H. M., and Van Uitert, L. G., "Laser Oscillations in Nd-Doped Yttrium Aluminum, Yttrium Gallium, and Gadolinium Garnets," Appl. Phys. Letters, *4,* No. 10 (May 15, 1964), pp. 182–184.
10. Geusic, J. E., Hensel, M. L., and Smith, R. G., "A Repetitively Q-Switched, Continuously Pumped YAG: Nd Laser," Appl. Phys. Letters, *6,* No. 9 (1965), pp. 175–177.
11. Smith, R. G. and Galvin, M. F., "Operation of the Continuously Pumped, Repetitively Q-Switched YAIG:Nd Laser," IEEE J. Quantum Elec. *QE-3,* No. 10 (October 1967), pp. 406–414.
12. Kaiser, H. R. and Castro, P. S., "Capacitance Between Thin-Film Conductors Deposited on a High-Dielectric-Constant Substrate," Report 6–59–61–2, Lockheed Missiles and Space Company, Microsystems Electronics Department, Sunnyvale, California.

Performance Degradation by Postdetector Nonlinearities

By GEORGE H. ROBERTSON

(Manuscript received October 17, 1967)*

This article gives the performance degradation found by using a computer program to calculate the effects of various processing techniques applied after envelope detection of narrowband Gaussian noise plus a CW signal. Nonlinear processes that we studied are: suppression of low output levels, hard limiting of high output levels, restricted dynamic range of output level, and quantizing the detector output.

By modifying the program, we produced performance curves for systems in which a steady CW component was present in addition to possible signals. We also adapted the program to generate performance curves for systems using a square-law detector so that a comparison could be made with the results obtained for an envelope detector. We found that the presence of a CW component, and all the nonlinear processes applied after envelope detection, caused a loss of sensitivity in detecting small CW signals. Quantizing even in as few as 8 levels, however, caused very little additional degradation.

I. INTRODUCTION

A computer program was developed that would draw receiver operating characteristic curves for a system using an envelope detector to search for CW signals in narrowband Gaussian noise.[1] The program computed the probability that the detector output would exceed a chosen threshold under two circumstances:

(*i*) No signals were present; this result gives the probability of false alarm, P_{FA}.

(*ii*) Signals at various S/N were present; this result gives the probability of detection, P_D.

S/N is the ratio of the CW signal power to the noise power ac-

* The U. S. Navy supported this work under contract N600(63133)64940.

companying it in the specified narrow band. These probabilities were calculated using formulas which described the appropriate distributions of the detector output.

The program also averaged independent samples of the detector output, and produced appropriate receiver operating characteristic curves by deriving a Gram-Charlier A series for the distribution of the sample average.

When the program was adapted to take into account the modifications of the detector output caused by various subsequent nonlinear processes, receiver operating characteristic curves were produced for systems in which such processes occurred. The performance degradation was determined as the change in S/N that would be required to make the detectability of a signal, for given probability of false alarm, the same as without the nonlinear process.

II. DISCUSSION

Since some types of nonlinear processing greatly reduce the labor of handling large amounts of data, and others are unavoidable, or their effects are costly to minimize, it is useful to know the penalties on performance that are incurred by their presence.

The types of nonlinear process studied are: suppressing low output levels, hard limiting of high output levels, restricting dynamic range of the output level, and quantizing the output of the detector.

Suppressing low output levels occurs in some kinds of recording equipment where the output must reach a minimum level before a record can be made. High output limiting always occurs because the power handling capacity of physical equipment is limited. Dynamic range is restricted when the first two processes occur together; this also represents a commonly used technique in which a fluctuating output is converted to a binary waveform with respect to a chosen threshold. Quantizing is particularly useful when a digital computer is used to implement the statistical analysis. With this application in mind quantizing was evaluated in combination with restricted dynamic range

For simplicity in programming, we assumed that the unit used to process the output of the envelope detector had an appropriate transfer characteristic from the set shown in Fig. 1. It can be seen that over part of its operating range this unit has a linear characteristic. Quantizing would cause the slope of the linear region in Fig. 1(c) to be replaced by a staircase.

In addition to the changes needed to allow investigation of the

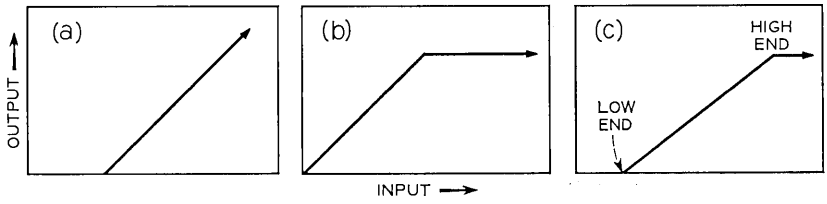


Fig. 1 — Transfer characteristics of post detector processor. (a) Suppression of low levels, (b) hard limiting of high levels, (c) restricted dynamic range.

above processing techniques, we made alterations in the program so that it would produce receiver operating characteristic curves for systems using a square-law detector, and for systems in which an unwanted CW component occurred in the specified narrow band in addition to the one representing a signal.

The latter case is typical of situations where a large, relatively stable, CW component tends to mask the presence of a smaller stable signal in the resolved band of a spectrum analyser. The results show how effective in such cases is a decision process similar to that described in Ref. 1, for detecting a small signal.

The distortions produced by some of these processes on the detector output distributions are so radical that many samples must be averaged to get an adequate Gram-Charlier representation even using up to 31 moments, the limit available in the program. Consequently a large number (8192) of samples were averaged before deriving the Gram-Charlier series from which the receiver operating characteristic curves could be produced. We made some checks where it was possible to produce curves for fewer samples averaged. We marked these on the figures; they show only slight deviations from the results obtained by averaging 8192 samples.

Since the normal distribution becomes a satisfactory approximation as the number of samples averaged grows large it may be inferred that the degradation caused by nonlinear processes could be estimated by averaging enough samples and using the normal approximation. For some of the processes described here the receiver operating characteristic curves for averaging 8192 samples showed perceptible variation from the results obtained by a normal approximation at P_{FA} values less than 10^{-4}. Thus, to ensure the validity of a normal approximation it might be necessary to average a very large number of samples indeed.

When the results for one of the simulated nonlinear processes were plotted on the appropriate chart of Ref. 1, the variation of the degra-

dation from the average was never found greater than 10 per cent and usually within a few per cent. Consequently the average degradation shown in the accompanying figures represents a good approximation within the range covered by the appropriate chart of Ref. 1.

Fig. 2 shows the degradation in performance when low output levels are suppressed, and when high levels are hard-limited. The position at which the levels are truncated is given along the abscissa as a multiple of the mean value of the ideal detector output when only noise is present (Rayleigh mean). The ordinate gives the change in S/N required with the distorting process to give the same performance as can be achieved without it.

Fig. 3 shows the degradation sustained when the dynamic range factor of the detector output is 10 (solid line) and 2 (broken line). The dynamic range factor is the ratio of the high end of the linear range to the low end for a transfer characteristic like that in Fig. 1(c). Fig. 3 also gives the degradation sustained when the output is virtually converted to a binary signal by limiting the dynamic range factor to 1.001 (dotted line). The lower truncation point is shown on the abscissa and the ordinate gives the degradation in dB.

III. QUANTIZING

When a digital computer is used to implement the statistical analysis of the detector output, it is important to know the maximum number of bits that need to be used to encode the output level. Three

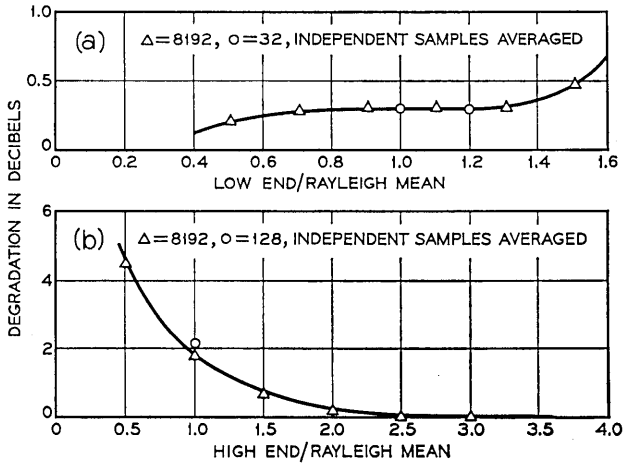


Fig. 2 — Degradation caused by (a) suppressing low output levels and (b) hard limiting high output levels.

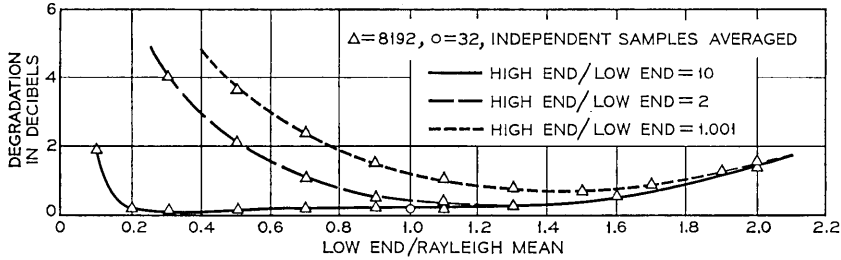


Fig. 3 — Degradation resulting from limited dynamic range.

bits allows coding into 8 levels, and four bits allows coding into 16 levels, and so on. If three bits are used, for example, the range of possible detector outputs must be divided into 8 regions. The highest region will include all levels that would equal or exceed the high end of a characteristic such as that in Fig. 1c, and the lowest region will include all levels below the low end. The remaining six regions will be between the two truncation points. The computer simulation was carried out by dividing the detector output in this way and assigning the value of the lower limit of any region to all the values in it less than the upper limit.

From Fig. 3 it can be seen that if a dynamic range factor of 10 is used and the low end is placed at three tenths of the Rayleigh mean, very little degradation in performance is suffered. When this range was quantized into 6 equal steps it was found that the total degradation was only about 0.01 dB more than the value shown in Fig. 3.

Two other ways of dividing the range into 6 unequal steps were tried, but the results were poorer than that for equal steps. In the first of these the range between the low end and high end was divided into 6 intervals in such a way that the change in probability density was the same between each. The degradation in performance was about 0.1 dB greater than that using 6 equal steps. The other way of quantizing the range divided it into 6 intervals of constant change in cumulative probability. The degradation in this case was 0.39 dB greater than that using 6 equal steps.

IV. CW INTRUDER

Fig. 4 shows that the system performance falls off rapidly owing to the presence of a CW component as the level of such a component increases from −10 dB with respect to the noise in the narrow band. The distribution curves corresponding to the no-signal hypothesis in this case were those for Gaussian noise plus a CW component at the

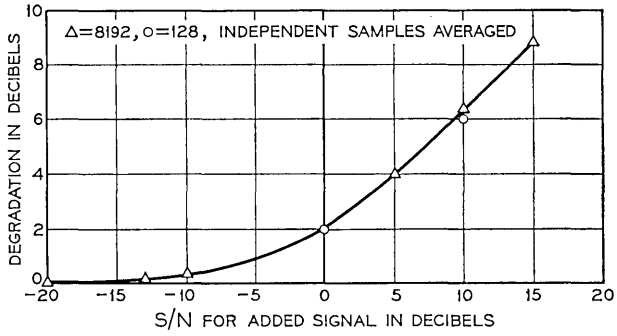
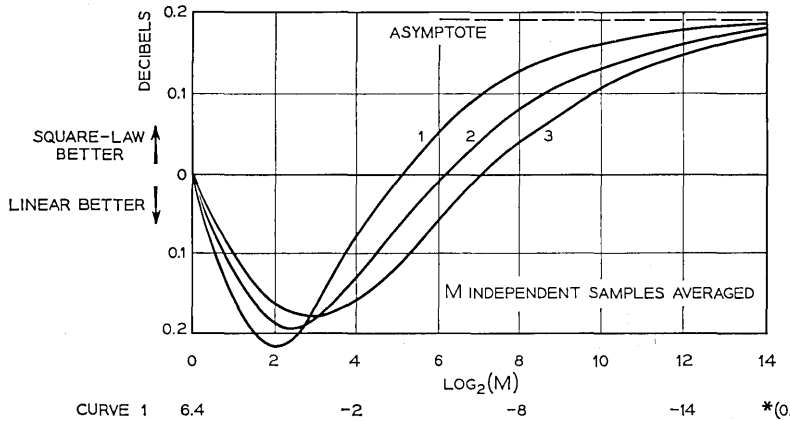


Fig. 4 — Degradation resulting from added CW signal.

S/N specified on the abscissa. The curves corresponding to the hypothesis that a signal was present were those for Gaussian noise plus a resultant sine wave component. The power of the resultant equalled the power of the small signal plus the power of a large CW component at the level specified on the abscissa. The results thus represent the average for all possible phase relationships between the CW component and small signal assuming that all phase values are equally likely, and many samples of the detector output are averaged.

V. SQUARE-LAW DETECTOR

It was found that satisfactory Gram-Charlier series approximations to the output distribution of a square-law detector could be produced only when a few hundred or more samples were averaged. Consequently subprograms were written to compute values for Chi-square and noncentral Chi-square distributions over the desired ranges of S/N and samples averaged. Using these and the results reported in Ref. 1 it was possible to produce the curves shown in Figs. 5, 6, and 7. which enable the receiver operating characteristic curves of Ref. 1 to be used to estimate the performance for a square-law detector with quite good accuracy. Fig. 5 compares the performance of linear and square-law detectors at a false-alarm probability of 10^{-6} and three different detection probabilities when the number of samples averaged ranges from 1 to 8192. The approximate S/N corresponding to a P_D, P_{FA} pair is also given over the range of samples averaged. Curve 1 in Figs. 6 and 7 compares the detector performance for the same S/N as curve 1 in Fig. 5, but when the P_{FA} is 5×10^{-4}, and 0.09, respectively, and curve 2 in Figs. 6 and 7 do the same for the S/N that applies to curve 2 in Fig. 5.



CURVE 1	6.4	-2	-8	-14	*(0.01)
CURVE 2	11	2	-5	-11	*(0.5)
CURVE 3	14.6	5	-2.5	-9	*(0.995)

Fig. 5 — Comparison of linear and square-law detectors. S/N (dB) in filter bandwidth, for $P_D = (*)$, $P_{FA} = 10^{-6}$.

VI. CONCLUSION

We have given curves which show that the sensitivity for detecting small signals using an envelope detector is degraded when any of several common kinds of nonlinear processes occur between the detector output and the averager used prior to the decision threshold. When only one sample of the detector output is used to form a decision, the system will be oblivious to the presence of the nonlinear process as long as the decision threshold lies within the linear range.

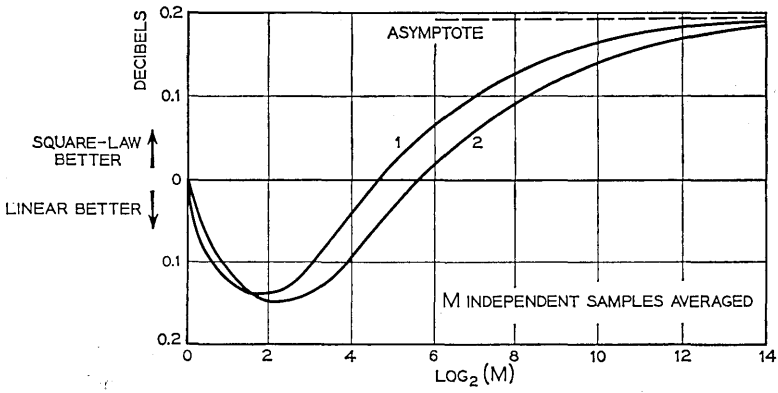


Fig. 6 — Comparison of linear and square-law detectors, $P_{FA} = 5.10^{-4}$.

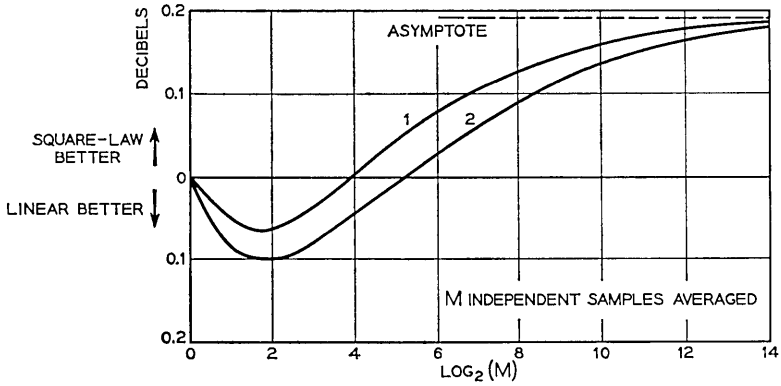


Fig. 7 — Comparison of linear and square-law detectors, $P_{FA} = 0.09$.

The system performance in such a case is thus restricted to threshold variations within this range.

It is interesting that J. V. Harrington has analyzed the detection of repeated signals in noise using binary integration.[2] The system he assumed corresponds closely to the one discussed here for which Fig. 3 shows the results. Harrington deduced that the optimum position for the quantizing threshold was about 1.44 times the Rayleigh mean, in which case binary integration would give results 0.77 dB poorer than if ideal processing with a linear detector had been used. These numbers are in excellent agreement with the minimum degradation condition shown in Fig. 3.

We have shown that the presence of an unwanted CW signal, even −10 dB with respect to the narrow-band noise, causes some degradation of the sensitivity for detecting small CW signals, and the degradation increases rapidly as the level of the unwanted component rises from there.

Computation shows that there is little performance difference to be expected from the use of envelope or square-law detectors. Envelope detectors would thus seem preferable in view of their greater simplicity, especially when a considerable dynamic range is to be covered.

REFERENCES

1. Robertson, G. H., "Operating Characteristics for a Linear Detector of CW Signals in Narrowband Gaussian Noise," B.S.T.J., *46*, No. 4 (April 1967), pp. 755–774.
2. Harrington, J. V., "An Analysis of the Detection of Repeated Signals in Noise by Binary Integration," I.R.E. Trans. Inform. Theory, *IT-1*, No. 1 (March 1955), pp. 1–9.

Phase Principle for Measuring Location or Spectral Shape of a Discrete Radio Source

By A. J. RAINAL

(Manuscript received May 19, 1967)

This paper describes a phase principle for measuring the location or the spectral shape of a discrete radio source. The phase principle is relatively simple to implement and leads to a measurement of location or spectral shape which is insensitive to receiver gain fluctuations. For measuring the location of a weak, discrete radio source, the theoretical accuracy is slightly better than the theoretical accuracy resulting from the Ryle interferometer. For measuring the spectral shape of a weak, discrete radio source, the theoretical accuracy is slightly better than the theoretical accuracy resulting from either the Ryle interferometer or the Dicke radiometer. Furthermore, the implementation of the phase principle doesn't require input switching. Also, the calibration curve associated with the phase principle is independent of changes in the average receiver gains of the two receivers.

I. INTRODUCTION

In many branches of science and technology observations of a discrete radio source provide fundamental knowledge. In the field of radar the illuminated target serves as the discrete radio source. In the field of space exploration the radio transmitter on-board the space vehicle serves as the discrete radio source. In the field of radio astronomy the "radio star" serves as the discrete radio source.

The "radio star" is a remarkable example of a discrete radio source. In the past twenty years radio astronomers have discovered that nature provided many discrete radio sources or radio stars at certain locations in the sky. What are the locations of these radio stars? What is the power spectrum of the observed radiation from a particular radio star? Answers to such questions are of fundamental importance in the field of radio astronomy.[1]

In order to measure relatively small values of radiated power from a discrete radio source, one must compete with the inevitable background noise and the inevitable radio receiver noise. It is well known that one requires a method of measurement which is relatively insensitive to receiver gain fluctuations. The papers by Dicke[2] and Ryle[3] discuss this important point in more detail. In fact, the present day method for measuring relatively small values of radiated power from a discrete radio source makes use of some form of the Dicke[2] radiometer or the Ryle[3] interferometer.

The purpose of this paper is to describe a phase principle for measuring the location or the spectral shape of a discrete radio source. We shall see that the phase principle is relatively simple to implement and leads to a measurement of location or spectral shape which is insensitive to receiver gain fluctuations. For measuring the location or spectral shape of a weak, discrete radio source, we shall see that the theoretical accuracy associated with the phase principle is slightly better than the theoretical accuracy associated with the Dicke radiometer or the Ryle interferometer. We shall also see that for measuring the location or spectral shape of a weak, discrete radio source using only phase information, the accuracy associated with the phase principle is essentially equal to the accuracy associated with the maximum likelihood principle.

II. MEASUREMENTS BASED ON THE PHASE PRINCIPLE

2.1 *Implementation*

Fig. 1 illustrates a simplified implementation of the phase principle for measuring the location or spectral shape of a discrete radio source. $S(t)$, $N_1(t)$, and $N_2(t)$ represent zero mean, independent, narrow-band, stationary Gaussian processes. $N_1(t)$ and $N_2(t)$ each represent the sum of background noise plus receiver noise. $N_1(t)$ and $N_2(t)$ are assumed to have equal variances. The spacing, d, between the two antennas is many wavelengths in order that $N_1(t)$ and $N_2(t)$ can be considered as independent processes. $S(t - \Delta t)$ and $S(t + \Delta t)$ are due to the presence of a discrete radio source located at a small angle θ with respect to boresight. We assume that the receivers preserve the phase difference between the antenna excitations.

$S(t - \Delta t) + N_1(t)$ and $S(t + \Delta t) + N_2(t)$ represent the outputs of the two receivers. η_i represents the ith independent sample of the phase difference between $S(t - \Delta t) + N_1(t)$ and $S(t + \Delta t) + N_2(t)$.

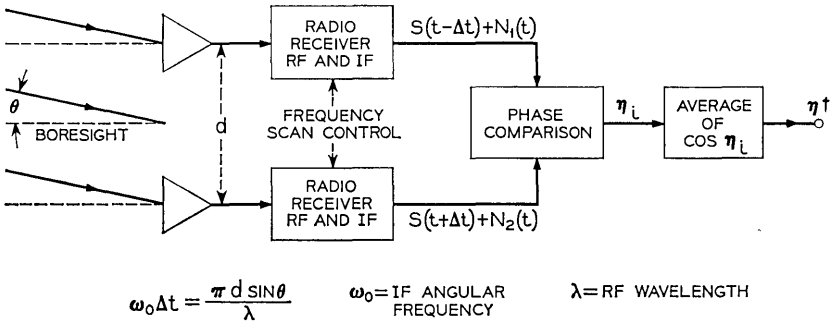


Fig. 1 — Simplified implementation of a phase principle for measuring location or spectral shape of a discrete radio source. When source is located at $d \sin \theta = \lambda/4$, $\eta^\dagger \doteq 0$. When source is located at $\theta = 0$, receivers scan in frequency and $\eta^\dagger$ traces out a measure of spectral shape.

η_i is taken to be in the primary interval $(-\pi, \pi)$. After n such samples the output $\eta^\dagger$ is given by

$$\eta^\dagger = \frac{1}{n} \sum_{i=1}^{n} \cos \eta_i \ , \tag{1}$$

where

$$\begin{aligned} n &= B\tau \\ B &= \text{IF bandwidth} \\ \tau &= \text{observation time.} \end{aligned}$$

We shall assume that n is relatively large like $n \geqq 10^4$, since we are primarily interested in observing relatively weak, discrete radio sources.

Fig. 2 illustrates a relatively simple method for generating $\eta^\dagger$ from the inputs $S(t - \Delta t) + N_1(t)$ and $S(t + \Delta t) + N_2(t)$. R_1 , ω_0 , and θ_1 represent the envelope, IF angular frequency, and phase angle, respec-

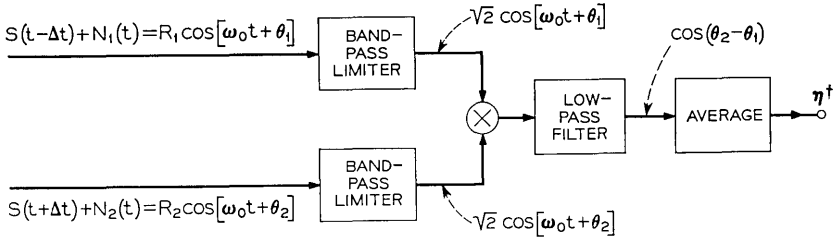


Fig. 2 — A method for generating $\eta^\dagger$ from the inputs $S(t-\Delta t)+N_1(t)$ and $S(t+\Delta t)+N_2(t)$. The band-pass limiters remove all amplitude information.

tively, of the narrow-band Gaussian process $S(t - \Delta t) + N_1(t)$. Similarly, R_2, ω_0, and θ_2 represent the envelope, IF angular frequency, and phase angle respectively of the narrow-band Gaussian process $S(t + \Delta t) + N_2(t)$. The band-pass limiters shown in Fig. 2 are well-known devices for removing all amplitude information and preserving the phase information as is indicated in Fig. 2. See Davenport and Root[4] for a discussion of the band-pass limiter.

As indicated in Fig. 2, if one takes the product of $\sqrt{2}\cos(\omega_0 t + \theta_1)$ and $\sqrt{2}\cos(\omega_0 t + \theta_2)$ and passes the result through a suitable low-pass filter the result is $\cos(\theta_2 - \theta_1) \equiv \cos \eta(t)$. By taking the average of this result, one can generate $\eta^\dagger$.

This method of generating $\eta^\dagger$ can also be used to help implement the phase principle described in Ref. 5 in order to detect the presence of a discrete radio source located at $\theta = 0$.

Fig. 2 indicates clearly that $\eta^\dagger$ is independent of receiver gain fluctuations or changes in the average receiver gain of each receiver. Also, the two receivers need not have the same average gain. Thus, an unusually long observation time τ is advantageous when using the phase principle.

Notice that if the band-pass limiters in Fig. 2 are shorted out, we have the well-known correlator configuration.[6,7,8,9]

We shall now show that a measurement of $\eta^\dagger$ leads to a measurement of location or spectral shape of the discrete radio source.

2.2 *Statistical Properties of* η_i

In order to simplify the analysis, we shall always assume that the discrete radio source is at a small angle θ with respect to boresight. To begin, we shall state some known statistical properties of the angle η_i.

Equation (34) of Ref. 10 gives the probability density $p_2(\eta)$ of each independent sample η_i as

$$p_2(\eta) = \frac{1 - l^2}{2\pi}(1 - \beta_2^2)^{-\frac{3}{2}}\left[\beta_2 \sin^{-1}\beta_2 + \frac{\pi\beta_2}{2} + \sqrt{1 - \beta_2^2}\right], \qquad (2)$$

where

$$\beta_2 = l\cos(\eta - \eta_\theta)$$

$$l = \frac{a}{1 + a}$$

$$a = \frac{\text{Var } S(t)}{\text{Var } N_1(t)} = \frac{\text{Var } S(t)}{\text{Var } N_2(t)}$$

$$\text{Var} = \text{Variance}$$

$$\eta_\theta/2 = \frac{\pi}{\lambda} d \sin\theta = \omega_0 \Delta t$$

$$-\frac{\pi}{2} \leqq \sin^{-1}\beta_2 \leqq \frac{\pi}{2}.$$

The Fourier series development of (2) follows from Middleton's[11] equation (9.33)

$$p_2(\eta) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} \frac{l^n \Gamma^2(n/2+1)}{n!} {}_2F_1\left(\frac{n}{2}, \frac{n}{2}; n+1; l^2\right) \cos n(\eta - \eta_\theta), \tag{3}$$

where ${}_2F_1$ is the Gaussian hypergeometric function

$${}_2F_1(\alpha, \beta; \gamma; x) \equiv 1 + \frac{\alpha\beta}{\gamma} x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)} \frac{x^2}{2!} + \cdots$$

and Γ is the gamma function.

The expectations $E \cos \eta_i$ and $E \cos^2 \eta_i$ follow from (3)

$$E \cos \eta_i = \frac{\pi l}{4} {}_2F_1(\tfrac{1}{2}, \tfrac{1}{2}; 2; l^2) \cos \eta_\theta \tag{4}$$

$$E \cos^2 \eta_i = \frac{1}{2} + \frac{l^2}{4} {}_2F_1(1, 1; 3; l^2) \cos 2\eta_\theta \,. \tag{5}$$

We shall see that the phase principle for measuring location or spectral shape of a discrete radio source is based on (4). Equation (4) should be compared with (4) of Ryle,[3] the equation which characterizes the output of a Ryle interferometer. Both equations are proportional to $\cos \eta_\theta$.

2.3 *Measurement of Location*

Let us first consider the problem of measuring the location of a discrete radio source whose true location is some small positive angle θ. From (4) we see that $E \cos \eta_i = 0$ when $\eta_\theta = \pi/2$ or $d \sin\theta = \lambda/4$. This suggests that we observe $\eta^\dagger$ and conclude that $d \sin\theta = \lambda/4$ when $\eta^\dagger \doteq 0$. How accurately can we form an estimate $\hat{\theta}$ of θ in this manner?

For η_θ near $\pi/2$ let the estimate $\hat{\eta}_\theta$ of η_θ be determined from the linear equation

$$\eta^\dagger = \frac{\pi l}{4} {}_2F_1(\tfrac{1}{2}, \tfrac{1}{2}; 2; l^2)\left(\frac{\pi}{2} - \hat{\eta}_\theta\right). \tag{6}$$

Thus,

$$\text{Var } \hat{\eta}_\theta = (2\pi)^2\left(\frac{d}{\lambda}\right)^2 \text{Var } \hat{\theta} = \left[\text{Var } \eta^\dagger\right]\left[\left(\frac{\pi l}{4}\right)_2F_1(\tfrac{1}{2}, \tfrac{1}{2}; 2; l^2)\right]^{-2} \quad (7)$$

or, in a more suitable form,

$$n\left(\frac{ad}{\lambda}\right)^2 \text{Var } \hat{\theta} = \frac{4}{\pi^4}(1 + a)^2(E \cos^2 \eta_i)_2F_1^{-2}(\tfrac{1}{2}, \tfrac{1}{2}; 2; l^2), \quad (8)$$

where $E \cos^2 \eta_i$ is given by (5) with $\eta_\theta = \pi/2$. Equation (8) characterizes the theoretical accuracy associated with the phase principle for measuring the location of a discrete radio source and is plotted in Fig. 3.

2.4 *Measurement of Spectral Shape*

Now let us consider the problem of measuring the spectral shape of a discrete radio source located at $\theta = 0$. We shall assume that the variances of the background noises and receiver noises are invariant over the frequency region of interest. Under these conditions the estimate $\hat{a}$ of the signal-to-noise power ratio "a" can serve as an estimate of spectral shape by using the well-known frequency scan technique

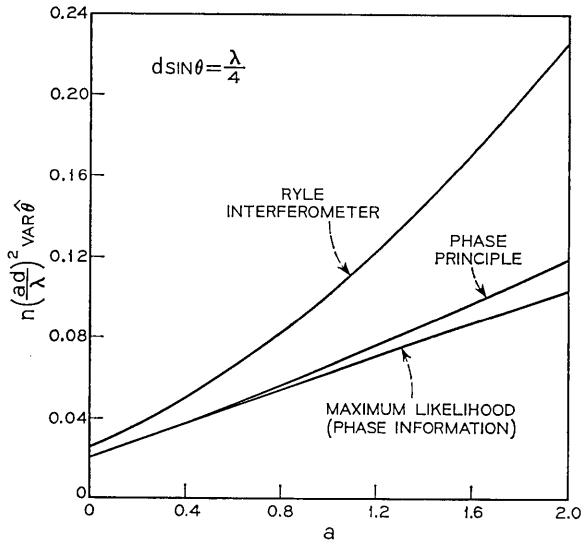


Fig. 3 — Theoretical accuracies of Ryle interferometer, phase principle, and maximum likelihood principle (using only phase information) for measuring the angular location of a discrete radio source.

indicated in Fig. 1. How accurately can we form an estimate $\hat{a}$ of "a" in this manner?

Equation (4) with $\theta = 0$ or $\eta_\theta = 0$ defines "a" as an implicit function of $E \cos \eta_i$ which we shall indicate by

$$a = H(E \cos \eta_i). \tag{9}$$

Equation (9) suggests that we form an estimate $\hat{a}$ of "a" from the equation

$$\hat{a} = H(\eta^\dagger). \tag{10}$$

A plot of (10) is presented in Fig. 4. This figure can be considered as a theoretical calibration curve. Notice that the theoretical calibration curve is independent of changes in the average receiver gain of each receiver. This is indeed unusual. One measures $\eta^\dagger$ and reports the corresponding value of $\hat{a}$. Assuming that Var N_1 and Var N_2 are invariant with frequency over the frequency range of interest, $\hat{a}$ will then trace out the spectral shape of the discrete radio source as the receivers scan in frequency. We shall now characterize the accuracy of the estimate $\hat{a}$.

For large n, the only case of interest in this paper, Cramér's[12] work shows that the estimate $\hat{a}$ is characterized, approximately, by a Gaus-

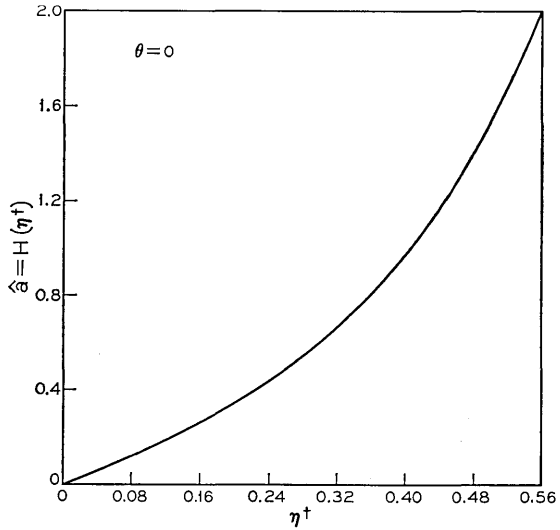


Fig. 4 — The theoretical calibration curve associated with the phase principle for measuring spectral shape of a discrete radio source. The receivers scan in frequency and â traces out the spectral shape.

sian probability density having the following expectation and variance:

$$E\hat{a} = H(E \cos \eta_i) + O(n^{-1}) \doteq a \tag{11}$$

$$\text{Var } \hat{a} = H_1^2 \text{ Var } \eta^{\dagger} + O(n^{-\frac{3}{2}}) \doteq H_1^2 \text{ Var } \eta^{\dagger}, \tag{12}$$

where

$$H_1 = \frac{d\hat{a}}{d\eta^{\dagger}}\bigg|_{E \cos \eta_i} = \left[\frac{dE \cos \eta_i}{da}\bigg|_0\right]^{-1}$$

$$= \frac{4(1+a)^2}{\pi}\left[\frac{l^2}{4}\,{}_2F_1(\tfrac{3}{2}, \tfrac{3}{2}; 3; l^2) + {}_2F_1(\tfrac{1}{2}, \tfrac{1}{2}; 2; l^2)\right]^{-1}.$$

This last result follows by differentiating (4) with respect to "a", setting $\eta_\theta = 0$, and then taking its reciprocal. Equation (11) implies that the estimate $\hat{a}$ is essentially unbiased for large n.

From (12) and (1) we have

$$n \text{ Var } \hat{a} \doteq H_1^2(E \cos^2 \eta_i - E^2 \cos \eta_i), \tag{13}$$

where $E \cos \eta_i$ and $E \cos^2 \eta_i$ are given by (4) and (5) with $\eta_\theta = 0$.

Equation (13) characterizes the theoretical accuracy associated with the phase principle for measuring the signal-to-noise power ratio "a" or the spectral shape of the discrete radio source and is plotted in Fig. 5.

III. MEASUREMENTS BASED ON THE RYLE INTERFEROMETER OR DICKE RADIOMETER

3.1 *Measurement of Location*

When θ is small and $\eta_\theta = \pi/2$ or $d \sin \theta = \lambda/4$, the theoretical accuracy associated with the Ryle interferometer for measuring the location of the discrete radio source was derived by Manasse.[13] In our notation Manasse's[13] (60) becomes

$$n\left(\frac{ad}{\lambda}\right)^2 \text{Var } \hat{\theta} = (2\pi)^{-2}(1+a)^2. \tag{14}$$

Equation (14) characterizes the theoretical accuracy associated with the Ryle interferometer for measuring the location of the discrete radio source and is plotted in Fig. 3.

3.2 *Measurement of Spectral Shape*

One can measure the spectral shape of the discrete radio source located at $\theta = 0$ by using the Ryle[3] interferometer or the Dicke[2] radi-

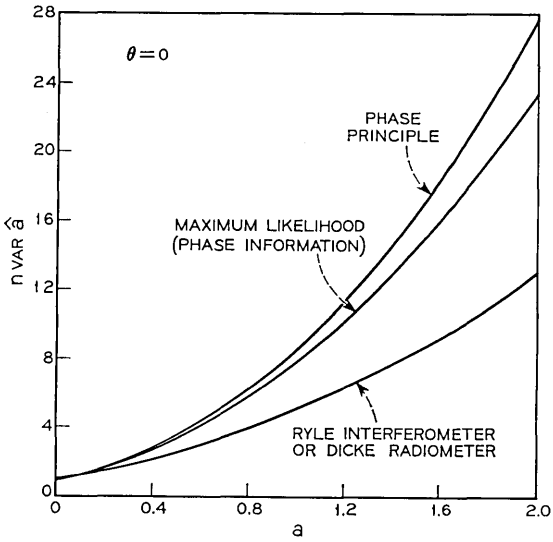


Fig. 5 — Theoretical accuracies of Ryle interferometer, Dicke radiometer, phase principle, and maximum likelihood principle (using only phase information) for measuring the spectral shape of a discrete radio source. Both the Ryle interferometer and the Dicke radiometer require some amplitude information.

ometer. In fact, these methods are at present the accepted methods. We shall go on to characterize the accuracy associated with these methods of measuring spectral shape.

In order to simplify the notation of this section, let $P_s = \text{Var } S(t)$ and $P_N = \text{Var } N_1(t) = \text{Var } N_2(t)$. Then,

$$a = \frac{P_s}{P_N} \quad \text{and} \quad \hat{a} = \frac{\hat{P}_s}{P_N}. \tag{15}$$

$\hat{P}_s$ denotes the unbiased estimate of P_s and P_N is regarded as a parameter.

The Ryle[3] interferometer utilizes a phase reversing switch to produce, periodically, the following inputs to a square-law detector

$$\frac{2S(t) + N_1(t) + N_2(t)}{2} \tag{16}$$

or

$$\frac{N_1(t) + N_2(t)}{2}. \tag{17}$$

Thus, using some of Rice's[14] results, the mean values at the output of the square-law detectors are, periodically,

$$\frac{2P_s + P_N}{2} \tag{18}$$

or

$$\frac{P_N}{2}. \tag{19}$$

The difference in the outputs of the square-law detector is taken as the unbiased estimate $\hat{P}_s$. Thus, $E\hat{P}_s = P_s$, and $E\hat{a} = a$.

Also, using some of Rice's[14] results, the variances at the output of the square-law detector are, periodically,

$$\frac{[2P_s + P_N]^2}{4(n/2)} \tag{20}$$

or

$$\frac{[P_N]^2}{4(n/2)}. \tag{21}$$

The factor $n/2$ appears because the output is in either position only one half of the time. Since the difference in the outputs of the square-law detector is taken as the unbiased estimate $\hat{P}_s$, the variance of $\hat{P}_s$ is given by the sum of expressions (20) and (21):

$$\operatorname{Var} \hat{P}_s = \frac{[2P_s + P_N]^2 + [P_N]^2}{2n} \tag{22}$$

or

$$n \operatorname{Var} \hat{a} = nP_N^{-2} \operatorname{Var} \hat{P}_s = 2a^2 + 2a + 1. \tag{23}$$

Notice that the Ryle interferometer can be considered as a Dicke[2] radiometer switching between the two inputs given by expressions (16) and (17). Thus, (23) characterizes the accuracy of both the Ryle interferometer and the Dicke radiometer for measuring the spectral shape of the discrete radio source. Equation (23) is plotted in Fig. 5.

IV. MEASUREMENTS BASED ON THE MAXIMUM LIKELIHOOD PRINCIPLE USING ONLY PHASE INFORMATION

4.1 *Measurement of Location*

If one uses the maximum likelihood principle[12] to process a large number n of independent samples of the phase difference η_i in order

to estimate the location of the discrete radio source when $\eta_\theta = \pi/2$ or $d \sin \theta = \lambda/4$, one finds

$$n\left(\frac{ad}{\lambda}\right)^2 \operatorname{Var} \hat{\theta} = \frac{n}{(2\pi)^2} a^2 \operatorname{Var} \hat{\eta}_\theta = \frac{1}{(2\pi)^2} \left\{ \int_{-\pi}^{\pi} \left[\frac{1}{p_2} \left(\frac{1}{a} \frac{\partial p_2}{\partial \eta_\theta} \right)^2 \right]_{\pi/2} d\eta \right\}^{-1}, \tag{24}$$

where p_2 is given by (2) and the integrand of the integral in equation (24) is to be evaluated at $\eta_\theta = \pi/2$. For various values of "a", the definite integral appearing in (24) was evaluated numerically by using a digital computer and Simpson's rule. The resulting curve is plotted in Fig. 3. Incidentally, this curve applies for all values of η_θ.

As $a \to 0$ we find that (8) and (24) both yield

$$\lim_{a \to 0} \left[n\left(\frac{ad}{\lambda}\right)^2 \operatorname{Var} \hat{\theta} \right] = \frac{2}{\pi^4} \doteq 0.02053. \tag{25}$$

Thus, as $a \to 0$, the phase principle and the maximum likelihood principle using only phase information are essentially equivalent.

4.2 *Measurement of Spectral Shape*

If one uses the maximum likelihood principle[12] to process a large number n of independent samples of the phase difference η_i in order to estimate the signal-to-noise power ratio "a" or the spectral shape of a discrete radio source when $\theta = 0$, one finds

$$n \operatorname{Var} \hat{a} = \left\{ 2 \int_0^{\pi} \frac{1}{p_2} \left[\frac{\partial p_2}{\partial a} \right]^2 d\eta \right\}^{-1}, \tag{26}$$

where p_2 is given by equation (2) with $\eta_\theta = 0$. For various values of "a", the definite integral appearing in (26) was evaluated numerically by using a digital computer and Simpson's rule. The resulting curve is plotted in Fig. 5. This curve also applies for all values of η_θ.

As $a \to 0$ we find that (13) and (26) both yield

$$\lim_{a \to 0} [n \operatorname{Var} \hat{a}] = \frac{8}{\pi^2} \doteq 0.81057. \tag{27}$$

Thus, as $a \to 0$ the phase principle and the maximum likelihood principle using only phase information are essentially equivalent.

V. COMPARISONS OF THEORETICAL ACCURACIES

5.1 *Measurements of Location*

When using the Ryle interferometer to measure the location of a weak, discrete radio source located at $d \sin \theta = \lambda/4$, we have, from (14),

$$\lim_{a \to 0} \left[n \left(\frac{ad}{\lambda} \right)^2 \operatorname{Var} \hat{\theta} \right] = \frac{1}{(2\pi)^2} \doteq 0.02533. \tag{28}$$

Whereas, when using the phase principle or the maximum likelihood principle to measure the location, we have, from (25),

$$\lim_{a \to 0} \left[n \left(\frac{ad}{\lambda} \right)^2 \operatorname{Var} \hat{\theta} \right] = \frac{2}{\pi^4} \doteq 0.02053. \tag{29}$$

Thus, the phase principle and the maximum likelihood principle are essentially equivalent, and they are both slightly more accurate than the Ryle interferometer.

See Fig. 3 for a comparison of the theoretical accuracies at other values of "a".

5.2 *Measurements of Spectral Shape*

When using the Ryle interferometer or the Dicke radiometer to measure the signal-to-noise power ratio "a" or the spectral shape of a weak, discrete radio source located at $\theta = 0$, we have, from (23),

$$\lim_{a \to 0} [n \operatorname{Var} \hat{a}] = 1. \tag{30}$$

Whereas, when using the phase principle or the maximum likelihood principle to measure the signal-to-noise power ratio "a" or the spectral shape, we have, from (27),

$$\lim_{a \to 0} [n \operatorname{Var} \hat{a}] = \frac{8}{\pi^2} \doteq 0.81057. \tag{31}$$

Again, the phase principle and the maximum likelihood principle are essentially equivalent, and they are both slightly more accurate than either the Ryle interferometer or the Dicke radiometer.

See Fig. 5 for a comparison of the theoretical accuracies at other values of "a".

For values of "a" away from zero, Fig. 5 shows that the Ryle interferometer or Dicke radiometer are more accurate than the maximum likelihood principle using only phase information. Thus, one must conclude that the Ryle interferometer or the Dicke radiometer require

some amplitude information. Consequently, their accuracy is subject to deterioration by gain variations.

Notice that (29) divided by (28) equals $8/\pi^2$, and (31) divided by (30) also equals $8/\pi^2$. Thus, for measuring the location or the spectral shape of a weak, discrete radio source, Var $\hat{\theta}$ and Var $\hat{a}$ associated with the phase principle are lower, by the same factor $8/\pi^2$, than the corresponding variances associated with the Ryle interferometer.

VI. CONCLUSIONS

For measuring the location or the spectral shape of a discrete radio source, the phase principle leads to a measurement which is insensitive to receiver gain fluctuations.

For measuring the location or the spectral shape of a weak, discrete radio source, the accuracy associated with the phase principle is slightly better than the accuracy associated with the Ryle interferometer or the Dicke radiometer. Also, the accuracy associated with the phase principle is essentially equal to the accuracy associated with the maximum likelihood principle using only phase information.

The phase principle is relatively simple to implement, and the implementation doesn't require input switching.

The calibration curve associated with the phase principle is independent of changes in the average receiver gain of each receiver. The two receivers need not have the same average gain.

An unusually long observation time is advantageous when using the phase principle.

For measuring spectral shape, both the Ryle interferometer and the Dicke radiometer require some amplitude information. Consequently, their accuracy is subject to deterioration by gain variations.

VII. ACKNOWLEDGMENT

The author is indebted to Miss A. T. Seery for programming a digital computer to produce Figs. 3, 4, and 5.

REFERENCES

1. Radio Astronomy Issue, Proc. IRE, *46,* No. 1, January, 1958. (Entire Issue)
2. Dicke, R. H., The Measurement of Thermal Radiation at Microwave Frequencies, Rev. Sci. Instr., *17,* July, 1946, pp. 268–275.
3. Ryle, M., A New Radio Interferometer and its Application to the Observation of Weak Radio Sources, Proc. Royal Society, A, *211,* March, 1952, pp. 351–375.
4. Davenport, W. B., Jr. and Root, W. L., *Random Signals and Noise,* McGraw-Hill Book Company, Inc., New York, New York, 1958.

5. Rainal, A. J., Phase Principle for Detecting Narrow-Band Gaussian Signals, B.S.T.J., *45,* January, 1966, pp. 143–148.
6. Goldstein, S. J., Jr., A Comparison of Two Radiometer Circuits, Proc. IRE, *43,* November, 1955, pp. 1663–1666.
7. Tucker, D. G., Graham, M. H., and Goldstein, S. J., Jr., A Comparison of Two Radiometer Circuits, Proc. IRE, *45,* March, 1957, pp. 365–366.
8. Blum, E.-J., Sensibilité Des Radiotélescopes Et Récepteurs A Corrélation, Annales D'Astrophysique, *22,* No. 2, March-April, 1959, pp. 140–163.
9. Tiuri, M. E., Radio Astronomy Receivers, IEEE Trans. Ant. Prop., *AP-12,* December, 1964, pp. 930–938.
10. Rainal, A. J., Monopulse Radars Excited by Gaussian Signals, IEEE Trans. Aerospace Electron. Systems, *AES-2,* No. 3, May, 1966.
11. Middleton, D., *Introduction to Statistical Communication Theory,* McGraw-Hill Book Company, Inc., New York, New York, 1960.
12. Cramér, H., *Mathematical Methods of Statistics,* Sections 27.7, 28.4, and 32.3, Princeton University Press, Princeton, N. J., 1946.
13. Manasse, R., Maximum Angular Accuracy of Tracking a Radio Star, IRE Trans. Ant. Prop., *AP-8,* January, 1960, pp. 50–56.
14. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., *24,* Section 4.1. January, 1945.

Some Considerations of Broadband Noise Performance of Optical Heterodyne Receivers

By V. K. PRABHU

(Manuscript received September 8, 1967)

We derive an explicit expression in this paper for the spot noise factor of a perfectly aligned optical heterodyne receiver consisting of a semiconductor photodiode followed by an IF amplifier. We show that this noise factor F_R, which is a function of the admittance of the diode, varies as a function of the modulation frequency. We obtain constraints imposed by the photodiode on the broadband noise performance of the optical receiver for any arbitrary lossless interstage network. The integral form of the constraint shows that the noise factor F_R cannot be made equal to its optimum value F_{RO} over any nonzero band of frequencies. We give explicit expressions for the amount of tolerance of broadband noise performance obtained with lossless interstage networks. We show that for certain types of approximations, and for a certain transistor IF amplifier usually used in practice, the interstage network which achieves broadband signal performance for the receiver also obtains broadband noise performance. The theory of broadband noise performance we present for optical heterodyne receivers can also be applied to the study of broadband noise performance of other linear systems normally encountered in practice.

I. INTRODUCTION

Semiconductor photodiodes like Schottky barrier diodes or conventional p-n or p-i-n diodes are increasingly being used for detection in optical heterodyne (double detection) receivers.[1-12] They normally are fast and efficient, converting up to 70 per cent of the photons of the light beam into photoelectric current.[11] Because of the intensity of the light beam, almost all photodiodes used in optical detection give an output proportional to the intensity of light.[1] However, this output is normally so small that further amplification is required, and

so any practical receiver consists of a photodiode followed by a high-gain low-noise IF amplifier.

We have already considered in detail the signal performance of such receivers.[13] In this paper we shall deal only with the noise performance of the receiver. The output of the diode is usually corrupted by noise generated within the diode and elsewhere in the system. We discuss briefly the characteristics of the photodiode in Section II and give its equivalent circuit. In Section III we discuss the noise performance of the IF amplifier and show that its noise factor is a function of its source admittance.[14] We show that the noise factor F_R of the optical receiver is a function of frequency in spite of the fact that the IF amplifier has a broadband noise performance characteristic.

In Section IV we discuss the role of the lossless interstage network in achieving broadband noise performance of the optical receiver and show that it is impossible to make the noise factor F_R equal to its optimum value F_{RO} over any nonzero band of frequencies.

Section V shows that Butterworth and Chebyshev approximations to F_R are realizable, and we obtain the tolerance of broadband noise performance for these approximations. We show that, for the photodiodes normally encountered in practice, this tolerance ϵ^2 is a monotonically increasing function of the complexity of the interstage network but decreasing for Chebyshev approximations.

We show in Section VI that to obtain broadband signal and noise performance characteristics from the optical receiver two separate lossless interstage networks are necessary. However, we also show that for certain types of approximations and for a certain transistor IF amplifier, the interstage network which achieves broadband signal performance also obtains broadband noise performance for the optical receiver.

The theory of broadband noise performance presented in this paper for an optical heterodyne receiver can also be applied to obtain broadband noise performance of other linear systems normally encountered in practice.

II. AVAILABLE SIGNAL AND NOISE OUTPUT POWERS

Fig. 1 shows the optical heterodyne receiver that we discuss. It consists of a photodiode followed by a lossless interstage network, and an IF amplifier of center frequency Ω_0 and a semibandwidth W.† The

† The amplifier may, depending on the frequency of modulation, use vacuum tubes, transistors, masers, parametric amplifiers, tunnel diodes, or other active devices.

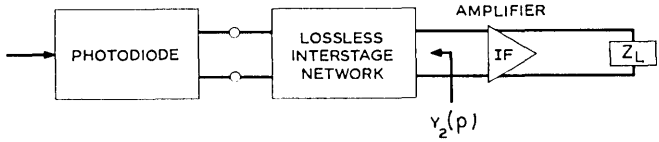


Fig. 1 — A double detection optical receiver. The input to the photodetector is the sum of the local oscillator beam and the incoming signal beam.

geometrical center frequency ω_o of the IF amplifier is defined as

$$\omega_o = \{(\Omega_o - W)(\Omega_o + W)\}^{1/2}. \tag{1}$$

The available gain G_o of the IF amplifier and its optimum noise factor F_o are assumed to be independent of the frequency ω for $\Omega_o - W \leqq \omega \leqq \Omega_o + W$.‡

In this paper we shall not consider any effects on the noise performance of the optical receiver of beam misalignment, nonuniformity of the surface of the diode, or distortion from transmission through a nonhomogeneous atmosphere.[20] Because background noise has been shown to be of almost negligible consideration, we shall assume that this noise has no effect on the broadband noise performance of the optical receiver.[21, 22]

The diode is normally so arranged that the junction or portions of it close to the junction are illuminated by the sum of the local oscillator beam and the incoming signal beam. The electron-hole pairs thus created by the incoming photons give rise to a small signal current.[1] In operation, a reverse bias V_o is put on the diode, and the characteristics of the device[12] for small excursions around V_o are that of a signal current generator I_s, and a direct current generator I_o, in parallel with a capacitance $C(V_o)$, and this combination in series with a parasitic series resistance $R(V_o)$.

The time-average current I_o is caused both by the time-average illumination and by the electrons and holes that are generated at or near the junction. The signal current I_s is caused by that portion of the illumination at or near the signal frequency of interest. In general, $C(V_o)$ and $R(V_o)$ are functions of the bias voltage. Assuming that, in practical cases, the excursions around the bias point are small, we shall henceforth assume that $C(V_o)$ is a constant capacitance C, and the series resistance $R(V_o)$ is a constant resistance R.

‡ Since G_o and F_o are real and even functions of ω, we shall only consider $\omega \geq 0$. The case in which G_{of} and B_{of} are functions of frequency ω is very complicated and will not be discussed in this paper.

To account for any thermal noise generated in the photodiode, we use an equivalent noise-voltage generator e_n with mean-squared value†

$$\overline{|e_n|^2} = 4kT_dR\,\Delta f, \tag{2}$$

where T_d is the temperature of the diode, k is Boltzmann's constant, and Δf is the spot frequency band about which we are concerned. In addition, there is another source of noise, shot noise, present in the photodiode. This can be accounted for[12] by placing in parallel with I_s and I_o a shot noise current generator i_n with mean-squared value

$$\overline{|i_n|^2} = 2qI_o\,|f(\omega\tau)|^2\,\Delta f, \tag{3}$$

where q is the electronic charge, and $f(\omega\tau)$ is a transit time reduction factor, τ being some effective transit time. We assume in this paper that $f(\omega\tau)$ which always satisfies the inequality

$$|f(\omega\tau)| \leqq 1, \tag{4}$$

can be considered independent of ω in $\Omega_o - W \leqq \omega \leqq \Omega_o + W$, and that

$$|f(\omega\tau)| = 1. \tag{5}$$

The equivalent circuit of the photodiode which describes its terminal signal and noise characteristics is shown in Fig. 2. This equivalent

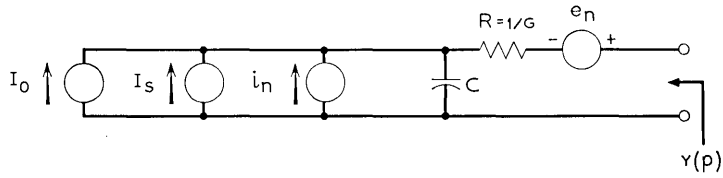


Fig. 2—Equivalent circuit of photodiode. The physical sources of noise that are present in the diode are also shown. The conductance G_p which appears in parallel with C is usually so small (around 10^{-7} mho) that it can be neglected for all practical purposes if $\omega \gg 0$.

circuit very well describes the behavior of the diode provided the lowest frequency of the signal occurring in the system is very far from zero, or $\omega \gg 0$.

The peak photoelectric current I_s, and dc current I_a for a double detection optical receiver can be shown to be given by[23,24]

$$I_s = \frac{2\eta q}{h\nu}\sqrt{P_oP_s}, \tag{6}$$

† The horizontal bar denotes an average.

and

$$I_o = \frac{\eta q}{h\nu}(P_o + P_s), \tag{7}$$

where η is the quantum efficiency of the photodiode,† h is Planck's constant, ν is the optical frequency, P_s is the signal power, and P_o is the local oscillator power.‡

The available signal and noise powers are easily determined from Fig. 2. The signal available power S_{pd} and noise available power N_{pd} can be written as§

$$S_{pd} = \frac{|I_s|^2}{8\omega^2 C^2 R}, \tag{8}$$

and

$$N_{pd} = \frac{\overline{|e_n|^2}}{4R} + \frac{\overline{|i_n|^2}}{4\omega^2 C^2 R}. \tag{9}$$

An important quantity that characterizes the noise performance of the photodiode is its signal-to-noise ratio S_{pd}/N_{pd}. According to equations 2, 3, and 5 through 9, this is given by¶

$$\frac{S_{pd}}{N_{pd}} = \left(\frac{\eta q}{h\nu}\right)^2 \frac{P_o P_s}{2kT_d G\left(\frac{\omega}{\omega_c}\right)^2 \Delta f + q\left(\frac{\eta q}{h\nu}\right) P_o \Delta f}, \tag{10}$$

where

$$\omega_c = \frac{1}{RC}. \tag{11}$$

The signal-to-noise ratio at the input to the diode will be defined as the best possible signal-to-noise ratio which an ideal detector could

† A quantum efficiency of greater than 70% has been obtained [11] for Schottky barrier photodiodes.

‡ In practice, a fraction $k \leqq 1$ of incident photons are absorbed in the active region of the diode. To account for this effect, I_s and I_o are usually multiplied by a factor k. We assume that $k = 1$. I_s is also usually multiplied by a reduction factor similar to the shot-noise function, $f(\omega\tau)$, determined by the signal frequency, optical wavelength, and device construction. We assume that this factor is unity.

To account for any mismatch between the local oscillator beam and signal beam, I_s and I_o are also multiplied by a beam matching factor β where $\beta \leqq 1$. We assume that $\beta = 1$.

§ We can argue from the physics of the diode that the shot noise source and thermal noise source shown in Fig. 2 are uncorrelated.

¶ We can assume that $P_o \gg P_s$, so that $P_o + P_s \approx P_o$.

achieve. This can be shown to be given by†

$$(S/N)_{in} = \frac{P_s}{h\nu\ \Delta f}. \tag{12}$$

III. IF AMPLIFIER NOISE FACTOR

The terminal characteristics of any IF amplifier used in the optical receiver (see Figs. 3 and 4) can normally be described by

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} + \begin{bmatrix} n_{i1} \\ n_{i2} \end{bmatrix} \tag{13}$$

Fig. 3 — Separation of twoport with internal noise sources into a source-free twoport.

or

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} + \begin{bmatrix} n_v \\ n_i \end{bmatrix} \tag{14}$$

where n_{i1} and n_{i2}, or n_v and n_i characterize all physical sources of noise present in the IF amplifier.[14] we assume that the IF amplifier has ideal broadband signal and noise performance characteristics and that the amplifier remains stable for all linear passive input and output terminations.[26–28]

By definition, the spot noise factor at a specified frequency of any linear twoport network (such as an IF amplifier) is given by the ratio of the total output noise power per unit bandwidth exchangeable‡ at the output port to that portion of that power which is engendered by the input termination at the standard temperature T_o.[14] To derive

† Ref. 25 shows that it is impossible to measure amplitude and phase of an incoming optical signal with a better signal-to-noise ratio than given by equation 12.

‡ Exchangeable power, exchangeable gain, etc. coincide[29] with available power and available gain when the output impedance of the amplifier is positive-real. They are the logical generalizations of the available power and available gain when the output impedance has a negative-real part. Since the amplifier is assumed to be absolutely stable, we may substitute the word "available" for the word "exchangeable" wherever it appears in this paper.

the noise factor of the IF amplifier we are considering, let us connect the amplifier to a statistical source comprising an internal admittance Y_s, and a noise current generator I_{ns} (see Fig. 5). It can be shown easily that the noise factor F is given by

$$F = 1 + \frac{\overline{|\, n_i + n_v Y_s \,|^2}}{\overline{|\, I_{ns} \,|^2}}. \tag{15}$$

We notice that the mean-square source noise current is related to the source conductance G_s by the Nyquist formula

$$\overline{|\, I_{ns} \,|^2} = 4kT_o G_s \, \Delta f. \tag{16}$$

Also, we can express the noise voltage fluctuation $\overline{|\, n_v \,|^2}$ in terms of an equivalent noise resistance R_n as

$$\overline{|\, n_v \,|^2} = 4kT_o R_n \, \Delta f \tag{17}$$

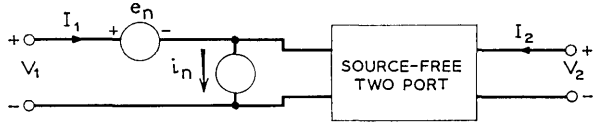


Fig. 4 — The Rothe-Dahlke noise model for a linear twoport network.

and the noise current fluctuation $\overline{|\, n_i \,|^2}$ can be expressed in terms of an equivalent noise conductance G_u where

$$\overline{|\, n_i \,|^2} = 4kT_o G_u \, \Delta f. \tag{18}$$

Let us also write

$$\overline{n_v n_i^*} = 4kT_o \rho \sqrt{R_n G_u} \, \Delta f, \tag{19}$$

where ρ is a complex number. It can be shown that

$$|\, \rho \,| \leqq 1. \tag{20}$$

From equations 15 through 19, the formula for the noise factor becomes§

$$F = 1 + \frac{G_u + 2\sqrt{R_n G_u}\,\{G_s \operatorname{Re} \rho - B_s \operatorname{Im} \rho\} + (G_s^2 + B_s^2)R_n}{G_s}. \tag{21}$$

As we can see from equation 21, the noise factor F is a function of the source conductance G_s and also of its susceptance B_s. We can

§ Re a and Im a denote the real and imaginary parts, respectively, of the complex number a.

show that F attains its optimum value¶

$$F_o = 1 + 2\sqrt{R_n G_u}\,\{\text{Re}\,\rho + [1 - (\text{Im}\,\rho)^2]^{\frac{1}{2}}\} \tag{22}$$

for a certain source admittance $Y_{of} = G_{of} + jB_{of}$ where

$$G_{of} = \frac{\sqrt{R_n G_u}}{R_n}\,\{1 - (\text{Im}\,\rho)^2\}^{\frac{1}{2}} \tag{23}$$

and

$$B_{of} = \frac{\sqrt{R_n G_u}}{R_n}\,\text{Im}\,\rho. \tag{24}$$

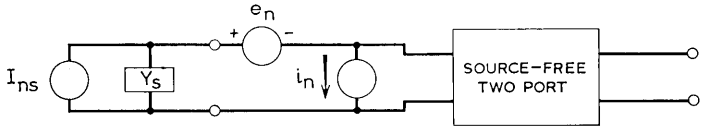


Fig. 5 — Network for noise factor computation.

Using equations 21 through 24 (see Ref. 14), we can show that F can be written in the form†

$$F = F_o + \frac{R_n}{G_s}\,|\,Y_s - Y_{of}\,|^2, \tag{25}$$

where the value of R_n is as given in equation 17.

Let $Y_2(p)$ be the admittance of the photodiode as seen by the IF amplifier (see Fig. 1). We shall now compute the over-all signal-to-noise ratio for the optical receiver, and its overall noise factor F_R. In defining F_R, we shall use the concept of noise factor as originally introduced by Friis and Fränz.[31] This noise factor F_R is defined as

$$F_R = \frac{(S/N)_{\text{in}}}{(S/N)_{\text{out}}}, \tag{26}$$

where $(S/N)_{\text{in}}$ and $(S/N)_{\text{out}}$ are the input and output exchangeable signal-to-noise power ratios of the optical receiver. From equations 10, 12, and 26 we can write

¶ It can be shown that Re ρ + [1−(Im ρ)^2]$^{1/2}$ is always a nonnegative quantity.

† It can be shown (see Ref. 30) that equation 25 can also be written in the form $(G_s - G_{or})^2 + (B_s - B_{or})^2 = F^2{}_p$ where G_{or}, B_{or}, and F_p can be determined from Equations 21-25.

$$(S/N)_{\text{out}} = \frac{\eta P_s}{h\nu\,\Delta f} \frac{1}{1 + \frac{2kT_o}{qI_o}\frac{1}{R}\left(\frac{\omega}{\omega_c}\right)^2\left\{F_o + \frac{R_n}{\text{Re } Y_2} \mid Y_2 - Y_{of} \mid^2 + \frac{T_d - T_o}{T_o}\right\}}, \tag{27}$$

and

$$F_R = \frac{1}{\eta}\left[1 + \frac{2kT_o}{qI_o}\frac{1}{R}\left(\frac{\omega}{\omega_c}\right)^2 \cdot\left\{F_o + \frac{R_n}{\text{Re } Y_2} \mid Y_2 - Y_{of} \mid^2 + \frac{T_d - T_o}{T_o}\right\}\right]. \tag{28}$$

We now note from equations 27 and 28 that $(S/N)_{\text{out}}$ and F_R are functions of the frequency of the detected signal, and for $Y_2 = Y_{of}$, $(S/N)_{\text{out}}$, and F_R attain their optimum values $(S/N)_O$ and F_{RO} where

$$(S/N)_O = \frac{\eta P_s}{h\nu\,\Delta f} \frac{1}{1 + \frac{2kT_o}{qI_o}\frac{1}{R}\left(\frac{\omega}{\omega_c}\right)^2\left\{F_o + \frac{T_t - T_o}{T_o}\right\}}, \tag{29}$$

and

$$F_{RO} = \frac{1}{\eta}\left[1 + \frac{2kT_o}{qI_o}\frac{1}{R}\left(\frac{\omega}{\omega_c}\right)^2\left\{F_o + \frac{T_d - T_o}{T_o}\right\}\right]. \tag{30}$$

If optimum noise performance of the heterodyne receiver at a finite number of signal frequencies is desired, it can be shown[16] that suitable lossless interstage networks can be designed so that $(S/N)_O$ in equation 29 and F_{RO} in Equation 30 can be realized at the respective signal frequencies. If the band of frequencies of interest is continuous and nonzero, it can be shown that we cannot make F_R equal to F_{RO} over the whole band.[15, 16] The question arises whether there are any constraints to be satisfied by F_R, imposed by the diode or any other components of the system, and what lossless interstage networks must be used to make F_R as close to F_{RO} as possible. We shall discuss these two topics in the rest of this paper.

IV. DERIVATION OF INTEGRAL CONSTRAINTS†

We shall use the results obtained in the theory of broadband matching of linear systems[13,13–19] to derive the expressions which relate

† The methods of derivation of most of the results in this section are very similar to those in Ref. 13.

the noise factor of the optical heterodyne receiver to other parameters of the system. The integral relation shows that it is impossible to make F_R equal to F_{RO} over any nonzero band of frequencies. Since the IF amplifier is assumed to have ideal broadband signal and noise performance characteristics, it follows that F_o and Y_{of} are independent of frequency ω for $\Omega_o - W \leqq \omega \leqq \Omega_o + W$.‡ It also follows that $B_{of} = 0$.§ Without any loss of generality we shall normalize all admittances with respect to G_{of}.

Let us look at Fig. 6 and define two reflection coefficients $\rho_1(p)$ and

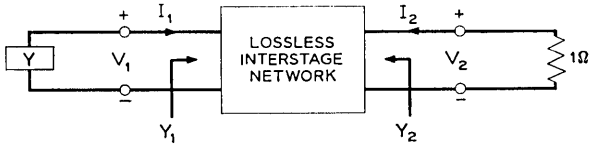


Fig. 6 — Lossless interstage network used in equalizing the noise performance of the optical receiver. $Y(p)$ is the admittance of the photodiode as shown in Fig. 2.

$\rho_2(p)$, and an all-pass function $\beta(p)$ where

$$\rho_1(p) = \frac{Y_1(p) - Y(-p)}{Y_1(p) + Y(p)}, \tag{31}$$

$$\rho_2(p) = \frac{Y_2(p) - 1}{Y_2(p) + 1}, \tag{32}$$

$$\beta(p) = \prod_{r=1}^{m} \frac{p - \alpha_r}{p + \alpha_r^*}, \tag{33}$$

and $p = \sigma + j\omega$ is the complex frequency variable. $Y(p)$ is the admittance of the photodiode as seen by the lossless interstage network, and $\alpha_1, \alpha_2, \cdots, \alpha_m$ are the poles of $Y(-p)$ in Re $p > 0$. Since the interstage network is lossless,[33] it can be shown that

$$| \rho_1(j\omega) | = | \rho_2(j\omega) |. \tag{34}$$

Also from Equations 25, 32, and 34

$$\frac{1}{| \rho_1(j\omega) |^2} = \frac{1}{| \rho_2(j\omega) |^2} = 1 + \frac{4R_nG_{of}}{F - F_o}. \tag{35}$$

‡ The study of the case in which F_o and Y_{of} may be functions of frequency is very complicated and beyond the scope of this paper.

§ For any realizable admittance Y_{of}, $B_{of}(\omega)$ must be an odd function of ω.

From equations 31 and 33, we can also write

$$\beta(p)\{1 - \rho_1(p)\} = \beta(p)\frac{Y(p) + Y(-p)}{Y_1(p) + Y(p)}. \tag{36}$$

Equation 36 shows that regardless of the lossless interstage network used in the receiver, every zero of $\gamma(p) = \frac{1}{2}[1+Y(-p)/Y(p)]$ in Re $p \geq 0$ must also be a zero of

$$\zeta(p) = \beta(p)\{1 - \rho_1(p)\}. \tag{37}$$

A zero p_o of $\gamma(p)$ in Re $p \geq 0$ of multiplicity k is said to be a zero of transmission of the admittance $Y(p)$ of order k. Youla distinguishes four kinds of such zeros.[17] Class 1 contains all those in the strict right-half plane. Class 2 contains all those on the real-frequency axis which are simultaneously zeros of $Y(p)$. Class 3 contains all those on the real-frequency axis for which $0 < | Y(p_o) | < \infty$, and class 4 contains all those for which $| Y(p_o) | = \infty$. The restrictions imposed on $\rho_1(p)$ through equation 36 are formulated[13, 15–19] most compactly in terms of coefficients of the power series expansions of the following quantities:

$$s(p) = \beta(p)\rho_1(p) = \sum_{k=0}^{\infty} S_k(p - p_o)^k \tag{38}$$

$$\ln s(p) = \sum_{k=0}^{\infty} s_k(p - p_o)^k \tag{39}$$

$$\beta(p) = \sum_{k=0}^{\infty} B_k(p - p_o)^k \tag{40}$$

$$\ln \beta(p) = \sum_{k=0}^{\infty} b_k(p - p_o)^k \tag{41}$$

$$F(p) = \beta(p)[Y(p) + Y(-p)] = \sum_{k=0}^{\infty} F_k(p - p_o)^k \tag{42}$$

$$\frac{1}{\pi}\frac{p}{p^2 + \omega^2} = \sum_{k=0}^{\infty} f_k(p - p_o)^k \tag{43}$$

$$g(p) = \frac{F(p)}{2\beta(p)} = \sum_{k=0}^{\infty} \theta_k(p - p_o)^k. \tag{44}$$

Also, let $\eta(p)$ be a regular all-pass network such that

$$\eta(p) = \prod_{l=1}^{\nu} \frac{p - \mu_l}{p + \mu_l^*}, \tag{45}$$

and

$$\ln \eta(p) = \sum_{k=0}^{\infty} \eta_k (p - p_o)^k. \tag{46}$$

μ_i's are a set of points in the right half of the complex plane. Let

$$\frac{1}{s_o(p)s_o(-p)} = 1 + \frac{4R_nG_{of}}{F - F_o}, \tag{47}$$

and let $s_o(p)$ be such that all its zeros and poles are in the left-half plane.† It is now clear‡ that if

$$\beta(p)\rho_1(p) = s(p) = \eta(p)s_o(p), \tag{48}$$

$$\frac{1}{|\, s(j\omega)\,|^2} = \frac{1}{|\, \rho_1(j\omega)\,|^2} = \frac{1}{|\, s_o(j\omega)\,|^2}$$

$$= 1 + \frac{4R_nG_{of}}{F - F_o}. \tag{49}$$

We may now show[17-19] that $Y_1(p)$ is a positive-real admittance if and only if:

(*i*) At every class 1 transmission zero p_o of order k,

$$S_r = B_r, \qquad 0 \leqq r \leqq k - 1; \tag{50}$$

or

$$b_o = \epsilon\pi j + \eta_o - \int_0^{\infty} f_o \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega,$$

$$\epsilon = 0, \quad \text{if} \quad \frac{d}{d\omega}\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} \neq 0, \tag{51}$$

$$\epsilon = 1, \quad \text{if} \quad \frac{d}{d\omega}\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} = 0,$$

and

$$b_r = \eta_r - \int_0^{\infty} f_r \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega, \qquad 1 \leqq r \leqq k - 1. \tag{52}$$

† It will be recognized that $s_o(p)$ is a minimum-phase function.[34]

‡ Multiplication of $\rho_1(p)$ by $\beta(p)$ is necessary to make it analytic in the right-half plane. This multiplication makes $s(p)$ a bounded real scattering coefficient.[17] Multiplication of $s_o(p)$ by $\eta(p)$ introduces right-half plane zeroes. This is sometimes necessary[16] and is done so that $s(p)$ can satisfy all the constraints imposed by $Y(p)$.

(*ii*) At every class 2 transmission zero $j\omega_o$ of order k,

$$S_r = B_r, \qquad 0 \leqq r \leqq k - 1, \tag{53}$$

and

$$\frac{S_k - B_k}{F_{k+1}} \leqq 0; \tag{54}$$

or

$$b_r = \eta_r - \int_0^\infty f_r \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega, \qquad 0 \leqq r \leqq k - 1, \tag{55}$$

and§

$$\frac{b_k - s_k}{\theta_{k+1}} \geqq 0. \tag{56}$$

If $| \omega_o | = 0$, or ∞, equation 56 may be replaced by

$$\frac{b_k - \eta_k + \int_0^\infty f_k \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega}{\theta_{k+1}} \geqq 0. \tag{57}$$

(*iii*) At every class 3 transmission zero of order k,

$$S_r = B_r, \qquad 0 \leqq r \leqq k - 2, \tag{58}$$

and

$$\frac{S_{k-1} - B_{k-1}}{F_k} \leqq 0; \tag{59}$$

or

$$b_r = \eta_r - \int_0^\infty f_r \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega, \qquad 0 \leqq r \leqq k - 1, \tag{60}$$

and

$$\frac{b_{k-1} - s_{k-1}}{\theta_k} \geqq 0, \tag{61}$$

with equality if and only if the matching network is nondegenerate. If $| \omega_o | = 0$ or ∞, equation 61 may be replaced by

$$\frac{b_{k-1} - \eta_{k-1} + \int_0^\infty f_r \ln\left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega}{\theta_k} \geqq 0. \tag{62}$$

§ If the interstage network is nondegenerate[16,17], equation 56 becomes an equality. If $| Y(j\omega_o) | \neq \infty$, the network is said to be nondegenerate if and only if $Y_1(j\omega_o) + Y(j\omega_o) \neq 0$. If $| Y(j\omega_o) | = \infty$, the network is said to be nondegenerate if and only if $| Y_1(j\omega_o) | \neq \infty$.

(*iv*) At every class 4 zero of order k,

$$S_r = B_r , \qquad 0 \leqq r \leqq k - 1, \tag{63}$$

and

$$\frac{F_{k-1}}{S_k - B_k} \leqq a_{-1} ; \tag{64}$$

where a_{-1} is the residue of $Y(p)$ at $p_o = j\omega_o$;
or

$$b_r = \eta_r - \int_0^\infty f_r \ln \left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega, \qquad 0 \leqq r \leqq k - 1, \tag{65}$$

and

$$\frac{2\theta_{k-1}}{b_k - s_k} \geqq a_{-1} , \tag{66}$$

with equality if and only if the matching network is nondegenerate. If $| \omega_0 | = 0$ or ∞, equation 66 may be replaced by

$$\frac{2\theta_{k-1}}{b_k - \eta_k + \int_0^\infty f_k \ln \left\{1 + \frac{4R_nG_{of}}{F - F_o}\right\} d\omega} \geqq a_{-1} . \tag{67}$$

If equations 50 through 67 are satisfied, $Y_1(p)$ is a positive-real admittance. If $Y_1(p)$ is positive-real, the Darlington method can be used to obtain the lossless twoport interstage network needed in the receiver.

Let us now use the theory of broadband noise performance that we have presented to derive the constraints imposed by the photodiode on the noise performance of the optical receiver. The normalized admittance $Y(p)$ of the photodiode shown in Fig. 2 can be written as†

$$Y(p) = \frac{1}{RG_{of}} \frac{p}{p + \omega_c} . \tag{68}$$

From equation 36 we can show that the only transmission zero of $Y(p)$ lies at $p_o = 0$, and is of order 1. Also from equation 33,

$$\beta(p) = \frac{p - \omega_c}{p + \omega_c} \tag{69}$$

$$= -1 + \frac{2p}{\omega_c} - \frac{2p^2}{\omega_c^2} + \cdots . \tag{70}$$

† The equivalent circuit shown in Fig. 2 for the photodiode is valid for frequencies $\omega \gg 0$. Also, without any loss of generality, we normalize all admittances with respect to G_{of} .

From equation 42 we can write

$$F(p) = \frac{1}{G_{of}} \frac{2p^2C}{\omega_c(1 + p/\omega_c)^2} \tag{71}$$

$$= \frac{1}{G_{of}} \left[\frac{2Cp^2}{\omega_c} - \frac{4Cp^3}{\omega_c^2} + \cdots \right]. \tag{72}$$

Since the transmission zero is of class 2, we can write from equations 53 and 54 that

$$S_o = -1, \tag{73}$$

and

$$S_1 \leqq 2/\omega_c , \tag{74}$$

where

$$s(p) = \pm\eta(p)s_o(p), \tag{48}$$

and

$$\frac{1}{s_o(p)s_o(-p)} = 1 + \frac{4R_nG_{of}}{F - F_o} . \tag{47}$$

Also from equations 55 through 57 it follows that

$$\frac{1}{\pi} \int_0^\infty \frac{1}{\omega^2} \ln \left\{ 1 + \frac{4R_nG_{of}}{F - F_o} \right\} d\omega \leqq \frac{2}{\omega_c} - \sum_{l=1}^{\nu} \frac{\mu_l}{\mu_l^*} \left\{ \frac{1}{\mu_l} + \frac{1}{\mu_l^*} \right\}, \tag{75}$$

where μ_l's are a set of points in the right-half plane. Since Re $(1/\mu_l) \geq 0$ for all l, we put $\eta\ (p) = 1$. We, therefore, have

$$\frac{1}{\pi} \int_0^\infty \frac{1}{\omega^2} \ln \left\{ 1 + \frac{4R_nG_{of}}{F - F_o} \right\} d\omega \leqq \frac{2}{\omega_c}. \tag{76}$$

Also, from equations 25, 28, and 30, we can write equation 76 as

$$\frac{1}{\pi} \int_0^\infty \frac{1}{\omega^2} \ln \left[1 + \frac{8kT_o}{\eta q I_o} \frac{R_nG_{of}}{R} \frac{(\omega/\omega_c)^2}{F_R - F_{RO}} \right] d\omega \leqq \frac{2}{\omega_c}. \tag{77}$$

We must notice that R_n and G_{of} are completely determined by the IF amplifier used in the system, and if we assume that the signal power P_s remains constant at all frequencies of interest, equation 77 shows that F_R cannot be made equal to F_{RO} over any nonzero band of frequencies in spite of the fact that any arbitrary linear lossless interstage network may be used in the receiver. This is one of the important results of this paper. We must notice that the equivalent

circuit shown in Fig. 2 has been assumed in deriving equation 77. This equivalent circuit very well describes the behavior of the diode provided that the lowest frequency of the signal occurring in the system is very far from zero.[12] Also we must observe from equation 77 that F_R can be made equal to F_{RO} at a finite number of discrete frequencies.[16]

V. RATIONAL FUNCTION APPROXIMATIONS

We have shown that F_R cannot be equal to F_{RO} over any nonzero interval $\Omega_o - W \leqq \omega \leqq \Omega_o + W$ of the frequency spectrum.† We shall, therefore, make some rational function approximations to a flat noise performance characteristic of the optical receiver. If these rational function approximations satisfy all the constraints of Section IV, a finite linear lumped lossless network can be found which realizes this kind of noise factor for the optical receiver.[34] A complete treatment of this problem is beyond the scope of this paper; but let us consider certain kinds of approximations widely used in network theory.

5.1 *Butterworth Approximations*

The problem at hand is to approximate F_R as close to F_{RO} as possible over the range $\Omega_o - W \leqq \omega \leqq \Omega_o + W$. A set of polynomials which can be used for this purpose are Butterworth polynomials.[34,35] Let

$$F_R = \frac{1}{\eta} + \left(F_{RO} - \frac{1}{\eta}\right)\left\{1 + \epsilon^2\left(\frac{\omega^2 - \omega_o^2}{2\omega W}\right)^{2n}\right\}, \tag{78}$$

where n is the order of complexity of the interstage network to be used in obtaining broadband performance from the optical receiver, and n is also the order of the Butterworth polynomial. It may be verified that F_R approximates F_{RO} in a maximally flat manner. The behavior of F_R as a function of ω is shown in Fig. 7. Since it can be shown that F_R in equation 78 can be made to satisfy equations 73 through 77 by properly choosing ϵ^2 for all values of n, the approximation of equation 78 is realizable.

From equation 47,

$$s_o(p)s_o(-p) = \frac{1}{1 + \dfrac{4R_nG_{of}M_o}{\epsilon^2\left(\dfrac{\omega^2 - \omega_o^2}{2\omega W}\right)^{2n}}}, \tag{79}$$

† Since the noise factor is a real and even function of ω, we shall only consider the behavior of F_R for $\omega \geqq 0$.

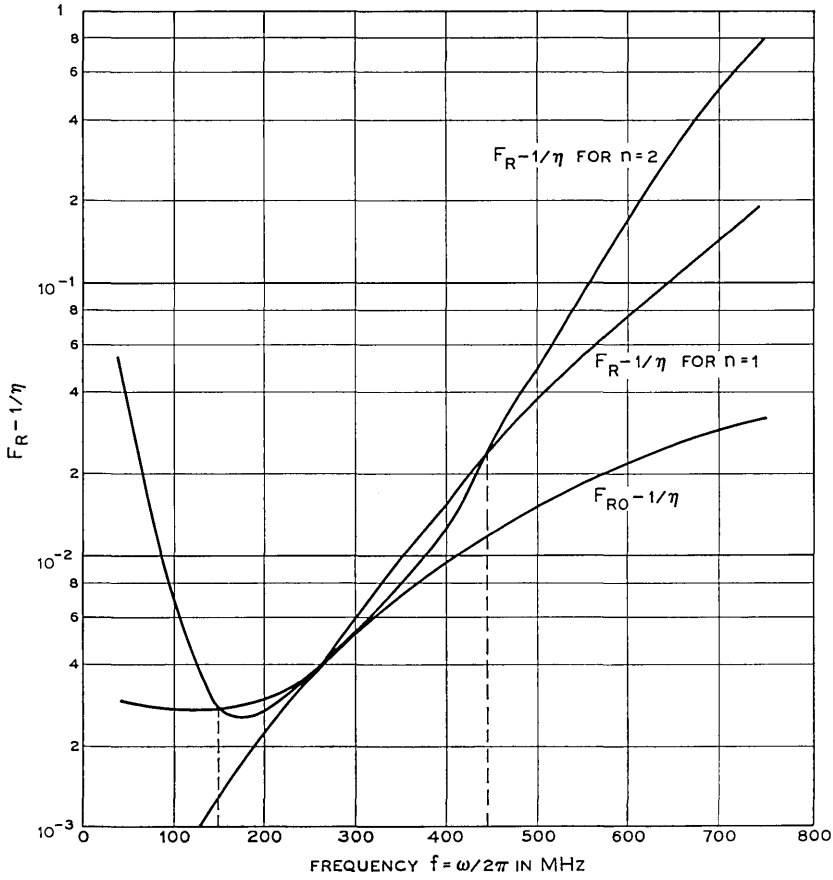


Fig. 7 — Butterworth approximations of order $n = 1,2$. It is assumed that $f_c = 31.83$ GHz, $\eta = 0.70$, $T_a = 290°$K, $I_o = 500$ $\mu\alpha$, $F_o = 2.0833$, and $\epsilon^2 = 1$.

where

$$M_o = \frac{1}{F_o}. \tag{80}$$

It can be shown[35] that

$$s_o(p) = \pm \frac{(p^2 + \omega_o^2)^n}{(p^2 + \omega_o^2)^n + a_{n-1}(2pW)(p^2 + \omega_o^2)^{n-1} + \cdots + a_o(2pW)^n}, \tag{81}$$

where

$$a_{n-1} = \frac{\left(\frac{4R_nG_{of}M_o}{\epsilon^2}\right)^{1/2n}}{\sin\frac{\pi}{2n}}. \tag{82}$$

We can now expand equation 81 into a Taylor series‡ about $p = 0$. We have

$$s_o(p) = -1 + a_{n-1}\frac{2W}{\omega_o^2}p - \cdots . \tag{83}$$

Now

$$\eta(p) = \prod_{l=1}^{\nu}\frac{p-\mu_l}{p+\mu_l^*}$$

$$= (-1)^{\nu}\left[1 - p\sum_{l=1}^{\nu}\left\{\frac{1}{\mu_l} + \frac{1}{\mu_l^*}\right\}\frac{\mu_l}{\mu_l^*} + \cdots\right]. \tag{84}$$

From equations 48, 74, 83, and 84, we have

$$a_{n-1}\frac{2W}{\omega_o^2} + \sum_{l=1}^{\nu}\frac{\mu_l}{\mu_l^*}\left\{\frac{1}{\mu_l} + \frac{1}{\mu_l^*}\right\} \leqq \frac{2}{\omega_c}. \tag{85}$$

Since $\{(1/\mu_l) + (1/\mu_l^*)\} \geqq 0$ for all l, let us put $\eta(p) = 1$. We can then write from equations 82 and 85 that§

$$\epsilon^2 \geqq \frac{4R_nG_{of}M_o}{\left(\frac{\omega_o}{\omega_c}\frac{\omega_o}{W}\sin\frac{\pi}{2n}\right)^{2n}}. \tag{86}$$

A typical value of $f_c = \omega_c/2\pi$ for a photodiode is about 31.83 GHz.¶ For this value of f_c, $f_o = \Omega_o/2\pi = 300$ MHz, and $2W/\Omega_o = 100$ percent, we have plotted in Fig. 8 $\epsilon_{\min}^2/4R_nG_{of}M_o$ as a function of n. It may be seen from the plot that $\epsilon_{\min}^2$ is a monotonically increasing function of n. This behavior of $\epsilon_{\min}^2$ can be explained by the fact that Butterworth polynomials approximate the ideal broadband noise performance characteristic of the optical receiver in a maximally flat fashion[35].

Since no useful purpose is served by using higher values of n, we

‡ We choose negative sign for $s_o(p)$ to satisfy equation 73. This does not entail any loss in generality.[17]

§ Equation 86 can also be obtained by using equation 77.

¶ Typical values of R and C for a photodiode are $C = 1\mu\mu f$, and $R = 5$ ohms.[36]

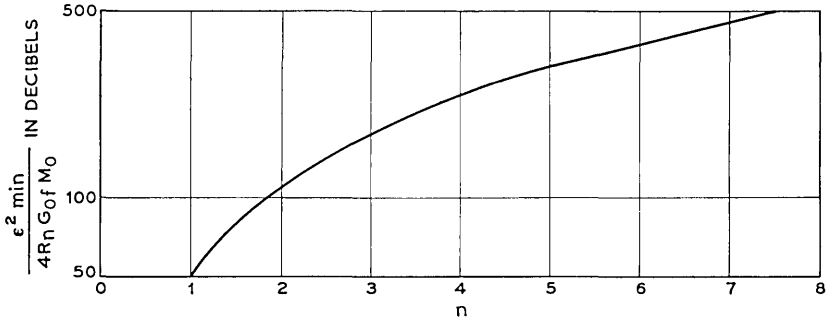


Fig. 8 — A typical plot when Butterworth polynomials are used to approximate the ideal noise performance characteristic. Even though n is a discrete variable the plot is given for all $n \geq 1$.

shall only consider the case $n = 1$.* For $n = 1$,

$$\epsilon^2_{\min} = \frac{4R_n G_{of} M_o}{\left(\frac{\omega_o}{\omega_c}\frac{\omega_o}{W}\right)^2}, \tag{87}$$

and

$$s(p) = s_o(p) = -\frac{p^2 + \omega_o^2}{p^2 + 2p\frac{\omega_o^2}{\omega_c} + \omega_o^2}. \tag{88}$$

Now from equations 31, 48, and 88, it can be shown that

$$Y_1(p) = \frac{1}{RG_{of}} \frac{\frac{\omega_o^2}{\omega_c}}{p + \frac{\omega_o^2}{\omega_c}}. \tag{89}$$

Fig. 9 shows the circuit to realize $Y_1(p)$. Remember that ω_o^2 is the geometric mean of the band of frequencies of interest, and

$$L = \frac{1}{C(\Omega_o - W)(\Omega_o + W)}, \tag{90}$$

$$t = \sqrt{G_{of}R}. \tag{91}$$

This circuit agrees very well with our physical intuition.

* Since we are interested in minimum value of ϵ^2 we have used the equality sign in equation 77

5.2 *Chebyshev Approximations*

Of the various means of approximating a given function, the Chebyshev method is one of the most interesting and important. It can be shown that given a set of n parameters, a function $f(\omega^2)$ approximates $g(\omega^2)$ in the Chebyshev sense if the parameters are determined in such a way that the largest value of $| g(\omega^2) - f(\omega^2) |$ in a given interval is minimum.* Since for a given complexity of the structure the maximum amount of tolerance for a Chebyshev approximation is the same through the band, this type of approximation seems to be the most desirable in the broadband noise performance of optical receivers.

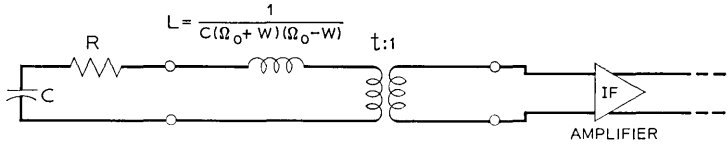


Fig. 9—Lossless interstage network for a Butterworth approximation of order $n = 1$. The ideal transformer ratio t is given by $t = \sqrt{RG_{of}}$.

Let us try to approximate F_R by

$$F_R = \frac{1}{\eta} + \left(F_{RO} - \frac{1}{\eta}\right)\left[1 + \epsilon^2 T_n^2\left(\frac{\omega^2 - \omega_o^2}{2\omega W}\right)\right], \tag{92}$$

where $T_n(x)$ is an n^{th} degree Chebyshev polynominal given by[34, 35, 37]

$$T_n(x) = \cos\,(n \cos^{-1} x). \tag{93}$$

The behavior of F_R as a function of ω for $n = 1, 2$ is shown in Fig. 10. The equiripple behavior of F_R is evident from equation 93. We can also show that the approximation of the type given in equation 92 can be made to satisfy equations 73 through 77. It can also be shown that

$$s_o(p) = -\frac{(p^2 + \omega_o^2)^n + b_{n-2}(2pW)^2(p^2 + \omega_o^2)^{n-2} + \cdots}{(p^2 + \omega_o^2)^n + a_{n-1}(2pW)(p^2 + \omega_o^2)^{n-1} + \cdots}, \tag{94}$$

where

$$a_{n-1} = \frac{\sinh\left[\frac{1}{n}\sinh^{-1}\frac{2\sqrt{R_n G_{of} M_o}}{\epsilon}\right]}{\sin \pi/2n}. \tag{95}$$

* The Chebyshev approximating function has the equiripple property.[34, 35, 37]

If we expand $s_o(p)$ about $p = 0$, we can show that

$$s_o(p) = -1 + pa_{n-1}\frac{2W}{\omega_o^2} + \cdots . \tag{96}$$

We again put $\eta(p) = 1$ to obtain minimum ϵ^2. From equations 74 and 96

$$\epsilon^2_{\min} = \frac{4R_nG_{of}M_o}{\sinh^2\left[n\sinh^{-1}\left\{\left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right)\sin\frac{\pi}{2n}\right\}\right]}. \tag{97}$$

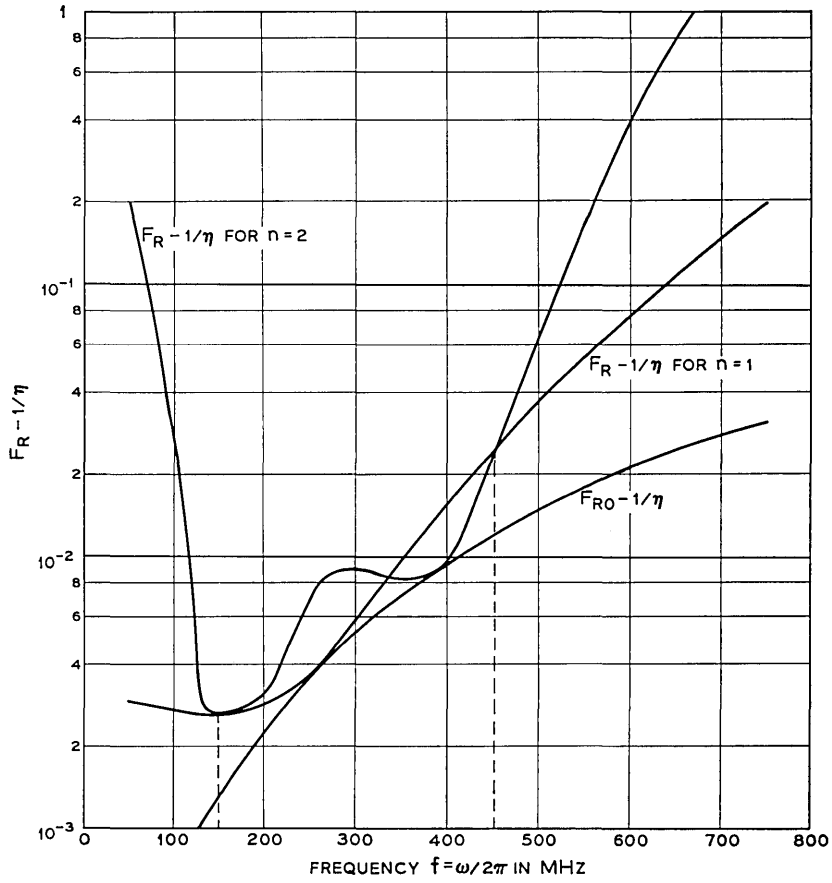


Fig. 10 — Noise factor F_R as a function of ω for Chebyshev approximations of order $n = 1,2$. It is assumed that $f_c = 31.83$ GHz, $\eta = 0.70$, $T_o = 290°$K, $I_o = 500\ \mu a$, $F_o = 2.0833$, and $\epsilon^2 = 1$.

In practical cases, assuming that $\omega_0/\omega_c \ll 1/\sqrt{15}$, $(\omega_0/\omega_c)(\omega_0/W)$ sin $\pi/2n \ll 1$, for $\omega_0/W \geqq \sqrt{15}$.† We can then write

$$\epsilon^2_{\min} = \frac{16R_nG_{of}M_o}{\pi^2\left(\frac{\omega_o}{\omega_c}\right)^2\left(\frac{\omega_o}{W}\right)^2\left(\frac{\sin \pi/2n}{\pi/2n}\right)^2}. \tag{98}$$

A normalized plot of $\epsilon^2_{\min}$ for $f_o = \Omega_o/2\pi = 300$ MHz, $f_c = 31.83$ GHz, and $2W/\Omega_o = 100$ percent is given in Fig. 11. We notice that

$$\frac{[\epsilon^2_{\min}]_{n=1}}{[\epsilon^2_{\min}]_{n=\infty}} = \frac{\pi^2}{4} \approx 2.5, \tag{99}$$

and

$$\frac{[\epsilon^2_{\min}]_{n=2}}{[\epsilon^2_{\min}]_{n=\infty}} = \frac{\pi^2}{8} \approx 1.25. \tag{100}$$

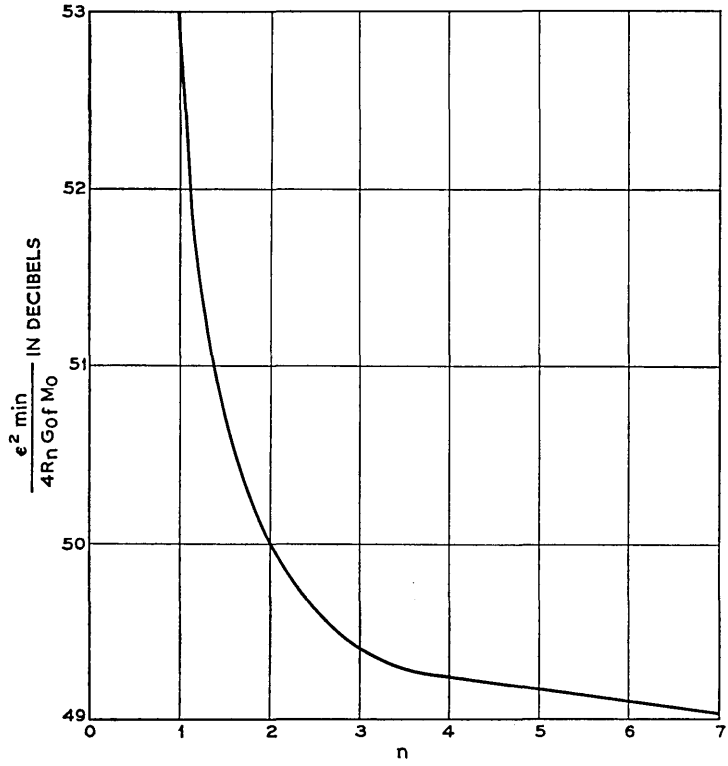


Fig. 11 — A typical plot when Chebyshev polynomials are used to approximate the ideal noise performance characteristic. Even though n is a discrete variable the plot is given for all $n \geq 1$.

§ It can be shown from equation 1 that $\omega_o/W \leqq \sqrt{15}$ for $2W/\Omega_o \geqq \frac{1}{2}$.

From equations 99 and 100 and Fig. 11, we conclude that $\epsilon^2_{\min}$ is a monotonically decreasing function of n, but that no great improvement in the value of $\epsilon^2_{\min}$ is obtained by using very high values of n. Since the complexity of the network increases with n, we shall only consider the cases $n = 1, 2$. For $n = 2$, $\epsilon^2_{\min}$ attains 1.25 times the minimum possible value. For $n = 1$ the Butterworth and Chebyshev approximations are the same. For $n = 2$, from equation 98

$$\epsilon^2_{\min} = 2R_nG_{of}M_o\left(\frac{\omega_c}{\omega_o}\right)^2\left(\frac{W}{\omega_o}\right)^2, \tag{101}$$

and from equations 94, 95, and 101, we can write

$$s_o(p) = -\frac{(p^2 + \omega_o^2)^2 + \frac{1}{2}(2pW)^2}{(p^2 + \omega_o^2)^2 + \left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right)(2pW)(p^2 + \omega_o^2) + \frac{1}{2}(2pW)^2}. \tag{102}$$

Equations 31, 48, and 102 show that

$$Y_1(p) = \frac{1}{RG_{of}}\frac{\omega_o^2}{\omega_c}\frac{p^2 + \omega_o^2}{p^3 + p^2\frac{\omega_o^2}{\omega_c} + p(\omega_o^2 + 2W^2) + \frac{\omega_o^4}{\omega_c}}. \tag{103}$$

The lossless interstage network realizing $Y_1(p)$ in equation 103 is shown in Fig. 12.

Similar methods can be used to determine the lossless interstage networks when $n > 2$. We have shown however that no great improvement can be obtained by using very high values of n.

5.3 *Approximations with Greater than Optimum Noise Factor*

In the preceding parts of this section we used Butterworth and Chebyshev polynomials to approximate the ideal broadband noise performance characteristic of the IF amplifier in such a way that the

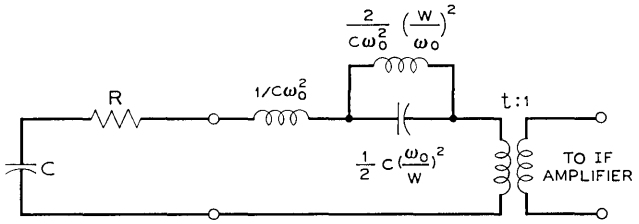


Fig. 12—Lossless interstage network for a Chebyshev approximation of order $n = 2$. The ideal transformer ratio t is given by $t = \sqrt{RG_{of}}$.

minimum passband noise factor is F_{RO}. These polynomials also can be used[17, 18] in a manner in which the minimum passband noise factor is KF_{RO} where

$$K \geqq 1. \tag{104}$$

Such approximations are given by

$$F_R = \frac{1}{\eta} + K\left(F_{RO} - \frac{1}{\eta}\right)\left[1 + \left(\frac{\omega^2 - \omega_o^2}{2\omega W}\right)^{2n}\right], \tag{105}$$

and

$$F_R = \frac{1}{\eta} + K\left(F_{RO} - \frac{1}{\eta}\right)\left[1 + \epsilon^2 T_n^2\left(\frac{\omega^2 - \omega_o^2}{2\omega W}\right)\right], \tag{106}$$

where $T_n(x)$ is an n^{th} degree Chebyshev polynomial.

We can seen from equations 105 and 106 that

$$[F_R]_{\min} = KF_{RO} - \frac{1}{\eta}(K - 1) \geqq F_{RO}\ . \tag{107}$$

If equations 73 through 77 are to be satisfied, it can be shown from equation 105 that

$$\left[1 - \frac{1}{K} + \frac{4R_nG_{of}M_o}{K}\right]^{1/2n} - \left[1 - \frac{1}{K}\right]^{1/2n} \leqq \left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right)\sin\frac{\pi}{2n}. \tag{108}$$

Also, if F_R in equation 106 is to be realizable, it can be shown that the following constraint must be satisfied:

$$\sinh\left[\frac{1}{n}\sinh^{-1}\frac{\left(1 - \frac{1}{K} + 4R_nG_{of}M_o/K\right)^{\frac{1}{2}}}{\epsilon}\right] - \sinh\left[\frac{1}{n}\sinh^{-1}\frac{\left(1 - \frac{1}{K}\right)^{\frac{1}{2}}}{\epsilon}\right] \leqq \left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right)\sin\frac{\pi}{2n}. \tag{109}$$

In general, for arbitrary n, equations 108 and 109 can only be solved numerically. The numerical solution of these two equations requires that the value of $R_nG_{of}M_o$, ω_o/ω_c, and ω_o/W be known. For any specific IF amplifier, the values of K and ϵ^2 can be determined from equations 108 and 109 and the interstage network can then be synthesized. Since we do not propose to go into the characteristics of the IF amplifier, we shall not consider these two equations any more in this paper.

Minimum average noise factor approximation, least-squares approxi-

mation,[23] and the like also can be used in the theory of broadband noise performance of the optical receiver. If these approximations satisfy the restrictions which are imposed by the photodiode, and which are given in equations 73 through 77, the methods given in Section IV can be used to obtain a positive-real $Y_1(p)$. This $Y_1(p)$ enables us to determine the lossless interstage network required in the broadband noise performance of the optical receiver.

VI. GAIN AND NOISE FACTOR

It has been shown for an optical heterodyne receiver[13] that the available output power P_{oa} must satisfy the constraint given by

$$\frac{1}{\pi}\int_0^\infty \frac{1}{\omega^2}\ln\left\{1+\left(\frac{h\nu}{\eta q}\right)^2 \frac{8R_fG_{og}}{RP_oP_s}\frac{(\omega/\omega_c)^2}{\dfrac{1}{P_{oa}}-\dfrac{1}{P_{o\max}}}\right\}d\omega \leqq \frac{2}{\omega_c} \tag{110}$$

where†

$$R_f = \frac{\operatorname{Re} y_{22}}{|\,y_{21}\,|^2} \tag{111}$$

$$G_{og} = \frac{|\,y_{12}y_{21}\,|}{2\operatorname{Re} y_{22}}\sqrt{\lambda^2-1} \tag{112}$$

$$\lambda = \frac{2\operatorname{Re}(y_{11})\operatorname{Re}(y_{22}) - \operatorname{Re}(y_{12}y_{21})}{|\,y_{12}y_{21}\,|} \tag{113}$$

$$P_{o\max} = \varphi_o R\left(\frac{\omega_c}{\omega}\right)^2 \tag{114}$$

$$\varphi_o = \tfrac{1}{2}G_{a\max}\left(\frac{\eta q}{h\nu}\right)^2 P_oP_s \tag{115}$$

$$G_{a\max} = \left|\frac{y_{21}}{y_{12}}\right| \frac{1}{\lambda+\sqrt{\lambda^2-1}}. \tag{116}$$

Equation 110 is identical in form to equation 77, and it can be shown from Ref. 13 that obtaining the broadband signal and noise performance characteristics of the optical receiver are analogous problems. It can also be shown from Ref. 13 that if Butterworth and Chebyshev approximations of the form given by

$$P_{oa} = P_{o\max}\frac{K'}{1+\left(\dfrac{\omega^2-\omega_o^2}{2\omega W}\right)^{2n}}, \qquad 0 < K' \leqq 1 \tag{117}$$

† For an IF amplifier which is absolutely stable, it can be shown[26-28] that $\lambda \geq 1$.

and

$$P_{oa} = P_{o\max} \frac{K''}{1 + \epsilon^2 T_n^2\left(\frac{\omega^2 - \omega_o^2}{2\omega W}\right)}, \qquad 0 < K'' \leqq 1 \tag{118}$$

are used for the available output power of the optical receiver† realizability by lossless interstage networks requires that

$$[1 - K' + 4K'R_f G_{og} G_{a\max}]^{1/2n} - [1 - K']^{1/2n} \leqq \left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right) \sin\frac{\pi}{2n} \tag{119}$$

and

$$\sinh\left[\frac{1}{n}\sinh^{-1}\frac{(1 - K'' + 4K''R_f G_{og} G_{a\max})^{\frac{1}{2}}}{\epsilon}\right] - \sinh\left[\frac{1}{n}\sinh^{-1}\frac{(1 - K'')^{\frac{1}{2}}}{\epsilon}\right] \leqq \left(\frac{\omega_o}{\omega_c}\right)\left(\frac{\omega_o}{W}\right)\sin\frac{\pi}{2n}. \tag{120}$$

We now notice that equation 119 is similar in form to equation 108 and 120 is similar to 109. However, it can be shown that the element values of the lossless interstage network obtained by solving either equation 119 or 120 will not be identical to those obtained by solving either 108 or 109. This shows that the problem of broadband noise performance, in general, requires a network different from that required for obtaining the broadband signal performance of the optical receiver. But for $K = K' = K'' = 1$, and for $G_{og} = G_{of}$, it can be shown‡ from equations 108, 109, 119, and 120 that the network which achieves broadband signal performance for the optical receiver also achieves broadband noise performance.

For a single stage common emitter transistor IF amplifier (see Fig. 13), we can show that the source conductance G_{of} for minimum noise factor is approximately equal to the source conductance G_{og} for maximum available gain.§ We can then say that a common emitter transistor IF amplifier can be used with advantage in obtaining simultaneously broadband signal and noise performance from the optical receiver.

---

† We can compare equation 117 to equation 105 and equation 118 to equation 106.

‡ We have assumed $B_{og} = B_{of} = 0$ for the IF amplifier.

§ In fact it can be shown (see Ref. 39) that for reasonable transistor parameters and frequencies below $(1 - \alpha_o)f_\alpha$, G_{of} is always within a factor of $\sqrt{2}$ of the common emitter G_{og}. α_o is the low frequency alpha of the transistor and f_α is the alpha cutoff frequency.

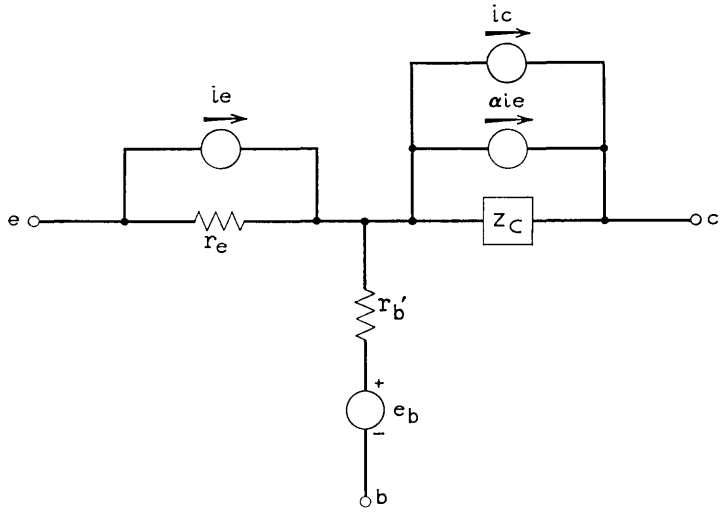


Fig. 13 — Simplified signal and noise equivalent circuit for the transistor.

VII. RESULTS AND CONCLUSIONS

A theory of obtaining broadband noise performance from an optical heterodyne receiver is presented in this paper. It is shown that the following constraint must be satisfied by any lossless interstage network used for obtaining broadband noise performance from the optical receiver:

$$\frac{1}{\pi}\int_0^\infty \frac{1}{\omega^2}\ln\left\{1+\frac{8kT_o}{\eta qI_o}\frac{R_nG_{of}}{R}\frac{(\omega/\omega_c)^2}{F_R-F_{RO}}\right\}d\omega \leqq \frac{2}{\omega_c}. \tag{77}$$

This equation 77 shows that it is impossible to make F_R equal to F_{RO} for any nonzero band of frequencies and for any realizable lossless interstage networks.

We then consider certain types of rational function approximations to an ideal noise performance characteristic of the optical receiver. We show that Butterworth approximations to an ideal characteristic are realizable, but that the broadband noise performance of the receiver deteriorates with increasing values of n, the order of complexity of the interstage network. By approximating the ideal characteristic by Chebyshev polynomials, it can be shown that the performance improves with n, but no great improvement can be obtained by using very high values of n. We have shown that the performance for $n = 2$ is slightly worse than for $n = \infty$. Realizations of networks for $n = 1, 2$ are given.

We also consider the problem of obtaining simultaneously both signal and noise broadband performance from the optical heterodyne receiver and show that, in general, these two problems require two separate lossless interstage networks. We then show that for a common emitter transistor IF amplifier, and for certain types of Butterworth and Chebyshev approximations, these two networks turn out to be identical.

We give design methods and equations for any kind of rational function approximations to an ideal broadband noise performance characteristic of the optical receiver, and explicitly state the constraints to be satisfied by these approximations.

As is evident from Section IV, the theory of broadband noise performance presented in this paper for an optical heterodyne receiver can be applied to any other linear twoport network driven by a source whose internal admittance is a function of frequency.

VIII. ACKNOWLEDGMENT

The author is indebted to Clyde L. Ruthroff for pointing out the existence of this problem, and for his constructive suggestions and comments.

REFERENCES

1. Anderson, L. K., "Photodiode Detection," Proc. Symp. Opt. Masers, Polytechnic Press (April 16–19, 1963).
2. Saito, S., Kurokawa, K., Fujii, Y., Kamura, T., and Uno, Y., "Detection and Amplification of the Microwave Signal in Laser Light by a Parametric Diode," Proc. IRE, *50,* No. 11 (November 1962), pp. 2369–2370.
3. Sommers, H. S., Jr., "Demodulation of Low-Level Broad-Band Optical Signals with Semiconductors," Proc. IEEE, *51,* No. 1 (January 1963), pp. 140–146.
4. DiDomenico, M., Jr. and Svelto, O., "Solid-State Photodetection: A Comparison between Photodiodes and Photoconductors," Proc. IEEE, *52,* No. 2 (February 1964), pp. 136–144.
5. Anderson, L. K., "Measurement of the Microwave Modulation Frequency Response of Junction Photodiodes," Proc. IEEE, *51,* No. 5 (May 1963), pp. 846–847.
6. Riesz, R. P., "High-Speed Semiconductor Photodiodes," Rev. Sci. Inst., *33* (September 1962), pp. 994–998.
7. Inaba, H., and Siegman, A. E., "Microwave Photomixing of Optical Maser Outputs with a P-I-N Junction Photodiode," Proc. IRE, *50,* No. 8 (August 1962), p. 1823.
8. Kibler, L. U., "A High-Speed Point Contact Photodiode," Proc. IRE, *50,* No. 8 (August 1962), pp. 1834–1835.
9. Sharpless, W. M., "Cartridge-Type Point-Contact Photodiode," Proc. IEEE, *52,* No. 2 (February 1964), pp. 207–208.
10. DiDomenico, M., Jr., Sharpless, W. M., and McNicol, J. J., "High-Speed Photodetection in Germanium and Silicon Cartridge-Type Point-Contact Photodiodes," Appl. Opt., *4* (June 1965), pp. 677–682.

11. Schneider, M. V., "Schottky Barrier Photodiodes with Antireflection Coating," B.S.T.J., *45,* No. 9 (November 1966), pp. 1611–1638.
12. Penfield, P., Jr. and Sawyer, D. E., "Photoparametric Amplifier," Proc. IEEE, *53,* No. 4 (April 1965), pp. 340–347.
13. Prabhu, V. K., "A Theory of Broadband Matching for Optical Heterodyne Receivers," Scheduled for Appl. Opt.
14. Haus, H. A., et al., "Representation of Noise in Linear Twoports," Proc. IRE, *48,* No. 1 (January 1960), pp. 69–74.
15. Bode, H. W., Network Analysis and Feedback Amplifier Design, D. Van Nostrand and Co., Inc., Princeton, N. J., 1945.
16. Fano, R. M., "Theoretical Limitations on the Broadband Matching of Arbitrary Impedances," J. Franklin, Inst. 249 (January–February 1950), pp. 57–83, 139–154.
17. Youla, D. C., "A New Theory of Broad-Band Matching," IEEE Trans. Circuit Theory, *CT-11* (March 1964), pp. 33–50.
18. Chan, Y. T. and Kuh, E. S., "A General Matching Theory and Its Application to Tunnel Diode Amplifiers," IEEE Trans. Circuit Theory, *CT-13* (March 1966), pp. 6–17.
19. Prabhu, V. K., unpublished work.
20. Fried, D. L., "Optical Heterodyne Detection of an Atmospherically Distorted Signal Wavefront," Proc. IEEE, *55,* No. 1 (January 1967), pp. 57–67.
21. Fried, D. L. and Seidman, J. B., "Heterodyne and Photon-Counting Receivers for Optical Communications," Appl. Optics, *6,* No. 2 (February 1967), pp. 245–250.
22. Mandel, L., "Heterodyne Detection of a Weak Light Beam," J. Opt. Soc. Amer., *56,* No. 9 (September 1966), pp. 1200–1206.
23. Lucovsky, G., Lasser, M. E., and Emmons, R. B., "Coherent Light Detection in Solid-State Photodiodes," Proc. IEEE, *51,* No. 1 (January 1963), pp. 166–172.
24. Champagne, B. P., "Optimization of Optical Systems," Applied Optics, *5,* No. 11 (November 1966), pp. 1843–1845.
25. Oliver, B. M., "Thermal and Quantum Noise," Proc. IEEE, *53,* No. 5 (May 1965), pp. 436–454.
26. Llewellyn, F. B., "Some Fundamental Properties of Transmission Systems," Proc. IRE, *40,* No. 3 (March 1952), pp. 271–283.
27. Youla, D. C., "A Note on the Stability of Linear Nonreciprocal N-ports," Proc. IRE, *48* (1960), pp. 121–122.
28. Ku, W. H., "Unilateral Gain and Stability Criterion of Active Two-Ports in Terms of Scattering Parameters," Proc. IEEE, *54,* No. 11 (November 1966), pp. 1617–1618.
29. Haus, H. A. and Adler, R. B., *Circuit Theory of Linear Noisy Networks,* Cambridge, Mass.: The Technology Press, and New York: John Wiley and Sons, Inc., 1959.
30. Fukui, H., "Available Power Gain, Noise Figure, and Noise Measure of Two-Ports and Their Graphical Representation," IEEE Trans. Circuit Theory, *CT-13,* No. 2 (June 1966), pp. 137–141.
31. Friis, H. T., "Noise Figure of Radio Receivers," Proc. IRE, *32,* No. 7 (July 1944), pp. 419–422.
32. Fränz, K., "Messung der Empfängerempfindlichkeit bei kurzen elektrischen wellen," Z. Elect. Electroak, *59* (1942), pp. 105–111.
33. Carlin, H. J. and Giordano, A. B., *Network Theory,* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964.
34. Guillemin, E. A., "Synthesis of Passive Networks," New York: John Wiley and Sons, Inc., 1962.
35. Weinberg, L., *Network Analysis and Synthesis,* New York: McGraw-Hill Book Co., Inc., 1962.
36. Schneider, M. V., private communication.
37. Balabanian, N., *Network Synthesis,* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1958.
38. Nielsen, E. G., "Behavior of Noise Figure in Junction Transistors," Proc. IRE, *45,* No. 7 (July 1957), pp. 957–963.
39. Prabhu, V. K., unpublished work.

Contributors to This Issue

M. I. COHEN, B.S., 1957, M.S., 1958, Massachusetts Institute of Technology; Ph.D., 1964, Rensselaer Polytechnic Institute; United Aircraft Corporation, 1958—1961; Bell Telephone Laboratories, 1964—. Dr. Cohen is a member of the Mechanical Engineering Department, Electron Device Laboratory, and has worked on applications of lasers to microelectronic fabrication. Member, ASME, Sigma Xi, Pi Tau Sigma.

PATRICK J. MARINO, BEE, 1960, MEE, 1962, New York University; D. Phil., 1966, Oxford University; Bell Telephone Laboratories, 1960—63, 1965—. As a member of the Data and Private Telephone Systems Laboratory, Mr. Marino's early work at Bell Laboratories included the design of digital circuits and theoretical studies of store-and-forward systems. Sequential circuit theory is his current interest. Member, Eta Kappa Nu, Tau Beta Pi.

JOHN F. MILKOSKY, Bell Telephone Laboratories, 1949—. His first assignment was in assembly, tooling, and tool design for vacuum tubes and semiconductors. From 1960—1965, with Electron Device Mechanical Engineering Department, he worked on *Telstar*® satellite and thin film projects. Since 1965, he has been working on laser applications for electronics.

VASANT K. PRABHU, B. E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S. M., 1963, Sc. D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu is a member of the Radio Research Laboratory, and his areas of interest include systems theory, solid-state microwave devices, noise theory, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

ATTILIO J. RAINAL, Univerity of Alaska, University of Dayton, 1950—52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel Institute of Technology; Dr. Eng., 1963, Johns Hopkins University; Bell Telephone Laboratories, 1964—. Mr. Rainal has been engaged in research on noise theory with application to radar theory. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

G. H. Robertson, B.Sc., 1943, University of Glasgow; after three years in the Royal Navy as an Air Radio Officer he returned to the University of Glasgow for two years and obtained a Post Graduate Certificate in Natural Philosophy; Bell Telephone Laboratories, 1948—. Until 1958 Mr. Robertson was engaged in electronics research and a variety of electron tube development projects. Since 1958 he has been working on signal propagation and processing studies in the Underwater Research and Systems Departments. Associate member, IEEE; member, AAAS.

H. L. Schneider, B.S.E.E., 1949, Purdue University; M.S.E.E., 1955, Pennsylvania State University; Ph.D., 1961, Carnegie Institute of Technology; Bell Telephone Laboratories, 1961—. He had been with the Mobile Radio Research Department and is now in the Military Transmission Systems Department. Senior member, IEEE; member, Sigma Xi

James W. Smith, B.E.S., 1956, Dr. Eng., 1963, John Hopkins University; Bell Telephone Laboratories, 1963—. He has been concerned with analysis problems in the areas of analog and digital data transmission. Member, IEEE, Tau Beta Pi, Sigma Xi and Eta Kappa Nu.

Lee C. Thomas, B.S.E.E., 1962, University of Texas; S.M., E.E., 1963, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1963—. Mr. Thomas has been engaged in the simulation and analysis of a variety of electronic circuits and devices. He supervises a group concerned with passive network design, the development of programs to automate this design, and the development of active network design procedures. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Simulation Council.

Burton A. Unger, BSME, 1954, Purdue University; MSME, 1960, Newark College of Engineering; Bell Telephone Laboratories, 1960—. He participated in the thermal test program of the Telstar® satellite and a development program on a satellite attitude control system. He later worked on the design and development of thin film components and circuits. He is engaged in studying the application of lasers in the fabrication of thin film circuits. Member IEEE.

B. S. T. J. BRIEFS

Approximate and Exact Results Concerning Zeros of Gaussian Noise

By A. J. RAINAL

I. INTRODUCTION

Let τ denote the interval between two successive zeros of a stationary gaussian process having zero-mean and one-sided power spectral density $W(f)$. We shall refer to such an interval as a zero-crossing interval. This brief is concerned with these probability functions:

(*i*) $P_o(\tau)$ = Probability density of a zero-crossing interval.

(*ii*) $F_o(\tau)$ = Probability that a zero-crossing interval lasts longer than τ.

Thus, $F_o(\tau)$ and $P_o(\tau)$ are related by

$$F_o(\tau) = \int_\tau^\infty P_o(x)\,dx = 1 - \int_0^\tau P_o(x)\,dx. \tag{1}$$

An exact, explicit solution for $P_o(\tau)$ or $F_o(\tau)$ in terms of arbitrary $W(f)$ is at present unknown.

In a very interesting paper, E. Wong[1] presented exact, explicit solutions for both $P_o(\tau)$ and $F_o(\tau)$ for the special case when

$$W(f) = \frac{16\sqrt{3}/3}{(\omega^2 + 3)(\omega^2 + \frac{1}{3})} \tag{2}$$

where $\omega = 2\pi f$.

The corresponding autocorrelation function $\rho(\tau)$ is given by

$$\rho(\tau) = \int_0^\infty W(f)\cos 2\pi f\tau\,df = \tfrac{3}{2}e^{-|\tau|/\sqrt{3}}(1 - \tfrac{1}{3}e^{-2|\tau|/\sqrt{3}}). \tag{3}$$

Wong's exact, explicit solutions are in terms of complete elliptic integrals, and they stemmed from a recent result in the theory of Brownian motion.

The purpose of this brief is to compare Wong's exact results with the approximate results of McFadden.[2] McFadden's approximate results stem from the numerical solution of an integral equation, and they are based on the assumption of "quasi-independence" which assumes that a given zero-crossing interval is statistically independent

of the sum of the previous $(2m+2)$ zero-crossing intervals for all non-negative integral m.

We shall see that McFadden's approximate results compare well with Wong's exact results over a significant range of τ.

II. COMPARISON OF APPROXIMATE AND EXACT RESULTS

Figure 1 compares McFadden's approximate result $\hat{P}_o(\tau)$ with Wong's exact result $P_o(\tau)$. The exact first moment of $P_o(\tau)$ follows from Rice's work[3] and is indicated in Figure 1 as $E(\tau) = \pi$. Thus, the approximate and exact results for $P_o(\tau)$ compare well over a significant range of τ.

Figure 2 compares McFadden's approximate result $\hat{F}_o(\tau)$ with Wong's exact result $F_o(\tau)$. From Wong's equation 31 we have that as $\tau \to \infty$, $F_o(\tau) \sim Ce^{-\tau/(2\sqrt{3})}$ where C is a known constant. The semilog plot in Figure 2 shows this asymptotic exponential decay of $F_o(\tau)$.

III. CONCLUSION

McFadden's approximate results $\hat{P}_o(\tau)$, $\hat{F}_o(\tau)$ compare well with Wong's exact results $P_o(\tau)$, $F_o(\tau)$ over a significant range of τ.

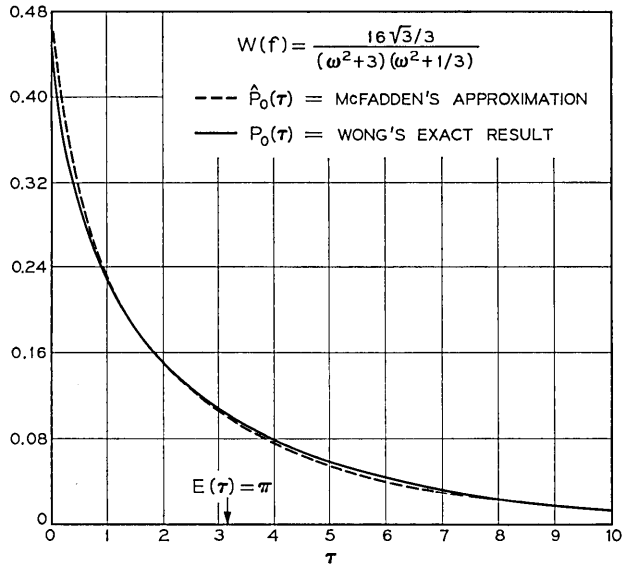


Fig. 1 — Comparison of approximate and exact results for $P_o(\tau)$, the probability density of a zero-crossing interval of gaussian noise having power spectral density $W(f)$.

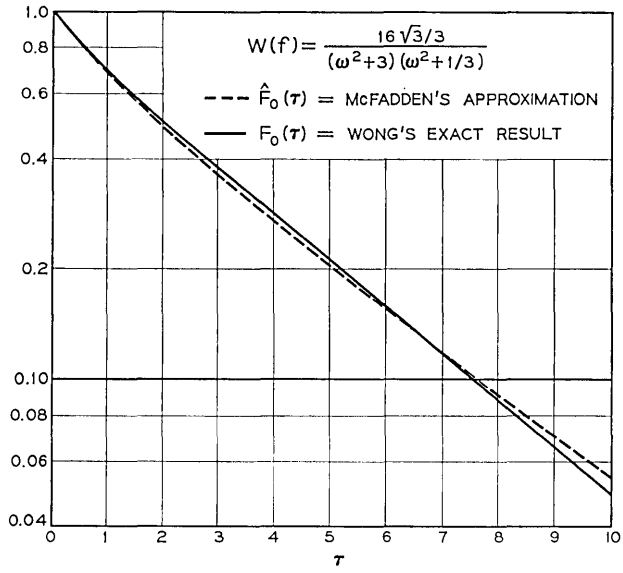


Fig. 2 — Comparison of approximate and exact results for $F_0(\tau)$, the probability that a zero-crossing interval lasts longer than τ.

IV. ACKNOWLEDGMENTS

The author wishes to thank S. O. Rice for suggesting the publication of this information. The author is indebted to Miss A. T. Seery for programming a digital computer to produce the figures.

REFERENCES

1. Wong, E., "Some Results Concerning the Zero-Crossings of Gaussian Noise," SIAM J. Appl. Math., *14*, No. 6 (November 1966), pp. 1246–1254.
2. McFadden, J. A., "The Axis-Crossing Intervals of Random Functions—II," IRE Trans. Inform. Theory, *IT-4* (March 1958), pp. 14–24.
3. Rice, S. O., "Distribution of the Duration of Fades in Radio Transmission," B.S.T.J., *37* (May 1958), pp. 581–635.

Erratum

On page 205 of the February 1968 *Bell System Technical Journal*, the drawings of Figs. 10 and 11 were inadvertently transposed. Fig. 10 is the drawing with the gate electrode marked −100V, and Fig. 11 is the drawing with the gate electrode marked +100V.