DECEMBER 1981
VOL. 60, NO. 10

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May–June and July–August combined issues, by the American Telephone and Telegraph Company, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; F. A. Hutson, Jr., Secretary. Editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, 101 J. F. Kennedy Parkway, Short Hills, N.J. 07078. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

ISSN0005-8580

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 60 December 1981 Number 10

Source Coding for Multiple Descriptions II: A Binary Source

By H. S. WITSENHAUSEN and A. D. WYNER

(Manuscript received March 19, 1981)

A uniformly distributed (iid) binary source is encoded into two binary data streams at rates R_1 and R_2, respectively. These sequences are such that by observing either one separately, a decoder can recover a good approximation of the source (at average error rates D_1, D_2, respectively), and by observing both sequences, a decoder can obtain a better approximation of the source (at average error rate D_0). In this paper a "converse" theorem is established on the set of achievable quintuples (R_1, R_2, D_0, D_1, D_2). For the special case $R_1 = R_2 = 1/2$, $D_0 = 0$, and $D_1 = D_2 = D$, our result implies that $D \geq 1/5$.

I. INTRODUCTION

Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of independent drawings of the binary random variable X, where $\Pr\{X = 0\} = \Pr\{X = 1\} = 1/2$. Assume that this sequence appears at a rate of 1 symbol per second as the output of a data source. (Refer to Fig. 1.) An encoder observes this sequence and emits two binary sequences at rates R_1, $R_2 \leq 1$. These sequences are such that by observing either one, a decoder can recover a good approximation to the source output, and by observing both sequences, a decoder can obtain a better approximation to the source output. Letting D_1, D_2, and D_0 be the error rates which result when the streams at rate R_1, rate R_2, and both streams are used by a decoder, respectively, our problem is to determine (in the usual Shannon sense) the set of achievable quintuples (R_1, R_2, D_0, D_1, D_2). Our main result is a "converse" theorem which gives a necessary condition on the achiev-

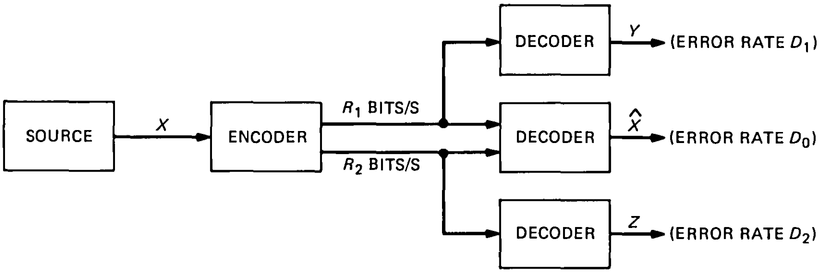


Fig. 1—Communication system.

able quintuples. This paper extends a previous one on the same subject.[1] This paper, however, is self-contained.

This problem is an idealization of the situation in which it is desired to

(*i*) send information over two separate channels, as in a packet communication network, and

(*ii*) recover as much of the original information as possible, should one of the channels break down.

To fix ideas, let us say that $R_1 = R_2 = 1/2$, $D_0 = 0$, and $D_1 = D_2 = D$. Thus, the source sequence at rate 1 is to be encoded into two sequences of rate 1/2 each, such that the original sequence can be recovered from these two encoded sequences with approximately zero error rate (i.e. $D_0 = 0$). Our question then becomes: How well can we reconstruct the source sequence from one of the encoded streams—that is, what is the minimum D? A simple-minded approach would be to let the encoded streams consist of alternate source symbols, which will allow $D_0 = 0$. In this case, $D = 1/4$, since by observing every other source symbol a decoder will make an error half the time on the missing symbol. Is it possible to do better? El Gamal and Cover[2] have looked at this problem and have a theorem which can be used to show that we can make $D = (\sqrt{2} - 1)/2 \approx 0.207$. In a previous paper[1] it was shown that (with $R_1 = R_2 = 1/2$, $D_0 = 0$) $D \geq 1/6$. The new result given here specializes to $D \geq 1/5 = 0.200$. The exact determination of the best D remains an open problem.*

II. FORMAL STATEMENT OF PROBLEM AND RESULTS

Let $\mathbf{B} = \{0, 1\}$, and let $d_H(\mathbf{x}, \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbf{B}^N$, be the Hamming distance between the binary N-vectors $\mathbf{x}$, $\mathbf{y}$; i.e., $d_H(\mathbf{x}, \mathbf{y})$ is the number of positions in which $\mathbf{x}$ and $\mathbf{y}$ do not agree. A code with parameters

* In Ref. 3, Witsenhausen proved a closely related result which encourages the conjecture that $D = 0.207$ is, in fact, the best possible.

$(N, M_1, M_2, D_0, D_1, D_2)$ is a quintuple of mappings $(f_1, f_2, g_0, g_1, g_2)$ where,

$$f_\alpha\colon \mathbf{B}^N \to \{1, \cdots, M_\alpha\}, \alpha = 1, 2 \tag{1a}$$

$$g_\alpha\colon \{1, 2, \cdots, M_\alpha\} \to \mathbf{B}^N, \alpha = 1, 2 \tag{1b}$$

$$g_0\colon \{1, 2, \cdots, M_1\} \times \{1, 2, \cdots, M_2\} \to \mathbf{B}^N. \tag{1c}$$

The source output is a random vector $\mathbf{X}$ uniformly distributed on $\mathbf{B}^N$. Define

$$\mathbf{Y} = g_1 \circ f_1(\mathbf{X}), \tag{2a}$$

$$\mathbf{Z} = g_2 \circ f_2(\mathbf{X}), \tag{2b}$$

and

$$\hat{\mathbf{X}} = g_0[f_1(\mathbf{X}), f_2(\mathbf{X})]. \tag{2c}$$

Then the average error rates are

$$D_1 = \frac{1}{N} Ed_H(\mathbf{X}, \mathbf{Y}), \tag{3a}$$

$$D_2 = \frac{1}{N} Ed_H(\mathbf{X}, \mathbf{Z}), \tag{3b}$$

$$D_0 = \frac{1}{N} Ed_H(\mathbf{X}, \hat{\mathbf{X}}). \tag{3c}$$

We say that a quintuple $(R_1, R_2, d_0, d_1, d_2)$ is *achievable* if, for arbitrary $\epsilon > 0$, there exists, for N sufficiently large, a code with parameters $(N, M_1, M_2, D_0, D_1, D_2)$, where $M_\alpha \le 2^{(R_\alpha+\epsilon)N}$, $\alpha = 1, 2$, and $D_\alpha \le d_\alpha + \epsilon$, $\alpha = 0, 1, 2$. The relationship of this formalism to the system of Fig. 1 should be clear. Our problem is the determination of the set of achievable quintuples, and our main result is a converse theorem.

Before stating our result, let us take a moment to state a positive theorem by El Gamal and Cover[2] as it specializes to our problem.

Theorem 1: *The quintuple* $(R_1, R_2, d_0, d_1, d_2)$ *is achievable if there exists a quadruple of random variables* $X, \hat{X}, Y, Z$, *which take values in* $\mathbf{B}$, *such that* $\Pr\{X = 0\} = \Pr\{X = 1\} = 1/2$, *and*

$$Ed_H(X, \hat{X}) \le d_0, \tag{4a}$$

$$Ed_H(X, Y) \le d_1, \tag{4b}$$

$$Ed_H(X, Z) \le d_2, \tag{4c}$$

and

$$R_1 \ge I(X; Y), \tag{5a}$$

$$R_2 \geq I(X; Z), \tag{5b}$$

$$R_1 + R_2 \geq I(X; \hat{X}, Y, Z) + I(Y; Z), \tag{5c}$$

where $I(\cdot;\cdot)$ is the usual Shannon information.

For the special case of $R_1 = R_2 = 1/2$, $d_0 = 0$, it can be shown that $d_1 = d_2 = (\sqrt{2} - 1)/2 \approx 0.207$ is achievable.

We now state our converse result.

Theorem 2: If $(R_1, R_2, d_0, d_1, d_2)$ is achievable, then in all cases

$$R_1 + R_2 \geq 1 - h(d_0), \tag{6a}$$

furthermore, if $2d_1 + d_2 \leq 1$, then

$$R_1 + R_2 \geq 2 - h(d_0) - h\left(2d_1 + d_2 - \frac{2d_1^2}{1 - d_2}\right), \tag{6b}$$

and if $d_1 + 2d_2 \leq 1$, then

$$R_1 + R_2 \geq 2 - h(d_0) - h\left(d_1 + 2d_2 - \frac{2d_2^2}{1 - d_1}\right), \tag{6c}$$

where

$$h(\lambda) = \begin{cases} 0, & \lambda = 0, \\ -\lambda \log_2 \lambda - (1 - \lambda)\log_2(1 - \lambda), & 0 < \lambda \leq 1/2. \\ 1, & \lambda \geq 1/2. \end{cases}$$

All logarithms in this paper are taken to the base 2. As (6a) is obvious, and (6c) follows from (6b) by symmetry, we need only prove (6b).

In the special case of $R_1 = R_2 = 1/2$, $d_0 = 0$, and $d_1 = d_2 = d$, inequality (6b) implies that

$$h\left(3d - \frac{2d^2}{1 - d}\right) \geq 1,$$

or

$$3d - \frac{2d^2}{(1 - d)} = \frac{3d(1 - d) - 2d^2}{(1 - d)} \geq \frac{1}{2},$$

which implies that $d \geq 1/5 = 0.200$.

III. PROOF OF THEOREM 2

We start from the standard identity

$$I(U_1; U_2 U_3) = I(U_1; U_2) + I(U_1; U_3 | U_2), \tag{7}$$

for arbitrary random variables U_1, U_2, U_3. We say that U_1, U_2, U_3 is a "Markov chain" if U_1, U_3 are conditionally independent given U_2; i.e., U_3 depends on U_1, U_2 only through U_2. If U_1, U_2, U_3 is a Markov chain then $I(U_1; U_3 | U_2) = 0$, and from (7)

$$I(U_1; U_3) \leq I(U_1; U_2\, U_3) = I(U_1; U_2). \tag{8}$$

Note that the hypothesis for (8) holds when U_3 is a function of U_2. A sequence $\{U_n\}$ is a Markov chain if, for all n,

$$(\cdots U_{n-2}, U_{n-1}), U_n, (U_{n+1}, U_{n+2}, \cdots)$$

is a Markov chain.

Let us now suppose that we are given a code $(f_1, f_2, g_0, g_1, g_2)$ with parameters $(N, M_1, M_2, D_0, D_1, D_2)$. We can write

$$\begin{aligned} \log M_1 + \log M_2 &\geq H(f_1(\mathbf{X})) + H(f_2(\mathbf{X})) && (9)\\ &= I(f_1(\mathbf{X}); f_2(\mathbf{X})) + H(f_1(\mathbf{X}) f_2(\mathbf{X})) \\ &= I(f_1(\mathbf{X}); f_2(\mathbf{X})) + I(\mathbf{X}; f_1(\mathbf{X}) f_2(\mathbf{X})) && (10)\\ &\geq I(f_1(\mathbf{X}); f_2(\mathbf{X})) + I(\mathbf{X}; \hat{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}), && (11) \end{aligned}$$

where (9) follows from the fact that $f_i(\mathbf{X})$ takes its values in a set of cardinality M_i, (10) holds because the pair $f_1(\mathbf{X})\, f_2(\mathbf{X})$ is determined by $\mathbf{X}$ and (11) holds because $\hat{\mathbf{X}}$, $\mathbf{Y}$, and $\mathbf{Z}$ depend on $\mathbf{X}$ only through $f_1(\mathbf{X})\, f_2(\mathbf{X})$ so that (8) applies.

Now (11) is getting close to (5c) in the direct theorem. In fact, using (8) twice, we can underbound $I[f_1(\mathbf{X}); f_2(\mathbf{X})]$ by $I(\mathbf{Y}; \mathbf{Z})$. Now the components of the source vector $\mathbf{X}$ are independent, and if the components of either $\mathbf{Y}$ or $\mathbf{Z}$ were also independent, we could make use of standard techniques to establish the necessity of (5c). But alas, we cannot assume that either the $\{Y_n\}$ nor the $\{Z_n\}$ are independent, so that another tactic is required. The key idea is the definition of another random vector $\mathbf{V} = (V_1, \cdots, V_n)$ the components of which are in fact independent.

For $1 \leq k \leq M_1$, define the set

$$A_k = \{\mathbf{x}: f_1(\mathbf{x}) = k\} = f_1^{-1}(k). \tag{12}$$

Let the cardinality of A_k be μ_k. Let the random vector $\mathbf{V}$ be defined by its conditional distribution given $\mathbf{X}$:

$$\Pr\{\mathbf{V} = \mathbf{v} | \mathbf{X} = \mathbf{x}\} = \begin{cases} [\mu_{f_1(\mathbf{x})}]^{-1}, & \mathbf{v} \in A_{f_1(\mathbf{x})}, \\ 0, & \text{otherwise}. \end{cases} \tag{13}$$

Thus, given $\mathbf{X} \in A_k$, $\mathbf{V}$ is uniformly distributed on A_k. It follows that the unconditional distribution for $\mathbf{V}$ is

$$\Pr\{\mathbf{V} = \mathbf{v}\} = 2^{-n}, \qquad \mathbf{v} \in \mathbf{B}^n,$$

and the components of $\mathbf{V}$ are independent, as desired.* Furthermore, $\mathbf{Z}$, $f_2(\mathbf{X})$, $\mathbf{X}$, $f_1(\mathbf{X})$, $\mathbf{V}$ is a Markov chain, so that, using (8),

* In effect, $\mathbf{V}$ is obtained from $f_1(X)$ by a channel with transition probabilities $\Pr\{\mathbf{X} = \mathbf{x} | f_1(\mathbf{X}) = k\}$ so that the distribution of $\mathbf{V}$ is the same as that of $\mathbf{X}$, hence, iid. This is valid for any distribution of X.

$$I(f_1(\mathbf{X}); f_2(\mathbf{X})) = I(\mathbf{V}, f_1(\mathbf{X}); f_2(\mathbf{X}), \mathbf{Z}) \geq I(\mathbf{V}; \mathbf{Z}). \tag{14}$$

Combining (11) and (14), we obtain

$$\begin{aligned}\frac{1}{N}\log M_1 + \frac{1}{N}\log M_2 &\geq \frac{1}{N} I(\mathbf{V}; \mathbf{Z}) + \frac{1}{N} I(\mathbf{X}; \hat{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \\ &\geq \frac{1}{N} I(\mathbf{V}; \mathbf{Z}) + \frac{1}{N} I(\mathbf{X}; \hat{\mathbf{X}}) \\ &\overset{(1)}{\geq} 2 - h\left[\frac{Ed_H(\mathbf{V}; \mathbf{Z})}{N}\right] - h\left[\frac{Ed_H(\mathbf{X}, \hat{\mathbf{X}})}{N}\right] \\ &= 2 - h(\Delta) - h(D_0), \end{aligned} \tag{15a}$$

where

$$\Delta = \frac{Ed_H(\mathbf{V}; \mathbf{Z})}{N}. \tag{15b}$$

Step (1) follows from the "rate-distortion bound" which states that if $\mathbf{U}$ is a random vector uniformly distributed in $\mathbf{B}^N$ (as are $\mathbf{V}$ and $\mathbf{X}$), and $\hat{\mathbf{U}}$ is an arbitrary binary random vector, then $I(\mathbf{U}; \hat{\mathbf{U}}) \geq 1 - h\left[\frac{1}{N} Ed_H(\mathbf{U}, \hat{\mathbf{U}})\right]$. (See Ref. 4.)

We will now obtain an upper bound on Δ in terms of D_1 and D_2. As a "warm up," let us observe that from the triangle inequality,

$$\Delta = \frac{1}{N} Ed_H(\mathbf{V}, \mathbf{Z}) \leq \frac{1}{N}[Ed_H(\mathbf{V}, \mathbf{Y}) + Ed_H(\mathbf{Y}, \mathbf{X}) + Ed_H(\mathbf{X}, \mathbf{Z})].$$

Now

$$Ed_H(\mathbf{Y}, \mathbf{X}) = D_1 N, \qquad Ed_H(\mathbf{Z}, \mathbf{X}) = D_2 N, \tag{16}$$

Furthermore,

$$Ed_H(\mathbf{V}, \mathbf{Y}) = \sum_{\mathbf{v}} E[d_H(\mathbf{v}, \mathbf{Y}) \,|\, \mathbf{V} = \mathbf{v}] \Pr\{\mathbf{V} = \mathbf{v}\}.$$

Now suppose that we are given $\mathbf{V} = \mathbf{v} \in A_k$. Then, $\mathbf{Y} = g_1(k)$. Since $\Pr\{\mathbf{V} = \mathbf{v}\} = 2^{-N}$,

$$\begin{aligned} Ed_H(\mathbf{V}, \mathbf{Y}) &= \sum_{k=1}^{M_1} \sum_{\mathbf{v} \in A_k} 2^{-N} d_H[\mathbf{v}, g_1(k)] \\ &= \sum_{k=1}^{M_1} \sum_{\mathbf{x} \in A_k} \Pr\{\mathbf{X} = \mathbf{x}\} d_H[\mathbf{x}, g_1(k)] = N D_1. \end{aligned} \tag{17}$$

Thus,

$$\Delta \leq 2D_1 + D_2. \tag{18}$$

Substitution of (18) into (15a) yields that for achievable $(R_1, R_2, d_0, d_1, d_2)$

$$R_1 + R_2 \geq 2 - h(2d_1 + d_2) - h(d_0), \tag{19}$$

which is the result reported in Ref. 1.

We will now establish a tighter bound on Δ, namely, for $D_2 + 2D_1 \leq 1$,

$$\Delta = \frac{1}{N} Ed_H(\mathbf{V}, \mathbf{Z}) \leq D_2 + 2D_1 - \frac{2D_1^2}{(1 - D_2)}, \tag{20}$$

so that (15) yields that for achievable $(R_1, R_2, d_0, d_1, d_2)$,

$$R_1 + R_2 \geq 2 - h\left(d_2 + 2d_1 - \frac{2d_1^2}{1 - d_2}\right) - h(d_0), \tag{21}$$

which is (6b), the inequality required for Theorem 2.

Upper bound on Δ: We establish inequality (20) as follows. Let k, $1 \leq k \leq M_1$ be fixed. Let A_k be as defined in (12), and let its cardinality $\mu_k = \mu$. Let the members of A_k be the N-vectors $\{\mathbf{x}_m\}_{m=1}^{\mu}$. Let $\mathbf{y} = g_1(k)$. Thus, when $\mathbf{X} = \mathbf{x} \in A_k$, then $\mathbf{Y} = \mathbf{y}$. Now, form a $\mu \times N$ array, $\mathbf{A}$, with mth row

$$\mathbf{a}_m = (a_{m_1}, a_{m_2}, \cdots, a_{mN}) = \mathbf{x}_m \oplus \mathbf{y}, \tag{22}$$

where $\oplus$ denotes modulo 2 vector addition. Thus, $a_{mn} = 1$, when the nth coordinates of $\mathbf{x}_m$ and $\mathbf{y}$ are different, and $a_{mn} = 0$, otherwise. Note that

$$\begin{aligned}
&\frac{1}{N} E[d_H(\mathbf{X}, \mathbf{Y}) \,|\, f_1(\mathbf{X}) = k] \\
&= \frac{1}{N} \frac{1}{\mu} \sum_{m=1}^{\mu} d_H(\mathbf{x}_m, \mathbf{y}) = \frac{1}{N} \frac{1}{\mu} \sum_{m=1}^{\mu} \sum_{n=1}^{N} a_{mn} \\
&= \frac{1}{N} \sum_{n=1}^{N} s_n,
\end{aligned} \tag{23a}$$

where for $1 \leq n \leq N$,

$$s_n = \frac{1}{\mu} \sum_{m=1}^{\mu} a_{mn} \tag{23b}$$

is the fraction of 1's in column n of $\mathbf{A}$.

Next, for $1 \leq m \leq \mu$, let $\mathbf{z}_m = g_2 \circ f_2(\mathbf{x}_m)$ be the value of $\mathbf{Z}$ which results when $\mathbf{X} = \mathbf{x}_m$. Let $\mathbf{B}$ be the $\mu \times N$ array with mth row

$$\mathbf{b}_m = (b_{m1}, b_{m2}, \cdots, b_{mN}) = \mathbf{z}_m \oplus \mathbf{y}. \tag{24}$$

Then,

$$
\begin{aligned}
\frac{1}{N} E[d_H(\mathbf{X}, \mathbf{Z}) \,|\, f_1(\mathbf{X}) = k] \\
&= \frac{1}{N}\frac{1}{\mu} \sum_{m=1}^{\mu} d_H(\mathbf{x}_m, \mathbf{z}_m) \\
&= \frac{1}{N\mu} \sum_{m} d_H(\mathbf{a}_m, \mathbf{b}_m) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{1}{\mu} \sum_{m=1}^{\mu} [a_{mn}(1 - b_{mn}) + b_{mn}(1 - a_{mn})] \right\}, \qquad (25)
\end{aligned}
$$

where the last equality follows from the fact that for a, $b \epsilon \{0,1\}$, $a(1 - b) + b(1 - a) = 0$ or 1 according as $a = b$ or $a \neq b$. Let τ_n be the expression in braces in (25). Then,

$$
\begin{aligned}
\tau_n &= \frac{1}{\mu} \sum_{m=1}^{\mu} [a_{mn} + b_{mn} - 2a_{mn}b_{mn}] \\
&= \begin{cases}
\dfrac{1}{\mu} \sum_{m} [a_{mn} - b_{mn} + 2b_{mn}(1 - a_{mn})] \\
\qquad \geq \dfrac{1}{\mu} \sum_{m} (a_{mn} - b_{mn}) = s_n - t_n \\
\dfrac{1}{\mu} \sum_{m} [b_{mn} - a_{mn} + 2a_{mn}(1 - b_{mn})] \\
\qquad \geq \dfrac{1}{\mu} \sum_{m} (b_{mn} - a_{mn}) = t_n - s_n,
\end{cases} \qquad (26a)
\end{aligned}
$$

where

$$
t_n = \frac{1}{\mu} \sum_{m=1}^{\mu} b_{mn}, \qquad 1 \leq n \leq N, \qquad (26b)
$$

and s_n is given by (23b). We conclude that $\tau_n \geq |t_n - s_n|$, so that (25) yields

$$
\frac{1}{N} E[d_H(\mathbf{X}, \mathbf{Z}) \,|\, f_1(\mathbf{X}) = k] \geq \frac{1}{N} \sum_{n=1}^{N} |t_n - s_n|. \qquad (27)
$$

Finally, consider

$$
\begin{aligned}
\frac{1}{N} E[d_H(\mathbf{V}, \mathbf{Z}) \,|\, f_1(\mathbf{X}) = k] \\
&= \frac{1}{N} \sum_{m=1}^{\mu} \frac{1}{\mu} E[d_H(\mathbf{V}, \mathbf{Z}) \,|\, \mathbf{X} = \mathbf{x}_m] \\
&= \frac{1}{N\mu} \sum_{m=1}^{\mu} E[d_H(\mathbf{V}, \mathbf{z}_m) \,|\, \mathbf{X} = \mathbf{x}_m] \\
&= \frac{1}{N\mu} \sum_{m=1}^{\mu} \sum_{\mathbf{v}} d_H(\mathbf{v}, \mathbf{z}_m) \Pr\{\mathbf{V} = \mathbf{v} \,|\, \mathbf{X} = \mathbf{x}_m\}. \qquad (28)
\end{aligned}
$$

Now, from (13) which defines $\mathbf{V}$,

$$\Pr\{\mathbf{V} = \mathbf{v} \mid \mathbf{X} = \mathbf{x}_m\} = \begin{cases} \mu^{-1}, & \mathbf{v} \in A_k, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, (28) yields,

$$\frac{1}{N} E[d_H(\mathbf{V}, \mathbf{Z}) \mid f_1(\mathbf{X}) = k]$$

$$= \frac{1}{N\mu^2} \sum_{m=1}^{\mu} \sum_{m'=1}^{\mu} d_H(\mathbf{x}_{m'}, \mathbf{z}_m)$$

$$= \frac{1}{N\mu^2} \sum_{m,m'} d_H(\mathbf{a}_{m'}, \mathbf{b}_m)$$

$$= \frac{1}{N\mu^2} \sum_{m,m'} \sum_{n} [a_{m'n}(1 - b_{mn}) + (1 - a_{m'n})b_{mn}]$$

$$= \frac{1}{N} \sum_{n=1}^{N} [s_n(1 - t_n) + t_n(1 - s_n)]. \tag{29}$$

Now make the dependence of s_n and t_n on k explicit by writing s_{nk} and t_{nk}, respectively. Then, on averaging over k, (23b) becomes

$$D_1 = \frac{1}{N} E\{d_H(\mathbf{X}, \mathbf{Y})\} = \sum_{k=1}^{M_1} \Pr\{f_1(\mathbf{X}) = k\} \frac{1}{N} \sum_{n=1}^{N} s_{nk}$$

$$= \sum_{k=1}^{M_1} \sum_{n=1}^{N} P(n, k) s_{nk}, \tag{30a}$$

where $P(n, k) = \Pr\{f_1(\mathbf{X}) = k\} \times \frac{1}{N}$. Similarly, (27) becomes

$$D_2 = \frac{1}{N} E\{d_H(\mathbf{X}, \mathbf{Z})\} \geq \sum_{k=1}^{M_1} \sum_{n=1}^{N} P(n, k) \,|\, t_{nk} - s_{nk} \,|. \tag{30b}$$

Finally, (29) becomes

$$\Delta = \frac{1}{N} E\{d_H(\mathbf{V}, \mathbf{Z})\}$$

$$= \sum_{k=1}^{M_1} \sum_{n=1}^{N} P(n, k)[s_{nk}(1 - t_{nk}) + t_{nk}(1 - s_{nk})]. \tag{30c}$$

We now apply the following inequality, the proof of which is given in Section IV:

Let S, T be random variables such that $0 \leq S$, $T \leq 1$, and $E\{S\} \leq D_1$, $E\{|T - S|\} \leq D_2$, with $2D_1 + D_2 \leq 1$. Then,

$$E\{S(1 - T) + T(1 - S)\} \leq D_2 + 2D_1 - \frac{2D_1^2}{1 - D_2}. \tag{31}$$

Let S, T be the random variables which take the value s_{nk}, t_{nk}, respectively, with probability $P(n, k)$, then (30) and (31) imply that, for $2D_1 + D_2 \leq 1$,

$$\Delta \leq D_2 + 2D_1 - \frac{2D_1^2}{1 - D_2}, \tag{32}$$

which, when substituted in (15) gives (6b), proving Theorem 2.

IV. PROOF OF THE INEQUALITY

Define $Q(D_1, D_2)$ as the supremum of $E\{S(1 - T) + T(1 - S)\}$ over all distributions of (S, T) on the unit square $[0, 1]^2$ for which $E\{S\} \leq D_1, E\{|T - S|\} \leq D_2$.

Theorem 3: (a) *For* $2D_1 + D_2 \leq 1$ *one has*

$$Q(D_1, D_2) = 2D_1 + D_2 - \frac{2D_1^2}{1 - D_2},$$

with $Q(0, 1) = 1$. (b) *For* $2D_1 + D_2 \geq 1$ *one has*

$$Q(D_1, D_2) \geq \frac{1}{2}.$$

To establish this, introduce for S, T, x, y in $[0, 1]$, $y \neq 1$, the function*

$$\begin{aligned} F(S, T, x, y) &= S(1 - T) + T(1 - S) \\ &+ \left(\frac{4x}{1 - y} - 2\right)(S - x) + \left[\frac{2x^2}{(1 - y)^2} - 1\right](|S - T| - y). \end{aligned} \tag{33}$$

Lemma: *For* $2x + y \leq 1$, $y < 1$ *the maximum of* $F(S, T, x, y)$ *over all* (S, T) *in* $[0, 1]^2$ *is* $2x + y - 2x^2/(1 - y)$.

Proof: For fixed S, the maximum of F over T must, by piecewise linearity, be at either $T = 0$, S, or 1.

(i) If $T = 0$, then

$$F = S + \left(\frac{4x}{1 - y} - 2\right)(S - x) + \left[\frac{2x^2}{(1 - y)^2} - 1\right](S - y),$$

and this is maximized over S at either $S = 0$ or $S = 1$.

(a) For $S = 0$

$$\begin{aligned} F &= 2x - \frac{4x^2}{1 - y} + y - \frac{2x^2y}{(1 - y)^2} \\ &= 2x + y - \frac{2x^2}{1 - y} - \frac{2x^2}{1 - y} - \frac{2x^2y}{(1 - y)^2} \leq 2x + y - \frac{2x^2}{1 - y}. \end{aligned}$$

* This choice of F comes from the duality theory of convex hull formation, as described, e.g., in Section VA of Ref. 5.

(b) For $S = 1$

$$F = 1 - 2 + 2x + \frac{4x(1-x)}{1-y} + \frac{2x^2}{1-y} - 1 + y$$

$$= 2x + y - \frac{2x^2}{1-y} - 2 + \frac{4x}{1-y} \leq 2x + y - \frac{2x^2}{1-y}, \text{ using } 2x + y \leq 1.$$

(*ii*) If $T = S$, then

$$F = 2S - 2S^2 + \left(\frac{4x}{1-y} - 2\right)(S - x) + y - \frac{2x^2y}{(1-y)^2}.$$

The maximum over S of this quadratic is at $S = x/(1-y)$ which is in $[0, 1]$ if $x \leq 1 - y$ and this holds as $x \leq (1-y)/2 \leq 1 - y$. Hence, the maximum is

$$\frac{2x}{1-y} - \frac{2x^2}{(1-y)^2} + \left(\frac{4x}{1-y} - 2\right)\left(\frac{x}{1-y} - x\right)$$

$$+ y - \frac{2x^2y}{(1-y)^2} = 2x + y - \frac{2x^2}{1-y}.$$

(*iii*) If $T = 1$, then

$$F = 1 - S + \left(\frac{4x}{1-y} - 2\right)(S - x) + \left[\frac{2x^2}{(1-y)^2} - 1\right](1 - S - y),$$

which is maximized by taking S either 0 or 1.

(a) For $S = 0$

$$F = 2x + y - \frac{4x^2}{1-y} + \frac{2x^2}{(1-y)^2} - \frac{2x^2y}{(1-y)^2}$$

$$= 2x + y - \frac{2x^2}{1-y}$$

(b) For $S = 1$,

$$F = 2x + y + \frac{4x}{1-y} - \frac{4x^2}{1-y} - 2 - \frac{2x^2y}{(1-y)^2}$$

$$= 2x + y - \frac{2x^2}{1-y} - \frac{2x^2}{1-y} + \frac{4x}{1-y} - 2 - \frac{2x^2y}{(1-y)^2}$$

$$\leq 2x + y - \frac{2x^2}{1-y},$$

since $4x/(1-y) \leq 2$.

Thus, the maximum is as stated and is attained for $T = S = x/(1-y)$ and for $T = 1$, $S = 0$. This completes the proof of the lemma.

Turning to the proof of Theorem 3, consider any distribution of (S, T) on the unit square for which

$$E\{S\} = x, \qquad E\{|T - S|\} = y, \qquad 2x + y \le 1. \tag{34}$$

If $y = 1$, then $x = 0$ from which it follows that $S = 0$, $T = 1$ almost surely, giving $E\{S(1 - T) + T(1 - S)\} = 1$.

If $y < 1$, one has, by the lemma,

$$E\{F(S, T, x, y)\} = E\{S(1 - T) + T(1 - S)\} \le 2x + y - \frac{2x^2}{1 - y}.$$

If one chooses the distribution $T = 1$, $S = 0$ with probability y and $T = S = x/(1 - y)$ with probability $1 - y$, equality is attained. This determines the maximum of $E\{S(1 - T) + T(1 - S)\}$ subject to (34). As $2x + y - 2x^2/(1 - y)$ is monotone increasing (for $2x + y \le 1$) in both x and y, the maximum is unchanged if one allows all x, y with $0 \le x \le D_1$, $0 \le y \le D_2$. This establishes part (a) of Theorem 3.

For part (b), it suffices to observe that Q is monotone nondecreasing in both arguments by its definition and that on the boundary, where $2D_1 + D_2 = 1$, one has $Q(D_1, D_2) = (1 + D_2)/2$. This establishes part (b). (It could easily be shown that $Q(D_1, D_2) = (1 + D_2)/2$ for all (D_1, D_2) in the unit square satisfying $2D_1 + D_2 \ge 1$.)

REFERENCES

1. J. K. Wolf, A. D. Wyner, and J. Ziv, "Source Coding for Multiple Descriptions," B.S.T.J., *59,* No. 8 (October 1980), pp. 1417-26.
2. A. A. El Gamal and T. M. Cover, "Information Theory of Multiple Descriptions," Technical Report No. 43, Department of Statistics, Stanford University, 1980.
3. H. S. Witsenhausen, "On Source Networks with Minimal Breakdown Degradation," B.S.T.J., *59,* No. 6 (July-August 1980), pp. 1038-87.
4. R. G. Gallager, *Information Theory and Reliable Communication,* New York: John Wiley, 1968, Theorem 4.3.2, p. 79.
5. H. S. Witsenhausen, "Some Aspects of Convexity Useful in Information Theory," IEEE Trans. Inform. Theory, *IT-26* (May 1980), pp. 265–71.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

Microprocessor Firmware Update Inventory Model

By S. M. BRECHER

(Manuscript received June 10, 1981)

Microprocessor-based systems are used in many applications of modern telecommunications. The controlling program of most microprocessor systems is stored in firmware which is usually coded into erasable programmable read-only memory chips (EPROM). As new services are implemented, there is a continuing need to update the firmware, a potentially expensive process. In this paper, a method is presented for determining the resources necessary for updating EPROM firmware from a centralized location by using a rotating inventory scheme.

I. INTRODUCTION

Microprocessor-based systems are used in many applications of modern telecommunications. The controlling program (firmware) of most microprocessor systems, is coded into either read-only memory (ROM), programmable read-only memory (PROM), or erasable programmable read-only memory (EPROM) chips, which are mounted on circuit boards. As new services are implemented in the microprocessor-based systems, there is a continuing need to update the firmware. Frequent updating of firmware is more effectively accomplished by the use of EPROM chips, because they can be repeatedly erased by exposure to ultraviolet light and reprogrammed by the use of a specially designed unit. Typical EPROM circuit packs may require 0.5 hours for erasing and 1.5 hours for programming.

One serious drawback to using EPROM firmware is that it cannot be altered by means of a data link. A change in firmware entails the removal and reinstallation of the memory circuit boards or packs. Because of this manual process, updating a large number of microprocessor systems may involve a long and costly procedure.

The object of this paper is to provide a quantitative method to determine the level of spare boards and programming units necessary to achieve a predeterminated time span for updating all the microprocessor-based systems.

1.1 Firmware update process

Typically, there are three conditions that could affect a firmware module: (*i*) Program updates because of new feature introduction, (*ii*) program changes caused by fixing of "bugs," and (*iii*) program changes because of hardware updates. In general, a firmware update process begins with a notice of a change sent by the microprocessor manufacturer to centralized programming sites where ultraviolet light programming units and a spare inventory of circuit packs are kept. Over a dial-up connection, the programming unit receives the latest program version and writes it onto spare circuit packs taken from inventory. When all spare circuit packs are rewritten they are shipped to specific distribution sites associated with a subset of the microprocessor systems. From each distribution site, craft persons are dispatched to install the updated boards. The removed circuit packs are then returned to the centralized programming site for updating. The recycling process continues until all the microprocessor systems in the field, plus their allocated maintenance spares, are updated.

1.2 Basic model

The flow and assumptions of the basic update model, for a typical process schedule, are shown in Fig. 1. The process of updating begins when a firmware change message is received at the centralized programming site. An initial period is used for administrative procedures, unpacking of inventoried circuit packs, erasing, and reprogramming, packing and crating, and shipping of the circuit packs to the distribution sites. Out of this initial time interval, it can be assumed that one day will be needed for reprogramming the spare EPROM circuit packs. If there are not enough programming units, a delay in the update process could be incurred.

Once the rewritten boards arrive at the distribution sites, the associated microprocessor systems are scheduled for update. The process of coordinating the installation forces necessary for the update is not immediate, and the following distribution can be assumed to characterize the update interval: a proportion, i_1, of the updated boards are installed and an equal amount of outdated boards are returned to the centralized location for reprogramming in one week; a proportion, i_2, of the updated boards are recycled in two weeks; a proportion, i_3, in three weeks, and so on until all the boards are recycled.

Once the outdated boards arrive back at the programming site, the process of reprogramming begins again. This procedure continues until all of the microprocessor systems in the field have been updated.

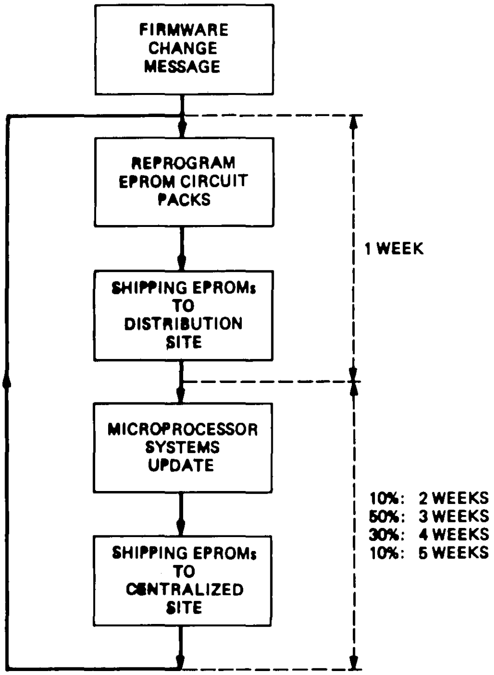

FIRMWARE CHANGE MESSAGE
REPROGRAM EPROM CIRCUIT PACKS
SHIPPING EPROMs TO DISTRIBUTION SITE
MICROPROCESSOR SYSTEMS UPDATE
SHIPPING EPROMs TO CENTRALIZED SITE
1 WEEK
10%: 2 WEEKS
50%: 3 WEEKS
30%: 4 WEEKS
10%: 5 WEEKS

Fig. 1—Typical firmware update process.

II. SYSTEM PARAMETERS

The system parameters to be determined are the update time, the total update spares, and the number of programming units. Update time, T, is defined as the interval in weeks between a change notice and the time at which all the systems in the field, including maintenance spares, have been updated. To determine T, it is necessary to define the concepts of update spare, total update spares, total systems, and spare ratio.

An update spare is defined as one complete set of EPROM circuit packs for one microprocessor system. Since an update spare is defined as one full set of boards, an update spare may be equated to a microprocessor system and vice versa. Total update spares is defined as the number of EPROM circuit packs at the centralized programming site available for initiating the update process. Total systems is defined as the total number of microprocessor-based systems in service and their maintenance spares to be updated when a program change is issued.

The spare ratio, S, is defined as the ratio of total update spares to total systems; that is,

$$S = \frac{\text{Total Update Spares}}{\text{Total Systems}}.$$

A spare ratio of one, $S = 1$, says that for each microprocessor system in the field and its maintenance spares there is one set of update spares at the programming site. In this case, T is the time required for one pass through the basic flow shown in Fig. 1. Similarly, if S is 0.5, one spare for every two systems in the field and their maintenance spares, T is the time required for approximately two passes through the basic flow. By following this reasoning, it becomes clear that T is a function of S.

III. MODEL FORMULATION

3.1 Update time

The update time can be determined by considering the following installation distribution mentioned in Section I and shown in Table I. In Table I, i_j = proportion of boards updated in week j, w = number of weeks required for returning all spares updated in week 0; $i_j = 0$, for $j = 0$ and $j > w$, i.e., no system can be updated in week 0 with spares reprogrammed in that week. Then,

$$\sum_{j=0}^{w} i_j = 1,$$

which implies that after w weeks all the spares reprogrammed in week 0 are returned. Define $q(n)$ as the quantity of spares updated in week n, with $n = 0$, week of change message, and $n = T$, week by which all systems are updated. The update distribution can be modeled by

$$\begin{aligned} q(0) &= \text{Total Update Spares} = S \times \text{Total Systems}, \\ q(1) &= i_1 q(0), \\ q(2) &= i_2 q(0) + i_1 q(1), \\ q(3) &= i_3 q(0) + i_2 q(1) + i_1 q(2), \\ &\vdots \qquad\qquad\qquad\qquad , \quad \text{and} \\ q(n) &= \sum_{j=0}^{n-1} i_{n-j} q(j). \end{aligned} \tag{1}$$

Equation (1) can be rewritten as:

$$q(n) = \sum_{j=0}^{w} i_j q(n-j), \tag{2}$$

since $i_o = 0$, and with initial conditions

$$\begin{aligned} q(0) &= \text{Total Update Spares} = S \times \text{Total Systems}, \\ q(n) &= 0, \qquad \text{for} \quad n < 0. \end{aligned} \tag{3}$$

Table I—Distribution of microprocessor systems update

Week of Update Process	Proportion Updated
1	i_1
2	i_2
3	i_3
⋮	⋮
w	i_w

Equations (2) and (3) are the difference equations describing the microprocessor update process. To obtain the solution of the difference equation, define the ratio of total update spares updated in week n:

$$a(n) = \frac{q(n)}{\text{Total Update Spares}}, \tag{4}$$

where

$$0 \leq a(n) \leq 1.$$

Substituting $q(n)$ from eq. (4) into eq. (2)

$$a(n) = \sum_{j=0}^{w} i_j a(n - j), \tag{5}$$

and

$$a(0) = 1. \tag{6}$$

Before solving eq. (5), notice that the update process ends when the total number of updated spares equals the number of total systems; that is,

$$\sum_{n=1}^{T} q(n) = \text{Total Systems}, \tag{7}$$

where T is the update time.

Substituting eq. (4) into eq. (7) and using the definition of spare ratio,

$$\sum_{n=1}^{T} a(n) = \frac{1}{S}. \tag{8}$$

Expressions (5), (6), and (8) are the difference equations for the weekly ratio of updated spares and T.

To obtain T, eqs. (5) and (8) must be solved. The solution can be obtained using a computer simulation, or a closed-form technique such as the z-transform. To use the z-transform method, recall the following properties:[1]

(*i*) *Translation property*:

$$Z[f(n+k)] = z^k F(z) - z^k \sum_{j=0}^{k-1} f(j) z^{-j},$$

where $F(z) = Z[f(n)]$.

(*ii*) *Summation property*:

$$Z\left[\sum_{n=-\infty}^{N} f(n)\right] = \frac{z}{z-1} F(z) + \frac{z}{z-1} \sum_{n=-\infty}^{-1} f(n).$$

(*iii*) *Final value property*:

$$\lim_{t\to\infty} f(t) = \lim_{z\to 1} (z-1) F(z).$$

Using property (*i*) in eqs. (5) and (6) gives:

$$A(z) = \frac{z^w}{z^w - \sum_{j=1}^{w} i_j z^{w-j}}, \tag{9}$$

where i and w were previously defined, and

$$A(z) = Z[a(n)].$$

Using property (*ii*) in eq. (8), and the initial conditions for $a(n)$, gives

$$Z\left[\sum_{n=0}^{T} a(n)\right] = \frac{z}{z-1} A(z). \tag{10}$$

Substituting eq. (9) into eq. (10) and considering that $i_o = 0$, gives

$$Z\left[\sum_{n=1}^{T} a(n)\right] = \frac{z}{z-1} \frac{\sum_{j=1}^{w} i_j z^{w-j}}{z^w - \sum_{j=1}^{w} i_j z^{w-j}}.$$

The inverse z-transform of eqs. (9) and (6) gives the following relationship between T and S:

$$\frac{1}{S} = Z^{-1} \left\{ \frac{z}{z-1} \frac{\sum_{j=1}^{w} i_j z^{w-j}}{z^w - \sum_{j=1}^{w} i_j z^{w-j}} \right\}. \tag{11}$$

The right-hand term of eq. (11) can be inverted using classical techniques and z-transform tables.

An example can be used to show an application of the z-transform technique. Consider the firmware update distribution shown in Fig. 1 and given in Table II.

Table II—Firmware update distribution for a six-week turnaround interval (w = 6)

Week of Update Process	Proportion Updated i_j
3	0.10
4	0.50
5	0.30
6	0.10

Substituting the above values of i_j in eq. (11),

$$\frac{1}{S} = Z^{-1}\left\{\frac{z}{z-1}\,\frac{0.1z^3 + 0.5z^2 + 0.3z + 0.1}{z^6 - 0.1z^3 - 0.5z^2 - 0.3z - 0.1}\right\}. \tag{12}$$

Using the partial-fraction expansion technique for inverting eq. (12), the closed-form expression for S as a function of T is

$$\begin{aligned}\frac{1}{S} &= 0.22727T - 0.36983\\ &+ 0.09374(-0.67495)^T\\ &+ 0.03228(0.45089)^T\sin(2.27567T + 1.23494)\\ &+ 0.32472(0.85369)^T\sin(1.41837T + 0.85769).\end{aligned} \tag{13}$$

A closed form solution to this transcendental equation in T is difficult to obtain. An approximate solution is obtained when T is large. In this case

$$\frac{1}{S} \approx 0.22727T - 0.36983, \tag{14}$$

or

$$T \approx \frac{4.4}{S} + 1.62725, \tag{15}$$

which is the expression for T. Equation (15) satisfies all update times with an error of less than 2 percent for $T \geq 10$, and with a maximum error of 7 percent occurring in week 7.

An integer solution can be obtained by making:

$$T = \left\lceil \frac{4.4}{S} + 1.6 \right\rceil, \tag{16}$$

with a maximum positive error of 1 week. The $\lceil\cdot\rceil$ operator denotes rounded-up approximation. Figure 2 shows the representation of the update time as a function of the spare ratio for this example.

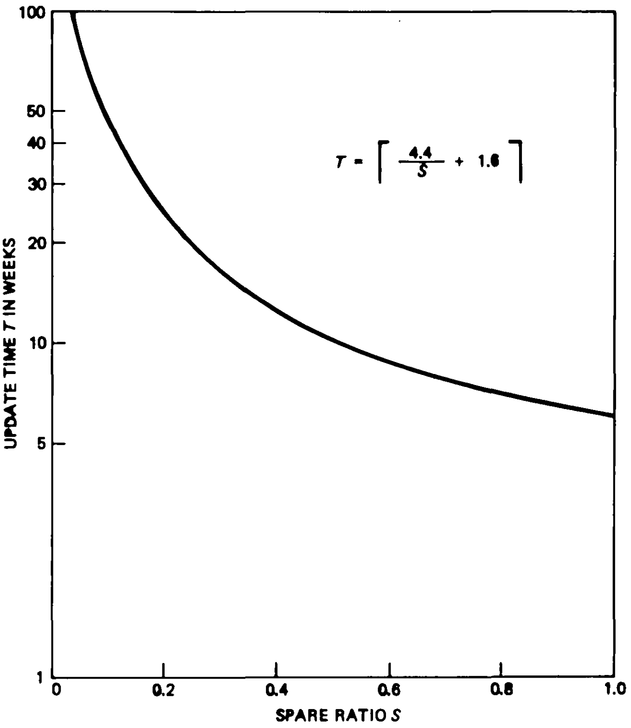


Fig. 2—Update time.

A more simplified approach was used for solving the original real-life update time problem. From the given installation distribution, the expected number of weeks required to program, ship, and install a set of spares, i.e. one pass through the basic flow of Fig. 1, is

$$\bar{T} = 0.1 \times 3 + 0.5 \times 4 + 0.3 \times 5 + 0.1 \times 6 = 4.4 \text{ weeks}.$$

This means that one reprogramming process for the average microprocessor system takes 4.4 weeks. The update time is determined by the number of times this process will be repeated, plus the error introduced by using the expected value approach. The number of times the process is repeated depends on the spare ratio. Therefore, the update time can be represented by the following equation:

$$T = \frac{4.4}{S} + \text{error},$$

which agrees with eq. (16) for an error of 1.6.

The update time algorithm shows the dependency between T, S, and the microprocessor's update distribution. For an illustrative purpose, assume that there are as many spare boards at the programming site as microprocessor systems in the field, i.e., $S = 1$. In this case, $T = 6$ weeks. This period is the minimum time possible under the given basic assumptions. On the other hand, if we have spare boards available for only 50 percent of microprocessor systems in the field, $S = 0.5$, the update time according to the algorithm is $T = 11$ weeks. The conclusion is that T is directly proportional to the installation distribution and inversely proportional to S.

3.2 Total update spare requirements

The number of update spares necessary at the centralized programming site is determined by the update time and the total number of microprocessor systems. The update time determines the spare ratio which, when combined with the number of systems, gives the number of update spares as indicated in Section II:

$$\text{Total Update Spares} = \lceil S \times \text{Total Systems} \rceil. \tag{17}$$

For the example of Section 3.1, the update spare requirements for implementing a firmware update system for 150 total systems in 20 weeks is computed as follows:

First, the spare ratio for 20 weeks from Fig. 2 is;

$$S = 0.24.$$

Then, the number of update spares is given by

$$\text{Total Update Spares} = \lceil 0.24 \times 150 \rceil = 36.$$

3.3 Programming units

The number of programming units necessary for a centralized operation can be determined as a function of their capacity. To determine the number of units, two concepts must be used: weekly spares and programming days.

Weekly spares, $q(n)$, already defined in Section III, is the number of spares returned each week to the site for reprogramming. The number of weekly spares is a function of the installation distribution. Programming days, P, is the number of days per week allocated to the erasing and programming of EPROM circuit packs.

The number of programming units is the following function of the weekly spares, the number of programming days, and the unit capacity:

$$\text{Programming Units} = \left\lceil \frac{q(n)}{P} \times \frac{1}{C} \right\rceil, \qquad \text{for all } n,$$

where C, microprocessor systems per day per unit, is the capacity factor of the unit. The number of programming units can be computed using the three following approaches.

An upper bound for the weekly spares can be used as a starting point for determining the number of programming units. The number of total update spares available at the site can be used for this purpose. In this case, the number of units is:

$$\text{Programming Units} = \left\lceil \frac{q(o)}{P \times C} \right\rceil . \tag{18}$$

This algorithm overestimates the number of programming units because only in the initial week of the update process are weekly spares equal to update spares. In the following weeks, the number of spares returning vary according to the assumed installation distribution.

A second approach for determining programming units is to make the weekly spares equal to the largest number of spares updated in any week after the initial week of the update process. In this case, the number of units is:

$$\text{Programming Units} = \max_{n \geq 1} \left\lceil \frac{q(n)}{P \times C} \right\rceil . \tag{19}$$

This algorithm reduces the number of units required and introduces a delay in the first week of the update process.

A third approach is to make the weekly spares equal to the expected number of spares returned each week for reprogramming. This is equivalent to the steady-state of the discrete time process. This approach satisfies the programming assumption of one day in 50 percent of the weeks and incurs a one-day delay during the other 50 percent. Therefore, on the average, programming will require 1.5 days per week. As a result, the nondelayed update time, which was derived assuming one day per week for programming, must be increased by 10 percent (one-half a day per five-day week). In this case, the number of units is:

$$\text{Programming Units} = \left\lceil \frac{\bar{Q}}{P \times C} \right\rceil , \tag{20}$$

where $\bar{Q}$ = expected number of spares returned each week. To evaluate $\bar{Q}$, the final value theorem can be used. Property (*iii*) of the z-transform methodology gives:

$$\begin{aligned} \bar{Q} = q(\infty) &= \lim_{z \to 1} (z-1) Q(z) \\ &= \text{Total Update Spares} \times \lim_{z \to 1} (z-1) A(z). \end{aligned} \tag{21}$$

Substituting eq. (9) into eq. (21) yields:

$$\bar{Q} = \text{Total Update Spares} \times \lim_{z \to 1} \frac{(z-1)z^{w}}{z^{w} - \sum_{j=0}^{w} i_j z^{w-j}} ,$$

which is indeterminate of the type 0 over 0, because $\sum_{j=0}^{w} i_j = 1$. The application of l'Hopital's rule solves this problem.

For the distribution given in the example of Fig. 1, the expected number of spares is:

$$\bar{Q} = q(\infty) = \text{Total Update Spares} \times \lim_{z \to 1} \frac{(z-1)z^6}{z^6 - 0.1z^3 - 0.5z^2 - 0.3z - 0.1}$$

$$= \text{Total Update Spares} \times 0.22727$$

$$= \text{Total Update Spares} \times \bar{A},$$

where $\bar{A}$ is the expected ratio of total update spares returned each week. It can be seen that the reciprocal of $\bar{A}$ represents the average number of weeks required for one pass through the basic flow of Fig. 1; that is,

$$\bar{T} = \frac{1}{\bar{A}} = \frac{1}{0.22727} = 4.4 \text{ weeks.}$$

For the example of Fig. 1, where one day per week is assumed for reprogramming, and the unit capacity is three systems per day, eqs. (18), (19), and (20) become, respectively:

$$\text{Programming Units} = \left\lceil \frac{\text{Total Update Spares}}{3} \right\rceil, \tag{22}$$

$$\text{Programming Units} = \left\lceil \frac{\text{Total Update Spares}}{6} \right\rceil, \quad \text{and} \tag{23}$$

$$\text{Programming Units} = \left\lceil \frac{\text{Total Update Spares}}{13.2} \right\rceil. \tag{24}$$

Comparing eqs. (22), (23), and (24), it can be seen that eq. (24) reduces the number of programming units by a factor of more than four, with respect to eq. (22), with the introduction of a 10-percent delay in the update time.

The algorithms discussed above are represented in Fig. 3. The number of required programming units are shown as a function of the number of total update spares. To contain the size of the plot, a scale factor, N, was introduced. Also included is a parameter D, which is the additional delay in the update time because of the reduced number of programming units at the centralized site.

3.4 Example

To have a better understanding of the determination of the resource requirements for the firmware update process, consider the following example.

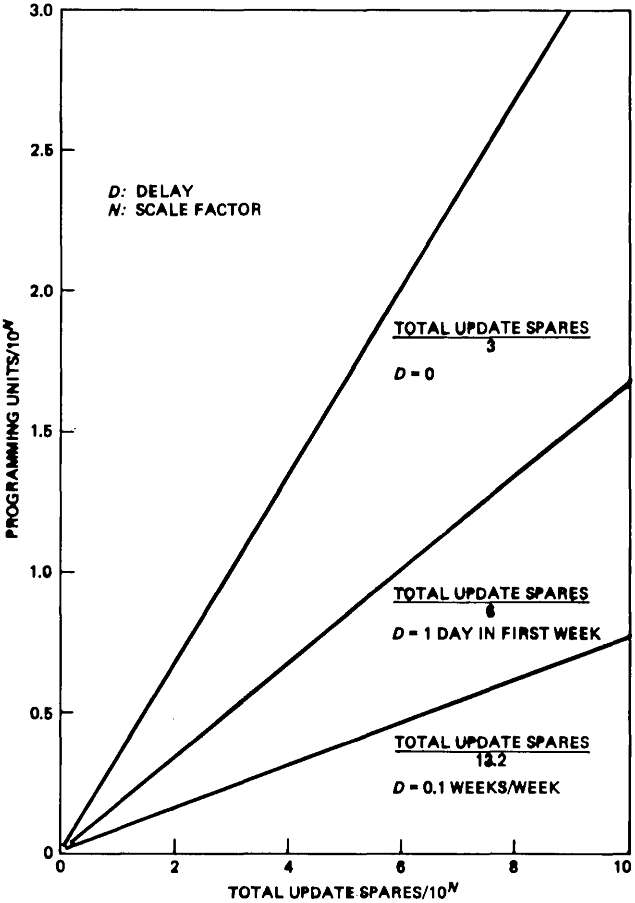


Fig. 3—Programming units.

For a projection of microprocessor systems, it is desired to determine the total update spares and programming unit requirements for the years 1982 and 1983, for an overall update time of 10 weeks in 1982 and 15 weeks in 1983, under the assumption of providing the fewest number of programming units. (See Tables III, IV, and V.)

IV. ECONOMIC IMPACT OF FIRMWARE UPDATE SCHEMES

4.1 Economic dependencies

Based on the obtained algorithms, an economic analysis of an update scheme can be performed if the capital and expense costs associated with the process are known.

The capital expenditures for the firmware update process are due to

Table III—Forecast of microprocessor systems

Year	Total Systems	Update Time Weeks
1982	26	10
1983	116	15

Table IV—Update spares from Fig. 2

Year	Nondelayed Update Time	Spare ratio	Total Systems	Total Update Spares
1982	9	0.59	26	15
1983	13	0.39	116	45

Table V—Programming units from Fig. 3 ($D = 0.1$)

Year	Programming Units
1982	2
1983	4

the total update spares and programming unit requirements. Therefore, the capital is a function of the desired update time, the total number of microprocessor systems, and the delay tolerated. If the capital budget is exceeded, the total update spares and programming units must be decreased to meet the dollar constraints, increasing the update time. The update time is independent of the number of microprocessor firmware changes; therefore, the capital is also independent of the change rate.

The update expenses are due to the labor costs associated with the programming unit operation, the installation efforts, and the costs of crating and shipping. Since these tasks are performed for each microprocessor system each time a program change occurs, the expenses are dependent on the system market and program change rate. The economic dependencies are shown in Fig. 4.

Using standard techniques, we can determine the economic feasibility of alternate memory schemes for microprocessor-based systems. One alternative, for example, could be to determine the benefits of replacing EPROM memory with random access memory (RAM) and

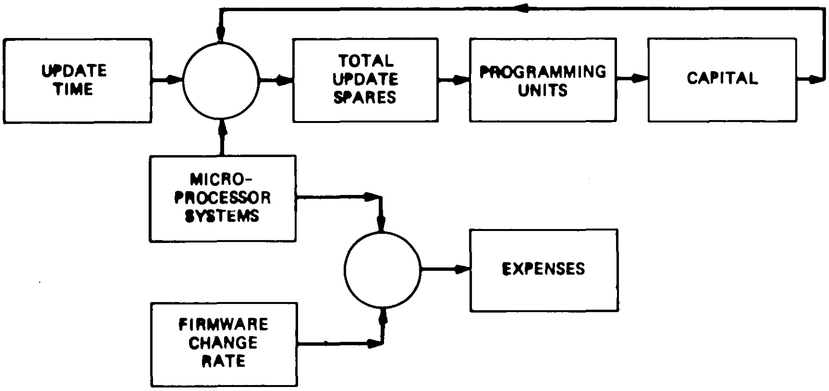


Fig. 4—Economic dependencies.

magnetic bubble store as backup. This possibility, even though initially more expensive, has the capability of being updated electrically via a data link, and may offer savings over the life cycle of the system.

V. SUMMARY

Algorithms related to the EPROM firmware update process for microprocessor systems have been developed in this paper. They give the update time, the number of update spares, and the numer of programming units required at a centralized site for a given firmware installation distribution. The algorithms presented do not render policy decisions, but enable the user to determine the economics and responsiveness of alternative administrative schemes.

REFERENCE

1. E. I. Jury, "Theory and Application of the z-Transform Method," New York: John Wiley, 1964.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

On The Performance of Phase-Shift-Keying Systems

By V. K. PRABHU and J. SALZ

(Manuscript received June 18, 1981)

Coherent phase-shift keying (CPSK) and differential phase-shift keying (DPSK) are widely used modulation methods in digital communications. Bandwidth efficiency, good noise immunity, constant envelope, and simplicity of implementation make these schemes particularly attractive for use over the satellite, terrestrial radio and voiceband telephone channels. While system analyses abound in the literature, treatment is usually restricted to the additive Gaussian channel. Important issues determining ultimate performance, such as the joint effect of intersymbol interference and the acquisition of carrier phase have not been adequately addressed. The main purpose of this paper is to develop analytical tools that can be used to assess system performance under practical operating conditions. Pure coherent demodulation schemes such as CPSK are ideals which are rarely achieved in practice, and carrier phase must be estimated prior to and/or during data transmission. This requires start-up time, as well as added equipment, and the fidelity of the phase estimate ultimately determines performance. In contrast, DPSK is independent of carrier phase, since decisions are made on phase differences. However, this comes at a price, and it is known that ideal multiphase DPSK suffers an asymptotic performance penalty of 3 dB in signal-to-noise ratio (s/n) over ideal CPSK. We develop a new rigorous method for calculating the error rates of both CPSK and DPSK, under a variety of operating conditions. In particular, we find that the intersymbol interference penalty for quaternary DPSK is about 1 dB worse in s/n than for CPSK. We demonstrate that the detection efficiency of CPSK approaches the ideal, provided that the s/n of the phase-recovery circuit is about 10 dB more than that at the receiver input. Alternatively, for the same s/n, a 10-baud phase-locked loop integration time is required to achieve near-ideal performance.

I. INTRODUCTION

Coherent phase-shift keying (CPSK) and differential phase-shift keying (DPSK) are two techniques often used in digital communications over channels such as satellite, terrestrial radio, and voiceband telephone. The literature abounds in analyses of their performance under a variety of conditions. A sample collection of some of this literature may be found in Ref. 1. The chief reasons for the widespread use of these techniques are simplicity of implementation, superior performance over the additive Gaussian noise channel, minimal bandwidth occupancy, and minimal envelope variation.

The relative performance of CPSK and DPSK systems is well understood only in the presence of additive Gaussian noise. In this case, the detection efficiency of DPSK is known to be about 1 dB (in s/n) below that of CPSK for binary modulation and this degradation approaches 3 dB for multilevel systems. In applications where a 3-dB loss in s/n is important, such as in down-link satellite, space communications, and terrestrial radio under deep fading conditions, CPSK is the preferred method. In CPSK, however, the generation and extraction of a local carrier-phase reference at the receiver is required. A coherent phase estimate is usually obtained by using phase-locked loop (PLL) techniques, and because of frequency instabilities and phase jitter inherent in transmitter and receiver systems, carrier recovery loop bandwidths cannot be made arbitrarily small. Consequently, in practice a noisy phase estimate is obtained and only partial coherent reception can be claimed. The reason for using DPSK is its immunity from slow carrier-phase fluctuations; therefore, the phase recovery problem inherent in CPSK is avoided. However, the detection efficiency of DPSK may approach that of CPSK under noisy phase estimation conditions and intersymbol interference (ISI). The need to understand this phenomenon on a fundamental level is the principal objective of this paper.

As bandwidth occupancy is always important, the effects of ISI generated by the use of band-limiting filters must be taken into account in any analysis of these systems. Because of the linear nature of the demodulation process in CPSK, the effect of ISI has been treated in great detail. Since DPSK demodulation is inherently nonlinear, the analysis of performance is very difficult and no adequate analytical methods are currently available. Also, the combined effects of imperfect phase estimation and ISI on CPSK must be determined so that the relative detection efficiencies of band-limited DPSK and CPSK can be fairly assessed.

In Section II of this paper, we describe a technique for determining the degradation in M-ary CPSK operating in the presence of ISI, additive Gaussian noise, and imperfect carrier phase. In Section III, we consider the performance of M-ary DPSK subject to ISI and additive Gaussian noise.

II. COHERENT DETECTION

2.1 System description of CPSK

Figure 1 shows the M-ary CPSK system that we consider. The signal, $s(t)$, before the transmit filter can be represented as

$$s(t) = \mathrm{Re}\{Ax(t)\exp[i(2\pi f_c t + \mu)]\},\ i = \sqrt{-1}, \tag{1}$$

where the baseband modulation signal is

$$x(t) = \sum_{k=-\infty}^{\infty} \exp(i\alpha_k)\mathrm{rect}[(t - kT)/T], \tag{2}$$

and the constants A, f_c, and μ are the carrier amplitude, frequency (in Hz), and phase, respectively. Also, rect$(\cdot)$ is the rectangular window function, T the signaling interval, and the sequence of discrete phases $[\alpha_k]$ corresponds to the data sequence to be transmitted. Without loss of generality, we assume that the M phase values of α_k are uniformly distributed with equal probability between $(-\pi, \pi]$. So, α_k takes on value in the set $\alpha_k \epsilon \Lambda$,

$$\Lambda \triangleq \left[\frac{\pi}{M}, \frac{3\pi}{M}, \cdots, \frac{2M-1}{M}\pi\right].$$

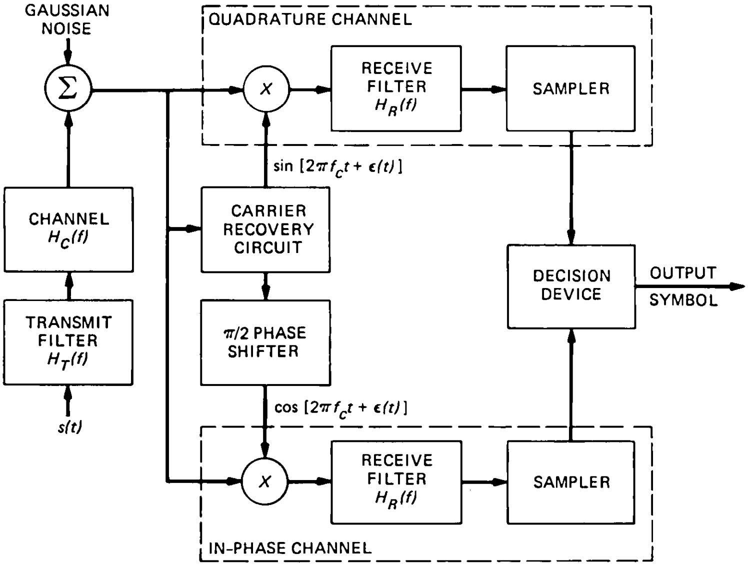


Fig. 1—M-ary CPSK receiver, $M > 2$.

We also assume that the data phases in different time slots are statistically independent.

In our model, the transmit filter, transmission channel, and receive filter are linear and time invariant. Therefore, the complex envelope, $y(t)$, at the output of the receive filter may be written as

$$y(t) = x(t) \circledast h(t) + n(t) + i\hat{n}(t)$$
$$h(t) = h_T(t) \circledast h_C(t) \circledast h_R(t),$$

where $h_T(t)$, $h_C(t)$, and $h_R(t)$ are, respectively, the impulse response of the transmit filter, the channel, and the receive filter. The symbol $\circledast$ denotes convolution. Also, $n(t) + i\hat{n}(t)$ is the complex envelope of the Gaussian noise passed through the receive filter. For symmetrical filters, $n(t)$ and $\hat{n}(t)$ are independently and identically distributed (iid) Gaussian random variables with mean zero, and variance

$$\sigma^2 = N_0 \int_{-\infty}^{\infty} |H_R(f)|^2 df,$$

where N_0 is the double-sided spectral density of the original white noise and $H_R(f)$ is the baseband equivalent transfer function of the receive filter.

2.1.1 Detection in CPSK

Assuming that the recovered carrier is $\exp[i(2\pi f_c t + \hat{\mu})]$, where $\hat{\mu}$ is an estimate of μ in eq. (1), the detector operates on the signal, $w(t)$, represented as

$$w(t) = \sum_{k=-\infty}^{\infty} z(t - kT)\exp[i(\alpha_k + \epsilon)] + \xi + i\eta, \qquad (3)$$

where ξ and η are iid gaussian random variables with mean zero, and variance σ^2,

$$\epsilon = \mu - \hat{\mu},$$

is the phase error, and

$$z(t) = h_T(t) \circledast h_C(t) \circledast h_R(t) \circledast \operatorname{rect}\left(\frac{t}{T}\right).$$

To estimate the transmitted phase, $\alpha_0 = \Phi \epsilon \Lambda$ at $t = 0$, an ideal CPSK detector measures the phase θ of $w(t)$ at $t = t_0$, and a correct decision results when

$$\Phi - \frac{\pi}{M} < \theta < \Phi + \frac{\pi}{M},$$
$$\theta = \text{phase angle of } w(t_0),$$
$$w(t_0) = w(t)\bigg|_{\substack{t=t_0 \\ \alpha_0=\Phi}}. \qquad (4)$$

2.1.2 Error rate for M-ary CPSK

Here, we briefly review some known results for CPSK and then develop new results applicable to our more general model.

Error-rate calculations for ideal CPSK in added Gaussian noise can be found in Refs. 2 to 7. References 8 and 9 provide numerical methods for calculating the probability of error in the presence of ISI. Reference 10 takes into account ISI and demodulation phase error, but the results are restricted to only binary and quaternary systems. We now generalize these results.

Using the union bound and the representation of the received signal, eq. (3), it follows from eq. (4) that the probability of error, $\mathrm{Pe}(|\Phi)$, given that the phase Φ is transmitted, is

$$\max(P_1, P_2) \le \mathrm{Pe}(|\Phi) \le P_1 + P_2,$$

where

$$\begin{aligned} P_1 &= \Pr\left[\sin\left(\theta - \Phi + \frac{\pi}{M}\right) < 0\right] \\ &= \Pr\left\{\operatorname{Im} w(t_0)\exp\left[-i\left(\Phi - \frac{\pi}{M}\right)\right] < 0\right\}, \\ P_2 &= \Pr\left[\sin\left(\theta - \Phi - \frac{\pi}{M}\right) > 0\right] \\ &= \Pr\left\{\operatorname{Im} w(t_0)\exp\left[-i\left(\Phi + \frac{\pi}{M}\right)\right] > 0\right\}. \end{aligned} \tag{5}$$

Note that the average symbol probability of error Pe is

$$\mathrm{Pe} = \frac{1}{M}\sum_{\Phi\epsilon\Lambda} \mathrm{Pe}(|\Phi).$$

But, since the signal constellation is assumed to be circularly symmetric, $\mathrm{Pe}(|\Phi)$ is independent of Φ.

For convenience, we shall now assume that $\Phi = \pi/M$. Hence,

$$\begin{aligned} P_1 = \Pr\Bigg(\operatorname{Im}\Bigg\{ & r_0\exp\left[i\left(\beta_0 + \frac{\pi}{M} + \epsilon\right)\right] \\ & + \textstyle\sum' r_k\exp\left[i\left(\beta_k + \alpha_k + \epsilon\right)\right]\Bigg\} + \eta < 0\Bigg), \end{aligned} \tag{6}$$

where

$$r_k\exp(i\beta_k) = z(t_0 - kT)$$

and $\sum'$ denotes the exclusion of the term $k = 0$. A similar expression can be written for P_2.

Accurate estimations of P_1 and P_2 are easy to obtain in the presence of only Gaussian noise, but are more difficult when ISI is added and are even more tedious when the distribution of carrier-phase error, ϵ, must be taken into account.

In the next section we derive an exponentially tight upper bound on these quantities for a fixed carrier-phase error and then perform asymptotic [large signal-to-noise ratio (s/n)] analyses on these upper bounds for a given distribution of carrier-phase error.

2.1.3 Bounds on the error rate

We begin by writing eq. (6) as

$$P_1 = < \Pr\left[I(\epsilon) + \eta < -r_0 \sin\left(\frac{\pi}{M} + \beta_0 + \epsilon\right) \Big| \, \right]_{\epsilon} >_{\epsilon}, \tag{7}$$

where $\langle \, \rangle_\epsilon$ denotes expectation with respect to ϵ, and where,

$$I(\epsilon) = \textstyle\sum' r_k \sin(\beta_k + \alpha_k + \epsilon). \tag{8}$$

Before we can proceed with eq. (7), we need specific information on the probability density function (pdf) of the demodulating phase error ϵ. We shall assume that the phase reference is derived from a pure tone by a first-order PLL. It is well known[1] that the resulting pdf for the phase error, ϵ, is

$$p_\epsilon(\epsilon) = \frac{\exp(\lambda \cos \epsilon)}{2\pi I_0(\lambda)}, \; |\epsilon| \leq \pi, \tag{9}$$

where λ is the s/n at the input to the PLL multiplied by the reciprocal of the PLL bandwidth,

$$\lambda = \frac{G}{N_p B_L}. \tag{10}$$

In eq. (10), G is the average power in the carrier, N_p is the double-sided noise spectral density, and B_L is the noise bandwidth of the linearized PLL. Also, in eq. (9), $I_0(x)$ represents the modified Bessel function of the first kind and of order 0. For a second-order PLL, the pdf of ϵ is also approximately given by eq. (9). We shall use this density to obtain bounds on P_1.

Since ϵ is a symmetric random variable, eq. (7) yields

$$P_1 = \frac{1}{2} \langle V(\epsilon) + V(-\epsilon) \rangle_\epsilon,$$

where

$$V(\epsilon) \triangleq \frac{1}{2} \operatorname{erfc}\left[\frac{r_0 \sin[(\pi/M) + \beta_0 + \epsilon] + \sum' r_k \sin(\alpha_k + \beta_k + \epsilon)}{\sqrt{2}\,\sigma} \right]. \tag{11}$$

Using upper bounding techniques and Laplace's method,[11] we show in Appendix A that

$$P_1 \leq J_{1a} + J_{2a},$$

where

$$J_{1a} \approx \frac{1}{2}\left[\frac{\exp\{-\rho^2[\sin^2[(\pi/M) + \beta_0 - \epsilon_0] + D(1 - \cos \epsilon_0)]\}}{\{\cos \epsilon_0 + (2/D) \cos 2[(\pi/M) + \beta_0 - \epsilon_0]\}^{1/2}} + \frac{\exp\{-\rho^2 \sin^2[(\pi/M) + \beta_0]\}}{\{1 + (2/D) \cos 2[(\pi/M) + \beta_0]\}^{1/2}}\right], \rho^2 = \frac{r_0^2}{2(\sigma^2 + \sigma_I^2)} \gg 1,$$

$$\sigma_I^2 = \sum{}' r_k^2,\ D = \frac{\lambda}{\rho^2},$$

$$\epsilon_0 = \frac{1}{D} \sin 2\left(\frac{\pi}{M} + \beta_0\right)\left[1 - \frac{2}{D} \cos 2\left(\frac{\pi}{M} + \beta_0\right) + \cdots\right], D \gg 1.$$

and

$$J_{2a} \approx \left(1 - \frac{\delta}{\pi}\right) \sqrt{2\pi D\rho^2} \exp[-D\rho^2(1 - \cos \delta)],\ \delta = \frac{\pi}{M} + \beta_0.$$

Note that ρ^2 is the s/n of the system. Also, D can be regarded as the ratio of s/n in the phase recovery circuit to that in the PSK system or the integration time in bauds.

Similarly, we can show that

$$P_2 \leq J_{1a} + J_{2a}.$$

In summary, the average symbol probability of error, Pe, for M-ary CPSK system can be upper bounded by

$$\text{Pe} \leq \frac{\exp\{-\rho^2[\sin^2[(\pi/M) + \beta_0 - \epsilon_0] + D(1 - \cos \epsilon_0)]\}}{\{\cos \epsilon_0 + (2/D) \cos 2[(\pi/M) + \beta_0 - \epsilon]\}^{1/2}} + \frac{\exp\{-\rho^2 \sin^2[(\pi/M) + \beta_0]\}}{\{1 + 2/D \cos 2(\pi/M + \beta_0)\}^{1/2}} + 2\left[1 - \frac{(\pi/M) + \beta_0}{\pi}\right] \sqrt{2\pi D\rho^2} \times \exp\left\{-D\rho^2\left[1 - \cos\left(\frac{\pi}{M} + \beta_0\right)\right]\right\}, \rho^2 \gg 1,$$

$$\epsilon_0 = \frac{1}{D} \sin 2\left(\frac{\pi}{M} + \beta_0\right)\left[1 - \frac{2}{D} \cos 2\left(\frac{\pi}{M} + \beta_0\right) + \cdots\right], D \gg 1.$$

This upper bound becomes

$$P \leq 2 \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0\right)\right], \tag{12}$$

when phase estimation is perfect, $D \to \infty$. Equation (12) is the well-known Chernoff bound for M-ary CPSK.[12]

If the observation interval of the PLL is large, $D \gg 1$, and if $M \gg 1$,

$$\epsilon_0 \approx \frac{1}{D} \sin 2\left(\frac{\pi}{M} + \beta_0\right)\left[1 - \frac{2}{D} \cos 2\left(\frac{\pi}{M} + \beta_0\right)\right],$$

and

$$\begin{aligned} P \leq{} & \exp\left(-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0\right)\right. \\ & \left.\times\left\{1 - \frac{2\cos^2[(\pi/M) + \beta_0]}{D}\left[1 - \frac{2}{D}\cos 2\left(\frac{\pi}{M} + \beta_0\right)\right]\right\}\right) \\ & + \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0\right)\right] \\ & \sim \exp\left(-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0\right)\right. \\ & \left.\times\left\{1 - \frac{2\cos^2[(\pi/M) + \beta_0]}{D}\left[1 - \frac{2}{D}\cos 2\left(\frac{\pi}{M} + \beta_0\right)\right]\right\}\right), \\ \rho^2 ={} & \frac{r_0^2}{2(\sigma^2 + \sigma_I^2)}, \rho^2 \gg 1, M > 2, D \gg 1. \end{aligned} \tag{13}$$

Comparing eqs. (12) and (13), we see that the degradation in s/n because of imperfect phase estimate for multiphase CPSK systems is asymptotically given by

$$G = \left[\left\{1 - \frac{2}{D}\cos^2\left(\frac{\pi}{M} + \beta_0\right)\right\}\left\{1 - \frac{2}{D}\cos 2\left(\frac{\pi}{M} + \beta_0\right)\right\}\right]^{-1},$$

where $G \to 1$ as $D \to \infty$ as it should.

2.2 Example of quaternary (M = 4) CPSK system

Let us consider a quaternary ($M = 4$) CPSK system and assume that the channel is ideal.

If 4-pole Butterworth transmit and receive filters are used, the resulting average symbol probability of error is plotted in Fig. 2. Note that the bound is fairly tight and when the s/n of the phase recovery circuit is about 10 dB more than at the receiver input, the detection efficiency of CPSK is essentially determined by ISI alone. Alternatively, we can say that, for the same s/n, a 10-baud PLL integration time is required to achieve this ISI-limited performance. For this filter, the ISI penalty is about 1 dB.

If $M > 2$, it is well known that the penalty in s/n because of Gaussian noise alone is asymptotically given by $1/[\sin^2(\pi/M)]$.

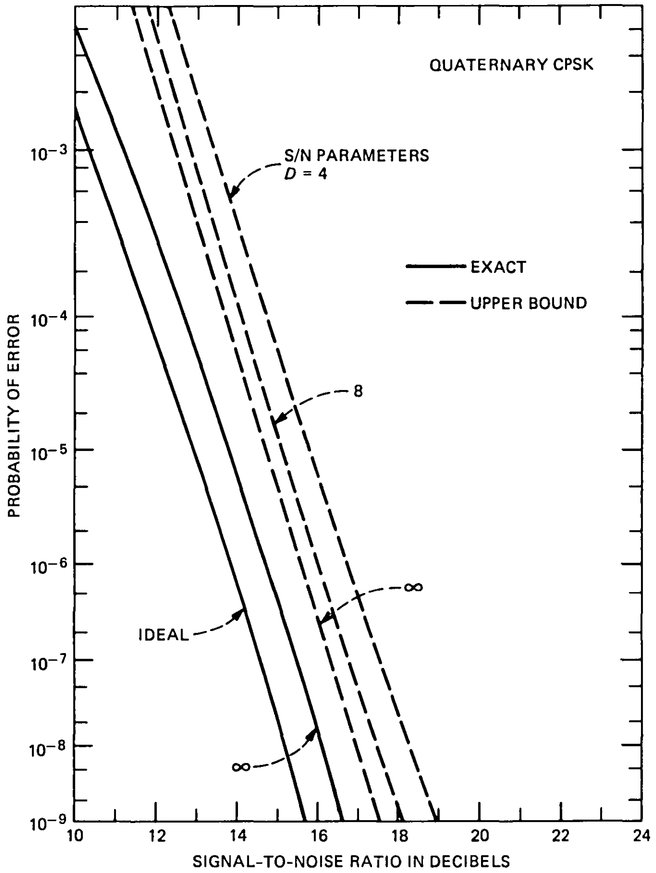


Fig. 2—Probability of error for quaternary phase-shift keying (QPSK) with rectangular signaling, noisy carrier-phase recovery, and 4-pole Butterworth transmit and receive filters. The s/n in decibels is defined as $10 \log_{10}[T/2N_0]$, where N_0 is the double-sided noise spectral density and the ideal received signal power has been normalized to unity. Parameter D is the ratio of s/n in the phase recovery loop to that in the PSK system. The double-sided 3-dB bandwidth of the transmit filter is $2/T$ and that of the receive filter is $1.06/T$. Sampling time is $1.74T$.

The upper bound in eq. (13) indicates that if the definition of s/n is modified to take into account the ISI power σ_I^2, the additional penalty, because of imperfect phase estimation, is

$$G_t = \left[\sin^2\left(\frac{\pi}{M} + \beta_0\right) \times \left\{ 1 - \frac{2}{D} \cos^2\left(\frac{\pi}{M} + \beta_0\right)\left[1 - \frac{2}{D} \cos 2\left(\frac{\pi}{M} + \beta_0\right)\right]\right\}\right]^{-1}.$$

This quantity is plotted in Fig. 3. We observe that the s/n penalty

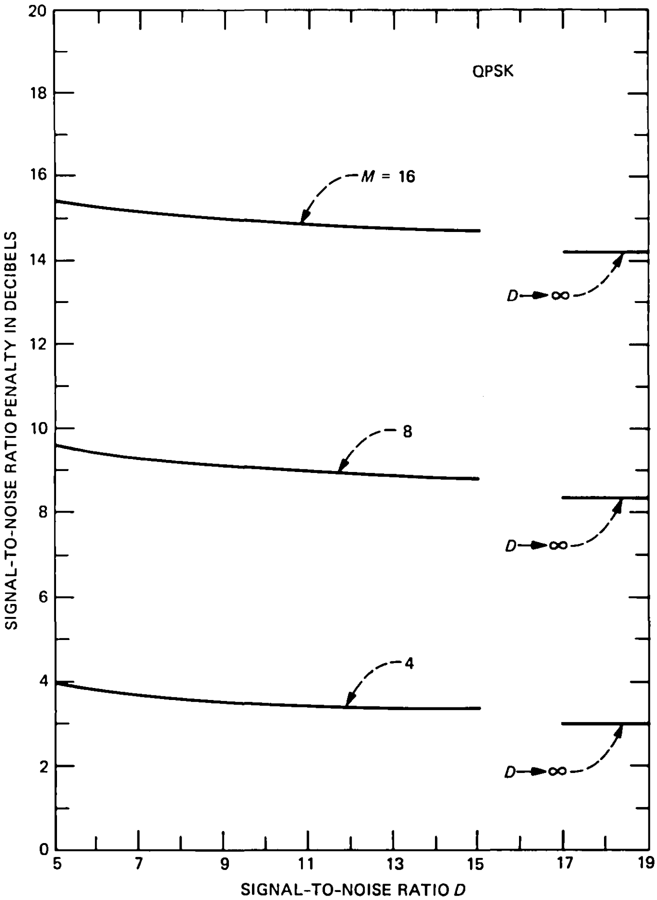


Fig. 3—Signal-to-noise ratio penalty for M-ary CPSK with imperfect phase estimation. Parameter D is the ratio of s/n in the phase recovery loop to that in the PSK system.

because of ISI is independent of M. In Fig. 3, also note that $G_t \to 1/[\sin^2(\pi/M)]$ as $D \to \infty$.

III. DIFFERENTIAL DETECTION

3.1 System description of DPSK

The M-ary DPSK system is shown in Fig. 4. As before, the baseband modulated DPSK signal can be represented as

$$x(t) = \sum_{k=-\infty}^{\infty} \exp(i\alpha_k)\text{rect}[(t - kT)/T]. \qquad (2)$$

Here, however, the sequence of phases $[\beta_k] = [\alpha_{k+1} - \alpha_k]$ corresponds to the data sequence to be transmitted. Again, we assume that M

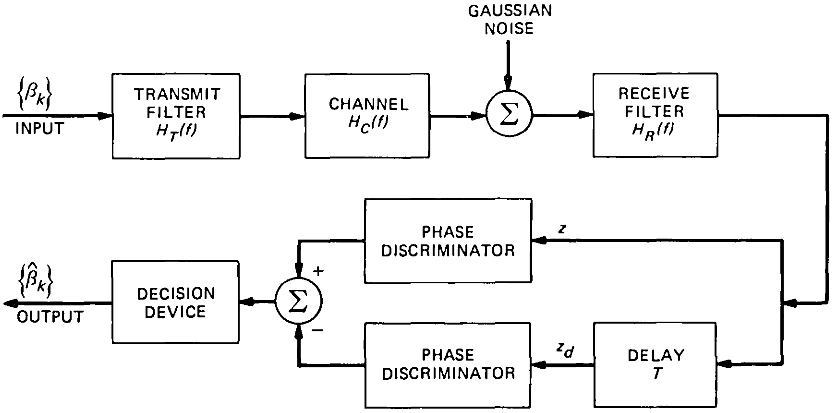


Fig. 4—M-ary DPSK system.

phase values of β_k are equally distributed over the interval $[0, 2\pi)$ and choose

$$\beta_k = (2l - 1)\frac{\pi}{M}, \; 1 \leq l \leq M, \text{ modulo } 2\pi.$$

As in CPSK, we represent the set

$$\left[\frac{\pi}{M}, \frac{3\pi}{M}, \cdots, (2M-1)\frac{\pi}{M}\right]$$

by $[\Lambda]$. Also, we shall assume that the phase symbols, β_k's, in different time slots are statistically iid.

If the received phasor at time t_0 is indicated by z and the one in the succeeding interval is indicated by z_d, the detected phase difference measured by an ideal differential detector is

$$\theta = \text{angle of } w, \; w \triangleq z^* z_d,$$

where $*$ represents the complex conjugate. For the system shown in Fig. 4,

$$z = \sum_{k=-\infty}^{\infty} (g_k + ip_k)\exp(i\alpha_k) + n_c + in_s, \tag{14}$$

and

$$z_d = \sum_{k=-\infty}^{\infty} (g_{k-1} + ip_{k-1})\exp(i\alpha_k) + n_{c_d} + in_{s_d}, \tag{15}$$

where g_k and p_k are real,

$$g_k + ip_k = \int_{-\infty}^{\infty} h(t_0 - \mu)\text{rect}\left(\frac{\mu - kT}{T}\right) d\mu. \tag{16}$$

As before, $h(t)$ is the overall impulse response of the system with transfer characteristic

$$H(f) = H_T(f)H_C(f)H_R(f). \tag{17}$$

In eqs. 14 and 15, n_c, n_s, n_{c_d}, and n_{s_d} are iid real Gaussian random variables with mean zero and variance

$$\sigma^2 = N_0 \int_{-\infty}^{\infty} |H_R(f)|^2 df,$$

where N_0, is the double-sided spectral density of the added white Gaussian noise. In eq. (17), H_T is the transfer function of the transmit filter, H_R, of the receive filter, and $H_C(f)$, the transfer function of the channel. The assumption that the Gaussian noise at t_0 is independent of the noise at $t_0 - T$ can be justified if the receive filter bandwidth is small compared with $1/T$. Most of our analysis can be extended if these two noise samples are correlated.

3.1.1 Probability of error for M-ary systems

If the transmitted symbol associated with the time index $k = 0$ is $\Phi \epsilon \Lambda$, a correct decision is made when the received phase difference θ is such that

$$\Phi - \frac{\pi}{M} < \theta < \Phi + \frac{\pi}{M}.$$

As before, the following bounds apply:

$$\max(P_1, P_2) \leq \text{Pe}(|\Phi) \leq P_1 + P_2,$$

where

$$P_1 = \Pr\left[\sin\left(\theta - \Phi + \frac{\pi}{M}\right) < 0\right],$$

$$P_2 = \Pr\left[\sin\left(\theta - \Phi - \frac{\pi}{M}\right) > 0\right]. \tag{18}$$

These statements are identical to the ones that apply to CPSK, but here θ represents a "differential phase" and, therefore, the estimation of these probabilities becomes extremely involved. Good estimates are only available when Gaussian noise is the sole source of impairment.

We proceed to analyze P_1 and observe that a bound on P_1 also provides a bound on P_2. Since the calculations are extremely tedious, we relegate the details to appendices and strive to develop only the main ideas here. Therefore, from eq. (18), we get

$$\begin{aligned} P_1(|\Phi) &= \Pr\left[\sin\left(\theta - \Phi + \frac{\pi}{M}\right) < 0\right] \\ &= \Pr\left\{\operatorname{Im} w \exp\left[-i\left(\Phi - \frac{\pi}{M}\right)\right] < 0\right\} \\ &= \Pr\left\{\operatorname{Im} z^* z_d \exp\left[-i\left(\Phi - \frac{\pi}{M}\right)\right] < 0\right\} \\ &= \Pr\left\{\operatorname{Re}\ (z)^* z_d \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right] < 0\right\} \\ &= \Pr(\operatorname{Re} z_1^* z_2 < 0), \end{aligned} \tag{19}$$

where

$$\begin{aligned} z_1 &= z \\ z_2 &= z_d \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right], \end{aligned}$$

and z and z_d are given in eqs. (14) and (15). Since

$$\operatorname{Re} z_1^* z_2 = \left|\frac{z_1 + z_2}{2}\right|^2 - \left|\frac{z_1 - z_2}{2}\right|^2, \tag{20}$$

eqs. (19) and (20) yield

$$P_1(|\Phi) = \Pr(|w_1| < |w_2|), \tag{21}$$

where

$$\begin{aligned} w_1 &= \frac{z_1 + z_2}{2} = \frac{z + z_d \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{2}\left\{(g_k + ip_k) + (g_{k-1} + ip_{k-1})\right. \\ &\qquad \left.\times \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]\right\} \exp(i\alpha_k) + \xi_+ + i\eta_+ \\ w_2 &= \frac{z_1 - z_2}{2} = \frac{z - z_d \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{2}\left\{(g_k + ip_k) - (g_{k-1} + ip_{k-1})\right. \\ &\qquad \left.\times \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]\right\} \exp(i\alpha_k) + \xi_- + i\eta_-, \end{aligned}$$

and where ξ_+, η_+, ξ_-, and η_- are Gaussian noise terms, given by

$$\xi_+ = \frac{n_c + \mathrm{Re}(n_{cd} + in_{sd}) \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2}$$

$$\eta_+ = \frac{n_s + \mathrm{Im}(n_{cd} + in_{sd}) \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2}$$

$$\xi_- = \frac{n_c - \mathrm{Re}(n_{cd} + in_{sd}) \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2}$$

and

$$\eta_- = \frac{n_s - \mathrm{Im}(n_{cd} + in_{sd}) \exp - i[\Phi - (\pi/M) + (\pi/2)]}{2}.$$

It can be verified that the above Gaussian random variables are iid with mean zero and variance $\sigma^2/2$.

3.1.2 Exact computation of probability of error

For a given symbol sequence, the conditional probability of error is seen from eq. (21) to be given by the probability that a particular Gaussian quadratic form exceeds another. This is a well-known problem and the answer can conveniently be expressed in terms of the tabulated Marcum Q function.[13] Thus, after some algebra, eq. (21) can be shown as[14]

$$\begin{aligned} &P_1\ (|\Phi, \text{symbol sequence}) \\ &= \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\left[1 - Q\left(\frac{a_+}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \frac{a_-}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)\right] \\ &\quad + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} Q\left(\frac{a_-}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \frac{a_+}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right), \end{aligned} \tag{22}$$

where

$$Q(a, b) = \int_b^\infty \exp\left(-\frac{a^2 + x^2}{2}\right) I_0(ax) x dx,$$

and $I_n(\cdot)$ is the modified Bessel function of the first kind and of order n,

$$a_+ = |\langle w_1 \rangle_{\xi_+, \eta_+}|,$$

$$a_- = |\langle w_2 \rangle_{\xi_-, \eta_-}|,$$

$$\sigma_1^2 = \langle \xi_+^2 \rangle = \langle \eta_+^2 \rangle = \sigma^2/2$$

and

$$\sigma_2^2 = \langle \xi_-^2 \rangle = \langle \eta_-^2 \rangle = \sigma^2/2.$$

The major difficulty at this point is clearly carrying out the averages in eq. (22) over all possible symbol sequences. In general, a_- and a_+ contain an infinite number of ISI terms and the averaging process is difficult. Clearly, for a small number of ISI terms, it can be carried out by enumeration. But, in general, the number of terms in computing the average explodes exponentially and enumeration becomes intractable. For example, for 10 ISI terms and a quaternary DPSK system, the number of terms is about a million! So, we obviously need more efficient methods of estimating these averages.[15]

In this paper, we assume that the number of dominant ISI terms contained in a_+ and a_- is not large and that they become insignificant when ISI samples are far away from the desired sample. Assuming that the same number N indicates the number of dominant preceding and succeeding ISI samples (total significant ISI terms is $2N$), our approach, then, is to obtain upper and lower bounds on P_1 as a function of N and demonstrate that these bounds coincide with $N \to \infty$.

For any N, the evaluation of these bounds requires M^{2N} computations. This can be carried out with modest effort on a high-speed digital computer. The error becomes smaller when N is increased.

We show in Appendix B that the error probability can be bounded as

$$\chi_1(N) \le P_1(|\Phi) \le \chi_2(N),$$

where

$$\begin{aligned}\chi_1(N) = {} & \frac{1}{1+(1-\Delta)^2} \\ & \times\left\{1-\left\langle Q\left[\frac{\sqrt{2}a_+}{\sigma\sqrt{1+(1-\Delta)^2)}},\frac{\sqrt{2}a_-(1-\Delta)}{\sigma\sqrt{1+(1-\Delta)^2}}\right]\right\rangle\right\} \\ & +\frac{(1-\Delta)^2}{1+(1-\Delta)^2}\left\langle Q\left[\frac{\sqrt{2}a_-(1-\Delta)}{\sigma\sqrt{1+(1-\Delta)^2}},\frac{\sqrt{2}a_+}{\sigma\sqrt{1+(1-\Delta)^2}}\right]\right\rangle \\ & -\left[1-\left\langle Q\left(\frac{\sqrt{2}a_-}{\sigma},\frac{\sqrt{2}p}{\sigma}\right)\right\rangle\right]-4\exp\left(-\frac{\Delta^2p^2}{2\sum\limits_{\substack{k<-N\\k>N}}G_k^2}\right) \\ & -4\exp\left(-\frac{\Delta^2p^2}{2\sum\limits_{\substack{k<-N\\k>N}}H_k^2}\right) \qquad (23)\end{aligned}$$

and

$$
\begin{aligned}
\chi_2(N) &= \frac{1}{1+(1+\Delta)^2} \\
&\quad \times \biggl\{ 1 - \biggl\langle Q\biggl[\frac{\sqrt{2}a_+}{\sigma\sqrt{1+(1+\Delta)^2}}, \frac{\sqrt{2}a_-(1+\Delta)}{\sigma\sqrt{1+(1+\Delta)^2}} \biggr] \biggr\rangle \biggr\} \\
&\quad + \frac{(1+\Delta)^2}{1+(1+\Delta)^2} \biggl\langle Q\biggl[\frac{\sqrt{2}a_-(1+\Delta)}{\sigma\sqrt{1+(1+\Delta)^2}}, \frac{\sqrt{2}a_-}{\sigma\sqrt{1+(1+\Delta)^2}} \biggr\rangle \biggr] \\
&\quad + \biggl[1 - \biggl\langle Q\biggl(\frac{\sqrt{2}a_-}{\sigma}, \frac{\sqrt{2}p}{\sigma} \biggr) \biggr\rangle \biggr] + 4\exp\Biggl(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} G_k^2} \Biggr) \\
&\quad + 4\exp\Biggl(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} H_k^2} \Biggr).
\end{aligned}
\tag{24}
$$

In eqs. (23) and (24), Δ and p are arbitrary, $0 \le \Delta \le 1$, and

$$
G_k = \left| \frac{1}{2} \left\{ (g_k + ip_k) + (g_{k-1} + ip_{k-1}) \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2} \right) \right] \right\} \right|,
$$

$$
H_k = \left| \frac{1}{2} \left\{ (g_k + ip_k) - (g_{k-1} + ip_{k-1}) \exp\left[-i\left(\Phi - \frac{\pi}{m} + \frac{\pi}{2} \right) \right] \right\} \right|.
$$

For any N, we can choose Δ and p by trial and error so that the difference between the upper and lower bounds is a minimum. Since this optimization is not critical to our method, we choose

$$
p = \left[\frac{2\sigma_R^2}{\Delta} \left(\frac{\langle a_-^2 \rangle}{\sigma^2} + \ln \frac{\sigma^2}{\sigma_R^2} \Delta \right) \right]^{1/2},
$$

where

$$
\sigma_R^2 = \max\Biggl(\sum_{\substack{k<-N \\ k>N}} G_k^2, \sum_{\substack{k<-N \\ k>N}} H_k^2 \Biggr).
$$

For $\Delta \ll 1$, the difference, Z, between the upper and the lower bounds can be shown as

$$
\begin{aligned}
Z &= 2\Delta \left[1 + \frac{a_{+\max}^2}{\sigma^2} - \frac{a_{-\min}^2}{\sigma^2} \right] P_1(|\Phi, N) \\
&\quad + 8\exp\left(\frac{-\langle a_-^2 \rangle}{\sigma^2} \right) \frac{\sigma_R^2}{\sigma^2 \Delta} \left(\frac{\langle a_-^2 \rangle}{\sigma^2} + \ln \frac{\sigma^2}{\sigma_R^2} \Delta \right) \\
&\quad + 8\exp\Biggl\{ - \frac{\Delta}{\sum\limits_{\substack{k<-N \\ k>N}} G_k^2} \left[\sigma_R^2 \left(\frac{\langle a_-^2 \rangle}{\sigma^2} + \ln \frac{\sigma^2}{\sigma_R^2} \Delta \right) \right] \Biggr\} \\
&\quad + 8\exp\Biggl\{ - \frac{\Delta}{\sum\limits_{\substack{k<-N \\ k>N}} H_k^2} \left[\sigma_R^2 \left(\frac{\langle a_-^2 \rangle}{\sigma^2} + \ln \frac{\sigma^2}{\sigma_R^2} \Delta \right) \right] \Biggr\},
\end{aligned}
\tag{25}
$$

where

$$
\begin{aligned}
&P_1(|\Phi, N) \\
&\quad = \frac{1}{2}\left[1 - \left\langle Q\left(\frac{a_+}{\sigma}, \frac{a_-}{\sigma}\right)\right\rangle\right] \\
&\quad + \frac{1}{2}\left\langle Q\left(\frac{a_-}{\sigma}, \frac{a_+}{\sigma}\right)\right\rangle,
\end{aligned}
$$

and a_+ and a_- contain only the first $2N$ significant terms. When $N \to \infty$, Z in eq. (25) can be seen to approach zero.

Since a_+ and a_- in eqs. (23) and (24) contain a finite number of ISI terms, we can use the direct method to evaluate the averages and then compute the bounds. We choose the initial N so that $\sigma_R^2 < 1$, $\Delta = \sqrt{\sigma_R}$. We then increase N so that the desired accuracy of computation is achieved.

3.1.3 Upper bound on the probability of error

Since the exact evaluation of $P_1(|\Phi)$ is difficult—though we have developed in the last section numerical techniques which can be used to compute P_1 with any desired accuracy—we attempt to derive an upper bound on P_1.

Although our bounding approach seems reasonable, the final bound that we obtain turns out to be loose. Our purpose in including this section is to alert readers about this approach and to emphasize the importance of the tedious, but necessary, computations outlined in Section 3.1.2.

To facilitate our bounding techniques, we need the following relations. For any two random variables x and y and any two real numbers a and Δ, we can show (see Fig. 5) that

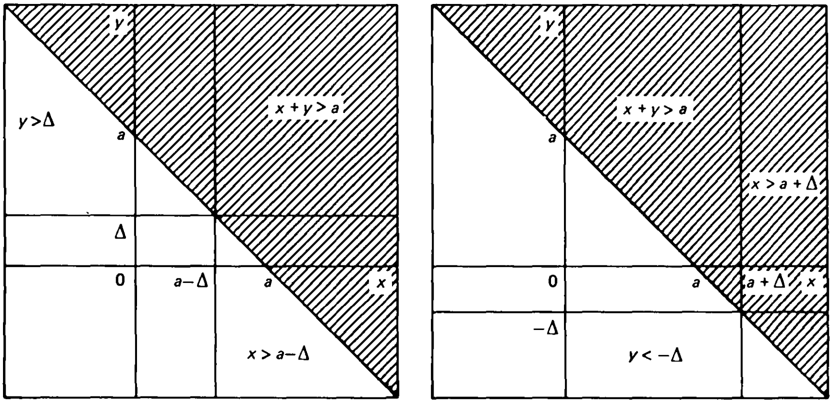
$$
\begin{aligned}
&\Pr(x > a + \Delta) - \Pr(y < -\Delta) \\
&\quad \le \Pr(x + y > a) \\
&\quad \le \Pr(x > a - \Delta) + \Pr(y > \Delta). \qquad (26)
\end{aligned}
$$

Equations (21) and (26) yield

$$
\begin{aligned}
P_1(|\Phi) &= \Pr(|w_2| - |w_1| > 0) \\
&\le \Pr(|w_2| > A) + \Pr(-|w_1| > -A) \\
&= \Pr(|w_2| > A) + \Pr(|w_1| < A), \qquad (27)
\end{aligned}
$$

where A is arbitrary. We choose $A > 0$ so that the upper bound in eq. (27) is a minimum. The method of choosing A will be discussed later.

Now, for any complex random variable $z = x + iy$, we can show (see Fig. 6) that



(a) (b)

Fig. 5—(a) Upper bound on Pr($x + y > a$), x and y, any two random variables and Δ is arbitrary. (b) Lower bound on Pr($x + y > a$), x and y, any two random variables and Δ is arbitrary.

$$\Pr(|z| > a) \\ \leq \Pr(|\mathrm{Re}\, z| > a_1) + \Pr(|\mathrm{Im}\, z| > \sqrt{a^2 - a_1^2}).$$

Hence,

$$\Pr(|w_2| > A) \\ \leq \Pr(|\mathrm{Re}\, w_2| > A_1) + \Pr(|\mathrm{Im}\, w_2| > A_2), \qquad A_1^2 + A_2^2 = A^2, \tag{28}$$

where

$$\mathrm{Re}\, w_2 = \xi_- + \sum_{k=-\infty}^{\infty} C_k \cos(\alpha_k + \lambda_k)$$

$$\mathrm{Im}\, w_2 = \eta_- + \sum_{k=-\infty}^{\infty} C_k \sin(\alpha_k + \lambda_k)$$

$$C_k = \left|\frac{1}{2}\left\{(g_k + ip_k) - (g_{k-1} + ip_{k-1})\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]\right\}\right|$$

and

$$C_k \exp(i\lambda_k) = \frac{1}{2}\left\{(g_k + ip_k) - (g_{k-1} + ip_{k-1})\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]\right\}.$$

Since the real and imaginary parts of w_2 are the sum of a Gaussian random variable and a set of interference terms, various methods given in Refs. 16 to 18 and 19 to 27 can be used to bound Pr($|\mathrm{Re}\, w_2| > A_1$) and Pr($|\mathrm{Im}\, w_2| > A_2$), $A_1^2 + A_2^2 = A^2$. Even though the other bounds are sometimes claimed to be tighter, we shall use the simpler Chernoff bounds.

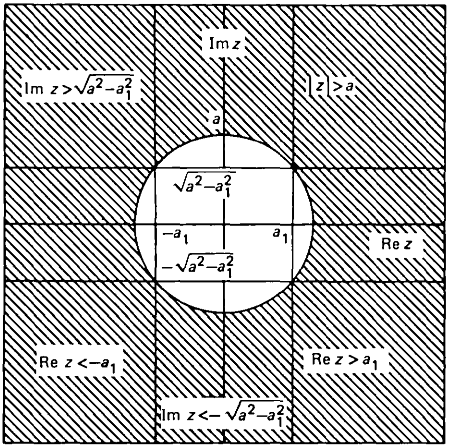


Fig. 6—Upper bound on $\Pr(|z| > a)$. Parameter Δ satisfies $0 \le \Delta \le a$.

Appendix C shows that

$$\Pr(|w_2| > A) \le 2 \exp\left[-\frac{(A_1 - C_{M1})^2}{\sigma^2 + \sigma_-^2}\right] + 2 \exp\left[-\frac{(A_2 - C_{M2})^2}{\sigma^2 + \sigma_-^2}\right],$$

where

$$A_1 = A \cos\left(\frac{\pi}{4} + \frac{\pi}{2M}\right)$$

$$A_2 = A \sin\left(\frac{\pi}{4} + \frac{\pi}{2M}\right)$$

$$C_{M1} = \max\{[C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)]\}$$

$$\alpha_1 - \alpha_0 = \frac{\pi}{M}$$

$$C_{M2} = \max\{[C_0\sin(\alpha_0 + \lambda_0) + C_1\sin(\alpha_1 + \lambda_1)]\}$$

$$\alpha_1 - \alpha_0 = \frac{\pi}{M}$$

$$\sigma_-^2 = \textstyle\sum'' C_k^2$$

and

$$\textstyle\sum'' \triangleq \sum_{\substack{k=-\infty \\ k\neq 0 \\ k\neq -1}}^{\infty} .$$

Also, for any complex random variable $z = x + iy$, we can show (see Fig. 7) that

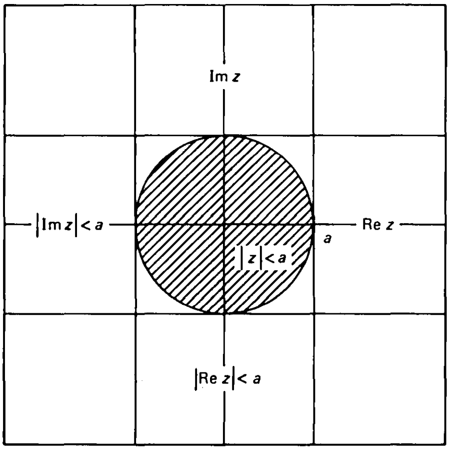


Fig. 7—Upper bound on $\Pr(|z| < a)$.

$$\Pr(|z| < a) \leq \Pr(|\mathrm{Re}\, z| < a, |\mathrm{Im}\, z| < a)$$
$$\Pr(|z| < a) \leq \Pr(|\mathrm{Re}\, z| < a)$$
$$\Pr(|z| < a) \leq \Pr(|\mathrm{Im}\, z| < a).$$

Hence,

$$\Pr(|w_1| < A) \leq \Pr(|\mathrm{Re}\, w_1| < A).$$

Since

$$\mathrm{Re}\, w_1 \leq |\mathrm{Re}\, w_1|,$$
$$\Pr(|\mathrm{Re}\, w_1| < A) \leq \Pr(\mathrm{Re}\, w_1 < A),$$

and

$$\Pr(|w_1| < A) \leq \Pr(\mathrm{Re}\, w_1 < A).$$

We write

$$\mathrm{Re}\, w_1 = \xi_+ + \sum_{k=-\infty}^{\infty} D_k \cos(\alpha_k + \delta_k)$$

$$\mathrm{Im}\, w_1 = \eta_+ + \sum_{k=-\infty}^{\infty} D_k \sin(\alpha_k + \delta_k),$$

$$D_k = \left| \frac{1}{2} \left\{ (g_k + ip_k) + (g_{k-1} + ip_{k-1}) \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2} \right) \right] \right\} \right|,$$

$$D_k \exp(i\delta_k) = \frac{1}{2} \left\{ (g_k + ip_k) + (g_{k-1} + ip_{k-1}) \exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2} \right) \right] \right\},$$

and

$$\sum_{k=-\infty}^{\infty} D_k \cos(\alpha_k + \delta_k)$$

$$= D_0 \cos(\alpha_0 + \delta_0) + D_1 \cos(\alpha_1 + \delta_1) + \sum'' D_k \cos(\alpha_k + \delta_k).$$

Using Chernoff bounding techniques, it can be shown that

$$\Pr(|w_1| < A) \le \exp\left[-\frac{(D_M - A)^2}{\sigma^2 + \sigma_+^2}\right],$$

where

$$D_M = \min[D_0 \cos(\alpha_0 + \delta_0) + D_1 \cos(\alpha_1 + \delta_1)]$$

$$\alpha_1 - \alpha_0 = \frac{\pi}{M}$$

$$\sigma_+^2 = \sum'' D_k^2 \cdot$$

The upper bound on $P_1(|\beta_0 = \pi/M)$ can, therefore, be written as

$$P_1\left(\middle|\beta_0 = \frac{\pi}{M}\right) \le 2\exp\left[-\frac{(A_1 - C_{M1})^2}{\sigma^2 + \sigma_-^2}\right] + 2\exp\left[-\frac{(A_2 - C_{M2})^2}{\sigma^2 + \sigma_-^2}\right] + \exp\left[-\frac{(D_M - A)^2}{\sigma^2 + \sigma_+^2}\right],$$

$$A_1 - C_{m1} \ge 0, A_2 - C_{M2} \ge 0, D_M - A \ge 0, A_1^2 + A_2^2 = A^2. \qquad (29)$$

The bound is minimum when A, A_1, and A_2 are chosen so that the derivative of eq. (29) is zero. This can be found by using well-known numerical methods.

3.2 Example of quaternary (M = 4) DPSK system

Let us consider a quaternary ($M = 4$) DPSK system and assume that the channel is ideal.

If 4-pole Butterworth transmit and receive filters are used, the bound given by eq. (29) is plotted in Fig. 8. The bound with zero ISI is plotted in Fig. 9. The exact probability of error with ISI is plotted in Fig. 10. With or without ISI, the bound is unfortunately not very tight. Actually, one can show that the penalty as predicted by the bound with zero ISI is about 4.6 dB worse than the actual penalty for a binary system, and 8.3 dB worse for a quaternary system. This is inherent in our techniques and not the result of using Chernoff bounding methods. In our opinion, obtaining tighter bounds is still an open problem. Comparing Figs. 2 and 10, we note that ISI penalty for quaternary DPSK is about 1 dB worse than for CPSK. We needed 9 ISI terms to compute Pe with 5 percent accuracy.

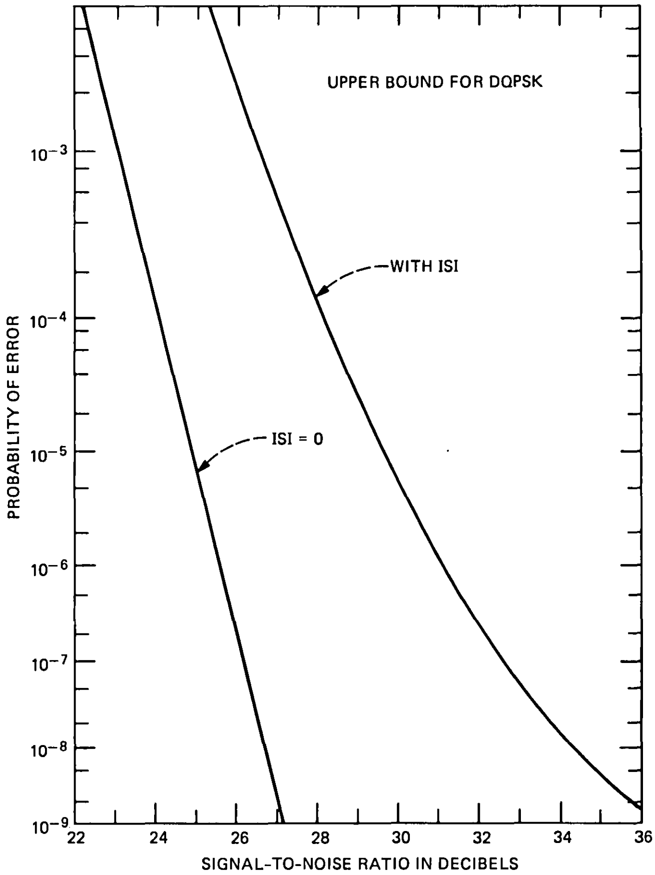


Fig. 8—Upper bound on the probability of error for differential QPSK (DQPSK) with the same transmit and receive filters as in Fig. 2. Other assumptions are as in Fig. 2.

IV. SUMMARY AND CONCLUSIONS

For multiphase M-ary CPSK, we develop an analytical procedure for determining detection efficiency when the system is subject to additive Gaussian noise, ISI, and imperfect carrier-phase estimation. For a large s/n, we provide a simple formula for calculating the combined penalty caused by ISI and noisy phase recovery. For multiphase DPSK, where the detection is inherently nonlinear, a rigorous method is developed for calculating the error rate in the presence of ISI and additive Gaussian noise. Using these analytical techniques, it is possible to compare the performance of CPSK and DPSK and examine various parameter trade-offs. Numerical examples are provided to illustrate our methods.

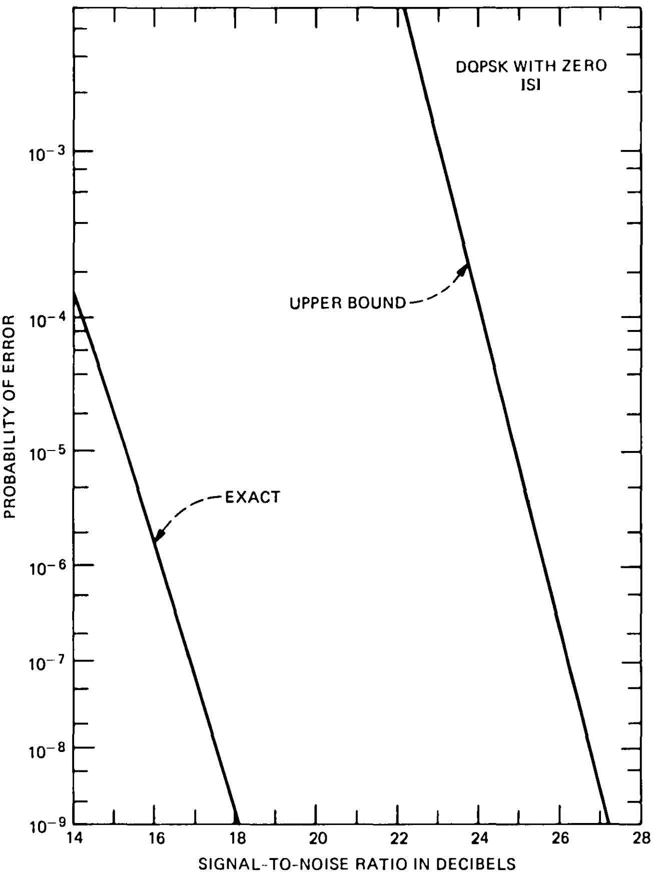


Fig. 9—Upper bound on the probability of error for DQPSK with zero ISI.

APPENDIX A

Chernoff Bound on the Probability of Error

From Section 2.1.3,

$$P_1 = \frac{1}{2}\int_{-\pi}^{\pi} [V(\epsilon) + V(-\epsilon)]p_\epsilon(\epsilon)d\epsilon$$

$$= \int_0^{\pi} [V(\epsilon) + V(-\epsilon)]p_\epsilon(\epsilon)d\epsilon$$

$$= J_1 + J_2,$$

where

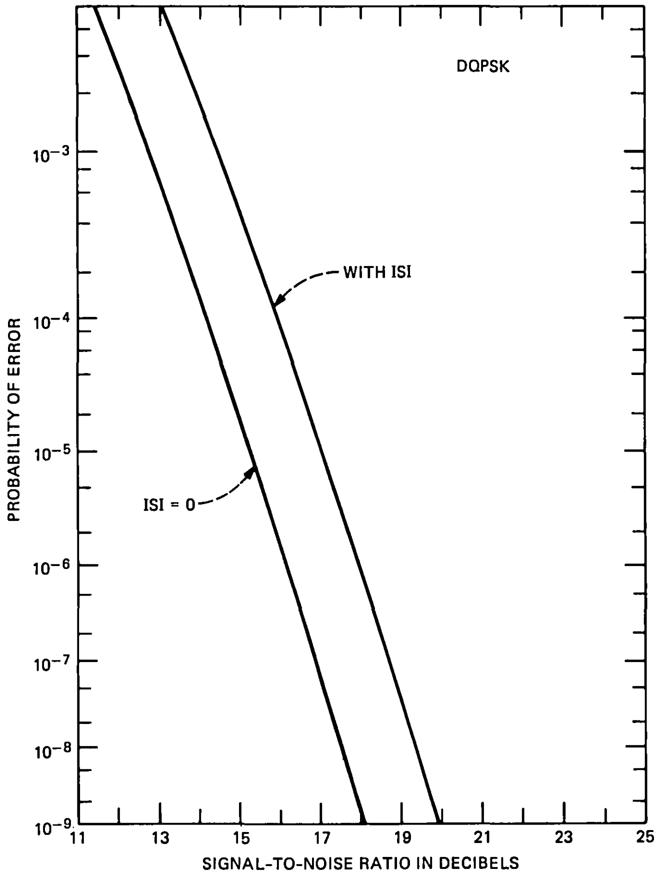


Fig. 10—Probability of error for DQPSK with the same transmit and receive filters as in Fig. 2. Other assumptions are as in Fig. 2. Note that 9 ISI terms were needed to compute Pe with 5 percent accuracy.

$$J_1 = \int_0^{\delta} [V(\epsilon) + V(-\epsilon)]p_\epsilon(\epsilon)d\epsilon$$

$$J_2 = \int_{\delta}^{\pi} [V(\epsilon) + V(-\epsilon)]p_\epsilon(\epsilon)d\epsilon, \tag{30}$$

and where $V(\epsilon)$ is given in eq. (11). Note that $\sin(\pi/M + \beta_0 - \epsilon) > 0$ for $0 \le \epsilon < \pi/M + \beta_0$; also, $\sin(\pi/M + \beta_0 + \epsilon) > 0$ for $0 \le \epsilon < \pi - (\pi/M + \beta_0)$. Hence, $\sin(\pi/M + \beta_0 + \epsilon) > 0$ for $0 \le \epsilon < \delta$, where

$$\delta = \min\left[\frac{\pi}{M} + \beta_0, \pi - \left(\frac{\pi}{M} + \beta_0\right)\right].$$

Since

$$0 \leq \operatorname{erfc}(x) \leq 2,$$

$$0 \leq [V(\epsilon) + V(-\epsilon)] \leq 2,$$

and, therefore,

$$J_2 \leq 2 \int_{\delta}^{\pi} p_{\epsilon}(\epsilon) d\epsilon = \Pr(|\epsilon| > \delta).$$

Now, from eq. (11),

$$V(\epsilon) = \Pr\left[-\eta - I(\epsilon) > r_0 \sin\left(\frac{\pi}{M} + \beta_0 + \epsilon \right) \right], \tag{31}$$

where η is a zero mean Gaussian random variable with variance σ^2 and $I(\epsilon)$ is given in eq. (8).

Using the Chernoff bound

$$\Pr[x > a] \leq \exp(-\mu a) \langle \exp(\mu x) \rangle, \qquad \mu \geq 0,$$

eq. (31) yields

$$V(\epsilon) \leq \exp\left[-\lambda r_0 \sin\left(\frac{\pi}{M} + \beta_0 + \epsilon \right) \right] \exp\{-\lambda[\eta + I(\epsilon)]\}. \tag{32}$$

For a given ϵ, I and η are independent, and since the data phases α_k in different time slots are iid,

$$\exp\{-\lambda[\eta + I(\epsilon)]\} = \exp \frac{\lambda^2 \sigma^2}{2} \prod{}' \langle \exp[-\lambda r_k \sin(\beta_k + \alpha_k + \epsilon)] \rangle_{\alpha_k}. \tag{33}$$

We shall now assume that M is an even number so that if $\Phi \epsilon \Lambda$, $(\pi + \Phi) \epsilon \Lambda$. Hence,

$$\begin{aligned} &\langle \exp[-\lambda r_k \sin(\beta_k + \alpha_k + \epsilon)] \rangle_{\alpha_k}, \qquad 0 \leq \alpha_k < 2\pi \\ &= \frac{1}{2} \langle \exp[-\lambda r_k \sin(\beta_k + \alpha_k + \epsilon)] \\ &+ \exp[-\lambda r_k \sin(\beta_k + \pi + \alpha_k + \epsilon)] \rangle_{\alpha_k}, \qquad 0 \leq \alpha_k < \pi \\ &= \langle \cosh \lambda r_k \sin(\beta_k + \alpha_k + \epsilon) \rangle_{\alpha_k}, \qquad 0 \leq \alpha_k < \pi. \end{aligned}$$

Since

$$\begin{aligned} &\cosh x \leq \exp(x^2/2), \\ &\langle \exp[-\lambda r_k \sin(\beta_k + \alpha_k + \epsilon)] \rangle_{\alpha_k}, \qquad 0 \leq \alpha_k < 2\pi \\ &\leq \langle \exp\left[\frac{\lambda^2 r_k^2}{2} \sin^2(\beta_k + \alpha_k + \epsilon) \right] \rangle_{\alpha_k}, \qquad 0 \leq \alpha_k < \pi \\ &\leq \exp\left(\frac{\lambda^2 r_k^2}{2} \right). \end{aligned} \tag{34}$$

From eqs. (32), (33), and (34),

$$V(\epsilon) \leq \exp\left[-\lambda r_0 \sin\left(\frac{\pi}{M} + \beta_0 + \epsilon\right)\right] \exp\left[\frac{\lambda^2}{2}(\sigma^2 + \sigma_I^2)\right], \qquad \lambda \geq 0,$$

where

$$\sigma_I^2 = \sum{}' r_k^2.$$

Similarly,

$$V(-\epsilon) \leq \exp\left[-\lambda r_0 \sin\left(\frac{\pi}{M} + \beta_0 - \epsilon\right)\right] \exp\left[\frac{\lambda^2}{2}(\sigma^2 + \sigma_I^2)\right].$$

Hence, for $0 \leq \epsilon < \pi/M + \beta_0$, $\sin(\pi/M + \beta_0 - \epsilon) > 0$, and

$$V(-\epsilon) \leq \exp\left[-\frac{r_0^2 \sin^2[(\pi/M) + \beta_0 - \epsilon]}{2(\sigma^2 + \sigma_I^2)}\right].$$

Also, for $0 \leq \epsilon < \pi - (\pi/M + \beta_0)$, $\sin(\pi/M + \beta_0 + \epsilon) > 0$, and

$$V(\epsilon) \leq \exp\left[-\frac{r_0^2 \sin^2[(\pi/M) + \beta_0 + \epsilon]}{2(\sigma^2 + \sigma_I^2)}\right].$$

Hence,

$$\begin{aligned} V(\epsilon) + V(-\epsilon) &\leq \exp\left[-\frac{r_0^2 \sin^2[(\pi/M) + \beta_0 + \epsilon]}{2(\sigma^2 + \sigma_I^2)}\right] \\ &\quad + \exp\left[-\frac{r_0^2 \sin^2[(\pi/M) + \beta_0 - \epsilon]}{2(\sigma^2 + \sigma_I^2)}\right], \\ &= \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0 + \epsilon\right)\right] \\ &\quad + \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0 - \epsilon\right)\right], \\ \rho^2 &= \frac{r_0^2}{2(\sigma^2 + \sigma_I^2)}, \qquad \sigma_I^2 = \sum{}' r_k^2, \\ 0 &\leq \epsilon < \delta = \min\left[\frac{\pi}{M} + \beta_0, \pi - \left(\frac{\pi}{M} + \beta_0\right)\right]. \end{aligned} \tag{35}$$

The parameter ρ^2 is the s/n of the system.

Note that β_0 is the phase angle of the complex overall impulse response evaluated at $t = t_0$. In a well-designed system, it is usually small, and eq. (35) shows that the optimum value of β_0 is zero. Also, since β_0 is usually small and we are interested in $M > 2$, δ is usually $\pi/M + \beta_0$.

Thus, from eqs. (9), (30), and (35), we conclude that

$$J_1 \leq J_{1a} = \frac{1}{2\pi I_0(D\rho^2)} \int_0^{\delta} \left\{ \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0 - \epsilon \right) \right] \right.$$
$$\left. + \exp\left[-\rho^2 \sin^2\left(\frac{\pi}{M} + \beta_0 + \epsilon \right) \right] \right\} \exp[D\rho^2 \cos\epsilon] d\epsilon, \qquad (36)$$

where the quantity D can be regarded as the ratio of s/n in the phase recovery circuit to that in the PSK system or the integration time in bauds.

We have not been able to evaluate eq. (31) in closed form but, if desired, numerical techniques can be used. To obtain physical insight, we shall assume that $\rho^2 \gg 1$ and use Laplace's method to evaluate eq. (36); the technique is an application of the following theorem: *If $h(t)$ is a real function of a real variable t, has a unique maximum at $t = a$, $\alpha_1 \leq a \leq \alpha_2$, and if x is a large positive variable, it can be shown that*[11]

$$f(x) = \int_{\alpha_1}^{\alpha_2} g(t) \exp[xh(t)] dt \approx g(a) \exp[xh(a)] \left(\frac{-\pi}{2xh''(a)} \right)^{1/2}.$$

From eq. (36),

$$J_{1a} = \frac{1}{2\pi I_0(D\rho^2)} \{J_b + J_c\}, \qquad (37)$$

$$J_b = \int_0^{\delta} \exp\left\{ \rho^2 \left[D\cos\epsilon - \sin^2\left(\epsilon - \frac{\pi}{M} - \beta_0 \right) \right] \right\} d\epsilon, \qquad (38)$$

$$J_c = \int_0^{\delta} \exp\left\{ \rho^2 \left[D\cos\epsilon - \sin^2\left(\epsilon + \frac{\pi}{M} + \beta_0 \right) \right] \right\} d\epsilon. \qquad (39)$$

The saddle point ϵ_0 at which the exponent in eq. (38) reaches its maximum in $(0, \delta)$ is given by the solution of

$$\frac{\sin\epsilon_0}{\sin 2[(\pi/M) + \beta_0 - \epsilon_0]} = \frac{1}{D}. \qquad (40)$$

The transcendental eq. (39) can only be solved numerically. However, we can obtain a series solution for ϵ_0 by using Lagrange's reversion formula.[27] If a function $f(z)$ is regular in a neighborhood of z_0 and if $f(z_0) = w_0$, $f'(z_0) \neq 0$, then it can be shown[28] that

$$f(z) = w$$

has a unique solution

$$z = z_0 + \sum_{n=1}^{\infty} \frac{(w - w_0)^2}{n!} \left\{ \frac{d^{n-1}}{dz^{n-1}} [\Phi(z)]^n \right\}_{z=z_0} , \tag{41}$$

where

$$\Phi(z) = \frac{z - z_0}{f(z) - w_0}.$$

Choosing $z_0 = 0$, eqs. (40) and (41) yield

$$\begin{aligned} \epsilon_0 = \frac{1}{D} \sin 2\left(\frac{\pi}{M} + \beta_0\right) & \\ \times\left[1 - \frac{2}{D} \cos 2\left(\frac{\pi}{M} + \beta_0\right) + \cdots\right], &\qquad D \gg 1, \qquad 0 \leq \epsilon_0 < \delta. \end{aligned}$$

From Laplace's formula and eq. (38)

$$\begin{aligned} J_b \approx \exp\left\{-\rho^2\left[\sin^2\left(\frac{\pi}{M} + \beta_0 - \epsilon_0\right) - D \cos \epsilon\right]\right\} & \\ \times\left\{\frac{\pi}{2\rho^2[D \cos \epsilon_0 + 2 \cos 2(\pi/M + \beta_0 - \epsilon_0)]}\right\}^{1/2} &. \end{aligned} \tag{42}$$

Similarly, it can be shown that the exponent in eq. (40) reaches its maximum in $(0, \delta)$ at $\epsilon = 0$ and

$$\begin{aligned} J_c \approx \exp\left\{-\rho^2\left[\sin^2\left(\frac{\pi}{M} + \beta_0\right) - D\right]\right\} & \\ \times\left[\frac{\pi}{2\rho^2[D + 2 \cos(\pi/M + \beta_0)]}\right]^{1/2} &. \end{aligned} \tag{43}$$

For $\rho^2 \gg 1$, $D > 1$,

$$I_0(D\rho^2) \approx \frac{\exp(D\rho^2)}{\sqrt{2\pi D\rho^2}}. \tag{44}$$

From eqs. (37) to (39) and (42) to (44)

$$\begin{aligned} J_{1a} \approx \frac{1}{2}\Bigg(\frac{\exp\{-\rho^2[\sin^2[(\pi/M) + \beta_0 - \epsilon_0] - D(1 - \cos \epsilon_0)]\}}{[\cos \epsilon_0 + (2/D)\cos 2[(\pi/M) + \beta_0 - \epsilon_0]]^{1/2}} & \\ + \frac{\exp[-\rho^2 \sin^2[(\pi/M) + \beta_0)]}{[1 + (2/D)\cos 2[(\pi/M) + \beta_0]]^{1/2}}\Bigg) &. \end{aligned}$$

Also,

$$
\begin{aligned}
J_2 \le J_{2a} &= 2\int_{\delta}^{\pi} \rho_{\epsilon}(\epsilon)dt \\
&= \frac{1}{\pi I_0(D\rho^2)}\int_{\delta}^{\pi} \exp(D\rho^2\cos \epsilon d\epsilon) \\
&\le \frac{\pi-\delta}{\pi}\frac{\exp(D\rho^2\cos\delta)}{I_0(D\rho^2)} \\
&\approx \left(1-\frac{\delta}{\pi}\right)\sqrt{2\pi D\rho^2}\,\exp[-D\rho^2(1-\cos\delta)] \\
&\qquad \rho^2 \gg 1, \qquad D > 1.
\end{aligned}
$$

APPENDIX B

Upper and Lower Bounds on the Probability of Error in DPSK

Let us write

$$
\begin{aligned}
z &= z_N + n_c + in_s + z_R \\
z_d &= z_{dN} + n_{c_d} + in_{s_d} + z_{dR},
\end{aligned}
$$

where

$$
\begin{aligned}
z_N &= \sum_{\substack{k\ge -N\\ k\le N}} (g_k + ip_k)\exp(i\alpha_k) \\
z_R &= \sum_{\substack{k< -N\\ k> N}} (g_k + ip_k)\exp(i\alpha_k) \\
z_{dN} &= \sum_{\substack{k\ge -N\\ k\le N}} (g_{k-1} + ip_{k-1})\exp(i\alpha_k)
\end{aligned}
$$

and

$$
z_{dR} = \sum_{\substack{k< -N\\ k> N}} (g_{k-1} + ip_{k-1})\exp(i\alpha_k).
$$

Note that z_N and z_{dN} contain a finite number of ISI terms, whereas z_R and z_{dR} contain an infinite number. Without loss of generality, we shall also assume that g_k and p_k are monotonic decreasing functions of $|k|$. N is an arbitrary positive number.

Now

$$
P_1(|\Phi) = \Pr(|w_{1N} + w_{1R}| < |w_{2N} + w_{2R}|), \tag{45}
$$

where

$$w_{1N} = \frac{z_N + z_{dN}\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]}{2} + \xi_+ + i\eta_+,$$

$$w_{1R} = \frac{z_R + z_{dR}\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]}{2},$$

$$w_{2N} = \frac{z_N - z_{dN}\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]}{2} + \xi_- + i\eta_-$$

and

$$w_{2R} = \frac{z_R - z_{dR}\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]}{2}.$$

Note that w_{1N} and w_{2N} contain a finite number of ISI terms, and $\xi_+ + i\eta_+$ and $\xi_- + i\eta_-$ are independent zero means complex Gaussian processes.

For any two complex numbers s_1 and s_2,

$$|s_1| - |s_2| \le |s_1 + s_2| \le |s_1| + |s_2|. \tag{46}$$

Hence, eqs. (45) and (46) yield

$$\begin{aligned}\Pr(|w_{1N}| < |w_{2N}| - |w_{1R}| - |w_{2R}|) &\le P_1(|\Phi)\\ &= \Pr(|w_{1N} + w_{1R}| < |w_{2N} + w_{2R}|)\\ &\le \Pr(|w_{1N}| < |w_{2N}| + |w_{1R}| + |w_{2R}|)\end{aligned}$$

or

$$\begin{aligned}\Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 - \frac{|w_{1R}| + |w_{2R}|}{|w_{2N}|}\right) &\le P_1(|\Phi)\\ &\le \Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 + \frac{|w_{1R}| + |w_{2R}|}{|w_{2N}|}\right).\end{aligned}$$

Using the bound in eq. (26), eq. (46) yields

$$\begin{aligned}\Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 - \Delta\right) - \Pr\left(\frac{|w_{1R}| + |w_{2R}|}{|w_{2N}|} > \Delta\right) &\le P_1(|\Phi)\\ \le \Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 + \Delta\right) &+ \Pr\left(\frac{|w_{1R}| + |w_{2R}|}{|w_{2N}|} > \Delta\right),\end{aligned} \tag{47}$$

where we choose $0 \le \Delta < 1$.

For any two random variables x and y and any $a > 0$,[10]

$$\Pr(xy > a) \le \Pr\left(|x| > \frac{a}{p}\right) + \Pr(|y| > p), \tag{48}$$

where $p > 0$ is arbitrary. From eqs. (47) and (48),

$$\begin{aligned}\Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 - \Delta\right) &- \Pr(|w_{2N}| < p) \\ &- \Pr(|w_{1R}| + |w_{2R}| > \Delta p) \le P_1(|\Phi) \\ &\le \Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 + \Delta\right) + \Pr(|w_{2N}| < p) \\ &+ \Pr(|w_{1R}| + |w_{2R}| > \Delta p). \end{aligned} \tag{49}$$

Since w_{1N} and w_{2N} are independent complex Gaussian processes, from Ref. 13,

$$\begin{aligned}\Pr\left(\frac{|w_{1N}|}{|w_{2N}|} < 1 \pm \Delta\right) &= \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2} \\ &\times\left\{1 - <Q\left[\frac{a_+}{\sqrt{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2}}, \frac{a_-(1 \pm \Delta)}{\sqrt{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2}}\right]>\right\}, \\ &+ \frac{\sigma_2^2(1 \pm \Delta)^2}{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2} \\ &\times <Q\left[\frac{a_-(1 \pm \Delta)}{\sqrt{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2}}, \frac{a_+}{\sqrt{\sigma_1^2 + \sigma_2^2(1 \pm \Delta)^2}}\right]>, \quad \sigma_1^2 = \sigma_2^2 = \frac{\sigma^2}{2}, \end{aligned} \tag{50}$$

where a_+ and a_- are the appropriate truncated values of a_+ and a_- defined in Section 3.1.2. Also, from Ref. 13,

$$\Pr(|w_{2N}| < p) = 1 - <Q\left(\frac{a_-}{\sigma_2}, \frac{p}{\sigma_2}\right)>. \tag{51}$$

Using Chernoff bounding techniques, we can also show that

$$\Pr(|w_{1R}| > \Delta p) \le 4 \exp\left(-\frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} G_k^2}\right) \tag{52}$$

$$\Pr(|w_{2R}| > \Delta p) \le 4 \exp\left(-\frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} H_k^2}\right), \tag{53}$$

where

$$G_k = \left|\frac{1}{2}\left\{(g_k + ip_k) + (g_{k-1} + ip_{k-1})\exp\left[-i\left(\Phi - \frac{\pi}{M} + \frac{\pi}{2}\right)\right]\right\}\right|,$$

and

$$H_k = \left| \frac{1}{2} \left\{ (g_k + ip_k) - (g_{k-1} + ip_{k-1}) \exp\left[-i \left(\Phi - \frac{\pi}{M} + \frac{\pi}{2} \right) \right] \right\} \right|.$$

Equations (49) to (53) yield

$$\chi_1(N) \le P_1(|\Phi) \le \chi_2(N),$$

where

$$\begin{aligned} \chi_1(N) = {} & \frac{1}{1 + (1 - \Delta)^2} \\ & \times \left\{ 1 - < Q \left[\frac{\sqrt{2}a_+}{\sigma\sqrt{1 + (1 - \Delta)^2}}, \frac{\sqrt{2}a_-(1 - \Delta)}{\sigma\sqrt{1 + (1 - \Delta)^2}} \right] > \right\} \\ & + \frac{(1 - \Delta)^2}{1 + (1 - \Delta)^2} < Q \left[\frac{\sqrt{2}a_-(1 - \Delta)}{\sigma\sqrt{1 + (1 - \Delta)^2}}, \frac{\sqrt{2}a_+}{\sigma\sqrt{1 + (1 - \Delta)^2}} \right] > \\ & - \left[1 - < Q \left(\frac{\sqrt{2}a_-}{\sigma}, \frac{\sqrt{2}p}{\sigma} \right) > \right] \\ & - 4 \exp\left(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} G_k^2} \right) - 4 \exp\left(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} H_k^2} \right) \end{aligned}$$

and

$$\begin{aligned} \chi_2(N) = {} & \frac{1}{1 + (1 + \Delta)^2} \\ & \times \left\{ 1 - < Q \left[\frac{\sqrt{2}a_+}{\sigma\sqrt{1 + (1 + \Delta)^2}}, \frac{\sqrt{2}a_-(1 + \Delta)}{\sigma\sqrt{1 + (1 + \Delta)^2}} \right] > \right\} \\ & + \frac{(1 + \Delta)^2}{1 + (1 + \Delta)^2} < Q \left[\frac{\sqrt{2}a_-(1 + \Delta)}{\sigma\sqrt{1 + (1 + \Delta)^2}}, \frac{\sqrt{2}a_+}{\sigma\sqrt{1 + (1 + \Delta)^2}} \right] > \\ & + \left\{ 1 - < Q \left(\frac{\sqrt{2}a_-}{\sigma}, \frac{\sqrt{2}p}{\sigma} \right) > \right\} \\ & + 4 \exp\left(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} G_k^2} \right) + 4 \exp\left(- \frac{\Delta^2 p^2}{2 \sum\limits_{\substack{k<-N \\ k>N}} H_k^2} \right), \end{aligned}$$

where Δ and p are arbitrary.

APPENDIX C

Upper Bound on the Probability of Error

Now,

$$\Pr(|\mathrm{Re}\, w_2| > A_1) = \Pr(\mathrm{Re}\, w_2 > A_1) + \Pr(-\mathrm{Re}\, w_2 > A_1)$$

and

$$\Pr(\mathrm{Re}\, w_2 \geq A_1) \leq \exp(-\mu A_1)\langle \exp(\mu\, \mathrm{Re}\, w_2)\rangle$$
$$= \exp(-\mu A_1)\langle \exp(\mu\, Re\, w_2)\rangle.$$

Since the Gaussian random variable ξ_- is assumed to be independent of the interference

$$\langle \exp(\mu\, \mathrm{Re}\, w_2)\rangle = \exp(\mu^2\sigma^2/4) \quad \exp\left[\mu \sum_{k=-\infty}^{\infty} C_k \cos(\alpha_k + \lambda_k)\right].$$

Since $\beta_k = \alpha_{k+1} - \alpha_k$, and we assume that β_k's are iid, we can assume that α_k's are independent, and

$$\alpha_k \in \Lambda, \qquad k \text{ even}$$
$$\alpha_k \in \left[0, \frac{2\pi}{M}, \frac{4\pi}{M}, \cdots, (2M-2)\frac{\pi}{M}\right] \triangleq \Lambda_s, \qquad k \text{ odd};$$

that is, the signal constellation for odd (even) k can be obtained by a simple rotation of the constellation for even (odd) k.

Let us now assume that the transmitted symbol Φ is π/M so that

$$P_1(|\Phi) = P_1\left(|\beta_0 = \frac{\pi}{M}\right)$$
$$= \frac{1}{M}\sum P_1\left(|\alpha_1 - \alpha_0 = \frac{\pi}{M}\right).$$

Noting that

$$\sum_{k=-\infty}^{\infty} C_k \cos(\alpha_k + \lambda_k) = C_0 \cos(\alpha_0 + \lambda_0) + C_1 \cos(\alpha_1 + \lambda_1)$$
$$+ \sum'' C_k \cos(\alpha_k + \lambda_k),$$

$$\Pr\left(\mathrm{Re}\, w_2 > A_1 | \alpha_1 - \alpha_0 = \frac{\pi}{M}\right) \leq \exp\left\{-\mu A_1 + \frac{\mu^2\sigma^2}{4}\right.$$
$$\left. + \mu[C_0 \cos(\alpha_0 + \lambda_0) + C_1 \cos(\alpha_1 + \lambda_1)]\right\}$$
$$\times \langle \exp[\mu \sum'' C_k \cos(\alpha_k + \lambda_k)]\rangle.$$

We use the notation $''$ to indicate that $k = 0$, and $k = 1$ terms are not included.

Now,

$$\langle \exp[\mu \textstyle\sum'' C_k \cos(\alpha_k + \lambda_k)] \rangle = \prod'' \langle \exp[\mu C_k \cos(\alpha_k + \lambda_k)] \rangle$$
$$= \prod_{\text{even}}'' \langle \exp[\mu C_k \cos(\alpha_k + \lambda_k)] \rangle$$
$$\times \prod_{\text{odd}}'' \langle \exp[\mu C_k \cos(\alpha_k + \lambda_k)] \rangle .$$

Most often, $M = 2^L$, L an integer, and since this assumption simplifies our bound, we shall assume that M is an integer power of two. (A slightly more complex bound can be derived if $M \neq 2^L$.) Now,

$$\prod_{\text{even}}'' \langle \exp[\mu C_k \cos(\alpha_k + \lambda_k)] \rangle_{(\alpha_k | -\pi < \alpha_k \leq \pi)}$$
$$= \prod_{\text{even}}'' \langle \cosh[\mu C_k \cos(\alpha_k + \lambda_k)] \rangle_{(\alpha_k | 0 \leq \alpha_k < \pi)}$$
$$= \prod_{\text{even}}'' \frac{1}{2} < \cosh[\mu C_k \cos(\alpha_k + \lambda_k)]$$
$$+ \cosh\left[\mu C_k \cos\left(\frac{\pi}{2} + \alpha_k + \lambda_k\right)\right] >_{\left(\alpha_k | 0 < \alpha_k \leq \frac{\pi}{2}\right)}$$
$$= \prod_{\text{even}}'' < \cosh\left[\mu C_k \cos\frac{\pi}{4} \cos\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right)\right]$$
$$\times \cosh\left[\mu C_k \sin\frac{\pi}{4} \sin\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right)\right] >_{\left(\alpha_k | 0 < \alpha_k \leq \frac{\pi}{2}\right)} .$$

Since

$$\cosh x \leq \exp(|x|),$$

and

$$\cosh x \leq \exp(x^2/2),$$

$$< \cosh\left[\mu C_k \cos\frac{\pi}{4} \cos\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right]\right.$$
$$\times \cosh\left[\mu C_k \cos\frac{\pi}{4} \sin\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right)\right] >_{\left(\alpha_k | 0 < \alpha_k \leq \frac{\pi}{2}\right)} \leq \exp(\mu C_k)$$

and

$$< \cosh\left[\mu C_k \cos\frac{\pi}{4} \cos\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right)\right]$$
$$\times \cosh\left[\mu C_k \cos\frac{\pi}{4} \sin\left(\alpha_k + \lambda_k + \frac{\pi}{4}\right)\right] >_{(\alpha_k | 0 \leq \alpha_k \leq \pi/2)} \leq \exp\left(\frac{\mu^2 C_k^2}{4}\right).$$

Identical bounds can be derived when k is odd. Therefore, we have

$$\langle \exp[\mu \textstyle\sum'' C_k \cos(\alpha_k + \lambda_k)] \rangle \le \exp\left(\mu \displaystyle\sum_{k\in\Omega_1}'' C_k + \frac{\mu^2}{4} \sum_{k\in\Omega_1^c}'' C_k^2\right),$$

where Ω_1 is a subset of $[\cdots, -2, -1, 2, 3, \cdots]$. For simplicity, we choose Ω_1 to be the null set.

Choosing optimum μ,

$$\Pr\left(\mathrm{Re}\, w_2 > A_1 \,|\, \alpha_1 - \alpha_0 = \frac{\pi}{M}\right)$$

$$\le \exp\left[-\frac{\{A_1 - [C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)]\}^2}{\{\sigma^2 + \sum'' C_k^2\}}\right],$$

$$A_1 - [C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)] > 0.$$

Similarly, we can show that

$$\Pr\left(-\mathrm{Re}\, w_2 > A_1 \,|\, \alpha_1 - \alpha_0 = \frac{\pi}{M}\right)$$

$$\le \exp\left(-\frac{\{A_1 - [C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)]\}^2}{(\sigma^2 + \sum'' C_k^2)}\right),$$

$$A_1 + [C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)] \ge 0,$$

$$\Pr\left(|\mathrm{Im}\, w_1| > A_2 \,|\, \alpha_1 - \alpha_0 = \frac{\pi}{M}\right)$$

$$\le \exp\left(-\frac{\{A_2 - [C_0\sin(\alpha_0 + \lambda_0) + C_1\sin(\alpha_1 + \lambda_1)]\}^2}{(\sigma^2 + \sum'' C_k^2)}\right)$$

$$+ \exp\left(-\frac{\{A_2 + [C_0\sin(\alpha_0 + \lambda_0) + C_1\sin(\alpha_1 + \lambda_1)]\}^2}{(\sigma^2 + \sum'' C_k^2)}\right),$$

$$A_2 \pm [C_0\sin(\alpha_0 + \lambda_0) + C_1\sin(\alpha_1 + \lambda_1)] \ge 0, \qquad A_1^2 + A_2^2 = A^2,$$

and

$$\Pr(|w_2| > A) \le 2\exp\left(-\frac{(A_1 - C_{M1})^2}{\sigma^2 + \sigma_-^2}\right) + 2\exp\left(-\frac{(A_2 - C_{M2})^2}{\sigma^2 + \sigma_-^2}\right),$$

where

$$C_{M1} = \max\{[C_0\cos(\alpha_0 + \lambda_0) + C_1\cos(\alpha_1 + \lambda_1)]\}$$

$$\alpha_1 - \alpha_0 = \frac{\pi}{M},$$

$$C_{M2} = \max\{[C_0\sin(\alpha_0 + \lambda_0) + C_1\sin(\alpha_1 + \lambda_1)]\}$$

$$\alpha_1 - \alpha_0 = \frac{\pi}{M},$$

$$\sigma_-^2 = \sum{}'' C_k^2.$$

For zero ISI, it can be shown that optimum values of A_1 and A_2 are

$$A_2 = A\cos\left(\frac{\pi}{4} + \frac{\pi}{2M}\right),$$

$$A_2 = A\sin\left(\frac{\pi}{4} + \frac{\pi}{2M}\right).$$

Even when there is ISI, we shall use these values of A_1 and A_2.

REFERENCES

1. P. Stavroulakis, *Interference Analysis of Communication Systems*, New York: IEEE Press, 1980.
2. W. C. Lindsey, "Phase-Shift-Keyed Signal Detection with Noisy Reference Signals," IEEE Trans. Aerosp. and Electron. Syst., *AES-2* (July 1966), pp. 393–401.
3. S. A. Rhodes, "Effect of Noisy Phase Reference on Coherent Detection of Offset-QPSK Signals," IEEE Trans. Commun., *COM-22* (August 1974), pp. 1046–54.
4. V. K. Prabhu, "Effect of Imperfect Carrier Phase Recovery on the Performance of PSK Systems," IEEE Trans. Aerosp. and Electron. Syst., *AES-12* (March 1976), pp. 275–85.
5. J. M. Aein, "Coherency for the Binary Symmetric Channel," IEEE Trans. Commun. (August 1970), pp. 344–52.
6. K. Shibata, "Error Rate of CPSK Signals in the Presence of Coherent Carrier Phase Jitter and Additive Gaussian Noise," Trans. IECE (Japan), *58-A* (June 1975), pp. 388–95.
7. S. Kabasawa, N. Morinaga, and T. Namekawa, "Effect of Phase Jitter and Gaussian Noise on *M*-ary CPSK Signals," Electronics and Communications in Japan, *61-B* (January 1978), pp. 68–75.
8. O. Shimbo, R. J. Fang, and M. I. Celebiler, "Performance of *M*-ary PSK Systems in Gaussian Noise and Intersymbol Interference," IEEE Trans. Information Theory, *IT-9* (January 1973), pp. 44–58.
9. V. K. Prabhu, "Error Probability Performance of *M*-ary CPSK Systems with Intersymbol Interference," IEEE Trans. Commun., *COM-21* (February 1973), pp. 97–109.
10. V. K. Prabhu, "Imperfect Carrier Recovery Effect on Filtered PSK Signals," IEEE Trans. Aerosp. and Electron. Syst., *AES-14* (July 1978), pp. 608–15.
11. A. Erdelyi, *Asymptotic Expansions*, New York: Dover Publications, 1956, pp. 36–9.
12. V. K. Prabhu, "Performance of Coherent Phase-Shift Keyed Systems with Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-17* (July 1971), pp. 418–31.
13. S. Stein, "Unified Analysis of Certain Coherent and Noncoherent Binary Communication Systems," IEEE Trans. Inform. Theory, *IT-10* (January 1964), pp. 43–51.
14. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.
15. O. Shimbo, M. I. Celebiler, and R. Fang, "Performance Analysis of DPSK Systems in Both Thermal Noise and Intersymbol Interference," IEEE Trans. Commun., *COM-19* (December 1971), pp. 1179–88.
16. F. E. Glave, "An Upper Bound on the Probability of Error Due to Intersymbol

Interference for Correlated Digital Signals," IEEE Trans. Inform. Theory, *IT-8* (May 1972), pp. 356–63.

17. J. W. Mathews, "Sharp Error Bounds for Intersymbol Interference," IEEE Trans. Information Theory, *IT-19* (July 1973), pp. 440–7.
18. K. Yao and R. M. Tobin, "Moment Space Upper and Lower Bounds for Digital Systems with Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-22* (January 1976), pp. 65–74.
19. V. K. Prabhu, "Some Considerations of Error Bounds in Digital Systems," B.S.T.J., *50,* No. 10 (December 1971), pp. 3127–51.
20. B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application to Ideal Band-Limited Signaling," IEEE Trans. Inform. Theory, *IT-14* (July 1968), pp. 563–8.
21. E. Y. Ho and Y. S. Yeh, "A New Approach for Evaluating the Error Probability in the Presence of Intersymbol Interference and Additive Gaussian Noise," B.S.T.J., *49,* No. 9 (November 1970), pp. 2249–65.
22. Y. S. Yeh and E. Y. Ho, "Improved Intersymbol Interference Error Bounds in Digital Systems," B.S.T.J., *50,* No. 8 (October 1971), pp. 2585–98.
23. O. Shimbo and R. Fang, "The Probability of Error Due to Intersymbol Interference and Gaussian Noise in Digital Communication Systems," IEEE Trans. Commun., *COM-19* (April 1971), pp. 113–9.
24. R. Lugannani, "Intersymbol Interference Error Bounds with Applications to Ideal Bandlimited Signaling," IEEE Trans. Inform. Theory, *IT-15* (November 1969), pp. 682–8.
25. O. C. Yue, "Saddle Point Approximation for the Error Probability in PAM Systems with Intersymbol Interference," IEEE Trans. Commun., *COM-27* (October 1979) pp. 1604–9.
26. K. Yao and E. M. Biglieri, "Multidimensional Error Bounds for Digital Communication Systems," IEEE Trans. Inform. Theory, *IT-26* (July 1980), pp. 454–64.
27. V. K. Prabhu, unpublished work.
28. E. T. Copson, "Functions of a Complex Variable," London: Oxford University Press, 1960, pp. 123–5.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

On the Rapid Initial Convergence of Least-Squares Equalizer Adjustment Algorithms

By M. S. MUELLER

(Manuscript received May 11, 1981)

Adjustment algorithms for transversal equalizers derived from least-squares cost functions are known to converge extremely fast. While various simulation results confirming this fact abound in the literature, a theory explaining the fast convergence has been lacking. This paper reports on steps toward such a theory. For some commonly used start-up data sequences it was found that algebraic properties of the sampled signal vectors play a critical role in the transient behavior of these algorithms; namely, successive signal vectors are linearly independent for a large class of transmission channels in the absence of noise. After N iterations (N being the number of taps), the resulting coefficient vector is found to be related to well-known equalizer coefficient vectors. If a single pulse is used as a training signal, the zero forcing equalizer is obtained; if a pseudo random noise sequence, with a period in symbols equal to the number of coefficients is used, the steady-state solution of the cyclic equalization is obtained. Thus, after only N iterations, the least-squares algorithms yield a coefficient vector which is only asymptotically obtainable by gradient techniques.

I. INTRODUCTION

Adaptive equalizers are important building blocks in modems for digital data transmission over linear dispersive channels. They adaptively mitigate the adverse effects of intersymbol interference. A critical parameter in the start-up performance of modems is the speed of convergence of the equalizer adjustment algorithm. The overall data throughput depends on it and, consequently, a high convergence speed is desirable.

Various different equalizer structures are known at this time. In the following, we concentrate on the frequently used transversal filter structure.[1] Many equalizer update algorithms are based on the steepest descent, or gradient technique, which minimizes the mean-squared error (mse) between the equalizer output and the transmitted data symbols.[2] In particular, the stochastic approximation of the gradient algorithm with an mse criterion is used frequently. The convergence speed of this algorithm was analyzed in Refs. 3 and 4. It was found to be dependent on the number of coefficients used and, to a lesser degree, on the eigenvalue spread of the channel autocorrelation matrix.

Several methods to improve the convergence speed of the gradient algorithm were published in Refs. 5 to 8. In Ref. 5, prior knowledge of the transmission channel is assumed and a transformation of the received signal is proposed which reduces the effect of a large eigenvalue spread on the convergence speed, whereas in Ref. 6 a transformation of the correction vectors, yielding the same performance, is proposed. Used in conjunction with a stochastic gradient algorithm, these methods reduce the convergence time to the minimal time which is obtainable with an ideal channel having an eigenvalue spread of one. In Refs. 7 and 8 cyclic equalization was proposed as a means to speed up the convergence time. The number of iterations required for convergence of the algorithm is about the same as for the stochastic gradient algorithm, but theoretically, the convergence time can be reduced to the time required to fill the register of the equalizer. Practically, however, the convergence speed is limited by the available computational power of the implemented algorithm.

In Ref. 9, Godard cast the equalizer adjustment problem as an estimation of a stationary state vector in Gaussian noise—a classical Kalman filtering problem. This resulted in a new, powerful, and rapidly converging equalizer adjustment algorithm. While this algorithm was familiar in the area of stochastic approximation theory,[10] it was never applied to equalizers prior to Godard's work. It was shown by computer simulations,[9] that this coefficient adjustment algorithm converges considerably faster than the stochastic gradient algorithm and virtually independently of the channel used. After about N iterations, where N is the number of coefficients of the equalizer, the mse of the equalized signal is generally close to the minimal obtainable. This is an improvement by a factor of three to ten[9,11-16] compared with the performance of the stochastic gradient algorithm. The exact improvement factor depends on the channel involved and on the modulation scheme used. Godard showed further that under certain modeling assumptions, the excess mse converges asymptotically, as the inverse of the number of iterations.

More recently, various methods were published[12,13,14] that reduce the computational complexity in the implementation of the Godard algorithm. These methods exploit the fact that only one new element, which may be a vector, is introduced in the vector of the received signal at each iteration. They avoid the processing of large matrices which are required in the Godard algorithm. Accordingly, the number of arithmetic operations can be reduced. For the Godard algorithm those grow quadratically, while for the algorithms described in Refs. 12 to 14 they grow proportionally to the number of equalizer coefficients—a considerable reduction for long equalizers. In Refs. 15 and 16 all these so-called least-squares algorithms are extended to include fractionally spaced, complex equalizer structures. The least-squares lattice algorithm was also extended to a decision feedback equalizer in Ref. 17. Investigations regarding the implementation are reported in Ref. 18.

The objective of this paper is to provide insight, on a fundamental level, into the rapid initial convergence of the least-squares algorithms relative to the stochastic gradient algorithm. Godard's approach is probabilistic and aimed at minimizing the ensemble mse as rapidly as possible. He assumed that all the involved random variables have joint Gaussian distributions. For this case, the Kalman filtering technique gives the fastest possible convergence. Although these assumptions are not generally satisfied in data transmission applications, a very powerful algorithm emerged.

In Ref. 19, a general equivalence between Kalman filtering and least-squares estimation techniques was exhibited. For the problem at hand, it implies that the algorithm obtained, while not optimal in a probabilistic sense, is the solution of a deterministic least-squares problem. This fact was stated for equalizers in Ref. 12. The Godard algorithm, as well as the ones proposed in Refs. 12 to 17, minimize the sum of the squared equalizer output errors under the condition that the coefficient vector remains constant from the start of the session to the current time. Note that this particular cost function is not the mse. Therefore, it does not explain directly why the actual mse converges so rapidly.

One obvious reason for the fast convergence of the least-squares algorithms is that all the information available from the start of the equalizer adaptation is stored and exploited in the update procedure, while the stochastic gradient algorithm relies mostly on current information. In Ref. 11, the Godard algorithm was interpreted as a stochastic gradient algorithm, where the coefficient corrections are transformed by an estimate of the inverse of the channel autocorrelation matrix. The fast initial convergence was attributed there to "self-orthogonalization" of the equalizer adjustments. However, this can only account for part of the high speed. It cannot explain the fact that the least-squares algorithms converge considerably faster than the stochastic gradient algorithms under the best conditions, i.e., when an ideal channel is involved or, equivalently, when the channel autocorrelation matrix is known exactly.

Here, we investigate the initial convergence of the deterministic least-squares algorithms from an algebraic point of view and offer alternative interpretations for the fast initial convergence. In Section II, the problem is stated and certain algebraic properties of the commonly utilized pseudo random start-up sequences are derived. In Section III, we relate the coefficient vector that results after N iterations (N being the number of coefficients) to the zero forcing equalizer and to the coefficients resulting with cyclic equalization.[7] The influence of the added noise is taken into account in Section IV.

II. THE OPTIMAL EQUALIZER COEFFICIENTS

Let $[a(k)]$ denote the complex data sequence which is transmitted at a signaling rate of $1/T$ over a channel with sampled impulse response $h(nT) = h_n$. The samples $\xi(nT)$ of the received signal, can be expressed as

$$\xi(nT) = \sum_{k=-\infty}^{\infty} a(k)h|(n+M-k)T| + \nu(nT), \tag{1}$$

where $\nu(nT)$ denotes the channel noise, and the equalizer length $N = 2M + 1$ is an odd number.

Let $x(n)$ denote the complex vector of the past N received signal samples

$$x(n)^T = [\xi|(nT)|, \xi|(n-1)T|, \cdots \xi|(n-N+1)T|], \tag{2}$$

and let $c(k)$ denote the complex coefficient vector of the adaptive equalizer. Then the equalizer output $y(n)$ at time nT, using the coefficient vector $c(k)$, can be written as the scalar product

$$y(n) = c(k)^*x(n). \tag{3}$$

The * denotes conjugate transposed vectors (matrices) or conjugate complex scalars.

2.1 The mean-square criterion

For the mean-square criterion, the equalizer coefficients are to be chosen such that the equalizer output $y(n)$ approximates the transmitted data value $a(n)$ as closely as possible. The optimum coefficient minimizes the mse[1,2]

$$\epsilon(k) = E[|c(k)^*x(n) - a(n)|^2], \tag{4}$$

where $E[\cdot]$ denotes ensemble averaging over the sequence $[a(n)]$. The vector c_{opt} which minimizes eq. (4) is given by

$$c_{\text{opt}} = A^{-1}v, \tag{5}$$

where

$$A = E[x(n)x(n)^*] \tag{6}$$

is the autocorrelation matrix of the channel, and

$$v = E[x(n)a(n)^*] \tag{7}$$

is the cross-correlation vector between the signal vector and the transmitted data value. In eq. (5), c_{opt} is the coefficient vector which, on the average, will perform better than any other. Ref. 2 indicates that the solution of eq. (5) exists if the absolute square of the transfer function and the spectral density of the noise have no zeros.

2.2 The least-squares criterion

The least-squares algorithms[9,11-16] minimize the following cost function

$$z(n) = \sum_{k=1}^{n} |c(n)^* x(k) - a(k)|^2, \tag{8}$$

i.e., the equalizer coefficient vector $c(n)$ minimizes the sum of error squares which resulted if $c(n)$ was used from the beginning of the transmission to the present instant nT. Usually, this is performed iteratively, i.e., $c(n)$ is calculated recursively for $n = 1 \cdots \infty$. Note that for $n \rightarrow \infty$, time invariant channels, ergodic data sequences $[a(n)]$ and noise, the cost function approaches asymptotically the ensemble mse, $\epsilon(n)$. Also, the solution $c(n)$ converges then to c_{opt}.

Differentiating eq. (8) with respect to $c(n)$ and setting the derivative to zero yields the following set of linear equations for the coefficient vector $c(n)$:

$$\sum_{k=1}^{n} x(k)x(k)^* c(n) = \sum_{k=1}^{n} x(k)a(k)^*. \tag{9}$$

Although the least-squares algorithms do not attempt to minimize the mse, it was observed[9,11-12] that the mse

$$\epsilon(n-1) = E[|c(n-1)^* x(n) - a(n)|^2], \tag{10}$$

which results if the equalizer's coefficient vectors stemming from the previous iteration at $(n-1)T$ is used, converges roughly when n reaches N.

To find reasons for this interesting property, we examine closely in the following, the solution of eq. (9) after exactly N iterations and derive relations between the coefficient vector $c(N)$ and some well-known equalizer coefficient vectors, namely, the zero forcing and the cyclic equalizer.

In the noiseless case, we show that the vectors $x(1), \cdots x(N)$ are linearly independent of each other for the data sequences that are usually used for equalizer start-up. Under this hypothesis of linearly independent signal vectors $x(n)$, it can only be that the linear combination

$$\sum_{k=1}^{N} x(k)b(k) = 0, \tag{11}$$

if $b(k) = 0$ for all $k = 1 \cdots N$. Then, it follows from eqs. (9) and (11) that

$$x(k)^{*}c(N) = a(k)^{*} \qquad k = 1, \cdots N. \tag{12}$$

This means that $c(N)$ equalizes the first N received signal vectors ideally. All errors are zero and accordingly $z(N) = 0$. In the special case where the channel transfer function is of the all-pole type and of order $N - 1$, $z(n)$ will remain zero for $n > N$ and $c(n) = c(N)$. Therefore, the optimal coefficient vector is found exactly when $n = N$. Although this is not the case generally, $c(N)$ still has interesting properties.

III. ALGEBRAIC PROPERTIES OF SIGNAL VECTORS AND PARTICULAR EQUALIZER VECTORS WITHOUT NOISE

Using a notation which is similar to the one used in Ref. 4, it is possible to express the N-dimensional vector $x(k)$ in the noiseless case as follows:

$$x(k) = Bd(k), \tag{13}$$

$$\text{where} \quad B = \begin{vmatrix} \cdots h_0 & \cdots h_M & \cdots h_{N-1} \cdots \\ \vdots & \vdots & \vdots \\ \cdots h_{-M} & \cdots h_0 & \cdots h_M \cdots \\ \vdots & \vdots & \vdots \\ \cdots h_{1-N} & \cdots h_{-M} & \cdots h_0 \cdots \end{vmatrix} \tag{14}$$

and

$$d(k)^T = [\cdots, a(k+M), \cdots, a(k), \cdots, a(k-M), \cdots]. \tag{15}$$

In eq. (14), B is a stationary $N \times L$ matrix, where L is at least the sum of the channel memory plus $(N - 1)$. The center part of length $N = 2M + 1$ is shown in eq. (14). In eq. (15), $d(k)$ is a stationary L-dimensional vector.

The vectors $x(1) \cdots x(N)$ are linearly independent, if they span the N dimensional space. This is equivalent to the $N \times N$ matrix

$$[x(1)\,|\,x(2)\,|\,\cdots\cdot\,|\,x(N)] = B[d(1)\,|\,\cdots\,|\,d(N)] \tag{16}$$

having rank N. This can only be true if both matrices on the right-hand side have rank N. This is a necessary but not a sufficient condition. Although eq. (16) has a very particular form, namely Toeplitz, it is not easy to find sufficient conditions for linear independence. We, therefore, investigate two interesting special cases involving particular start-up data sequences; namely, a sequence consisting of just a single pulse, and the other a periodic pseudo random noise sequence. Before pursuing this line of attack, we need additional facts about the matrix B.

For B to have rank N, the row vectors of B must be linearly independent. A necessary and sufficient condition is Gram's criterion: N vectors $b_1, \cdots b_N$ are linearly independent if and only if

$$\det \left| \begin{array}{lll} b_1{}^*b_1 & b_1{}^*b_2\cdots & b_1{}^*b_N \\ b_2{}^*b_1 & b_2{}^*b_2\cdots & b_2{}^*b_N \\ \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots\cdots \\ b_N{}^*b_1 & \cdots\cdots\cdots & b_N{}^*b_N \end{array} \right| \neq 0. \tag{17}$$

We identify $b_1{}^*$ as the first row of B and in general $b_n{}^*$ as the nth row of B. Consequently, the matrix in eq. (17) is the autocorrelation matrix of the channel. Gram's criterion then requires that the autocorrelation matrix be nonsingular. This is the same condition as the one required for the existence of a solution of eq. (5) in the noiseless case. Thus, whenever an optimal coefficient vector (in the mean-square sense) exists, then the matrix B has full rank N.

3.1 Single training pulse

Consider now the transmission of a single pulse at $k = 1 + M$. Then, inserting eqs. (13) to (15) into eq. (12) yields the following set of equations for the equalizer coefficient vector after N iterations:

$$\left| \begin{array}{lll} h_0\cdot\cdot & h_M\cdot\cdot & h_{N-1} \\ \cdots\cdot & \cdots\cdot & \cdots\cdot \\ h_{-M}\cdot\cdot & h_0\cdot\cdot & h_M \\ \cdots\cdot & \cdots\cdot & \cdots\cdot \\ h_{1-N} & h_{-M}\cdot\cdot & h_0 \end{array} \right|^* c(N) = \left| \begin{array}{c} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{array} \right|. \tag{18}$$

This is precisely the equation which defines the zero forcing equalizer.[1] In case the peak distortion of the channel impulse response is smaller than one, i.e.,

$$D = \frac{1}{|h_0|} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} |h_k| \leq 1, \tag{19}$$

the resulting equalizer minimizes the peak distortion of the overall channel,[1] and Gershgorin's criterion guarantees a unique solution of eq. (18). This is a sufficient but not a necessary condition.

This example shows that the least-squares equalizer adjustment technique yields the zero forcing equalizer in N iterations [if a unique solution of eq. (18) exists], whereas the gradient technique only attains this asymptotically in the steady state and practically requires a multiple of N iterations to obtain a good approximation.

As the time instant approaches infinity, the n equations

$$x(k)^*c(n) = a(k)^* \quad \text{and} \quad k = 1, \cdots n \tag{20}$$

cannot be satisfied simultaneously any more. In this case, $c(n)$ will be determined from eq. (9). Inserting eqs. (13) to (15) into eq. (9) yields

$$\lim_{n \geq \infty} BB^*c(n) = B \begin{vmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix}. \tag{21}$$

Since BB^* is proportional to A as defined in eq. (6) and the right-hand side of eq. (21) is proportional to v as defined in eq. (7), with the same proportionality constant, it follows that

$$\lim_{n \geq \infty} c(n) = c_{\text{opt}}, \tag{22}$$

i.e., the least-squares algorithms converge in the noiseless case to the optimal coefficients in the mse sense.

3.2 Periodic pseudo random training sequence

While the technique of sounding the channel through isolated test pulses is technically possible, a different method has found wider use. In Refs. 7, 8, and 20, periodic pseudo random noise sequences (PRNSs) were proposed and analyzed for equalizer training purposes. When these sequences are used, the resulting equalizer coefficient vector in the steady state is found to be different from the optimal one when random data was used. Nevertheless, the former vector approaches the optimal solution very closely even for short periods. If a PRNS of period length $P = N$ is used, then the vectors of the sampled signal are periodic also and may be written as:

$$\tilde{x}(k) = \tilde{B}\tilde{d}(k), \tag{23}$$

$$\text{where} \quad \tilde{B} = \begin{vmatrix} \tilde{h}_0 \cdot\cdot & \tilde{h}_{N-1} \cdot\cdot \\ \cdots\cdot & \cdots\cdot \\ \cdots\cdot & \cdots\cdot \\ \tilde{h}_{1-N} & \tilde{h}_0 \end{vmatrix}, \tag{24}$$

$$\tilde{h}_n = \sum_{k=-\infty}^{\infty} h_{n+kN}, \tag{25}$$

and

$$\tilde{d}(k)^T = |\underline{a}_{k+M}, \cdots \underline{a}_k, \cdots \underline{a}_{k-M}|. \tag{26}$$

The rows of $\tilde{B}$ and the vector $\tilde{d}\ (k)$ have dimension N.

Analyzing eqs. (24) and (25) reveals that $\tilde{B}$ is a $N \times N$ circulant matrix. It is well known from Ref. 7 that the eigenvalues of circulant matrices are determined by the discrete Fourier transform of the first row. Since the first row of the mentioned $N \times N$ circulant is formed by samples of the periodically repeated channel impulse response, it is concluded that $\tilde{B}$ has full rank N, provided that the discrete Fourier transform of the periodically repeated channel impulse response has no zero values.

This condition is very similar to the one stated for the existence of an optimal coefficient vector in the mse sense,[2] which, in the noiseless case, requires that the absolute value of the channel transfer function have no zeros.

From eq. (26) and the fact that the data sequence is periodic, i.e., $a(k + N) = a(k)$, it follows that N successive data vectors $d(k) \cdots\cdot d(k + N - 1)$ form an $N \times N$ circulant matrix. Arguing as above, it follows in general that these vectors are linearly independent if the dft of the data sequence has no zero values.

For PRNSs of period N in particular it is known that

$$\begin{aligned} d(k)^* d(k) &= N \\ d(k)^* d(j) &= -1 \qquad k \neq j. \end{aligned} \tag{27}$$

The matrix in Gram's criterion (17) is then a circulant with eigenvalues

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_j &= N + 1 \qquad j = 2 \cdots N. \end{aligned} \tag{28}$$

The determinant is $(N + 1)^{N-1} \neq 0$; therefore, N successive data vectors of a PRNS with period N are always linearly independent. This, together with the fact that $\tilde{B}$ has rank N, implies that any N different $\tilde{x}$ vectors are linearly independent. Therefore, the coefficient vector after N iterations is given as the solution of

$$\tilde{x}(k)^* c(N) = \tilde{a}(k)^* \quad \text{for} \quad k = 1, \cdots N \tag{29}$$

and is guaranteed to exist. In Ref. 7 it was shown that for the particular case of $N = P$, i.e., when the equalizer length equals the period P of the PRNS, the solution of eq. (29) has an interesting interpretation in the frequency domain: it equalizes the channel transfer function of the periodically repeated impulse response at N equidistant points. Again, this solution is obtained by the least-squares algorithms after N iterations, whereas the gradient technique obtains this only asymptotically as the number of iterations becomes large.

If the equalizer length is smaller than the period of the PRNS, i.e., $N < P$ no specific information on the nature of $c(N)$ can be obtained. We, therefore, consider the solution $c(P)$ after P iterations. Inserting eq. (23) in eq. (9) and using eq. (27) yields

$$\tilde{B}(I - D)\tilde{B}^{*}c(P) = \tilde{B}\left|\begin{matrix} -1 \\ \vdots \\ -1 \\ P \\ -1 \\ \vdots \\ -1 \end{matrix}\right| \frac{1}{P+1}, \tag{30}$$

where D is a matrix containing identical elements

$$D_{i,j} = \frac{1}{P+1}. \tag{31}$$

Using the fact that all rows of $\tilde{B}$ have the same sum

$$\sum_{K=1}^{N} \tilde{h}_K = \sum_{K=-\infty}^{\infty} h_K,$$

it follows that

$$\tilde{B}\tilde{B}^{*}c(P) = \tilde{v} + q, \tag{32}$$

where q is a vector with identical elements

$$q_i = \frac{1}{P+1}\left| \left|\sum_{k=-\infty}^{\infty} h_k\right|^2 \sum_{k=1}^{N} c_k(P) - \sum_{k=-\infty}^{\infty} h_k \right| \tag{33}$$

and

$$\tilde{v}_i = \tilde{h}_{-i}. \tag{34}$$

Observe that

$$\sum_{k=-\infty}^{\infty} h_k$$

and

$$\sum_{k=1}^{N} c_k(P)$$ [†]

equal the *dc* value of the transfer function of the sampled channel and

[†] $c_k(P)$ denotes the kth element of $c(P)$.

of the equalizer, respectively. In the absence of noise and for $P \geq \infty$, their product will be one. Then, according to eq. (33), $q_i = 0$. Therefore, q_i will be small even for finite P, and $c(P)$ may be approximated as follows

$$c(P) \cong (\tilde{B}\tilde{B}^*)^{-1}\tilde{v}. \tag{35}$$

This result may be interpreted as the optimal solution in the mean-square sense for a channel with an impulse response of finite duration P which is identical to the periodically repeated impulse response in the base interval $|-PT/2, PT/2|$. If P is large enough to span the channel impulse response, then it follows that $c(P)$ is very close to the optimal coefficient vector after only P iterations.

IV. THE INFLUENCE OF NOISE

Generally, the channel noise is not negligible as assumed in the previous section. In the presence of additive noise the vector $w(k)$ of the sampled signal can be written as

$$w(k) = x(k) + r(k). \tag{36}$$

If only one single pulse is transmitted, $x(k)$ is defined in eq. (13) and $r(k)$ is the noise vector. In this case, a coefficient vector $c_a(n)$ is obtained after N iterations, which is the solution of the following equation

$$|H + R|c_a(N) = \begin{vmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix}, \tag{37}$$

where

$$H = \begin{vmatrix} h_0 & \cdot\ \cdot\ h_M & \cdot\ \cdot & h_{N-1} \\ & \cdot\ \cdot & & \cdot \\ & \cdot\quad\cdot & & \cdot \\ & \cdot\qquad\cdot & & \cdot \\ & \cdot\qquad\quad\cdot & & \cdot \\ & \cdot\qquad\qquad\cdot & & \\ h_{1-N} & \cdot\ \cdot\ h_{-M} & \cdot & \cdot\ h_0 \end{vmatrix}^* \tag{38}$$

and

$$R = \begin{vmatrix} \nu_1 & \nu_2 \cdot\cdot\cdot\cdot\cdot & \nu_N \\ \nu_0 & \nu_1 \cdot\cdot\cdot\cdot\cdot & \\ \vdots & \vdots & \vdots \\ \nu_{2-N} & \cdot\cdot\cdot\cdot\cdot & \nu_1 \end{vmatrix}^* . \tag{39}$$

The difference between this solution and the one given in eq. (18) equals

$$c_a(N) - c(N) = H^{-1}Rc_a(N). \tag{40}$$

If $c_a(N)$ is used instead of $c(N)$ to compute the value of the cost function defined in eq. (8), we obtain

$$z_a(N) = |c_a(N) - c(N)|^*H^*H|c_a(N) - c(N)|. \tag{41}$$

Substituting eq. (40) into eq. (41) and evaluating the expected value of the cost function yields

$$E|z_a(N)| = E|c_a(N)^*R^*Rc_a(N)| = c_a(N)^*E|R^*R|c_a(N), \tag{42}$$

where $E|R^*R|$ is N times the correlation matrix of the random noise. Using Parseval's theorem, eq. (42) can be expressed in terms of the transfer function $C_a(\omega)$ of the equalizer and of the power density spectrum $S_\nu(\omega)$ of the noise $\nu(n)$

$$E|z_a(N)| \cong TN/2\pi \int_{-\pi/T}^{\pi/T} |C_a(\omega)|^2 S_\nu(\omega)d\omega. \tag{43}$$

Thus, increase of the cost function is N times the average noise power after the equalizer. This means that the average squared error (12) per equalized symbol is equal to the noise variance after the equalizer.

We now examine the solution of eq. (9) for $n > N$. In this case, the equation for the equalizer coefficient vector becomes

$$\left| B(n)B(n)^* + \sum_{k=1}^{n} r(k)r(k)^* \right| c(n) = B(n) \begin{vmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix}, \tag{44}$$

where $B(n)$ is similar to eq. (14) with row length equal to n. This indicates that, as n becomes large, the influence of the noise grows proportionally. Since $B(n)B(n)^*$ converges to the channel correlation matrix $A = BB^*$, which is stationary, it can be concluded that for very large n the influence of the noise becomes dominant. Therefore, it is not advisable to use the sounding technique for more than N to $2N$ iterations.

If a PRNS with period P symbols is used during start-up and noise is present, the coefficient vector after P iterations is not determined anymore by eq. (35). Inserting eqs. (36) and (23) into eq. (9) yields instead

$$\left| \tilde{B}\tilde{B}^* + \frac{1}{P+1} \sum_{k=1}^{P} r(k)r(k)^* \right| c(P) \cong \tilde{v}, \tag{45}$$

where it was assumed that cross products of $r(k)$ and $x(k)$ may be neglected.

The matrix of eq. (45) will, in the mean, become the correlation matrix of the channel plus the noise. Therefore, the solution will, in the mean, be the optimal solution as given by eq. (5). Note that now there is no danger in letting the algorithm run for an indefinite time, since both terms of the matrix, as well as the right-hand side of the equation, grow proportionally.

V. CONCLUSION

The initial convergence of least-squares equalizer adjustment algorithms was analyzed to determine why the least-squares algorithms converge so much faster than the widely used stochastic gradient algorithms. The algebraic properties of the sampled signal vectors were found to be of crucial importance for the convergence behavior. In particular, it was found that, for a wide class of transmission channels and for commonly used data sequences, successive sampled signal vectors are linearly independent. This ensures a unique equalizer coefficient vector after exactly N iterations, where N is the dimension of the equalizer. In the noiseless case, this coefficient vector was found to correspond to particular equalizer coefficients which were reported and studied earlier. If a single pulse is transmitted, the zero forcing equalizer is obtained. If a pseudo random noise sequence with a period in symbols equal to the number of equalizer coefficients is used, the steady state solution of the cyclic equalization technique results. This explains why the least-squares adjustment algorithms converge much faster than the gradient techniques: the above-mentioned particular equalizer coefficients are obtained after only N iterations, whereas with the stochastic gradient techniques they are only approximated as the number of iterations becomes very large. The influence of the inevitable channel and measurement noises was evaluated. Approximations show that similar performance, as in the noiseless case, is obtainable.

REFERENCES

1. R. W. Lucky, J. Salz, and E. J. Weldon, Jr, *Principles of Data Communication*, New York: McGraw-Hill, 1968.
2. A. Gersho, "Adaptive Equalization of Highly Dispersive Channels," B.S.T.J., *48*, No. 1 (January 1969), pp. 55–70.
3. G. Ungerboeck, "Theory on the Speed of Convergence in Adaptive Equalizers for Digital Communication," IBM J. Res. and Dev., *16*, No. 6 (November 1972), pp. 546–55.

4. J. E. Mazo, "On the Independence Theory of Equalizer Convergence," B.S.T.J., *58,* No. 5 (May-June 1979), pp. 963-93.
5. R. W. Chang, "A New Equalizer Structure for Fast Start-Up Digital Communication," B.S.T.J., *50,* No. 6 (July-August 1971), pp. 1969-2014.
6. K. H. Mueller, "A New, Fast-Converging Mean-Square Algorithm for Adaptive Equalizers with Partial-Response Signaling," B.S.T.J., *54,* No. 1 (January 1975), pp. 143-53.
7. K. H. Mueller and D. A. Spaulding, "Cyclic Equalization—A New Rapidly Converging Equalization Technique for Synchronous Data Communication," B.S.T.J., *54,* No. 7 (February 1975), pp. 369-406.
8. S. U. H. Qureshi, "Fast Start-Up Equalization with Periodic Training Sequences," IEEE Trans. Inform. Theory, *IT-23,* No. 5 (Sept. 1977), pp. 553-63.
9. D. Godard, "Channel Equalization Using a Kalman Filter for Fast Data Transmission," IBM J. Res. Develop., *18,* No. (May 1974).
10. A. E. Albert and L. A. Gardner, Jr., *Stochastic Approximation and Nonlinear Regression,* Cambridge: M.I.T. Press, 1967, Chapter 7.
11. R. D. Gitlin and F. R. Magee, "Self-Orthogonalizing Adaptive Equalization Algorithms," IEEE Trans. Commun., *COM-25,* No. 7 (July 1977), pp. 666-72.
12. D. D. Falconer and L. Ljung, "Application of Fast Kalman Estimation to Adaptive Equalization," IEEE Trans. Commun., *COM-26,* No. 10 (October 1978), pp. 1439-46.
13. E. H. Satorius and T. D. Pack, unpublished work.
14. E. Shichor, unpublished work.
15. T. L. Lim and M. S. Mueller, "Rapid Equalizer Start-Up Using Least-Squares Algorithms," Conf. Rec., Int. Conf. Commun., 1980, Seattle, Wa., Paper No. 57.7.
16. M. S. Mueller, "Least-Squares Algorithms for Adaptive Equalizers," B.S.T.J., *60,* No. 8 (October 1981), pp. 193-213.
17. M. T. Shensa, "A Least-Squares Lattice Decision Feedback Equalizer," Conf. Rec., Int. Conf. Commun., 1980, Seattle, Wa., Paper No. 57.6.
18. D. D. Falconer, V. B. Lawrence, and S. K. Tewksbury, "Processor-Hardware Considerations for Adaptive Digital Filter Algorithms," Conf. Rec., Int. Conf. Commun., 1980, Seattle, Wa., Paper No. 57.5.
19. H. W. Sorenson, "Comparison of Kalman, Bayesian and Maximum Likelihood Estimation Techniques," *Theory and Applications of Kalman Filtering,* NATO, Agardograph, No. 139, 1970, Chapter 6.
20. R. W. Chang and E. Y. Ho, "On Fast Start-Up Data Communication Systems Using Pseudo-Random Training Sequences," B.S.T.J. *51,* No. 9 (November 1972), pp. 2013-27.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

Criteria for the Response of Nonlinear Systems to be L-Asymptotically Periodic

I. W. SANDBERG

(Manuscript received April 22, 1981)

We consider the behavior of a general type of system governed by an input-output operator G that maps each excitation x into a corresponding response r. Here excitations and responses are R^n-valued functions defined on a set T. To accommodate both continuous time and discrete time cases, T is allowed to be either $[0, \infty)$ or $\{0, 1, 2, \cdots\}$. We address the following question. Under what conditions on G and x is it true that the response r is L-asymptotically periodic in the sense that $r = p + q$, where p is periodic with a given period τ, and q has finite energy (i.e., is square summable)? This type of question arises naturally in many applications. The main results given (which include a necessary and sufficient condition) are basically "tool theorems." To illustrate how they can be used, an example is discussed involving an integral equation that is often encountered in the theory of feedback systems.

I. INTRODUCTION

In this paper we consider the behavior of a general type of system governed by an input-output operator G that maps each excitation x into a corresponding response r. Here excitations and responses are R^n-valued functions defined on a set T. (As usual, n is an arbitrary positive integer.) To accommodate both continuous time and discrete time cases, we allow T to be either $[0, \infty)$ or $\{0, 1, 2, \cdots\}$. As in L_2-stability theory,[1-6] each x is drawn from a family $E(L)$ of functions whose truncations belong to a set L of finite-energy (i.e., square summable) functions (the details are given in Section 3.1), and G is assumed to map $E(L)$ into $E(L)$.

We address, and give in Section 3.2 results concerning, the following question. Under what conditions on G and x is it true that the response r is *L-asymptotically periodic* in the sense that $r = p + q$, where p is periodic with a given period τ, and q has finite energy?† This type of question arises naturally in many applications. Our results are basically "tool theorems" which appear to be widely applicable. An example is given in Section 3.4.

II. MOTIVATION AND BACKGROUND MATERIAL

To provide motivation for considering an abstract input-output operator G, and also to describe some earlier results related to those of Section III, we begin by recalling that an important example of a type of equation that arises in the study of physical systems (such as feedback systems or networks containing linear lumped and/or distributed elements, as well as memoryless, possibly time-varying, nonlinear elements) is the integral equation

$$x(t) = r(t) + \int_0^t k(t-\sigma)\psi[r(\sigma), \sigma]d\sigma, \qquad t \geq 0 \tag{1}$$

in which x and r take values in R^n (whose elements we take to be column vectors), k is an $n \times n$ matrix-valued function, and ψ maps $R^n \times [0, \infty)$ into R^n. In eq. (1), typically x takes into account initial conditions as well as inputs, and r is the output (i.e., is the intermediate or final output) corresponding to x (see, for instance, Ref. 2, pp. 872–4 for a specific application). Discrete-time counterparts of eq. (1) (see Ref. 7, pp. 449–51, for example) also arise often in system studies.

Much is known about the properties of eq. (1), e.g., see Refs. 2, 8, and 9. In particular, if $n = 1$ and the "circle criterion" of Ref. 2, together with certain associated conditions concerning k, ψ, x, and r described in the reference, are met; if $\psi(\cdot, t)$ is periodic in t with some period τ; and if $x = x_1 + x_0$ with x_1 bounded and τ-periodic and with x_0 bounded and such that $x_0(t) \to 0$ as $t \to \infty$, then we have $r = p + q$ in which p is periodic with period τ, and $q(t) \to 0$ as $t \to \infty$ (see Ref. 2, Theorem 4).‡

This result generalized to arbitrary n is proved in Ref. 6 by first showing that there is a τ-periodic p defined on $(-\infty, \infty)$ such that, with x_1 extended periodically for negative values of t, the *auxiliary equation*

$$x_1(t) = p(t) + \int_{-\infty}^t k(t-\sigma)\psi[p(\sigma), \sigma]d\sigma, \qquad t > -\infty \tag{2}$$

† In contrast, it is standard (see, for example, Ref. 6, p. 195) to mean by r is *asymptotically periodic* that $r = p + q$ with p continuous and as indicated, and with q continuous and such that its values go to zero as time approaches infinity. It is often possible to show without much difficulty that r is asymptotically periodic if r is L-asymptotically periodic and some natural additional hypotheses are met (for an example, see Section 3.4).

‡ The earlier related circle criteria in Refs. 10 and 11 address different issues.

is satisfied. Then, using eq. (1) and the fact that eq. (2) gives

$$x_1(t) - \int_{-\infty}^{0} k(t-\sigma)\psi[p(\sigma), \sigma]d\sigma$$
$$= p(t) + \int_0^t k(t-\sigma)\psi[p(\sigma), \sigma]d\sigma, \qquad t \geq 0,$$

it is proved that when $x = x_1 + x_0$, we have $r(t) - p(t) \to \theta$ as $t \to \infty$, in which θ is the zero element of R^n. A similar proof shows that if $n = 1$ (for the sake of simplifying a statement of a result) and both x_0 and s, where $s(t) = \int_t^\infty |k(\sigma)| d\sigma$ for $t \geq 0$, have finite energy, then under the conditions indicated above, $r - p$ has finite energy [see Ref. 2, Corollary 1(a)]. The proofs in Ref. 2 are of a functional analytic nature. For material related in a general sense concerning systems of differential equations, and in which a Lyapunov-function approach is used, see Ref. 12, pp. 210–23. Concerning more recent material, a result along the same lines as the one described in the preceding paragraph for interconnected systems[6] governed by a somewhat different class of integral equations, is proved in Ref. 13. There, too, an auxiliary-equation approach is used.†

Under reasonable conditions on k and ψ (see Section 3.4), the set $E(L)$, previously described (and defined in Section III), contains exactly one solution r of eq. (1) for each $x \in E(L)$. We now introduce the typically trivial restriction that only solutions r of eq. (1) contained in $E(L)$ are of interest to us. Thus, under reasonable conditions, there is associated in a natural way with eq. (1) a map $G:E(L) \to E(L)$ such that $r = Gx$ for each $x \in E(L)$. Of course, many other examples can be given in which such a map G arises.

Assuming that $\psi(\cdot, \sigma)$ in eq. (1) is independent of σ and that $\psi(\theta, 0) = \theta$, notice that the G associated with eq. (1) has the property that it is *time invariant* in the usual sense that the response to a delayed input is the delayed response to the original input. (For a precise definition of time invariance, see Section 3.3.) This type of property of G, rather than the concept of an auxiliary equation, plays a central role in our approach in Section III.

III. L-ASYMPTOTIC PERIODICITY, TIME INVARIANCE, AND PERIODICALLY-VARYING SYSTEMS

3.1 Preliminary notation and definitions

Throughout the remainder of the paper the following notation and definitions are used.

† The method used in Ref. 13 to show the existence of a periodic solution of the auxiliary equation is very different from that in Ref. 2.

The symbol T denotes either $[0, \infty)$ or $\{0, 1, 2, \cdots\}$. Elements of R^n are taken to be column vectors, v' denotes the transpose of an arbitrary $v \in R^n$, and θ stands for the zero element of R^n.‡

If $T = [0, \infty)$, then L denotes the set of Lebesgue measurable functions v from T into R^n such that

$$\int_0^\infty v'(t)v(t)dt < \infty.$$

Alternatively, when $T = \{0, 1, 2, \cdots\}$, L stands for the set of maps v from T into R^n such that

$$\sum_{t=0}^{\infty} v'(t)v(t) < \infty.$$

The norm $\|v\|$ of an arbitrary element v of L is defined by

$$\|v\| = \left(\int_0^\infty v'(t)v(t)dt\right)^{1/2} \quad \text{if} \quad T = [0, \infty),$$

and

$$\|v\| = \left(\sum_{t=0}^{\infty} v'(t)v(t)\right)^{1/2} \quad \text{if} \quad T = \{0, 1, 2, \cdots\}.$$

With this norm, L is a Banach space of finite energy (i.e., square summable) functions.

For $v{:}T \to R^n$ and $\omega \in T$, $v_{(\omega)}$ denotes the map from T into R^n defined by $v_{(\omega)}(t) = v(t)$ for $t \in T$ with $t \leq \omega$, and $v_{(\omega)}(t) = \theta$ for $t \in T$ such that $t > \omega$. We use $E(L)$ to denote the "extended set" $\{v{:}T \to R^n \,|\, v_{(\omega)} \in L \text{ for } \omega \in T\}$, and θ_E stands for the zero element of $E(L)$. [Note that $E(L)$ is the set of *all* maps $v{:}T \to R^n$ when $T = \{0, 1, 2, \cdots\}$.]

We say that a map $H{:}E(L) \to E(L)$ is *causal* (see Ref. 2, p. 888) if we have $(Hv)_{(\omega)} = [Hv_{(\omega)}]_{(\omega)}$ for each $v \in E(L)$ and each $\omega \in T$.

For any $v \in E(L)$ and each $\tau_0 \in T$, $v(\cdot + \tau_0)$ denotes the element w of $E(L)$ defined by $w(t) = v(t + \tau_0)$, $t \in T$.

The symbol τ denotes a fixed positive element of T, and P stands for the set of periodic functions $\{v \in E(L) \,|\, v(t + \tau) = v(t) \text{ for } t \in T\}$.

A central role is played by the set S defined by $S = \{v \in L \,|\,$ there is a $v^* \in L$ with the property that

$$\sum_{k=1}^{K} v(\cdot + k\tau) \to v^*$$

as $K \to \infty\}$, where $\to v^*$ means convergence in norm to v^*.

‡ We have repeated the definitions of T and θ for the reader's convenience.

Finally, for each $\omega \in T$, the "delay map" $D_\omega : E(L) \to E(L)$ is defined by $(D_\omega v)(t) = v(t - \omega)$ for $t \geq \omega$, and $(D_\omega v)(t) = \theta$ for $t < \omega$.

3.2 L-Asymptotic periodicity

We shall use the following hypothesis:

H.1: G is a map from $E(L)$ into $E(L)$ such that for any $v \in E(L)$, we have $(GD_\tau v)(t) = (D_\tau Gv)(t)$ for $t \in T$ with $t \geq \tau$.

This hypothesis is satisfied whenever G is a causal map of $E(L)$ into itself that is either time invariant or periodically varying with period τ (see Section 3.3). Our main result is the following:

Theorem 1: Assume that H.1 holds. Let $x \in E(L)$, and let r denote Gx. Then r has the form $p + q$ with $p \in P$ and $q \in L$ if and only if $(Gx - GD_\tau x) \in S$.

Proof: Suppose first that $(Gx - GD_\tau x) = v$ for some $v \in S$. Let $v^* \in L$ be such that

$$\sum_{k=1}^{K} v(\cdot + k\tau) \to v^*$$

as $K \to \infty$. We shall use the proposition that

$$v^*(\cdot + \tau) + v(\cdot + \tau) = v^*(\cdot), \tag{3}$$

which follows from the inequality

$$\|v^*(\cdot + \tau) + v(\cdot + \tau) - v^*(\cdot)\|$$
$$\leq \left\| v^*(\cdot + \tau) - \sum_{k=2}^{K} v(\cdot + k\tau) \right\| + \left\| \sum_{k=1}^{K} v(\cdot + k\tau) - v^*(\cdot) \right\| \tag{4}$$

for $K \geq 2$, and the fact that the right side of inequality (4) approaches zero as $K \to \infty$.

Let p_0 denote $r + v^*$, which is clearly an element of $E(L)$. Since $r(t) = (GD_\tau x)(t) + v(t)$ for $t \in T$, by H.1, we have $r(t) = r(t - \tau) + v(t)$ for $t \geq \tau$. Therefore, for $t \in T$, $p_0(t + \tau) = r(t + \tau) + v^*(t + \tau) = r(t) + v(t + \tau) + v^*(t + \tau)$. On the other hand, using eq. (3), $r(t) + v(t + \tau) + v^*(t + \tau) = r(t) + v^*(t) = p_0(t)$ for all $t \in T$ if $T = \{0, 1, 2, \cdots\}$, and for almost all $t \in T$ if $T = [0, \infty)$. Therefore, with p the element of P defined by $p(t) = p_0(t)$ for $t \in [0, \tau) \cap T$, we have $p_0(t) - p(t) = \theta$ for all $t \in T$ if $T = \{0, 1, 2, \cdots\}$ and for almost all $t \in T$ if $T = [0, \infty)$, and clearly $r = p + (p_0 - p - v^*)$ in which $(p_0 - p - v^*) \in L$.

Suppose now that $r = p + q$ with $p \in P$ and $q \in L$, and let $u = (Gx - GD_\tau x)$. For $t \geq \tau$, $u(t) = r(t) - r(t - \tau) = p(t) + q(t) - p(t - \tau) - q(t - \tau) = q(t) - q(t - \tau)$ which, together with $u \in E(L)$, shows that $u \in L$.

Let $u^{(K)}(\cdot)$ in L be defined by

$$u^{(K)}(t) = \sum_{k=1}^{K} u(t + k\tau)$$

for $t \in T$ and any positive integer K, and let J be an integer such that $J > K$. Using $u(t + k\tau) = q(t + k\tau) - q[t + (k - 1)\tau]$ for $k \geq 1$ and $t \in T$, we have, for $t \in T$,

$$\begin{aligned} u^{(J)}(t) - u^{(K)}(t) &= \sum_{k=1}^{J} u(t + k\tau) - \sum_{k=1}^{K} u(t + k\tau) \\ &= \sum_{k=(K+1)}^{J} u(t + k\tau) \\ &= q(t + J\tau) - q(t + K\tau). \end{aligned}$$

Thus, $\| u^{(J)} - u^{(K)} \| \leq \| q(\cdot + J\tau) \| + \| q(\cdot + K\tau) \|$. Since $\| q(\cdot + K\tau) \| \to 0$ as $K \to \infty$, $\{u^{(K)}\}_1^\infty \subset L$ is a Cauchy sequence, and, by the completeness of L, there is a $u^* \in L$ such that $\| u^{(K)} - u^* \| \to 0$ as $K \to \infty$. This concludes the proof.

3.2.1 Comments

The following example shows that S is a *proper* subset of L. Let G be the identity operator on $E(L)$, take $n = 1$, and let $x \in E(L)$ be defined by $x(t) = \ln 2$ for $t \in [0, 2] \cap T$, and $x(t) = \ln t$ for $t \in (2, \infty) \cap T$. Let $\tau = 1$. Then, $(Gx - GD_\tau x)(t) = r(t) - r(t - 1) = \ln[t(t - 1)^{-1}]$ for $t \in [3, \infty) \cap T$. Using the inequality $\ln(1 + \sigma) \leq \sigma$ valid for $\sigma \geq 0$, we see that $\ln[t(t - 1)^{-1}] \leq (t - 1)^{-1}$ for $t \in [3, \infty) \cap T$, and therefore that v, defined by $v(t) = (Gx - GD_\tau x)(t)$ for $t \in T$, belongs to L. Since here Gx cannot be written as $p + q$ with $p \in P$ and $q \in L$, it follows from the theorem that $v \notin S$.

It is not difficult to verify that the proof given of the theorem can be modified to show that H.1 can be replaced with the following somewhat weaker hypothesis.

H.1′: $G{:}E(L) \to E(L)$ is a map such that for any $v \in E(L)$, there is an $s \in S$ such that $(GD_\tau v)(t) = (D_\tau Gv)(t) + s(t)$ for $t \in T \cap [\tau, \infty)$.

The simple example: $n = 1$, $(Gv)(t) = v(t) + e^{-t}$ for $t \in T$ and each $v \in E(L)$ is one for which H.1′, but not H.1, is met.

3.2.2 Corollaries (the use of weighting functions)

In this section, and in the Appendix, w denotes any function from T into R^1 such that there is a constant $\beta > 0$ for which $w(t) \geq (1 + \beta t)^2$ when $t \in T$, and such that w is measurable on T and bounded on bounded subsets of T if $T = [0, \infty)$. By wv, where $v \in E(L)$, we mean the element of $E(L)$ defined by $(wv)(t) = w(t)v(t)$ for $t \in T$.

Corollary 1: Suppose that H.1 is met, that $x \in E(L)$, and that

$w(Gx - GD_\tau x) \in L$. Then $Gx = p + q$ for some $p \in P$ and some $q \in L$.

Proof: Let $h = w(Gx - GD_\tau x)$, and let s denote $(Gx - GD_\tau x)$. Observe that $s \in L$. For any positive integers J and K with $J > K$,

$$\begin{aligned}\left\| \sum_{k=1}^{J} s(\cdot + k\tau) - \sum_{k=1}^{K} s(\cdot + k\tau) \right\| &= \left\| \sum_{k=K+1}^{J} s(\cdot + k\tau) \right\| \\ &\leq \sum_{k=K+1}^{J} \| s(\cdot + k\tau) \| \\ &\leq \sum_{k=K+1}^{J} (1 + k\beta\tau)^{-2} \| h(\cdot + k\tau) \| \\ &\leq \sum_{k=K+1}^{J} (1 + k\beta\tau)^{-2} \| h \|,\end{aligned}$$

which shows that

$$\left\| \sum_{k=1}^{J} s(\cdot + k\tau) - \sum_{k=1}^{K} s(\cdot + k\tau) \right\| \to 0$$

as J and K approach infinity. By the completeness of L, we have $s \in S$ and the corollary follows.

In Corollary 2, below, $w(\cdot + \tau)[x(\cdot + \tau) - x(\cdot)]$ denotes the element of $E(L)$ whose values are $w(t + \tau)[x(t + \tau) - x(t)]$.

Corollary 2: Assume that H.1 is met, and that there is a positive constant ρ such that

$$\| w(Gu - Gv)_{(\omega)} \| \leq \rho \| w(u - v)_{(\omega)} \| \tag{5}$$

for u and v in $E(L)$ and $\omega \in T$. If $x \in E(L)$ is such that $w(\cdot + \tau) \cdot [x(\cdot + \tau) - x(\cdot)] \in L$, then $Gx = p + q$ for some $p \in P$ and some $q \in L$.

Proof: We have $\| w(Gx - GD_\tau x)_{(\omega)} \| \leq \rho \| w(x - D_\tau x)_{(\omega)} \|$ for $\omega \in T$ and any $x \in E(L)$. When $w(\cdot + \tau)[x(\cdot + \tau) - x(\cdot)] \in L$, it follows that $\sup_{\omega \in T} \| w(x - D_\tau x)_{(\omega)} \| < \infty$; hence, $w(Gx - GD_\tau x) \in L$. By Corollary 1, $Gx = p + q$ with p and q as indicated.

3.2.3 Comments

The condition that $w(\cdot + \tau)[x(\cdot + \tau) - x(\cdot)] \in L$ is met if $\sup_{t \in T}[w(t + \tau)/w(t)] < \infty$ and $x = p_0 + q_0$ with $p_0 \in P$ and $wq_0 \in L$, and of course $\sup_{t \in T}[w(t + \tau)/w(t)] < \infty$ is satisfied if, for example, $w(t) = e^{\lambda t}$ for $t \in T$ or $w(t) = (1 + \lambda t)^2$ for $t \in T$, with λ a positive constant. Input-output stability theory techniques can frequently be used to show, in specific cases, that eq. (5), with an appropriate w, is met.

Regarding the case in which $T = \{0, 1, 2, \cdots\}$, since τ could have been taken to be unity, Theorem 1 and Corollaries 1 and 2 provide conditions under which r is L-asymptotically *constant* in the sense that $r = c + q$ with $q \in L$ and $c \in C$, where C is the set of constant R^n-valued functions $\{v \in P | v(t) = u \text{ for } t \in T \text{ and some } u \in R^n\}$. Corresponding results for $T = [0, \infty)$ are given in the Appendix.

3.3 Time invariance and periodically-varying systems

Hypothesis 1 plays a prominent role in Section 3.2. Here we give definitions which make precise the essentially self-evident proposition that H.1 is met if G is a causal map of $E(L)$ into itself that is either, in the usual sense, time invariant or periodically varying with period τ.

Let H be an arbitrary causal map of $E(L)$ into $E(L)$.

Definition 1: H is *time invariant* if (i) there is an element ν of R^n such that $(H\theta_E)(t) = \nu$ for $t \in T$, and (ii) for any $x \in E(L)$, we have

$$\begin{aligned}(HD_\omega x)(t) &= \nu, && t \in [0, \omega) \cap T\\ &= (D_\omega Hx)(t), && t \in [\omega, \infty) \cap T\end{aligned}$$

for each $\omega \in (T - \{0\})$.

Definition 2: H is *periodically varying* with period τ if (i) $H\theta_E = \nu$ for some $\nu \in P$, and (ii) for each $x \in E(L)$ and any positive integer k,

$$\begin{aligned}(HD_{k\tau} x)(t) &= \nu(t), && t \in [0, k\tau) \cap T\\ &= (D_{k\tau} Hx)(t), && t \in [k\tau, \infty) \cap T.\end{aligned}$$

Notice that H is "periodically varying" with period τ if H is time invariant. A related definition is the following:

Definition 2′: H is *periodically varying* with period τ if (i) $H\theta_E = \nu$ for some $\nu \in P$, and (ii) for any $x \in E(L)$, we have

$$\begin{aligned}(HD_\tau x)(t) &= \nu(t), && t \in [0, \tau) \cap T\\ &= (D_\tau Hx)(t), && t \in [\tau, \infty) \cap T.\end{aligned}$$

To see that Definitions 2 and 2′ are consistent, we observe the following: If H meets the conditions of Definition 2, then obviously H satisfies the conditions of Definition 2′. On the other hand, if H meets the conditions of Definition 2′, and $x \in E(L)$ is given, and if

$$\begin{aligned}(HD_{k\tau} x)(t) &= \nu(t), && t \in [0, k\tau) \cap T && \text{(6a)}\\ &= (D_{k\tau} Hx)(t), && t \in [k\tau, \infty) \cap T && \text{(6b)}\end{aligned}$$

for some k, then, by the conditions of Definition 2′ with x replaced with $D_{k\tau}x$,

$$\begin{aligned}(HD_{(k+1)\tau} x)(t) &= \nu(t), && t \in [0, \tau) \cap T\\ &= (D_\tau HD_{k\tau} x)(t), && t \in [\tau, \infty) \cap T.\end{aligned}$$

Since $HD_{k\tau}x$ has the values given by eqs. (6a) and (6b), we see that

$$\begin{aligned}(HD_{(k+1)\tau}x)(t) &= \nu(t), & t &\in [0, (k+1)\tau) \cap T\\ &= (D_{(k+1)\tau}Hx)(t), & t &\in [(k+1)\tau, \infty) \cap T,\end{aligned}$$

which shows that the conditions of Definition 2 are met.

Notice that our assumption that H is causal is not explicitly used. That assumption restricts the class of operators H so that the definitions given above are appropriate and natural.†

3.4 An example

Let $T = [0, \infty)$, and consider eq. (1) which is repeated below.

$$x(t) = r(t) + \int_0^t k(t-\sigma)\psi[r(\sigma), \sigma]d\sigma, \qquad t \geq 0. \tag{1}$$

Assume the following, in which L_1 denotes the set of functions from $[0, \infty)$ to R^1 that are summable over $[0, \infty)$.

A.1: $x \in E(L)$, k is a measurable real $n \times n$ matrix-valued function defined on $[0, \infty)$ such that each k_{ij} is bounded and belongs to L_1, and ψ is a map from $R^n \times [0, \infty)$ into R^n with the properties that $\psi(\theta, \sigma) = \theta$ for $\sigma \geq 0$, and

(*i*) there is a constant $c > 0$ such that $|\psi(u, t) - \psi(v, t)| \leq c|u - v|$ for all $u, v \in R^n$ and all $t \geq 0$, in which $|\cdot|$ is some norm on R^n, and

(*ii*) $\psi[z(\cdot), \cdot]$ is measurable on $[0, \infty)$ whenever $z \in E(L)$.

Since $x \in E(L)$ and each $k_{ij} \in L_1$, it follows that u defined by

$$u(t) = \int_0^t k(t-\sigma)\psi[x(\sigma), \sigma]d\sigma, \qquad t \geq 0$$

is an element of $E(L)$. Also, since each k_{ij} is bounded, there is a constant c_0 such that $|k(t-\sigma)[\psi(z_1, \sigma) - \psi(z_2, \sigma)]| \leq c_0|z_1 - z_2|$ for all nonnegative t and σ such that $t \geq \sigma$, and for all z_1 and z_2 in R^n. These two observations show that a proof given by Tricomi (see Ref. 8, pp. 42–7) can be modified to prove that $E(L)$ contains a unique solution r of eq. (1).‡

Let G be the map of $E(L)$ into $E(L)$ defined by the condition that for each $x \in E(L)$, $r = Gx$ is the solution in $E(L)$ of eq. (1). Since

† Although the concepts involved are obviously well known, it appears that Definitions 2 and 2′ have not actually been given earlier. Also, Definition 1 is not entirely standard. For example, in Ref. 4, p. 20, time invariance requires that $\nu = \theta$.

‡ The integral on the right side of eq. (1) can easily be shown to be an element of R^n for each t whenever $r \in E(L)$. Since the value of the integral for a given t is unchanged if r is replaced by any element of $E(L)$ that agrees with r almost everywhere, eq. (1) has a solution if there is an element $E(L)$ that satisfies the equation almost everywhere, and, moreover, any solution $r \in E(L)$ is unique and not merely essentially unique.

$\psi(\theta, t) = \theta$ for $t \geq 0$, it is easy to see that H.1 is met when $\psi(z, t+\tau) = \psi(z, t)$ for $t \geq 0$ and $z \in R^n$.

Now consider four additional assumptions.

A.2: $\psi(z, t) = \psi(z, t+\tau)$ for $t \geq 0$ and all $z \in R^n$.

A.3: For any z_a and z_b in $E(L)$, there is a measurable real $n \times n$ matrix-valued function D defined on $[0, \infty)$ such that (*i*) each D_{ij} is bounded on $[0, \infty)$, (*ii*) $\psi[z_a(t), t] - \psi[z_b(t), t] = D(t)[z_a(t) - z_b(t)]$ for $t \geq 0$, and (*iii*) the relation

$$x_0(t) = r_0(t) + \int_0^t k(t-\sigma)D(\sigma)r_0(\sigma)d\sigma, \qquad t \geq 0 \tag{7}$$

implies that we have $r_0 \in L$ whenever $r_0 \in E(L)$ and $x_0 \in L$. (See Ref. 2, pp. 876–8 for conditions under which A.3 holds when ψ has a certain important specific form.)

A.4: For each i and j, $t^p k_{ij} \in L_1$ for $p = 1, 2$.†

A.5: Concerning eq. (1), $x = u_1 + u_2$ with $u_1 \in P$ and $t^p u_2 \in L$ for $p = 0, 1, 2$.

We shall prove the following.

Theorem 2: If A.1 through A.5 hold, then $E(L)$ contains a unique solution r of eq. (1), *and we have $r = p + q$ for some $p \in P$ and some $q \in L$.*

Proof: As indicated earlier, A.1 implies that there is a unique solution of eq. (1) in $E(L)$. Let r and s denote Gx and $GD_\tau x$, respectively, and let D satisfy $\psi[r(t), t] - \psi[s(t), t] = D(t)[r(t) - s(t)]$, $t \geq 0$ with D such that (*i*) and (*iii*) of A.3 hold. Then, with $\Delta = r - s$ and $v = x - D_\tau x$,

$$v(t) = \Delta(t) + \int_0^t k(t-\sigma)D(\sigma)\Delta(\sigma)d\sigma, \qquad t \geq 0.$$

Note that $v(t) = x(t)$ for $t \in [0, \tau)$, and $v(t) = u_2(t) - u_2(t-\tau)$ for $t \geq \tau$, from which it easily follows that $(1+t)^p v \in L$ for $p = 0, 1, 2$.

By A.3, $\Delta \in L$. In addition, observe that we have

$$\begin{aligned}(1+t)v(t) = (1+t)\Delta(t) \\ + \int_0^t k(t-\sigma)D(\sigma)(1+\sigma)\Delta(\sigma)d\sigma \\ + \int_0^t (t-\sigma)k(t-\sigma)D(\sigma)\Delta(\sigma)d\sigma, \qquad t \geq 0.\end{aligned}$$

Since $tk_{ij} \in L_1$ for all i and j, I_1 given by

† By $t^p k_{ij}$ we mean, of course, the map from $[0, \infty)$ into R^1 whose value at t is $t^p k_{ij}(t)$.

$$I_1(t) = \int_0^t (t - \sigma)k(t - \sigma)D(\sigma)\Delta(\sigma)d\tau, \qquad t \geq 0$$

belongs to L. Thus, by A.3, $(1 + t)\Delta \in L$. Similarly,

$$\begin{aligned}(1 + t)^2 v(t) = (1 + t)^2\Delta(t) &+ \int_0^t k(t - \sigma)D(\sigma)(1 + \sigma)^2\Delta(\sigma)d\sigma \\ &+ 2\int_0^t (t - \sigma)k(t - \sigma)D(\sigma)(1 + \sigma)\Delta(\sigma)d\sigma \\ &+ \int_0^t (t - \sigma)^2 k(t - \sigma)D(\sigma)\Delta(\sigma)d\sigma, \qquad t \geq 0,\end{aligned}$$

together with the hypothesis that A.3 holds and that $t^2 k_{ij} \in L_1$ for all i and j, shows that $(1 + t)^2\Delta \in L$. By Corollary 1, $r = p + q$ with p and q as indicated.

3.4.1 Comments

Under the conditions of Theorem 2, it can be shown that the integral on the right side of eq. (1) depends continuously on t for $t > 0$. Thus, if u_1 and u_2 of Theorem 2 are continuous, then so is r.

Concerning the standard concept of asymptotic periodicity (see Ref. 6, p. 195 and refer to the footnote in Section I), arguments of the kind used in Ref. 14 show that r is asymptotically τ-periodic whenever x is asymptotically τ-periodic, the conditions of Theorem 2 are met, and k satisfies the additional assumption:

A.6: Each tk_{ij} is bounded on $[0, \infty)$.†

More specifically, let A.6 and the conditions of Theorem 2 be met, and let p and q be as described in Theorem 2. Then, with $\psi[p(\sigma), \sigma]$ defined on $(-\infty, 0)$ by periodically extending $\psi[p(\sigma), \sigma]$ on $[0, \tau)$, the integral

$$\int_{-\infty}^t k(t - \sigma)\psi[p(\sigma), \sigma]d\sigma$$

exists as an element of R^n for each t (see Ref. 14, pp. 2852-3). This integral is periodic in t, and it can be shown to be continuous in t. These facts can be used to verify that when A.1 through A.6 are satisfied,

$$\int_0^t k(t - \sigma)\psi[p(\sigma) + q(\sigma), \sigma]d\sigma,$$

† This hypothesis and A.5 imply that each $(1 + t)k_{ij}$ is square summable over $[0, \infty)$.

which is continuous in t for $t > 0$, can be written as $v_1 + v_2$ with v_1 continuous and τ-periodic and with $v_2(t) \to \theta$ as $t \to \infty$. The rest is obvious.

The proof of Theorem 2 involves the use of a quadratic weighting function w. A similar result can be proved using an exponential weighting function. Specifically, suppose that A.1 holds, that there is an $\alpha > 0$ such that A.3 is met with k replaced with $e^{\alpha t}k$, and that A.5 holds with the integrability conditions on u_2 replaced by the requirement that $e^{\alpha t}u_2 \in L$. Then, since

$$e^{\alpha t}v(t) = e^{\alpha t}\Delta(t) + \int_0^t e^{\alpha(t-\sigma)}k(t-\sigma)D(\sigma)e^{\alpha\sigma}\Delta(\sigma)d\sigma, \qquad t \geq 0$$

we have $e^{\alpha t}\Delta \in L$.†

IV. APPENDIX

Throughout this appendix, δ denotes an arbitrary positive constant, C stands for the subset of $E(L)$ whose elements are constant R^n-valued functions, $T = [0, \infty)$, $P(\omega)$ denotes the set of periodic functions $\{v \in E(L) \mid v(t+\omega) = v(t) \text{ for } t \in T\}$ for each $\omega > 0$, and $S(\omega)$ is defined by $S(\omega) = \{v \in L \mid$ there is a $v^* \in L$ with the property that

$$\sum_{k=1}^{K} v(\cdot + k\omega) \to v^*$$

as $K \to \infty\}$ for any $\omega > 0$.

Consider hypothesis H.2 below.

H.2: $T = [0, \infty)$, and G is a map of $E(L)$ into $E(L)$ with the following property: For each $v \in E(L)$ and each $\omega \in (0, \delta)$, we have $(GD_\omega v)(t) = (D_\omega Gv)(t)$ for $t \geq \omega$.

Theorem 3: Let H.2 hold, and let $x \in E(L)$. Then Gx has the form $c + q$ with $c \in C$ and $q \in L$ if and only if $(Gx - GD_\omega x) \in S(\omega)$ for $\omega \in (0, \delta)$.

Proof: By Theorem 1, $(Gx - GD_{\tau_0}x) \in S(\tau_0)$ for any $\tau_0 \in (0, \delta)$ when Gx has the form indicated.

On the other hand, suppose that $(Gx - GD_\omega x) \in S(\omega)$ for $\omega \in (0, \delta)$, and let $\tau_0 \in (0, \delta)$. By Theorem 1, $Gx = p_{\tau_0} + q_{\tau_0}$ with $p_{\tau_0} \in P(\tau_0)$ and $q_{\tau_0} \in L$. Similarly, for any integer $m > 0$, and with $\tau_1 = \tau_0/m$, we have $Gx = p_{\tau_1} + q_{\tau_1}$ for some $p_{\tau_1} \in P(\tau_1)$ and some $q_{\tau_1} \in L$. Notice that p_{τ_1} and, therefore, $(p_{\tau_0} - p_{\tau_1})$ belong to $P(\tau_0)$, and hence have Fourier series expansions. Since $(p_{\tau_0} - p_{\tau_1})$ also belongs to L, and $m > 0$ is arbitrary, it easily follows that there is a $u \in R^n$ such that $p_{\tau_0}(t) = u$ for almost all $t \geq 0$. This completes the proof.

† Both quadratic and exponential weighting functions have been used earlier for the different purpose of obtaining criteria for the boundedness of solutions of equations (see Refs. 2 and 7, and, for example, Refs. 5 and 6).

Theorem 3 and the material in Section 3.2.2 can be used to immediately obtain the following two results.

Corollary 3: Assume that H.2 is met, that $x \in E(L)$, and that $w(Gx - GD_{\omega}x) \in L$ for $\omega \in (0, \delta)$ (w is defined in Section 3.2.2). Then $Gx = c + q$ with $c \in C$ and $q \in L$.

Corollary 4: Suppose that H.2 is satisfied, that w (see Section 3.2.2) satisfies $\sup_{t \geq 0}[w(t + \omega)/w(t)] < \infty$ for $\omega \in (0, \delta)$, that there is a constant $\rho > 0$ such that $\| w(Gu - Gv)_{(\omega)} \| \leq \rho \| w(u - v)_{(\omega)} \|$ for u and v in $E(L)$ and $\omega > 0$, and that $x = c_0 + q_0$ with $c_0 \in C$ and $wq_0 \in L$. Then $Gx = c + q$ for some $c \in C$ and some $q \in L$.

REFERENCES

1. I. W. Sandberg, "On the L_2-Boundedness of Solutions of Nonlinear Functional Equations," B.S.T.J., *43* (July 1964), pp. 1581–99.
2. I. W. Sandberg, "Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations," B.S.T.J., *44* (May–June 1965), pp. 871–98.
3. G. Zames, "On the Input-Output Stability of Nonlinear Time-Varying Feedback Systems, Pt. I," IEEE Trans. Automatic Contr., *11,* No. 2 (April 1966), pp. 228–38.
4. J. C. Willems, *The Analysis of Feedback Systems*, Cambridge, Ma.: M.I.T. Press, Research Monograph No. 62, 1971.
5. C. A. Desoer and M. Vidyasagar, *Feedback Systems*: *Input-Output Properties*, New York: Academic Press, 1975.
6. A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*, New York: Academic Press, 1977.
7. I. W. Sandberg, "On the Boundedness of Solutions of Nonlinear Integral Equations," B.S.T.J., *44* (March 1965), pp. 439–53.
8. F. G. Tricomi, *Integral Equations*, New York: Interscience, 1957.
9. R. K. Miller, *Nonlinear Volterra Integral Equations*, Menlo Park: Benjamin, 1971.
10. I. W. Sandberg, "On the Response of Nonlinear Control Systems to Periodic Input Signals," B.S.T.J., *43* (May 1964), pp 911–26.
11. I. W. Sandberg, "On the Stability of Solutions of Linear Differential Equations with Periodic Coefficients," J. Soc. Indust. Appl. Math., *12,* No. 2 (June 1964), pp. 487–96
12. T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, New York: Springer, 1975.
13. R. K. Miller and A. N. Michel, "On the Response of Nonlinear Multivariable Interconnected Feedback Systems to Periodic Input Signals," IEEE Trans. Circuits and Systems, *27,* No. 11 (November 1980), pp. 1088–96.
14. I. W. Sandberg and V. E. Beneš, "On the Properties of Nonlinear Integral Equations that Arise in the Theory of Dynamical Systems," B.S.T.J., *43,* No. 6 (November 1964), pp. 2839–54.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

Study of a Time-Compression Technique for TV Transmission Using a Chirp Filter and Envelope Detection

By K. Y. ENG and B. G. HASKELL

(Manuscript received January 19, 1981)

We study a time-compression (or expansion) technique for possible application in communication signal processing, e.g., broadcast-quality TV *transmission through satellites. The method uses a linear chirp, a linear dispersive filter realized by surface acoustic wave devices and an envelope detector. This technique is heuristic and can be viewed as a quasistationary model of the* FM *wave involved. Numerical results show that excessive distortion is created, and its application to* TV *transmission is not suitable unless some kind of equalization is provided. One such form of equalization is the chirp transform processor which involves considerably more complexity. Simpler equalizations may be possible but do not seem to be straightforward.*

I. INTRODUCTION

We study a time-compression technique motivated by the longstanding interest in transmitting multiple broadcast-quality color TV signals through a single satellite transponder, i.e., a usable RF bandwidth of 36 MHz in a communications satellite such as COMSTAR. This can be done by the use of frequency division multiplexing (FDM). However, the nonlinearity of the transponder can cause serious intelligible crosstalk and intermodulation interference between the FM carriers unless the satellite power amplifier is backed off substantially. Such a backoff, in turn, leads to a reduction in the downlink carrier-to-noise ratio. As a result, there exists an optimum trade-off between the crosstalk and s/n's, which limits the overall system performance, and achieving broadcast-quality TV transmission becomes difficult.

It is possible to time compress each scan line of a color TV signal by the use of a linear chirp, a linear dispersive filter (LDF) and an envelope detector. Two or more time-compressed scan lines from different, but synchronized, TV signals can then be time multiplexed together in the time duration of an ordinary TV scan line. This concept of time-compression multiplexing (TCM) is not new,[1,2] but recent advances in fast analog-to-digital converters, digital-to-analog converters and charge-coupled devices have greatly facilitated the implementation of time compression or expansion. However, because of their present limitations on bandwidth and speed, time-compression factors for achievable TV signals are only 2 or 3. For large time compressions, LDFs realizable by surface acoustic wave (SAW) technology are promising candidates because of their large time-bandwidth property. In addition to high speed, the TV application also requires extremely high signal fidelity. There are other applications on the high-speed time expansion (or compression) of waveforms where the distortion requirement is less stringent than the TV transmission case. In the specific case of multiple TV transmissions through a single satellite transponder, there are many advantages in using TCM, e.g., higher transponder efficiency, no intermodulation, no crosstalk, possible compatibility with time division multiplex (TDM) operations, etc. The crucial question is, of course, how much distortion the compression/expansion process would introduce on the signals. This paper gives both analysis and numerical examples that illustrate the method.

The study revealed that considerable distortion is introduced by these operations, and its application to broadcast-quality TV transmission would require SAW filter performance beyond the present state of the art. However, if the distortion requirement can be relaxed somewhat, then the present approach is advantageous because of its simplicity. On the other hand, if high signal fidelity is required, then some kind of equalization is needed for the present technique. This has motivated the study of an extension of the present method which is capable of producing high signal quality with SAW filter requirements within the present state of the art even at compression factors of 10 or more, but at the expense of higher complexity. This latter development is not discussed here but is covered in Ref. 3. The remainder of the paper covers the theoretical analysis and computer simulation. However, the discussion of either subject by itself is not adequate for the complete understanding of the system. The theoretical analysis establishes that although the basic concept was derived heuristically through physical interpretations, it can be viewed as a quasi-stationary approximation to the time-compression process. The computer simulation, on the other hand, provides the quantitative results that lead to the conclusion that the resulting distortion is excessive for today's SAW filter parameters.

II. THEORETICAL ANALYSIS

In this section, we describe and analyze the proposed compression method using TV as an example. We first describe a heuristic argument of how the technique is supposed to work. We then derive the impulse response of a general LDF—the understanding of which is important to the subsequent analysis and simulation. A brief step-by-step analysis of the compression process is shown, and its result reveals that the technique can be interpreted as a quasi-stationary approximation of the chirp signal. The mathematical expressions describing the time compressor are complicated; thus, numerical results are obtained using a computer simulation discussed in Section III. Various other properties are also discussed.

2.1 A physical interpretation

A block diagram is shown in Fig. 1. The input signal $v(t)$ consists of successive scan lines, each with a duration of T_l seconds, and the voltage is biased to be positive. It is multiplied synchronously by a periodic chirp signal $c(t)$ with a center frequency f_0 and a chirp range of Δf_c. The instantaneous frequency of $c(t)$ sweeps linearly from $(f_0 - \Delta f_c/2)$ to $(f_0 + \Delta f_c/2)$ over each scan line duration T_l. The lowest frequency of the chirp signal is assumed to be much greater than the highest frequency in the TV signal. The input $x(t)$ to the LDF is then an amplitude-modulated chirp waveform. For simplicity, let us restrict our attention to a single scan line, say $(0, T_l)$. In this interval, $x(t)$ chirps from $(f_0 - \Delta f_c/2)$ to $(f_0 + \Delta f_c/2)$ with the TV signal as the

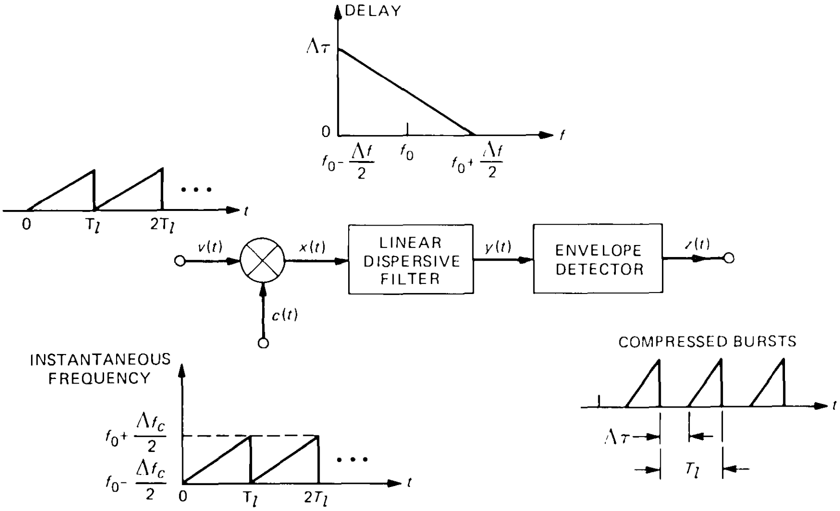


Fig. 1—Time-compression filter.

envelope modulation. As for the LDF, we assume that it has a bandwidth Δf centered at f_0, and $\Delta f \geq \Delta f_c$.

Again, for the purpose of a simple illustration, let us assume that $\Delta f = \Delta f_c$ and, over its passband, the LDF has a constant gain and a linear group delay characteristic decreasing from $\Delta\tau$ to zero, where $\Delta\tau$ is the delay dispersion of the LDF. At the time instant $t = 0$, $x(t)$ has an instantaneous frequency $(f_0 - \Delta f_c/2)$ and an envelope magnitude proportional to $v(0)$. This "envelope piece" is delayed by $\Delta\tau$ as it transits through the LDF. Similarly, at $t = T_l$, the instantaneous frequency of $x(t)$ is the highest chirp frequency which gives a zero delay for the passage through the LDF. In between these two end points, the envelope of $x(t)$ is delayed linearly from $\Delta\tau$ to zero. Equivalently, the envelope of $x(t)$ over the interval $(0, T_l)$ is time compressed into $(\Delta\tau, T_l)$ in the LDF output. An envelope detector is then used to retrieve the time-compressed TV signal. Similar compression occurs for all the other scan lines, and as a result, the envelope detector output consists of a sequence of compressed TV bursts. If we denote the duration of each burst by T_c, the ratio (T_l/T_c) is called the time-compression ratio (TCR).

It is easy to show that, in general, for a given set of T_l, $\Delta\tau$, Δf, and Δf_c,

$$T_c = T_l - \frac{\Delta\tau}{\Delta f}\Delta f_c, \tag{1}$$

where $0 \leq T_c \leq T_l$, $\Delta f_c \leq \Delta f$, and the compression ratio is

$$\text{TCR} = \left(1 - \frac{\Delta\tau}{\Delta f}\frac{\Delta f_c}{T_l}\right)^{-1}. \tag{2}$$

It is clear from the above description that time expansion is also possible by the use of an increasing delay characteristic in the LDF, and in such a case, the filter will become a time-expansion filter.

2.2 Impulse response of a general LDF

We consider the general case of an idealized LDF as shown in Fig. 2a. The bandwidth of the filter is Δf centered at f_0. The group delay varies linearly over the passband as shown in the diagram. The transfer function of the filter, using analytic-signal notation, is

$$H(f) = \begin{cases} \exp\left\{-j2\pi\left[(f-f_0)\tau_0 + \dfrac{(f-f_0)^2}{2\alpha} + \phi_1\right]\right\}, \\ \qquad -\dfrac{\Delta f}{2} \leq f - f_0 \leq \dfrac{\Delta f}{2}, \\ 0, \qquad \text{elsewhere,} \end{cases} \tag{3}$$

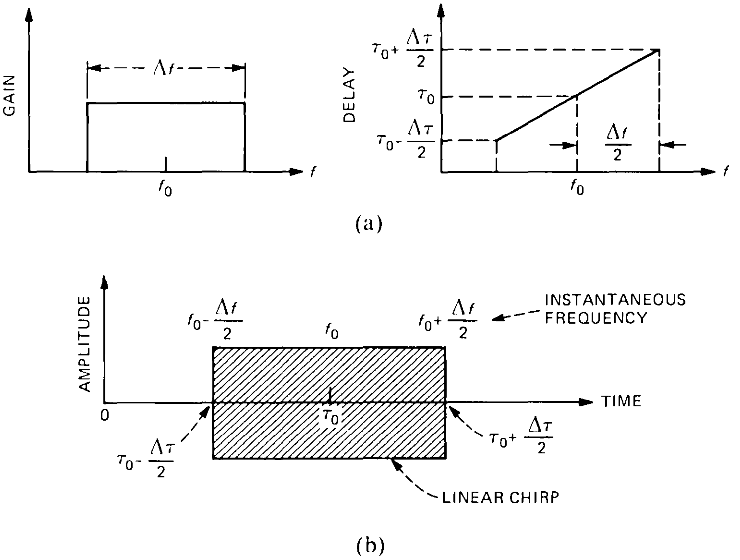

GAIN
Δf
f_0
f
DELAY
$\tau_0 + \frac{\Delta\tau}{2}$
τ_0
$\tau_0 - \frac{\Delta\tau}{2}$
$\frac{\Delta f}{2}$
f_0
f
(a)
AMPLITUDE
$f_0 - \frac{\Delta f}{2}$
f_0
$f_0 + \frac{\Delta f}{2}$
INSTANTANEOUS FREQUENCY
TIME
0
τ_0
$\tau_0 - \frac{\Delta\tau}{2}$
$\tau_0 + \frac{\Delta\tau}{2}$
LINEAR CHIRP
(b)

Fig. 2—Characteristics and impulse response of an LDF. (a) Transfer function. (b) Impulse response.

where

$$\alpha \triangleq \frac{\Delta f}{\Delta\tau}, \tag{4}$$

τ_0 is the group delay at f_0, and ϕ_1 is a constant. The inverse Fourier transform of $H(f)$ gives the impulse response $h(t)$. The derivation for $h(t)$ is given in Appendix A. Neglecting some small envelope and phase perturbations, a good approximation (sufficient for our present consideration) for $h(t)$ is

$$h(t) = \begin{cases} \cos 2\pi\left[\left(f_0 - \frac{\Delta f}{2}\right)t + \frac{\alpha}{2}t^2 + \phi_2\right], \\ \qquad \tau_0 - \frac{\Delta\tau}{2} \le t \le \tau_0 + \frac{\Delta\tau}{2}, \\ 0, \qquad \text{elsewhere}, \end{cases} \tag{5}$$

where the multiplying constant has purposely been dropped, and ϕ_2 is a constant. A sketch of $h(t)$ is shown in Fig. 2b. It is a linear chirp waveform starting at $\tau_0 - \Delta\tau/2$ and ending at $\tau_0 + \Delta\tau/2$, with the instantaneous frequency varying from $f_0 - \Delta f/2$ to $f_0 + \Delta f/2$ at a chirp rate of α.

2.3 Analysis of time compression

In this analysis, we again restrict our discussion to a single scan line for simplicity. We denote this input scan line by $A(t)$ in the interval $(0, T_l)$. Referring to Fig. 1, the linear chirp signal is given by

$$c(t) = \cos 2\pi\left[\left(f_0 - \frac{\Delta f_c}{2}\right)t + \frac{\beta}{2}t^2 + \phi_3\right], \qquad 0 \le t \le T_l, \tag{6}$$

where β is the chirp rate defined by

$$\beta = \frac{\Delta f_c}{T_l}, \tag{7}$$

and ϕ_3 is a constant. The parameters f_0 and Δf_c are the chirp center frequency and deviation, respectively. There are also two implicit assumptions: (*i*) $f_0 \gg \Delta f_c$; and (*ii*) $f_0 \gg$ the highest frequency in the TV signal.

The LDF input is the product of the TV input $A(t)$ and $c(t)$, i.e.,

$$x(t) = A(t)\cos 2\pi\left[\left(f_0 - \frac{\Delta f_c}{2}\right)t + \frac{\beta}{2}t^2 + \phi_3\right], \qquad 0 \le t \le T_l. \tag{8}$$

To obtain the LDF output, we can either use the time-domain approach by convolving $x(t)$ with the impulse response $h(t)$ of the LDF, or use frequency domain analysis by multiplying the Fourier transform of $x(t)$ by the LDF transfer function and then performing an inverse transform. The former method is much simpler and is presented here. The latter is tedious, offers no additional insight, and is, therefore, deleted for brevity.

The LDF is assumed to have an extended passband over $(f_0 - \Delta f/2, f_0 + \Delta f/2)$, where $\Delta f > \Delta f_c$. Its delay characteristic is shown in Fig. 3. Note that the delay slope is given by

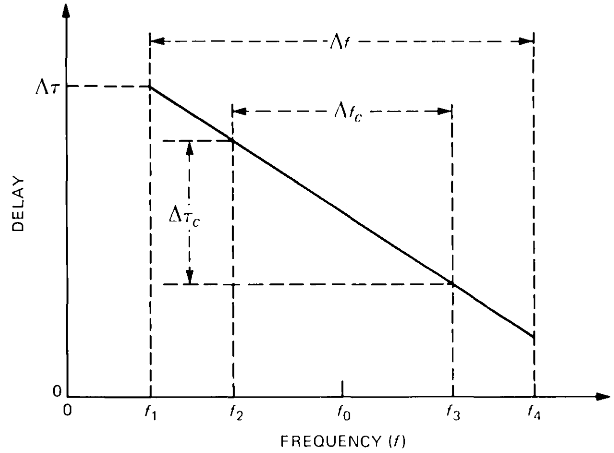


Fig. 3—Delay characteristic of the LDF.

$$\alpha^{-1} = \frac{\Delta\tau}{\Delta f}. \tag{9}$$

Also note in Fig. 3 that the delay at $f_0 + \Delta\tau/2$ is zero, as some constant delay through the device has been dropped for simplicity. Using the result of Section 2.2, the impulse response of the LDF is

$$h(t) = \begin{cases} \cos 2\pi\left[f_1 t - \frac{\alpha}{2} t^2 + \phi_4 \right], & 0 \le t \le \Delta\tau, \\ 0, & \text{elsewhere}, \end{cases} \tag{10}$$

where $f_1 \triangleq f_0 - \Delta f/2$, and ϕ_4 is a constant. The LDF output is then

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} x(\tau) h(t-\tau) d\tau \\ &= \int_{T_1}^{T_2} \frac{A(\tau)}{2} \\ &\quad \times \left\{ \cos 2\pi \left[\Phi(t) + (\alpha t - \Delta f_c - f_4 + f_3)\tau + (\beta - \alpha)\frac{\tau^2}{2} \right] + \right. \\ &\quad \left. \cos 2\pi \left[\Phi(t) + (2f_0 - \alpha t + f_4 - f_3)\tau + (\beta + \alpha)\frac{\tau^2}{2} \right] \right\} d\tau, \\ &\quad 0 \le t \le T_l + \Delta\tau, \end{aligned} \tag{11}$$

where the limits of integration are defined by

$$T_1 \triangleq \max(0, t - \Delta\tau), \tag{12}$$

$$T_2 \triangleq \min(t, T_l), \tag{13}$$

f_3 and f_4 are defined in Fig. 3, and

$$\Phi(t) \triangleq f_4 t - \frac{\alpha}{2} t^2 + \phi_4. \tag{14}$$

A constant phase term has been dropped in (eq. 11), and we will neglect all unimportant constant multipliers and phase shifts in subsequent discussions. The integral of the second cosine term in eq. (11) can be discarded because of the high frequency component $2f_0\tau$ in the integrand. Therefore, $y(t)$ becomes

$$\begin{aligned} y(t) = \int_{T_1}^{T_2} A(\tau) \cos 2\pi \Big[&\Phi(t) + (\alpha t - \Delta f_c - f_4 + f_3)\tau \\ &+ (\beta - \alpha)\frac{\tau^2}{2} \Big] d\tau, \qquad 0 \le t \le T_l + \Delta\tau. \end{aligned} \tag{15}$$

Let us examine the above integral expression carefully. Using analytic-signal notation, we rewrite it as

$$\begin{aligned} y(t) &= \int_{T_1}^{T_2} A(\tau) \exp\left\{ j2\pi \left[\Phi(t) + (\alpha t - \Delta f_c - f_4 + f_3)\tau \right.\right. \\ &\quad \left.\left. + (\beta - \alpha)\frac{\tau^2}{2} \right] d\tau \right\} \\ &= \exp j2\pi\Phi(t) \int_{T_1}^{T_2} A(\tau) \exp\left\{ j2\pi \left[(\alpha t - \Delta f_c - f_4 + f_3)\tau \right.\right. \\ &\quad \left.\left. + (\beta - \alpha)\frac{\tau^2}{2} \right] d\tau \right\}, \\ &\quad 0 \le t \le T_l + \Delta\tau. \end{aligned} \tag{16}$$

In this form, we can define a complex envelop for $y(t)$ as

$$A_y(t) \triangleq \int_{T_1}^{T_2} A(\tau) \exp\left\{ j2\pi \left[(\alpha t - \Delta f_c - f_4 + f_3)\tau + (\beta - \alpha)\frac{\tau^2}{2} \right] d\tau \right\}, \qquad 0 \le t \le T_l + \Delta\tau. \tag{17}$$

In the above, note that $A(\tau)$ is a slow-moving function while the exponential term contains a highly oscillatory chirp given by the derivative of the bracketed argument with respect to τ, i.e.,

$$f_i(\tau) = (\alpha t - \Delta f_c - f_4 + f_3) + (\beta - \alpha)\tau, \qquad T_1 \le \tau \le T_2. \tag{18}$$

This can also be obtained by simply taking the difference of the instantaneous frequencies of $x(\tau)$ and $h(t - \tau)$ in eq. (11). The convolution integral involved is illustrated in Fig. 4, where we show $x(\tau)$ and also the corresponding $f_i(\tau)$ at a fixed $t = t_1$. It can be seen that $A_y(t_1)$ is given by an integral over (T_1, T_2) of a linear chirp waveform at a chirp rate of $(\beta - \alpha)$ and with an envelope modulation $A(\tau)$. Furthermore, the chirp frequency f_i inside (T_1, T_2) may vanish at some τ as shown in Fig. 4. In such a case, the value of $y(t_1)$ is dominated by the integral eq. (15) over the small interval surrounding that τ, where f_i goes to zero. This is, of course, the well-known quasi-stationary approximation. The approximation is good if the chirp rate $(\beta - \alpha)$ and the interval (T_1, T_2) are large, and $A(\tau)$ variation is slow by comparison. Using this approximation, at $t = t_1$ inside the valid interval $(0, T_l + \Delta\tau)$,

$$A_y(t_1) \approx kA(\tau_1), \tag{19}$$

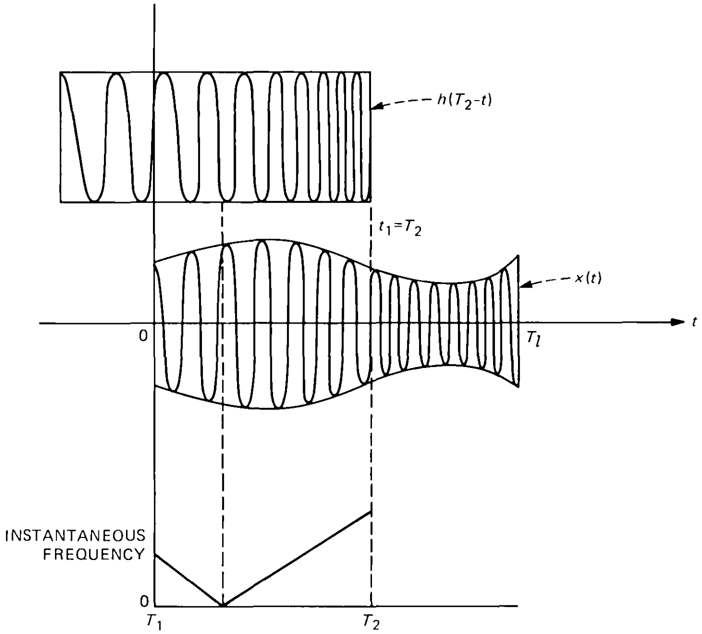


Fig. 4—Convolution integral.

where k is a constant, and τ_1 is obtained by solving eq. (18) with f_i set to zero and $t = t_1$, i.e.,

$$\tau_1 = \frac{\alpha t_1 - \Delta f_c - f_4 + f_3}{\beta - \alpha}. \tag{20}$$

This quasi-stationary approximation is indeed equivalent to the physical interpretation described in Section 2.1. As a check, let us derive the TCR from the approximation above. We know that $A(\tau_1)$ is nonzero only if $0 \leq \tau_1 \leq T_l$. The corresponding t_1 can be solved for using eq. (20), and the end points of t_1 constitute the interval T_c, i.e.,

$$\begin{aligned} T_c &= \frac{(\Delta f_c + f_4 - f_3) - T_l(\beta - \alpha)}{\alpha} - \frac{(\Delta f_c + f_4 - f_3)}{\alpha} \\ &= T_l\left(1 - \frac{\beta}{\alpha}\right). \end{aligned} \tag{21}$$

Therefore,

$$\text{TCR} = \left(1 - \frac{\beta}{\alpha}\right)^{-1}, \qquad \alpha \geq \beta, \tag{22}$$

which agrees with eq. (2) with the substitutions of α and β according to the definition eqs. (4) and (7), respectively. See Section 2.4 for a continued discussion of $A_y(t)$.

To finish our analysis of the time compression, we return to $y(t)$ in eq. (15) and expand the cosine term:

$$y(t) = y_c(t)\cos 2\pi\Phi(t) - y_s(t)\sin 2\pi\Phi(t), \tag{23}$$

where

$$y_c(t) \triangleq \int_{T_1}^{T_2} A(\tau)\cos 2\pi\left[(\alpha t - \Delta f_c - f_4 + f_3)\tau + (\beta - \alpha)\frac{\tau^2}{2}\right]d\tau \tag{24}$$

and

$$y_s(t) \triangleq \int_{T_1}^{T_2} A(\tau)\sin 2\pi\left[(\alpha t - \Delta f_c - f_4 + f_3)\tau + (\beta - \alpha)\frac{\tau^2}{2}\right]d\tau. \tag{25}$$

In eq. 25, $y_c(t)$ and $y_s(t)$ can be viewed as the in-phase and quadrature components of $y(t)$ after synchronous demodulation. The envelope detector output is then

$$z(t) = [y_c^2(t) + y_s^2(t)]^{1/2}. \tag{26}$$

2.4 Some fundamental properties

We have just derived mathematical expressions for the time-compression process. These expressions are too complicated for easy interpretation. However, we have demonstrated that the technique works if a quasi-stationary approximation is made for the chirp waveform. In other words, the instantaneous frequency of the chirp wave could be used as if it were a stationary carrier frequency in a steady state analysis.

Making such an assumption, it is seen from eq. (2) that infinite compression, TCR $= \infty$, results if $\Delta f = \Delta f_c$ and $\Delta\tau = T_l$ $(\alpha = \beta)$. But from eqs. (15), (24), and (25), it is obvious that the LDF output after synchronous detection (for $\alpha = \beta$) will actually become the Fourier transform of the input envelope $A(t)$. This can also be recognized as the well-known chirp transform or real-time transform[4,5] commonly used in chirp radar and SAW processors. Therefore, the quasi-stationary model is invalid for this case.

A case where the quasi-stationary approximation clearly holds is where $\Delta f_c \rightarrow \infty$ and $\alpha^{-1} \rightarrow 0$, i.e., the chirp range is very large, and the delay slope of the LDF is close to zero. The result is, of course, a very slight compression, i.e., the TCR is slightly larger than 1. Therefore, without doing any specific calculation, we see that the quasi-stationary assumption is valid, at best, for small TCRs, and it breaks down somewhere between TCR = 1 and ∞. Since our practical applications require TCR ≥ 2, the case of TCR $= 2$ becomes most interesting and is investigated in the simulation later.

Let us now assume the quasi-stationary model and see what kind of distortion would result. The input to the LDF can be expressed as a multitone AM signal:

$$\tilde{x}(t) = \left(1 + \sum_{j=1}^{N} m_j \cos\omega_j t\right)\cos\omega_c t, \tag{27}$$

where ω_j are the angular modulating frequencies, and ω_c is the angular carrier frequency at some fixed instant of time. Let τ_c be the delay of the LDF at $\omega = \omega_c$. The phase characteristic of the LDF is

$$\phi(\omega) = -(\tau_c + \omega_c/\alpha)\omega + \omega^2/(2\alpha) + c, \tag{28}$$

where c is a constant, and α is defined in eq. (9). The LDF output is

$$\begin{aligned} \tilde{y}(t) = {} & \cos\left[\omega_c t + (\tau_c + \omega_c/\alpha)(-\omega_c) + \frac{\omega_c^2}{2\alpha} + c\right] \\ & + \sum_{j=1}^{N} \frac{m_j}{2}\left\{\cos\left[\omega_c t + \omega_j t - \left(\tau_c + \frac{\omega_c}{\alpha}\right)(\omega_c + \omega_j) + \frac{(\omega_c + \omega_j)^2}{2\alpha} + c\right]\right. \\ & \left. + \cos\left[\omega_c t - \omega_j t - \left(\tau_c + \frac{\omega_c}{\alpha}\right)(\omega_c - \omega_j) + \frac{(\omega_c - \omega_j)^2}{2\alpha} + c\right]\right\}. \end{aligned} \tag{29}$$

In the above expression, various sidebands of the input AM signal are delayed asymmetrically about the carrier frequency. This would, of course, distort the signal. After some tedious manipulations, the envelope of $\tilde{y}(t)$ is found to be

$$\begin{aligned} \tilde{Y}(t) = {} & \left\{\left[1 + \sum_{j=1}^{N} m_j \cos\omega_j(t - \tau_c)\cos\left(\frac{\omega_j^2}{2\alpha}\right)\right]^2 + \right. \\ & \left. \left[\sum_{j=1}^{N} m_j \cos\omega_j(t - \tau_c)\sin\left(\frac{\omega_j^2}{2\alpha}\right)\right]^2\right\}^{1/2}. \end{aligned} \tag{30}$$

Comparing $\tilde{Y}(t)$ and $\tilde{x}(t)$, it is obvious that distortionless transmission results only if $\omega_j^2/(2\alpha) \ll 1$. More importantly, the distortion in $\tilde{Y}(t)$ is dependent on the input signal and, therefore, cannot be equalized easily. On the other hand, if synchronous demodulation is used in place of envelope detection, we obtain the first square bracket term in eq. (30), i.e.,

$$\tilde{Y}_s(t) = 1 + \sum_{j=1}^{N} m_j \cos\omega_j\,(t - \tau_c)\cos\left(\frac{\omega_j^2}{2\alpha}\right), \tag{31}$$

where the distortion shows up as $\cos[\omega_j^2/(2\alpha)]$ which is independent of the input signal, and equalization is thereby feasible. To do synchronous demodulation, we need to know the instantaneous frequency of the carrier in $y(t)$, the LDF output, which we have not addressed so far. Let us do so in the following.

Stating eq. (16) again,

$$y(t) = \exp[j2\pi\Phi(t)] \int_{T_1}^{T_2} A(\tau) \exp\left[j(\beta - \alpha) \frac{\tau^2}{2} \right] \cdot \exp[j2\pi(\alpha t - \Delta f_c - f_4 + f_3)\tau] d\tau. \quad (32)$$

In the above we have, in part, an output chirp frequency of $d\Phi/dt = (f_4 - \alpha t)$. The integral part, on the other hand, can be interpreted in two different ways: (*i*) It is a "quadratic" chirp transform of $A(\tau)$. As such, little is known about this transform. (*ii*) It is the Fourier transform of $A(\tau) \exp[j2\pi(\beta - \alpha)\tau^2/2]$, where $A(\tau)$ can be viewed as an envelope modulation on a chirp signal with frequency $(\beta - \alpha)\tau$. Equivalently, it is the convolution of the Fourier transform of $A(\tau)$ and that of the chirp signal. The result may very well contain high frequency components depending on the magnitude of $(\beta - \alpha)$ and the shape of $A(\tau)$. In fact, simulation results show that the instantaneous frequency of $y(t)$ can be quite different from $(f_4 - \alpha t)$. Therefore, frequency predictability is difficult in the general case, and synchronous detection as discussed above cannot be used easily.

The following is a summary of the analytic results:

(*i*) The mathematical expressions are complicated, and simulation is necessary to obtain numerical insights.

(*ii*) The compression operation can be justified by a quasi-stationary model. Under this model and with envelope detection, multitone AM leads to nonequalizable distortion. Furthermore, the quasi-stationary model is valid only for small TCRs.

(*iii*) Synchronous detection is very difficult because of frequency unpredictability.

III. SIMULATION

This section describes simulation results. These numerical results show that the proposed technique indeed time compresses the input signal, but the output distortion is probably unacceptable for TV transmission using today's SAW devices.

3.1 Preliminary set-up

A computer subroutine was written to simulate the time-compression operation. It accepts an arbitrary input $A(t)$ defined in the interval $(0, T_l)$ and outputs $y(t)$, $y_c(t)$, and $y_s(t)$, given by eqs. (15), (21), and (22), respectively. Note that $y(t)$ is the RF output of the LDF, and $y_c(t)$ and $y_s(t)$ are the in-phase and quadrature components after synchronous demodulation with $\Phi(t)$ so that the ideal envelope detector output $z(t)$ can be easily calculated from eq. (23). Also note that all three outputs have the same integral form, i.e.,

$$I = \int_{t_1}^{t_2} f(t)\cos[2\pi(at^2 + bt + c)]dt, \tag{33}$$

where we can assume $t_1 \geq 0$, $t_2 \geq t_1$, and $a > 0$. In the computer program, we break the interval $(0, T_l)$ into many small segments such that within each segment, a linear approximation of $A(t)$ is valid. After doing so, the integration over each of these segments can be done in the form of eq. (33) with

$$f(t) \approx mt + k. \tag{34}$$

The limits of integration t_1 and t_2 in eq. (33) are, of course, the end points of the small time segment. Using the linear representation eq. (34) for $f(t)$, I in eq. (33) can be integrated in closed form in terms of Fresnel integrals, and the result is

$$I = mI_2 + kI_1, \tag{35}$$

where

$$I_1 = [\pi/(2p)]^{1/2}\{[C(z_2) - C(z_1)]\cos B - [S(z_2) - S(z_1)]\sin B\}, \tag{36}$$

$$\begin{aligned} I_2 = {} & \frac{\cos B}{2p}\{\sin x_2^2 - \sin x_1^2 - q[\pi/(2p)]^{1/2}[C(z_2) - C(z_1)]\} \\ & + \frac{\sin B}{2p}\{\cos x_2^2 - \cos x_1^2 + q[\pi/(2p)]^{1/2}[S(z_2) - S(z_1)]\}, \end{aligned} \tag{37}$$

and

$$p = 2\pi a; \quad q = 2\pi b; \quad r = 2\pi c, \tag{38}$$

$$B = r - \left(\frac{q}{2\sqrt{p}}\right)^2, \tag{39}$$

$$x_1 = \sqrt{p}t_1 + \frac{q}{2\sqrt{p}}, \tag{40}$$

$$x_2 = \sqrt{p}t_2 + \frac{q}{2\sqrt{p}}, \tag{41}$$

$$z_1 = \frac{x_1}{\sqrt{x/2}}, \tag{42}$$

$$z_2 = \frac{x_2}{\sqrt{\pi/2}}. \tag{43}$$

The Fresnel integrals in eq. (36) are defined by

$$C(z) = \int_0^z \cos\left(\frac{\pi}{2}x^2\right) dx \tag{44}$$

and

$$S(z) = \int_0^z \sin\left(\frac{\pi}{2}x^2\right) dx. \tag{45}$$

There are also some simpler cases, e.g., $a = 0$ or $b = 0$, which are not shown for brevity. Since time expansion can be achieved by reversing the delay slope of the LDF, the subroutine for simulating the time compression can also be used to simulate the time-expansion filter. Examples of both time compression and expansion will be shown later.

In all subsequent simulations, the input is always

$$A(t) = \begin{cases} 1, & 0 \le t \le T_l \qquad (T_l = 64\ \mu\text{s}) \\ 0, & \text{otherwise} \end{cases}. \tag{46}$$

This rectangular pulse is, of course, not a representative of the video signal. However, except for edge "ringings," the system should compress the pulse properly, and peak-to-peak (p-p) ripples at the center portion of the output pulse should give an indication of the magnitude of distortion involved. To examine the outputs $y(t)$, $y_c(t)$, and $y_s(t)$, we have also developed a set of programs to estimate the instantaneous frequency at various time instants of $y(t)$, as well as the p-p ripples of the output pulse.

Because of the complexity of the equations and various computer routines, it is not easy to assure that there is no bug in the programs. However, we did make some runs for the special case $\alpha = \beta$ (Fourier transform case) where results are known. The waveforms $y_c(t)$, $y_s(t)$, and $z(t)$, and the output chirp frequency and offset frequency of the LDF given by eq. (31) were all verified carefully. Such a check assures the validity of the computer programs.

3.2 Examples of time compression

Four examples of time compression are discussed in this section. They all have TCR = 2 (i.e., 2:1 compression) and an input given by eq. (46) (see Fig. 5a). The chirp frequency range Δf_c for the input $x(t)$ to the LDF is (f_2, f_3), while the passband Δf of the LDF extends over (f_1, f_4). The delays at f_i are denoted by τ_i ($i = 1$ to 4), respectively. The delay dispersion of the LDF is defined by $\Delta\tau = \tau_1 - \tau_4$. The descriptions for these examples are as follows:

(*i*) The key parameters for the LDF are shown in Fig. 5b. In this example, we let $f_1 = f_2$, $f_3 = f_4$ and choose

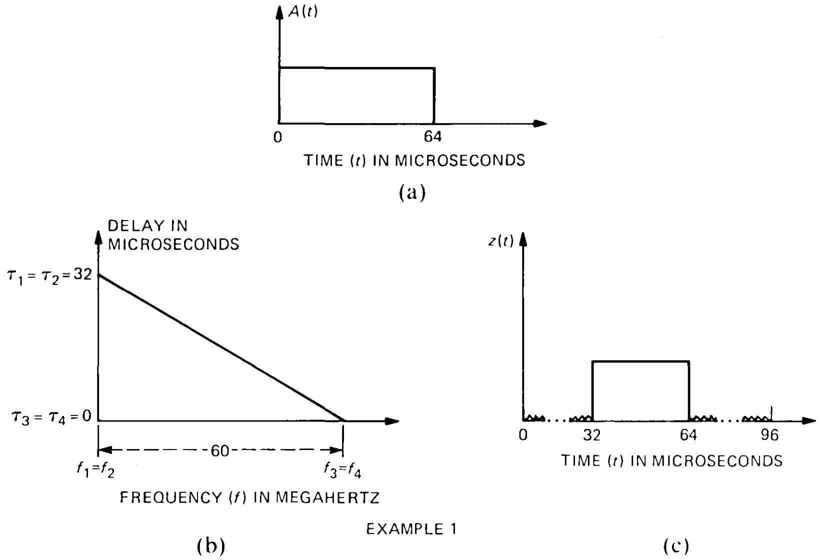


Fig. 5—Parameters for time-compression technique. Example 1—(a) Input pulse (t); (b) Delay of LDF; (c) Expected output (t).

$$\Delta f_c = \Delta f = 60 \text{ MHz}. \tag{47}$$

From eq. (2), TCR = 2 with $\Delta\tau = 32$ μs. The time-bandwidth product (BT) of the LDF is defined by

$$\text{BT} = (\Delta f)(\Delta\tau). \tag{48}$$

Here, BT = 1920. Bandwidth product $\approx$ 10,000 to 50,000 is considered as a practical range for SAW filters. The "expected" output is illustrated in Figure 5c where the output compressed pulse appears between t = 32 μs and 64 μs. The small ripples in the time intervals, 0 to 32 μs and 64 to 96 μs are to illustrate the nonzero output of the LDF in these regions.

(*ii*) The LDF parameters are shown in Fig. 6a and the expected output in Fig. 6b. Here, $\Delta f_c = 60$ MHz which is the same as in Example 1 (Fig. 5), but $\Delta\tau$ and Δf are both increased by a factor of 3. With $\Delta f = 180$ MHz and $\Delta\tau = 96$ μs, BT = 17,280.

(*iii*) The parameters are shown in Fig. 7a, and the expected output is the same as that of Example 2 (Fig. 6) because the delay dispersion has remained the same. We increased Δf to 600 MHz yielding BT = 57,600, which is probably a little beyond present-day state of the art. We used $\Delta f_c = 200$ MHz.

(*iv*) The parameters are shown in Figs. 6a and 6b. This is essentially the same as Example 3 in Fig. 7a, except Δf is increased to 1200 MHz,

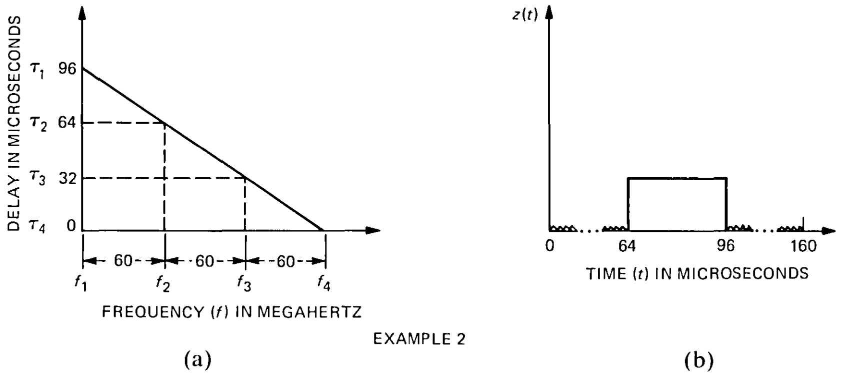


Fig. 6—Parameters for time-compression technique. Example 2—(a) Delay of LDF; (b) Expected output (t).

and BT = 115,200. This is probably not realizable in the near future. We used Δf_c = 400 MHz.

The above examples are arranged to illustrate the effect of increasing Δf (or the strengthening of the quasi-stationary model).

With the input being a pulse 64-μs long, the expected output for all examples is a compressed pulse, 32 μs in duration (TCR = 2). This is indeed so from the simulation results which show a compressed pulse approximately 32 μs long in the expected time slot. The output outside the compressed pulse duration is at least an order of magnitude lower. However, inside this compressed pulse, there are ripples created by the compression process, and these ripples are the resulting distortion that we are trying to estimate. The ripples are largest at the edges of the compressed pulse and smallest toward the center. Also, the ripples at the center are indications of the "best" performance of the time-compression filter. In our computer routines to estimate distortions,

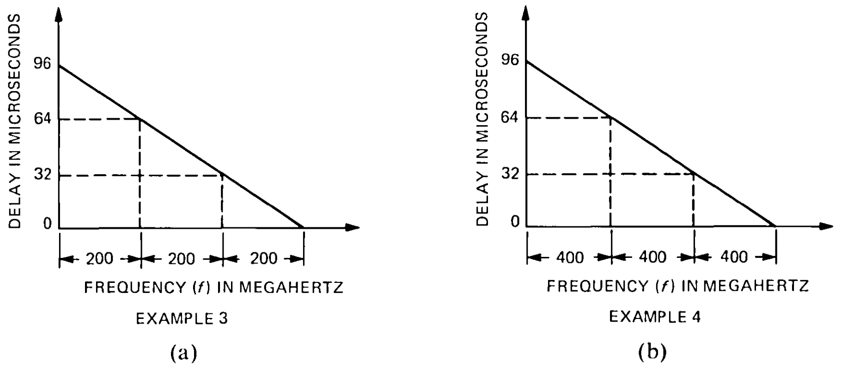


Fig. 7—Parameters for the time-compression technique. Example 3—(a) LDF (600 MHz). Example 4—(b) LDF (1200 MHz).

we take some symmetric time interval, say $\pm\Delta T_d$, about the center of the time-compressed pulse, and we record the maximum and minimum pulse magnitudes, denoted by $z_{\max}$ and $z_{\min}$, respectively. The *p-p* ripple distortion (over $\pm\Delta T_d$) is defined by

$$\mathrm{RD} \triangleq 20 \log \left[\frac{z_{\max} - z_{\min}}{(z_{\max} + z_{\min})/2} \right], \tag{49}$$

where RD is in dB. Both $z_{\max}$ and $z_{\min}$ are positive because $A(t)$ is a positive pulse. We plot the *p-p* ripple distortion versus ΔT_d in Fig. 8. The largest ΔT_d is 16 μs because that is the edge of the compressed pulse. Although it is difficult to translate the meaning of RD to a TV quality measure, it can be certain that a large RD (i.e., large ripple) would mean poor transmission quality. To make a TV system workable, an RD of less than −45 dB is probably necessary, although other applications may not require such a low distortion.

Referring to Fig. 8, it is obvious that the distortion gets progressively smaller from examples 1 to 4. But the lowest distortions, despite the large BTs involved, are still too excessive for high quality TV transmission. Some more examples are provided in Appendix B for additional insight into the problem of ripple distortion.

3.3 An example of time compression and expansion

In this example, we use practical design parameters involving both time compression and time expansion (Fig. 9). Figure 9a shows a 64-

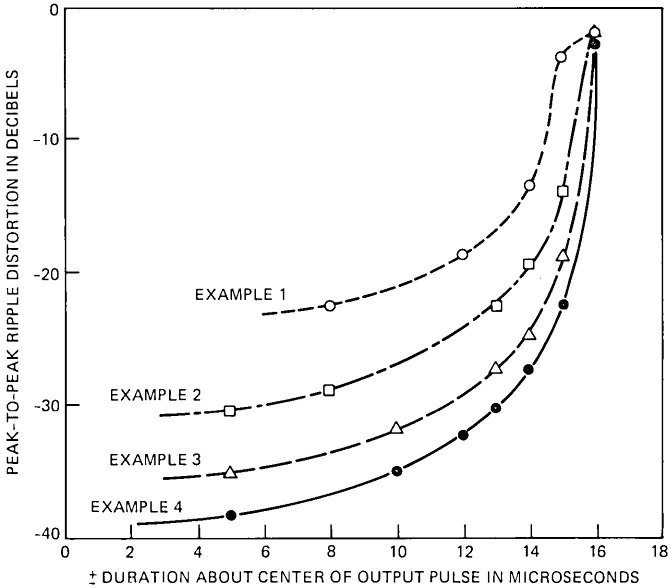


Fig. 8—Distortion results for time-compression examples.

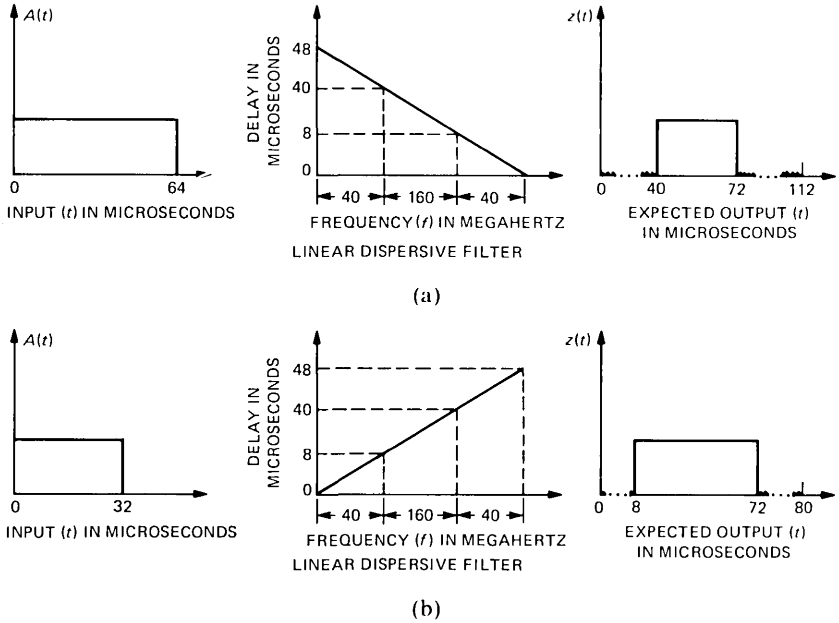

A(t)
0
64
INPUT (t) IN MICROSECONDS
DELAY IN MICROSECONDS
48
40
8
0
40
160
40
FREQUENCY (f) IN MEGAHERTZ
LINEAR DISPERSIVE FILTER
z(t)
0
40
72
112
EXPECTED OUTPUT (t) IN MICROSECONDS
(a)
A(t)
0
32
INPUT (t) IN MICROSECONDS
DELAY IN MICROSECONDS
48
40
8
0
40
160
40
FREQUENCY (f) IN MEGAHERTZ
LINEAR DISPERSIVE FILTER
z(t)
0
8
72
80
EXPECTED OUTPUT (t) IN MICROSECONDS
(b)

Fig. 9—Parameters for (a) time-compression and (b) time-expansion examples.

μs pulse as an input to the compressor, and the expected output. The actual ripple distortion (i.e., simulated result) over the 32-μs compressed pulse is shown in Fig. 10a. To check the operation of the expander alone, we assume an input of a 32-μs pulse to the expander. The expected result is shown in Fig. 9b, and the simulated ripple distortion is plotted in Fig. 10b. It is clear from Figs. 10a and 10b that both the compressor and the expander lead to considerable distortion.

To simulate a total system with compression and expansion, we used the distorted 32-μs pulse from the compressor as an input to the expander. The expanded output pulse is so distorted that it becomes meaningless to make a ripple distortion plot as before. Instead, we plot a small segment near the center of the expanded pulse in Fig. 10c to illustrate its excessive distortion. The performance of this total system is definitely not suitable for TV transmission. Some filterings were tried at the output of the expander, but little improvement was obtained.

IV. CONCLUSION

We have studied a time-compression technique based on a simple configuration of SAW filters. The technique was derived heuristically and can be viewed as a quasi-stationary model for the chirp signals. Numerical results show that excessive distortion is created, and its

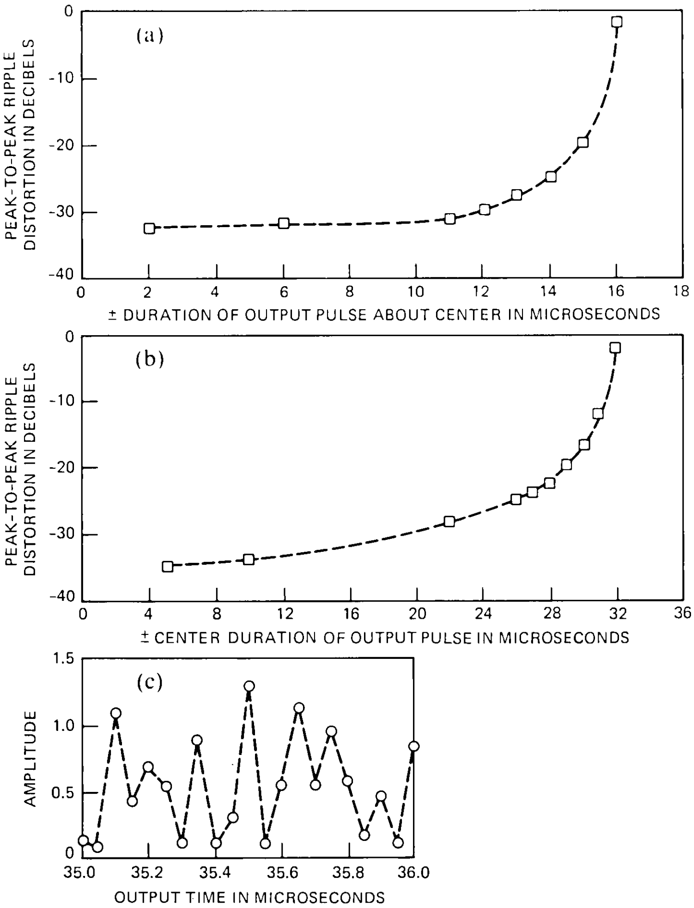


Fig. 10—Results for time-compression and time-expansion examples. (a) Ripple distortion of time-compression example. (b) Ripple distortion of time-expansion example. (c) Pulse ripple for time-compression and time-expansion examples.

application to TV transmission is not suitable unless some kind of equalization is provided. One such equalization is the chirp transform processor[3] which involves considerably more complexity. Simpler equalizations may be possible but do not seem to be straightforward. As for other applications where the distortion requirement is less stringent, this approach may be feasible.

V. ACKNOWLEDGMENTS

The authors wish to thank O.-C. Yue and L. A. Coldren for many helpful discussions. Some interesting discussions with T. Meeker at the early part of the study are also gratefully acknowledged.

APPENDIX A

Impulse Response of a Linear Dispersive Filter

A linear dispersive filter (LDF) is a filter having a linear delay characteristic. Using the analytic-signal notation, its transfer function can be written as

$$G(\omega) = \begin{cases} 1, & -\dfrac{\Delta\omega}{2} \le \omega - \omega_0 \le \dfrac{\Delta\omega}{2}, \\ 0, & \text{elsewhere} \end{cases} \tag{50}$$

and

$$D(\omega) = (\tau_0 - s\omega_0) + s\omega, \qquad -\frac{\Delta\omega}{2} \le \omega - \omega_0 \le \frac{\Delta\omega}{2}, \tag{51}$$

where $G(\omega)$ and $D(\omega)$ are the gain and group delay, respectively; ω_0 is the angular center frequency, and $\Delta\omega$ is the bandwidth in radians; τ_0 is the delay at ω_0; and $\Delta\tau$ is the delay dispersion over $\omega_0 + \Delta\omega/2$. The delay slope is defined by

$$s \triangleq \frac{\Delta\tau}{\Delta\omega}. \tag{52}$$

Integrating $D(\omega)$ with respect to ω, we obtain the phase,

$$\phi(\omega) = -(\tau_0 - s\omega_0)\omega - \frac{s}{2}\,\omega^2, \qquad -\frac{\Delta\omega}{2} \le \omega - \omega_0 \le \frac{\Delta\omega}{2}, \tag{53}$$

where the integration constant has been dropped for convenience. Using exponential notation and neglecting all unimportant multiplying constants and constant phase shifts in subsequent discussions, the impulse response of the LDF is

$$h(t) = \operatorname{Re} \int_{\omega_0 - \frac{\Delta\omega}{2}}^{\omega_0 + \frac{\Delta\omega}{2}} \exp\left\{-j\left[(\tau_0 - s\omega_0)\omega + \frac{s}{2}\,\omega^2\right]\right\} \exp(j\omega t)\,d\omega. \tag{54}$$

The above is simply an inverse Fourier transform of the transfer function. By a change of variable, it can be rewritten as

$$h(t) = \operatorname{Re} \exp(j\omega_0 t) \int_{\frac{\Delta\omega}{2}}^{\frac{\Delta\omega}{2}} \exp\left[-j\left(\tau_0\omega + \frac{s}{2}\,\omega^2\right)\right] \exp(j\omega t)\,d\omega. \tag{55}$$

There are two methods to solve for $h(t)$ according to the above: (*i*) extend the limits of integration to ±∞ and obtain a closed form solution easily; (*ii*) retain the finite integration limits and derive the result using Fresnel integrals. Both methods are shown briefly below.

Method I

Changing the limits of integration to ±∞ in eq. (55) and recognizing that the integration of the term $\exp(-j\tau_o\omega)$ leads to a delay of τ_0 in $h(t)$, we obtain an integral in the form of

$$J = \int_{-\infty}^{\infty} \exp(-jk\omega^2)\exp(j\omega t)d\omega, \tag{56}$$

where k is a constant. Putting the integrand in Gaussian form by completing the square, we get

$$J = \sqrt{\frac{\pi}{k}}\exp\left[j\left(\frac{t^2}{4k}-\frac{\pi}{4}\right)\right]. \tag{57}$$

Applying the above to eq. (55), we obtain the desired result

$$h(t) = \text{Re}\ \exp(j\omega_0 t)\exp\left[j\frac{(t-\tau_0)^2}{2s}\right], \tag{58}$$

or

$$h(t) = \cos\left[\omega_0 t + \frac{(t-\tau_0)^2}{2s}\right]. \tag{59}$$

However, because the original transfer function has a delay dispersion $\Delta\tau$ defined about τ_0, the valid range of t for eq. (59) is

$$\tau_0 - \frac{\Delta\tau}{2} \le t \le \tau_0 + \frac{\Delta\tau}{2}. \tag{60}$$

Method II

Setting $\omega_1 = \omega_0 - \Delta\omega/2$ and $\tau_1 = \tau_0 - \Delta\tau/2$, eq. (55) is written as

$$h(t) = \text{Re}\ \exp(j\omega_1 t)\int_0^{\Delta\omega} \exp\left[-j\left(\tau_1\omega + \frac{s}{2}\omega^2\right)\right]\exp(j\omega t)d\omega. \tag{61}$$

Again recognizing that $\exp(-j\tau_1\omega)$ leads to a delay of τ_1, we may substitute $\tau = t - \tau_1$ and

$$h(\tau) = \text{Re}\ \exp(j\omega_1\tau)\int_0^{\Delta\omega} \exp\left(-j\frac{s}{2}\omega^2\right)\exp(j\omega t)d\omega. \tag{62}$$

Completing the square in the integrand, we obtain

$$h(\tau) = \text{Re}\ \exp(j\omega_1\tau)\exp\left(j\frac{\tau^2}{2s}\right)\int_0^{\Delta\omega} \exp\left[-j\frac{s}{2}\left(\omega - \frac{\tau}{2}\right)^2\right]d\omega. \tag{63}$$

The above can be integrated using Fresnel integrals, and the result is

$$h(\tau) = \text{Re} \exp\left[j\left(\omega_1\tau + \frac{\tau^2}{2s}\right)\right]$$
$$\times\{[C(y_2) - C(y_1)] - j[S(y_2) - S(y_1)]\}, \quad (64)$$

where

$$y_1 \triangleq \sqrt{\frac{s}{\pi}}\Delta\omega\left(-\frac{\tau}{\Delta\tau}\right), \quad (65)$$

$$y_2 \triangleq \sqrt{\frac{s}{\pi}}\Delta\omega\left(1 - \frac{\tau}{\Delta\tau}\right). \quad (66)$$

$C(z)$ and $S(z)$ are the Fresnel integrals defined in eqs. (44) and (45). Consider the bracketed term:

$$[C(y_2) - C(y_1)] - j[S(y_2) - S(y_1)] \triangleq \rho(\tau)\exp[j\theta(\tau)]. \quad (67)$$

Neglecting small ripples, $\rho(\tau)$ is constant over $0 \le \tau \le \Delta\tau$ and vanishes outside this range. The phase term $\theta(\tau)$ is approximately constant over the same interval $0 \le \tau \le \Delta\tau$. Therefore, putting back the τ_1 delay, we have

$$h(t) \approx \text{Re} \exp\left\{j\left[\omega_1(t - \tau_1) + \frac{(t - \tau_1)^2}{2s}\right]\right\}. \quad (68)$$

APPENDIX B

Additional Examples On Ripple Distortion

Additional examples are provided here for more insight into the phenomenon of ripple distortion. The first case of interest is that of band-limiting effect on the input pulse. We use a 32-μs input pulse and low-pass filter it with a raised-cosine characteristic, where the gain is unity from zero to 8 MHz, 0.5 at 9 MHz, zero at 10 MHz, and the delay is constant over the passband. The output is, of course, a 32-μs pulse with ripples. The magnitude of these ripples are plotted in Fig. 11 in the manner similar to Figure 8.

The second case of interest is that of ripple distortion caused by the linear delay slope of the LDF. Three specific examples are provided to illustrate this effect:

Case A: The input to the time-compression filter is a 32-μs pulse. It is modulated by a CW frequency of f_0 = 1600 MHz. The LDF has a delay characteristic as shown in Fig. 1, where f_0 = 1600 MHz and Δf = 2400 MHz.

Case B: The conditions are identical to those of Case A, except f_0 = 1000 MHz and Δf = 1200 MHz.

Case C: The conditions are identical to those of Case A, except f_0 = 700 MHz and Δf = 600 MHz.

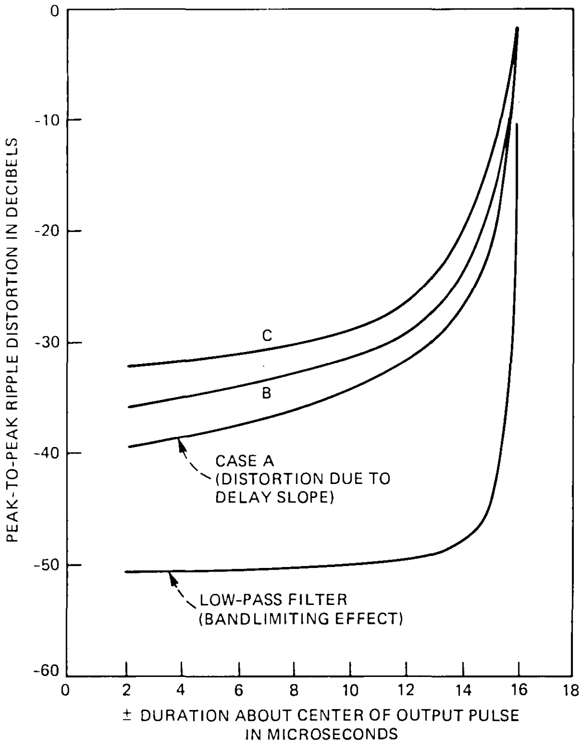


Fig. 11—Distortions because of bandlimiting and delay slope.

All three examples above illustrate the case of no time compression, but simply a constant delay through the LDF as viewed by the quasi-stationary model. Note that the magnitude of the delay slope increases by a factor of two from each example to the next and, hence, an increase in ripple distortion as seen from the results plotted in Fig. 11.

REFERENCES

1. J. E. Flood and D. I. Urguhart-Puller, "Time-Compression-Multiplex Transmission," Proc. IEE, *III*, No. 4 (April 1964), pp. 647–68.
2. D. H. Morgen and E. N. Protonotaries, "Time Compression Multiplexing For Loop Transmission of Speech Signals," IEEE Trans. Commun., *COM-22*, No. 12 (December 1974), pp. 1932–9.
3. Kai Y. Eng and On-Ching Yue, unpublished work.
4. J. R. Klauder et al., "The Theory and Design of Chirp Radars," B.S.T.J., *39*, No. 4 (July 1960), pp. 745–808.
5. M. A. Jack et al., "The Theory, Design, and Applications of Surface Acoustic Wave Fourier-Transform Processors," Proc. IEEE, *68*, No. 4 (April, 1980), pp. 450–68.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 10, December 1981
Printed in U.S.A.

Conformal Mapping and Complex Coordinates in Cassegrainian and Gregorian Reflector Antennas

By C. DRAGONE

(Manuscript received June 18, 1981)

In a Gregorian or Cassegrainian reflector antenna, the complex coordinate u' of an output ray is related to the corresponding input coordinate u by a bilinear transformation, $u' = (au + b)/(cu + d)$. We discuss the properties of this transformation, derive its coefficients a, b, c, d, and give explicitly the conditions that must be satisfied in order that symmetry be preserved. The conditions are expressed directly in terms of the parameters that specify the path of the principal ray, which is the ray corresponding to the feed axis. The results are directly related to well-known properties of stereographic projections, and they are shown to be useful in the design of multireflector antennas which minimize aberrations and cross-polarization.

I. INTRODUCTION

Gregorian and Cassegrainian reflector arrangements are needed for ground station and satellite antennas, and terrestrial radio relay systems.[1-8] In these antennas, a paraboloid of large aperture is combined with a smaller subreflector (an hyperboloid or an ellipsoid). The feed is placed at the antenna focal point, and it illuminates the subreflector with a spherical wave, which is then transformed by the two reflectors into a plane wave. Each input ray from the feed is thus transformed into an output ray parallel to the paraboloid axis.

This transformation can be represented by a stereographic projection.[9-11] Therefore, it is a conformal mapping—it transforms circles into circles, and it is described by the bilinear transformation

$$u' = \frac{au + b}{cu + d}, \tag{1}$$

where u is the "complex coordinate" of an input ray and u' the corresponding output coordinate. In this article, we discuss the properties of the bilinear transformation, derive its coefficients a, b, c, d, and give explicitly the conditions that must be satisfied to obtain circular symmetry, in which case

$$b = c = 0, \qquad d = 1. \tag{2}$$

The results are related to well-known properties of stereographic projections, and they generalize previous results in Refs. 12 to 18.

We first consider, in Section II, an ellipsoid illuminated by a spherical wave front S. We assume that S originates from one of the two foci of the ellipsoid, and determine the properties of a reflected wave front S', assuming geometric optics. We determine for each point P' of S' the corresponding point P of the incident wave front S and show that the correspondence $P \rightarrow P'$ is everywhere a conformal mapping. According to a well-known theorem of complex variables,[9] such a conformal mapping can be represented by the bilinear transformation (1), provided suitable complex variables u' and u are defined for the rays through P' and P. A suitable choice is obtained with two separate reference frames for P' and P using the familiar relations[9-13]

$$u = e^{j\phi}\tan\frac{\theta}{2} \qquad u' = e^{j\phi'}\tan\frac{\theta'}{2}, \tag{3}$$

where θ', ϕ' and θ, ϕ are spherical coordinates.

Since the two reference frames defining u and u' can be oriented arbitrarily, eq. (1) implies that an arbitrary rotation of the input frame must transform the input coordinate u according to a bilinear transformation.[10] The coefficients a, b, c, d of such a rotation are derived in Section V, where it is shown, for a rotation characterized by Euler angles α, β, γ, that

$$a = 1, \qquad b = -e^{j\alpha}\tan\frac{\beta}{2}, \tag{4}$$

$$c = e^{j\gamma}\tan\frac{\beta}{2}, \qquad d = e^{j(\alpha+\gamma)}. \tag{5}$$

Since an ellipsoid has an axis of symmetry, it is always possible to orient the input and output frames so as to reduce eq. (1) to the normal form

$$u' = au. \tag{6}$$

However, if the feed is centered around the z-axis of the input frame, then the reflected wave is blocked by the feed. For this reason, to avoid blockage, the z-axis must be tilted with respect to the ellipsoid axis.

Then, using eqs. (4) to (6) and properly orienting the input and output frames, it is shown in Section VI that eq. (1) assumes the form

$$u' = M\frac{u}{1 + (M - 1)u \tan i}, \tag{7}$$

where M and i are the magnification and angle of incidence for the principal ray corresponding to the feed axis (i.e., they ray $u = 0$).

The product of the two transformations given by eq. (6) and eqs. (4) to (5) gives the group of all possible transformations that can be obtained with an ellipsoid. One can show that this is the complete group of bilinear transformations. Thus, for any given values of the coefficients, a, b, c, d, it is always possible to find an ellipsoid (combined with suitable reference frames) which will produce the transformation (1) with the specified values of a, b, c, d. In Section VII, we consider an antenna consisting of N reflectors, each represented by a bilinear transformation. Obviously, the product of the N transformations is again a bilinear transformation and, therefore, the antenna can be represented by an equivalent ellipsoid.

Most antennas are focused at ∞. Then the equivalent ellipsoid becomes a paraboloid, and the antenna can be represented as shown in Fig. 1, showing a feed illuminating the equivalent paraboloid from its focus 0. The coefficients a, b, c, d in this case are obtained by letting $M \to 0$ in eq. (7), and it is shown in Section VI that we then obtain, for the complex coordinate $x' + jy'$ of an output ray,

$$x' + jy' = 2f\frac{u}{1 - u \tan i}, \tag{8}$$

where u is the input coordinate and $f = OI$ is the focal length of the equivalent paraboloid.

In Section VI, we derive a simple expression for the coefficient tan i in terms of the angles of incidence specifying the orientations of the various reflectors with respect to the principal ray. The value of tan i is needed in the design of a multibeam antenna to determine the aberrations caused by a small feed displacement from the focus. It is also needed to determine the output polarization, as shown in Section VIII.

Of special importance is the condition

$$\tan i = 0. \tag{9}$$

Then, using a corrugated feed[20] (or a feed with similar characteristics[21,22]) the output wave fronts become everywhere polarized in one direction. Furthermore, astigmatism is eliminated[19] for small feed displacements in a multibeam antenna. The above condition can

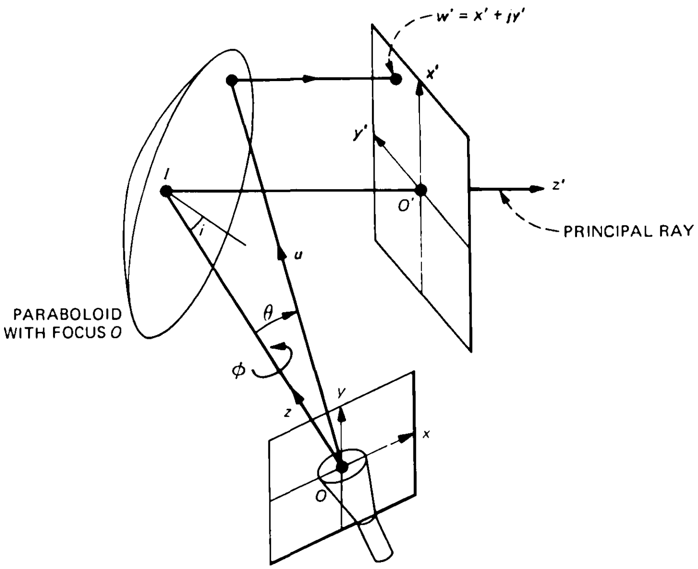


Fig. 1—A Cassegrainian or Gregorian reflector arrangement is represented by an equivalent paraboloid combined with a feed at 0. The principal ray corresponding to the feed axis is reflected at I with angle of incidence i.

always be satisfied by properly orienting the feed axis[15-17] as shown in Section VIII.

The above results are related to previous results by Brickell and Westcott,[13,14] Tanaka, Mizusawa,[15] Mizuguchi et al.,[16] Hanfling,[12] and the author.[17] There is a simple connection, pointed out in Section IX, between some of the expressions derived here and certain results in Refs. 13 and 14; the particular case, $N = 2$, is treated in those references. When only two reflectors are involved, there is always a common plane of symmetry, and the feed orientation can be derived geometrically as in Ref. 17. Our results differ from those of Refs. 15 and 16 in two respects: first, they apply also for $N > 2$; second, $\tan i$ is expressed directly in terms of the parameters (magnifications and angles of incidence) that specify the path of the principal ray. As pointed out in Ref. 19, an important application of our results is in the theory of aberrations.

The main results of this article are derived in Sections V through VII. Most of the results of Sections II through IV are well-known properties of ellipsoids, but their derivation is needed for Sections V through VII. In the following section, we discuss the transformation obtained when an ellipsoid is illuminated by a spherical wave front originating from one of the two foci. This transformation has the following basic property: it is a conformal mapping which gives correctly, not only the amplitude distribution of a reflected wave front, but also its polarization. All results of this article directly follow from this property.

II. CONFORMAL MAPPING, COMPLEX COORDINATES, AND THE BILINEAR TRANSFORMATION

Let a linearly polarized point source be placed at O, one of the two foci of an ellipsoid, and let O' be the other focus (in Fig. 2). Then for each ray from O, a corresponding ray through O' is obtained after reflection by the ellipsoid. To determine the properties of this correspondence, introduce at the two foci separate coordinate systems x, y,

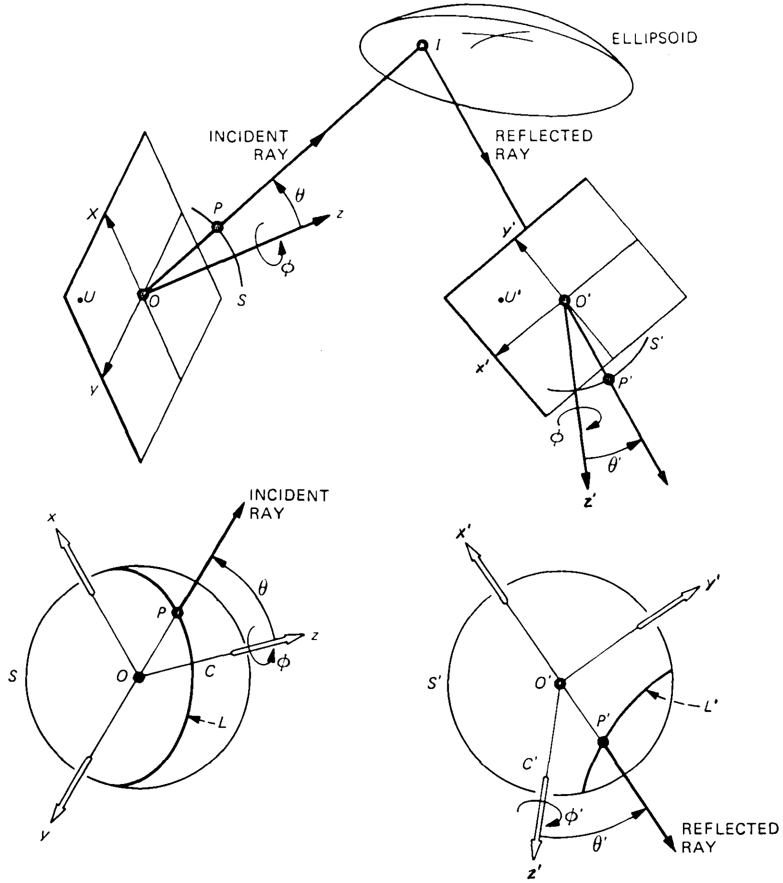


Fig. 2—A ray from one of the two foci of an ellipsoid determines, after reflection, a ray through the other focus. The correspondence $P \rightarrow P'$ between points of two wave fronts S and S' is everywhere as conformal mapping described by eq. (1).

z and x', y', z' oriented arbitrarily, as shown in Fig. 2. Consider an incident ray from O with spherical coordinates θ, ϕ, and let θ', ϕ' be the spherical coordinates of the corresponding ray through O'. It is convenient to introduce as in Refs. 12, 13 complex coordinates u and u' defined by eq. (3). Then, we show in this section that u' and u are related by a bilinear transformation. If both coordinate systems are right-handed, we shall see that the bilinear transformation does not relate u' directly to u, but to its complex conjugate u^*. To avoid this inconvenience, one of the two reference systems will be assumed to be left-handed as shown in Fig. 2.

To better visualize the one-to-one correspondence between rays through the two foci, consider in Fig. 2 two wave fronts S and S' centered at O and O', respectively. Then, for each ray from O, one obtains on S and S' two corresponding points P and P'. Furthermore, letting P be a variable point of a curve L of S, one obtains on S' a variable point P' describing a corresponding curve L' of S'. Since the point source is linearly polarized, the curve L can be drawn so that it is everywhere tangent to the magnetic field. Then one obtains a polarization line of S, and it is shown in Appendix A that the correspondence $P \rightarrow P'$ transforms polarization lines into polarization lines. That is,

if L is a polarization line,
then L′ is also a
polarization line. (10)

Another property is that angles are preserved and, therefore, the correspondence $P \rightarrow P'$ is a *conformal mapping.*

The above considerations apply also to an hyperboloid (in which case one of the two foci is behind the reflector), to a paraboloid, or to any combination of such reflectors. Thus, let the ellipsoid of Fig. 2 be combined with two paraboloids with foci at O and O'. Let the first paraboloid be centered around the z-axis, so as to map conformally the plane $z = 0$ onto the wave front S. Similarly, let the second paraboloid map S' onto the plane $z' = 0$. Then, the product of the above three reflections determines a one-to-one correspondence between points U and U' of the two planes $z = 0$ and $z' = 0$ in Fig. 2. This correspondence is everywhere conformal and, therefore, it implies a bilinear relation[2] between the complex coordinates $x + jy$ and $x' + jy'$ of two corresponding points U and U'. It is shown in the following section that the two paraboloids produce the transformations

$$x + jy = 2f_0 u, \qquad x' + jy' = 2f_0' u', \tag{11}$$

f_0 and f_0' being the focal lengths of the two paraboloids. Thus, the desired result, eq. (1), follows at once. As pointed out earlier, eq. (1) requires that one of the two reference systems in Fig. 2 be left-handed.

III. CONSTRUCTION OF A PARABOLOID BY A STEREOGRAPHIC PROJECTION

The mapping between two wave fronts S and S' in Fig. 2 can be represented as a product of two stereographic projections, as shown in Appendix B. In this section, we let O' go to ∞ on the z-axis. Then the ellipsoid degenerates into a paraboloid, S' becomes a plane, and only one stereographic projection is needed.[12]

Let the radius of S be chosen equal to the paraboloid focal length f_0, and let S' be the tangent plane $z = f_0$. Let a correspondence $P \rightarrow P'$ between points of S and S' be obtained as shown in Fig. 3, with a stereographic projection from the axial point $z = -f_0$. To show that this is the same correspondence determined by the rays reflected by the paraboloid, consider the reflected ray corresponding to P', and let I be its intersection with the incident ray OP. Since the triangle $PP'I$ is similar to the isosceles triangle NPO, $PI = P'I$ and, therefore,

$$OI = IP' = f_0, \tag{12}$$

which is the equation of a paraboloid.

Notice, if ρ is the radial distance of the reflected ray from the z-axis, then from the triangle $P'VN$ one has

$$\rho = 2f_0 \tan\frac{\theta}{2}. \tag{13}$$

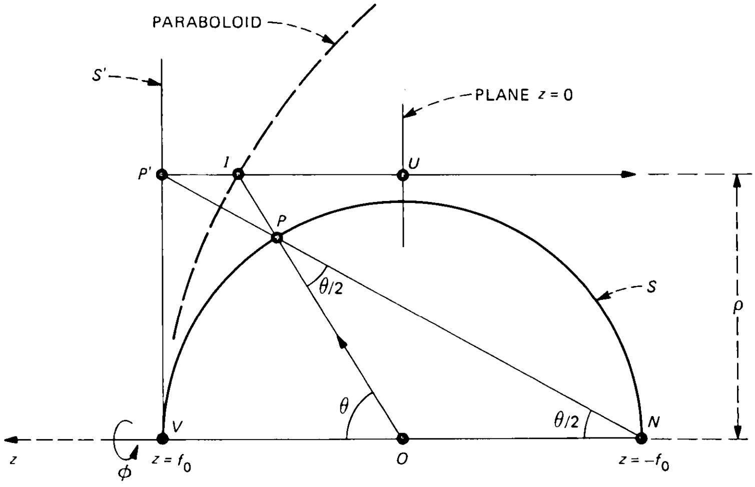


Fig. 3—Construction of a paraboloid by a stereographic projection.

Therefore, the reflected ray determines on the plane $z = 0$ a point U with the complex coordinate given by

$$x + jy = 2f_0 e^{j\phi} \tan \frac{\theta}{2}. \tag{14}$$

which gives eq. (11).

IV. TRANSFORMATION BY AN ELLIPSOID CENTERED AROUND THE *Z*-AXIS

The usual construction of an ellipse (Fig. 4) involves two fixed points A_1 and A_2 and a variable point A_3 which is varied, keeping the perimeter p of the triangle $A_1A_2A_3$ constant. Then, A_3 describes an ellipse with foci A_1 and A_2, as shown in Fig. 4. A simple relation among the angles of the triangle $A_1A_2A_3$ is given by the following theorem, which is derived in Appendix B with the help of two stereographic projections.

Theorem: Given a triangle $A_1A_2A_3$ with angles α_1, α_2, α_3, and perimeter p, its three sides d_1, d_2, d_3 are given by

$$2d_l = p\left(1 - \tan\frac{\alpha_m}{2} \tan\frac{\alpha_n}{2}\right), \tag{15}$$

where (l, m, n) is any permutation of (1, 2, 3) *and* $d_l = A_mA_n$.

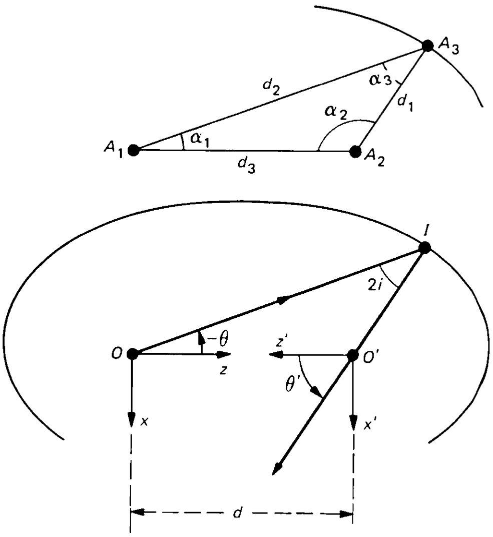


Fig. 4—The angles α_1, α_2, α_3 of a triangle $A_1A_2A_3$ are related by eq. (15), which implies a linear relation between tan $\theta/2$ and tan $\theta'/2$.

By letting A_1 and A_2 coincide with O and O' in Fig. 4, and letting A_3 be the point of incidence of a ray from O, one obtains from eq. (15) for $l = 3$ the well-known relation

$$1 + a_0 \tan \frac{\alpha_1}{2} \tan \frac{\alpha_2}{2} = 0, \tag{16}$$

where a_0 is a constant determined by the distance $d = d_3$ between the two foci and by the path length $t = OIO'$,

$$a_0 = -\frac{t + d}{t - d}. \tag{17}$$

If now the ellipsoid is centered around the z-axis as shown in Fig. 4, one has $\theta = -\alpha_1$. Furthermore, orienting the x', y', z'-axes as in Fig. 4, with the z'-axis opposite the z-axis, we have $\phi' = \phi$ and $\theta' = \pi - \alpha_2$; therefore, eq. (16) gives

$$u' = a_0 u. \tag{18}$$

Thus, for this particular orientation of the reference axes, eq. (1) assumes the normal form of eq. (6). If now arbitrary rotations are applied to the reference axes of Fig. 4, as we have seen in Section II, eq. (18) assumes the form of eq. (1). This implies that the above rotations transform u and u' according to bilinear transformations, whose coefficients are derived next.

V. ROTATIONS AND REFLECTIONS[10]

Consider Fig. 5a showing the x, y, z-axes oriented arbitrarily with respect to the x', y', z'-axes. We wish to determine, for a ray from O, the relationship between its coordinates θ', ϕ' and θ, ϕ with respect to the two coordinate systems. We have seen that the relationship can be written in the form (1), whose coefficients we now express in terms of the Euler angles α, β, γ specifying the orientation of the x', y', z'-axes with respect to the x, y, z-axes. Notice, for the purpose of determining the coefficients of eq. (1), consideration can be restricted to real values of u.

The x', y', z'-axes in Fig. 5a can be obtained from the x, y, z-axes by three successive rotations: a rotation around the z-axis through the Euler angle α, followed by a rotation around the y-axis involving the second Euler angle β and, finally, a rotation around the z-axis by the third Euler angle γ. The first and last rotations are described by the transformations

$$u' = ue^{-j\alpha}, \qquad u' = ue^{-j\gamma}. \tag{19}$$

To determine the second transformation, consider Fig. 5b which illus-

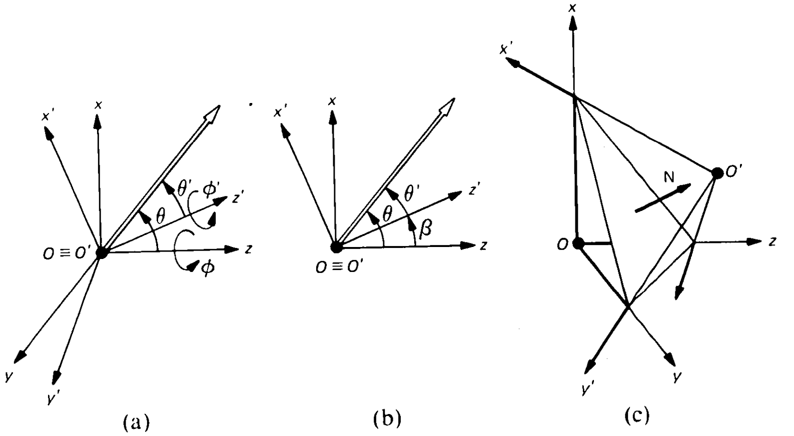


Fig. 5—The x', y', z'-axes are obtained from the x, y, z-axes through (a) an arbitrary rotation, (b) a rotation around the y-axis, and (c) a reflection by a plane.

trates a rotation around the y-axis by the angle β. Then, for a ray in the xz-plane, $\theta' = \theta - \beta$ and, therefore,

$$\tan\frac{\theta'}{2} = \frac{\tan\frac{\theta}{2} - \tan\frac{\beta}{2}}{1 + \tan\frac{\theta}{2}\tan\frac{\beta}{2}}, \tag{20}$$

which gives

$$u' = \frac{u + b_0}{1 - b_0 u}, \tag{21}$$

with

$$b_0 = -\tan\frac{\beta}{2}. \tag{22}$$

The product of the above three rotations can now be calculated straightforwardly. Letting

$$b = -\tan\frac{\beta}{2}\, e^{j\alpha}, \tag{23}$$

from eqs. (19), (21), and (22) we obtain

$$u' = e^{-j(\alpha+\gamma)}\frac{b + u}{1 - b^* u}, \tag{24}$$

which represents an arbitrary rotation. In the special case where the axis of rotation is orthogonal to the z-axis, we can verify that

$$\alpha + \gamma = 0 \tag{25}$$

and eq. (24) reduces to

$$u' = \frac{b + u}{1 - b^* u}. \tag{26}$$

Eqs. (24) and (26) can be considered generalizations of the trigonometric identity (20) to complex coordinates. They are directly related to the Cayley-Klein representation[10] of rotations by complex matrices.

5.1 Reflections

We now combine a rotation orthogonal to the z-axis with an inversion of the z-axis and obtain from eq. (26), replacing u with $1/u^*$,

$$u' = \frac{1 + bu^*}{u^* - b^*}, \tag{27}$$

representing a reflection by a plane shown in Fig. 5(c). The x', y', z'-axes in Fig. 5(c) are the reflected images of the x, y, z-axes, and we shall see in a moment that the coefficient b in eq. (27) is given by

$$b = \frac{N_x + jN_y}{N_z}, \tag{28}$$

where N_x, N_y, N_z are the x, y, z-components of a vector **N** orthogonal to the reflector.

Eqs. (23) through (28) give the transformation of u when a rotation (or a reflection) is applied to the reference axes. Suppose now the same rotation (or reflection) is applied to a ray with initial coordinate u, so as to obtain a new ray with coordinate u', as in Figs. 6(a) and (b). Then, if both u' and u are measured with respect to the x, y, z-axes, we find* that α, γ, β in eqs. (23) and (24) must be replaced with $-\alpha$, $-\gamma$, $-\beta$, whereas the coefficient b in eq. (27) is still given by eq. (28).

To derive eq. (28), let a reflection be applied to the ray $u = \infty$, as in Fig. 6(c). Then the angle formed by the z-axis and the reflected ray is bisected by **N**. Thus, since **N** is in the plane of incidence, and the angle of **N** with respect to the z-axis is $\theta'/2$,

$$u' = \frac{N_x + jN_y}{N_z}, \qquad \text{for} \quad u = \infty.$$

This gives the desired results of eq. (28).

If, instead of a plane, the ray is reflected by an arbitrary surface $z = f(x, y)$, then from eq. (28) one obtains

* To show this, first let u' be measured with respect to the x', y', z'-axes. Then one obtains the identity $u' = u$. Next apply to the x', y', z'-axes the inverse transformation of eq. (24) or (27). The inverse of eq. (24) is a rotation with Euler angles $-\alpha$, $-\gamma$, $-\beta$, whereas the inverse of eq. (27) is a reflection with the same coefficient b of eq. (28).

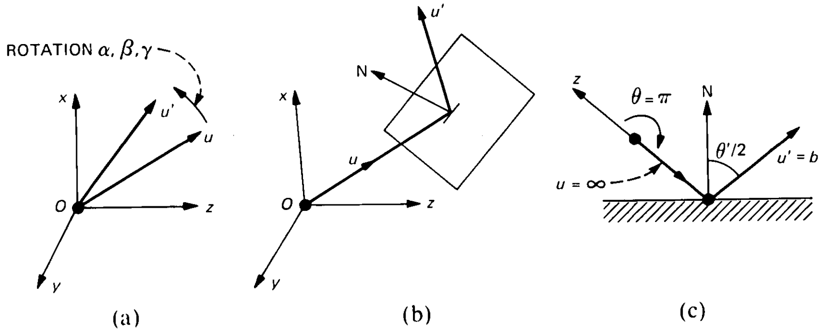


Fig. 6—A ray with initial coordinate u is subjected in (a) to a rotation and in (b) to a reflection by a plane. Notice in (c) $u' = b$ for $u = \infty$.

$$b = -\left(\frac{\partial}{\partial x} + j\frac{\partial}{\partial y}\right) f(x, y), \tag{29}$$

and eq. (27) gives a simple relation between the ray coordinates and the partial derivatives of $f(x, y)$. A similar result[13,14] is obtained if the equation of the reflector is specified in spherical coordinates, as shown in Appendix C.

Notice that eq. (27) applies not only to a ray, but also to its polarization in which case u and u' represent the directions of the incident and reflected polarization with respect to the x, y, z-axes

VI. TRANSFORMATION BY AN ELLIPSOID WHEN THE OUTPUT FRAME IS THE MIRROR IMAGE OF THE INPUT FRAME

In this section, we orient the z-axis in the direction of the principal ray. In a reflector antenna, this is the ray that corresponds to the feed axis. Thus, the principal ray determines the point of maximum illumination over the antenna aperture. To maximize aperture efficiency, the feed is usually oriented so that the principal ray $u = 0$ passes through the center of the aperture. The ray $u = \infty$, which leaves the feed in the direction opposite to the principal ray, will be called the *cardinal ray.*

Consider Fig. 7 which shows an ellipsoid with the principal ray incident at I with angle of incidence i. Let the x', y', z'-axes be the reflected images of the x, y, z-axes with respect to the tangent plane at I. Then, the principal ray after reflection has the direction of the z'-axis, whose complex coordinate $u = \lambda$ with respect to the x, y, z-axes is given by

$$\lambda = \frac{e^{j\psi}}{\tan i},$$

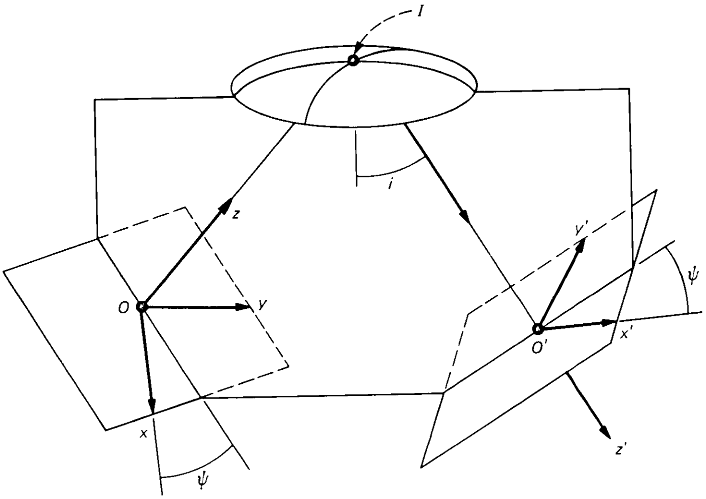


Fig. 7—Reference frames implied by eq. (34). Notice the plane of incidence for the principal ray rotated by ψ with respect to the xz-plane.

ψ being the angle of rotation of the plane of incidence with respect to the x-axis.

Initially, assume $\psi = 0$. Then the x-axis is in the plane of incidence, as shown in Fig. 8, and the same is true for the x'-axis. Thus, both reference systems can be obtained from those of Fig. 4 by suitable rotations around the y-axes, which have the same orientation in both cases. Taking into account that the coefficients of these rotations are real, they transform eq. (18) into

$$u' = a\frac{u}{1 + cu}, \tag{30}$$

where a and c are real coefficients which can be determined as follows. To determine a, consider a ray in the vicinity of the principal ray. Then θ and θ' are small and

$$\theta\ell \simeq -\theta\ell',$$

ℓ and ℓ' being the distances of I from the two foci. It follows that a is equal to the magnification M given by

$$M = -\frac{\ell}{\ell'}. \tag{31}$$

To determine c, let $u = \infty$. We then obtain in Fig. 8 the cardinal ray incident at i' with angle of incidence i'. From the triangle $II'O'$ of Fig. 8,

$$u' = \frac{1}{\tan(i + i')}, \qquad \text{for} \quad u = \infty.$$

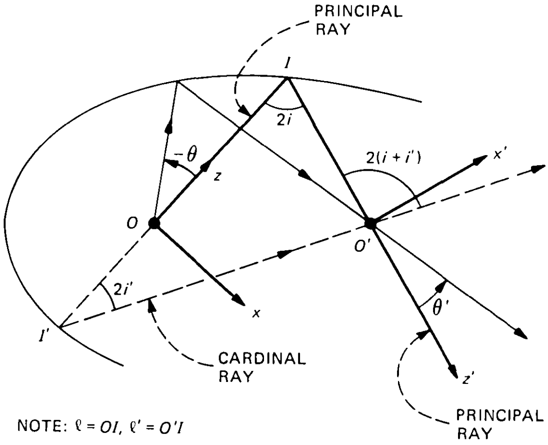


Fig. 8—To obtain eq. (33), each reference system must be oriented so that the z-axis is in the direction of the principal ray, and the x-axis is in the plane of incidence.

Furthermore, applying eq. (15) to the triangle $II'O'$ with perimeter $p = 2t = 2(\ell + \ell')$,

$$(\ell + \ell')\tan i = \ell \tan(i + i'), \tag{32}$$

which gives the desired result,

$$c = (M - 1)\tan i.$$

Finally,

$$u' = M\frac{u}{1 + (M - 1)u \tan i}, \tag{33}$$

which assumes the x-axis is in the plane of incidence, so that $\psi = 0$.

Now consider the general case $\psi \neq 0$. Then both reference systems in Fig. 8 must be rotated around the z-axes by $-\psi$, and from eq. (33) we obtain

$$u' = M\frac{u}{1 + u(M - 1)e^{-j\psi}\tan i}, \tag{34}$$

which applies in general when the plane of incidence is rotated by an arbitrary angle ψ with respect to the xz-plane.

Using this simple relation, we can now determine straightforwardly how the nonzero angle of incidence i in Fig. 7 affects the amplitude pattern of the reflected wave, its polarization, its symmetry, and the aberrations arising when the point source is slightly displaced from O. For $i = 0$, eq. (34) reduces to eq. (18). In this case the transformation has circular symmetry, since it is unaffected if identical rotations are applied to the reference systems around the z-axes. For $i \neq 0$, on the other hand, eq. (34) lacks this symmetry. We now show that, by properly combining several asymmetric transformations of the type (34), it is always possible to obtain the symmetric transformations (18). This was first shown in Refs. (15) and (16) for two reflectors with O' at ∞ and, in Ref. (17) under more general conditions.

VII. TRANSFORMATION BY A SEQUENCE OF ELLIPSOIDS

Replace the ellipsoid of Fig. 7 with a sequence of ellipsoids, with foci O_0, O_1, $\cdots$, O_N as shown in Fig. 9. Let the $(s+1)$th reference frame be the mirror image of the sth frame as in Section VI. Let M_s, i_s, ψ_s be the values of M, i, ψ for the sth ellipsoid. Then, the product of the N transformations of Fig. 9 gives eq. (34) with

$$M = M_1 \cdots M_N \qquad (35)$$

and

$$(M-1)e^{-j\psi}\tan i = (M_1-1)e^{-j\psi_1}\tan i_1 + (M_2-1)M_1e^{-j\psi_2}\tan i_2 + \cdots \qquad (36)$$

We have thus shown that eq. (34), derived in Section VI for the ellipsoid of Fig. 7, applies also to a sequence of N ellipsoids. It also

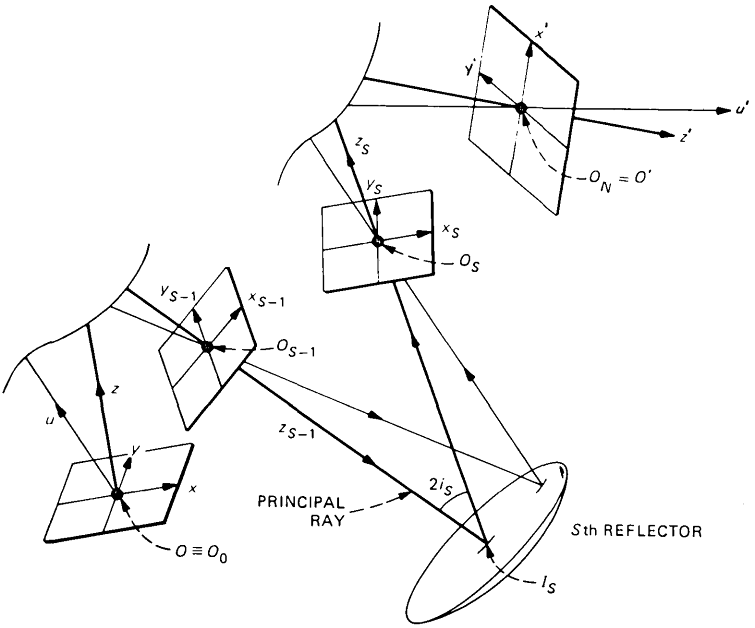


Fig. 9—An input ray with coordinated u is transformed by a sequence of N reflectors into an output ray with coordinate u' given by eq. (34). The principal ray for $u = 0$ is reflected by the s^{th} reflector at I_s with angle of incidence i_s.

applies to a hyperboloid, in which case $M < 0$ since then one of the two foci O and O' in Fig. 7 is behind the reflector and therefore either ℓ or ℓ' is negative. For

$$M \rightarrow \infty,$$

the ellipsoid of Fig. 7 degenerates into a paraboloid, shown in Fig. 1 for $\psi = 0$. Then, from eq. (34), letting

$$M \rightarrow 0, \qquad u' \rightarrow M \frac{w}{2\ell} \tag{37}$$

we obtain

$$w = 2f \frac{u}{1 - u \tan i \, e^{-j\psi}}, \tag{38}$$

where $f = \ell$, and

$$w = x' + jy' \tag{39}$$

is the complex coordinate intercepted in Fig. 1 by the reflected ray on the plane $z' = 0$.

Equation (38) also applies to the arrangement of Fig. 9, with the last ellipsoid replaced by a paraboloid, in which case ψ and $\tan i$ are given by eq. (36) with

$$M = M_N = 0. \tag{40}$$

Then f in eq. (38) is the *equivalent focal length* given by

$$f = M_e \ell_N, \tag{41}$$

where M_e is the magnification determined by the first $N - 1$ reflectors,

$$M_e = M_1 \cdots M_{N-1}, \tag{42}$$

and ℓ_N is the focal length of the last paraboloid.

As pointed out in the introduction, it is desirable in general that

$$\tan i = 0, \tag{43}$$

because then the transformation has circular symmetry with respect to the principal ray. From eq. (36) for $N = 2$, this requires

$$\psi_1 = \psi_2 \tag{44}$$

and

$$\tan i_1 (M_1 - 1) + \tan i_2 (M_2 - 1) M_1 = 0. \tag{45}$$

The first condition demands that the two planes of incidence (for the principal ray) coincide, in which case the two reflectors and the feed have a common plane of symmetry. In general, for arbitrary N, one finds that it is always possible to satisfy condition (43) by properly choosing one of the planes of incidence and one of the angles i_s for any arbitrary choice of the remaining i_s. The correct choice for i_s is obtained straightforwardly using eq. (36). In some cases, it may not be possible to satisfy exactly the requirement (43). For instance, the N reflectors may have to fit inside a satellite and, because of the limited available space, it may be convenient to choose $i \neq 0$. Then, the resulting aberrations and distortion of the polarization and amplitude illumination over the aperture are determined straightforwardly using eq. (34), as pointed out in Ref. 19.

VIII. GEOMETRICAL DERIVATION OF TAN *I* WHEN THE LAST REFLECTOR IS A PARABOLOID

Assume the final reflector is a paraboloid, and let the input point source be a corrugated feed,[20] or a feed with similar radiation characteristics.[21,22] Then, the spherical wave radiated by the feed has an axis of circular symmetry, and its polarization lines on a wave front S are given by a set of tangent circles as shown in Fig. 10. The contact point D for these circles is one of the two intersections of the feed axis with the wave front S. The other intersection C is the point of maximum illumination. Thus, C and D are determined by the principal ray ($u = 0$) and the cardinal ray ($u = \infty$), respectively. It is now recalled that the bilinear transformation (1) transforms circles into circles. This means that the polarization lines of an output wave front S' are also a set of tangent circles. Their contact point D' is determined by the

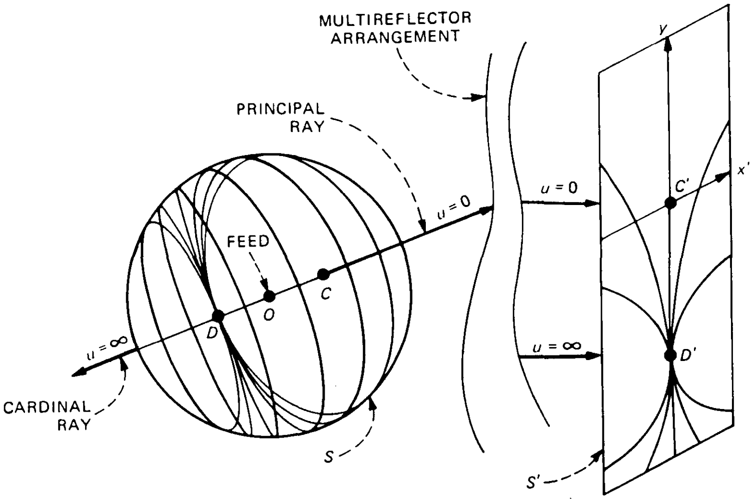


Fig. 10—The polarization lines produced by a corrugated feed are tangent circles with contact point determined by the cardinal ray.

ray $u = \infty$, and the point of maximum illumination C' is determined by the ray $u = 0$. From eq. (8) for $u \to \infty$, the distance of C' from D' is given by

$$CD' = \frac{2f}{\tan i} \tag{46}$$

and, therefore, for $\tan i \to 0$

$$D' \to \infty.$$

Then the circles degenerate into parallel lines, and the output wave front S' becomes everywhere polarized in one direction.[23]

The above considerations suggest a simple procedure for determining the feed axis orientation that corresponds to $\tan i = 0$. The feed must be oriented so that $D' \to \infty$. This means that the cardinal ray must be reflected at ∞ by the last reflector (a paraboloid). Thus, the cardinal ray after the first $N - 1$ reflections must pass through the paraboloid focus O_{N-1} with direction opposite to $O_{N-1}V$, where V is the paraboloid vertex. Therefore, one must orient the feed axis so that the cardinal ray ($u = \infty$) produces after $N - 1$ reflections the ray VO_{N-1}. Since the final direction of this ray is specified, the initial direction can be determined by retracing the ray in the reverse direction starting from O_{N-1}. Then, after ($N - 1$) reflections of the ray $O_{N-1}V$, one obtains through 0 the direction of OC characterized by $\tan i = 0$. This geometrical derivation is illustrated in Ref. 17.

IX. AMPLITUDE AND POLARIZATION OF THE OUTPUT WAVE

Consider two corresponding points P and P' on the two wave fronts S and S' of Fig. 2. Let the electric field at P be given by

$$\mathbf{E} = A\mathbf{e}\frac{e^{-jkr}}{r}, \tag{47}$$

where r is the distance from the focus O and $\mathbf{e}$ is a unit vector specifying the polarization of E. Similarly, for the field at P'

$$\mathbf{E} = -A'\mathbf{e}'\frac{e^{-jk(r'+t)}}{r'}, \tag{48}$$

where $t = \ell + \ell'$. Let

$$\mathbf{e} = \cos\phi_e\mathbf{i}_1 + \sin\phi_e\mathbf{i}_2, \tag{49}$$

where $\mathbf{i}_1$, $\mathbf{i}_2$ are unit vectors in the θ, ϕ-directions and ϕ_e is the angle of rotation of $\mathbf{e}$ with respect to $\mathbf{i}_1$. In this section, we show that the corresponding angle of rotation ϕ_e' for $\mathbf{e}'$ with respect to $\mathbf{i}_1'$ is simply given by

$$\phi_e' - \phi_e = \phi' - \phi, \tag{50}$$

where it is recalled that ϕ' and ϕ are the arguments of the coordinates u' and u, respectively. For the amplitude A' we show that

$$A' = \frac{1}{m} A, \tag{51}$$

where the magnification m is given by

$$m = \frac{1}{M} \frac{u'u'^*(1 + uu^*)}{uu^*(1 + u'u'^*)}. \tag{52}$$

These relations apply in general to an arbitrary sequence of ellipsoids, hyperboloids, and paraboloids arranged as in Fig. 9. If the first focus O is at infinity, then S is a plane and the spherical coordinates θ, ϕ must be replaced with polar coordinates ρ, ϕ. Then, $\mathbf{i}_1$ is a unit vector in the ρ-direction. Similar considerations apply if O' is at ∞.

To derive eqs. (50) and (52), it is convenient to combine the ellipsoid of Fig. 2 with two paraboloids, as in Section II. Then, one paraboloid maps conformally the sphere S onto the plane $z = 0$ and the other paraboloid the sphere S' onto the plane $z' = 0$. Let A_0, $\mathbf{e}_0$ and A_0', $\mathbf{e}_0'$ be the values of A, $\mathbf{e}$ produced on the two planes, and assume the two mappings are characterized by the transformations

$$u = x + jy, \qquad u' = x' + jy',$$

which imply $f_0 = f_0' = 1/2$ in eq. (11). These transformations do not affect the polarization angles ϕ_e and ϕ_e', while the amplitude A is transformed according to the well-known relation

$$A_0 = \frac{A}{1 + \tan^2 \dfrac{\theta}{2}} = \frac{A}{1 + uu^*}, \tag{53}$$

and similarly for A_0'. According to geometric optics, conservation of power requires

$$|A_0|^2 d\sigma_0 = |A_0'|^2 d\sigma_0', \tag{54}$$

where $d\sigma_0$ and $d\sigma_0'$ are the areas of two corresponding elements of the two planes. Since the mapping between the two planes is conformal,

$$\frac{d\sigma_0}{d\sigma_0'} = \left| \frac{du'}{du} \right|^2, \tag{55}$$

and using eqs. (34), (53), and (54), one obtains the desired result, eq. (52).

Next, we derive eq. (50). Consider two corresponding points Q and Q' of the two planes. Let the polarization line through Q be a straight line through the origin, as in Fig. 11. Then $\phi_e = 0$, and the corresponding polarization line through Q' is a circle. The circle must pass through the origin O', and also through the point D' of coordinate

$$u'_\infty = \frac{M}{M-1}\frac{1}{\tan i}e^{j\psi}, \tag{56}$$

which corresponds to the point at ∞ of the u-plane, as one can verify from eq. (34) letting $u \to \infty$. Now the angle made in Fig. 11 by the chord $D'Q'$ with the tangent $\mathbf{e}'_0$ is equal to the angle β subtended by the chord at O'. As a consequence, one can verify that the angle ϕ'_e in Fig. 11 is equal to the angle γ between $D'Q'$ and $D'O'$. Thus,

$$\gamma = \phi'_e - \phi_e = \angle\left[\frac{u' - u'_\infty}{-u'_\infty}\right] = \angle\left[1 - \tan i\,\frac{M-1}{M}\,u'e^{-j\psi}\right], \tag{57}$$

and one can verify that this agrees with eq. (50). Using eqs. (50) through (52), one can now derive straightforwardly the amplitude and polarization of the output wave in Fig. 9.

X. CONCLUSIONS

With simple geometric considerations, we have derived the coefficients of the transformation (1). Once the parameters i_s, M_s, ψ_s that specify the path of the principal ray are known, the coefficients can be derived straightforwardly using eqs. (34) and (36). For a corrugated feed, it has been shown in Section VIII that the circles describing the polarization of an output wave front can be determined straightfor-

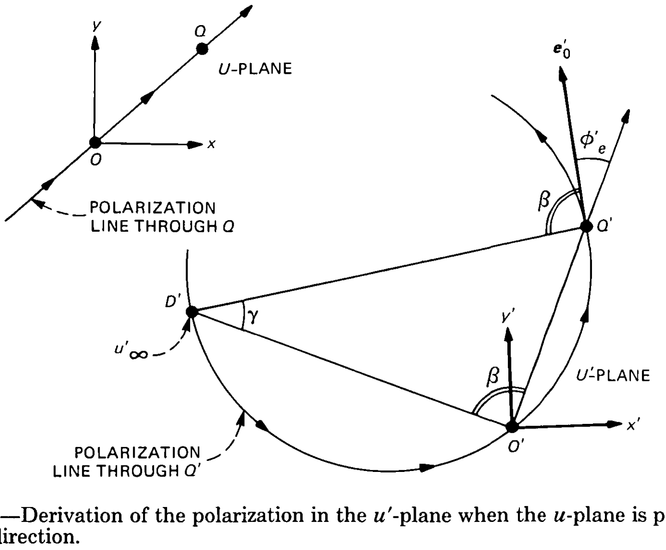


Fig. 11—Derivation of the polarization in the u'-plane when the u-plane is polarized in the ρ-direction.

wardly by tracing the cardinal ray. In Section V, it has been shown how the transformation (35) is affected by a rotation of the feed axis. The results will be useful to the design of reflector antennas as pointed out in Ref. 19. They also provide a simple interpretation for previous results of Refs. 13 and 19, as pointed out in Appendix C.

APPENDIX A

Consider Fig. 2. We wish to show that if L is everywhere tangent to the magnetic field, then this is true also for L'. Suppose initially that S and S' are both centered at O. Then, the mapping determined between S and S' by a ray from O is just a similarity, with magnification determined by the radii of S and S'. This means that each line element $\delta L'$ of L' is parallel* to the corresponding line element δL of L. Thus, if both S and S' are centered at O, or both at O', then the two curves L and L' certainly satisfy (10).

Next, let S and S' be centered at O and O', respectively, and let I be the point of incidence for the ray corresponding to δL. Then, the orientation of the corresponding line element $\delta L'$ is not affected if the ellipsoid in Fig. 2 is replaced by the tangent plane at I. Thus, we conclude that property (10) is true in general, even if S' and S are centered at different foci.

Notice (10) implies that the mapping between any two wave fronts S and S' preserves angles and, therefore, it is conformal.

APPENDIX B

We show that the mapping in Fig. 2 between the two wave fronts S and S' can be represented as a product of two stereographic projections. A variety of different representations can be obtained, depending on the radii r and r' of S and S'. For simplicity, here we choose the two radii so that the two spheres S and S' touch each other, as shown in Fig. 12. The contact point V is on the axis OO' of the ellipsoid. Let N and N' be the other intersections of the two spheres with the axis. Let P_1 be an arbitrary point of the tangent plane at V, and let two corresponding rays OP and $O'P'$ be obtained as shown in Fig. 12, with two stereographic projections from N and N', respectively. We now show that the intersection I of the two rays satisfies the condition

$$OI + IO' = r + r', \tag{58}$$

which is the equation of an ellipsoid. Let θ and θ' be the angles VOP and $VO'P'$, respectively. Then the isosceles triangle ONP has two of

* This property, is true *only if* S and S' are spherical wave fronts or if L is a geodesic line of S.

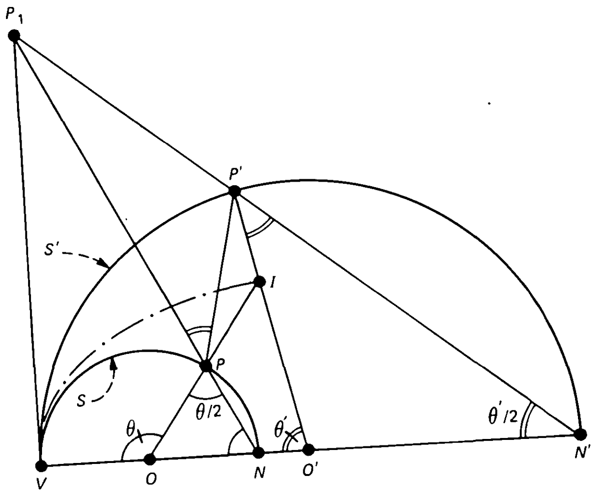


Fig. 12—A point I of an ellipse with foci O and O' is obtained with two stereographic projections from N and N'.

its angles equal to $\theta/2$, and similarly the triangle $O'N'P'$ has two angles equal to $\theta'/2$. Furthermore, since VP_1 is tangent to both spheres,

$$NP_1 \times P_1P = N'P_1 \times P_1P',$$

since both products must equal VP_1^2. Thus, the triangles PP_1P' and NP_1N' are similar and, therefore, $P_1PP' = \theta'/2$, $PP'P_1 = 180° - \theta/2$. The angles $PP'I$ and $P'PI$ can now be determined, and one finds they are both equal to $(\theta + \theta')/2$. Thus, $PI = P'I$, which gives eq. (58).

From the right triangles P_1VN and P_1VN', since they have one side in common,

$$r \tan \frac{\theta}{2} = r' \tan \frac{\theta'}{2}, \tag{59}$$

which gives eq. (16), and this implies the theorem of Section IV.

APPENDIX C

We now point out a simple connection between eqs. (28) and (29) and previous results by Brickell and Westcott.[13,14] Both u and u' will be measured with respect to the same reference frame, the x, y, z-axes.

Let an arbitrary reflector be illuminated by a spherical wave from the origin O of the x, y, z-axes, and let the reflecting surface be given in spherical coordinates by

$$\rho = \rho(u, u^*),$$

where ρ is the distance from the origin O, and $\rho(u, u^*)$ is an analytic function† of u, u^*. Consider a particular incident ray with coordinate $u = \lambda$, and let P be its point of incidence. To determine its reflected coordinate u' with respect to the x, y, z-coordinates, it is convenient to introduce temporarily a second reference frame with the z-axis through the point of incidence. Then, using the subscript $(\ \ )_0$ for the second frame, and applying eqs. (27) and (29) to the ray through P,

$$u_0' = \rho\left(\frac{\partial \rho}{\partial u_0}\right)^{-1}, \qquad \text{for} \quad u_0 = 0. \tag{60}$$

Next, we apply a suitable rotation to the x_0, y_0, z_0-axes, so as to transform u_0', u_0 into u', u,

$$u' = \frac{u_0' + \lambda}{1 - u_0'\lambda^*}, \qquad u = \frac{u_0 + \lambda}{1 - u_0\lambda^*}.$$

From the second expression for $u_0 = 0$,

$$\frac{\partial u}{\partial u_0} = (1 + uu^*), \qquad \frac{\partial u^*}{\partial u_0} = 0,$$

and, therefore,

$$\frac{\partial}{\partial u_0} = \frac{\partial u}{\partial u_0}\frac{\partial}{\partial u} + \frac{\partial u^*}{\partial u_0}\frac{\partial}{\partial u^*} = (1 + uu^*)\frac{\partial}{\partial u},$$

for $u_0 = 0$. From these relations, taking into account that $u = \lambda$ for $u_0 = 0$, one obtains the final result

$$u' = \frac{\rho + (1 + uu^*)u\dfrac{\partial \rho}{\partial u}}{(1 + uu^*)\dfrac{\partial \rho}{\partial u} - u^*\rho}, \tag{61}$$

valid for any u. This gives eq. (16) of Ref. 13. We have thus shown that this basic result can be considered a direct consequence of eqs. (27) and (29).

REFERENCES

1. C. Dragone and D. C. Hogg, "The Radiation Pattern and Impedance of Offset and Symmetrical Near-Field Cassegrainian and Gregorian Antennas," IEEE Trans. Ant. Prop., *AP-22*, No. 3 (May 1974), pp. 472–5.
2. T. S. Chu et al., "The Crawford Hill 7-Meter Millimeter Wave Antenna," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1257–88.
3. E. A. Ohm, "A Proposed Multiple-Beam Microwave Antenna for Earth Stations and Satellites," B.S.T.J., *53*, No. 8 (October 1974), pp. 1657–66.

† It is convenient to express ρ in terms of both u and u^*, since the real quantity ρ is not an analytic function of u. This is best understood by expanding ρ in a power series of u. Then, since ρ is real and u is complex, both powers of u and u^* must be considered.

4. R. A. Semplak, "100-GHz Measurements on a Multiple-Beam Offset Antenna," B.S.T.J., *56,* No. 3 (March 1977), pp. 385–97.
5. E. A. Ohm and M. J. Gans, "Numerical Analysis of Multiple-Beam Offset Cassegrainian Antennas," AIAA Paper, No. 76-301, AIAA/CASI 6th Commun. Satellite Systems Conf., Montreal, Canada, April 5–8, 1976.
6. P. J. B. Clarricoats and G. T. Poulton, "High-Efficiency Microwave Reflector Antennas—A Review," Proc. IEEE, *65,* No. 10 (October 1977), pp. 1470–504.
7. A. W. Rudge and N. A. Adatia, "Offset-Parabolic-Reflector Antennas: A Review," Proc. IEEE, *66,* No. 12 (December 1978), pp. 1592–618.
8. C. Dragone and M. J. Gans, "Imaging Reflector Arrangements to Form a Scanning Beam Using a Small Array," B.S.T.J., *58,* No. 2 (February 1979), pp. 501–15.
9. E. A. Guillemin, "The Mathematics of Circuit Analysis," New York: John Wiley, 1950.
10. H. Margenau and G. M. Murphy, "The Mathematics of Physics and Chemistry," Princeton: D. Van Nostrand, 1956.
11. H. S. M. Coxeter, *Introduction to Geometry*, New York: John Wiley, 1969.
12. J. D. Hanfling, "Aperture Fields of Paraboloidal Reflectors by Stereographic Mapping of Feed Polarization," IEEE Trans. Ant. Prop., *AP-18*, No. 3 (May 1970), pp. 392–6.
13. F. Brickell and B. S. Westcott, "Phase and Power Density Distributions on Plane Apertures of Reflector Antennas," J. Phys. A: Math. Gen., *11,* No. 4 (1978), pp. 777–89.
14. B. S. Westcott and F. Brickell, "Dual Offset Reflectors Shaped for Zero Cross-Polarization and Prescribed Aperture Illumination," J. Phys. D: Appl. Phys., *12,* No. 2 (1979), pp. 169–86.
15. H. Tanaka and M. Mizusawa, "Elimination of Crosspolarization in Offset Dual Reflector Antennas," Elec. Commun. (Japan), *58,* No. 12 (1975), pp. 71–8.
16. Y. Mizuguchi, M. Akagawa, and H. Yokoi, "Offset Gregorian Antenna," Elec. and Commun. in Japan, *61-B*, No. 3 (1978), pp. 58–66.
17. C. Dragone, "Off-Set Multireflector Antennas With Perfect Pattern Symmetry and Polarization Discrimination," B.S.T.J., *57,* No. 7 (September 1978), pp. 2663–84.
18. P. W. Hannan, "Microwave Antennas Derived From the Cassegrain Telescope," IRE Trans. Antennas Propagation, *AP-9,* No. 1 (March 1961), pp. 140–53.
19. C. Dragone, unpublished work.
20. V. H. Rumsey, "Horn Antennas with Uniform Power Patterns Around their Axes," IEEE Trans. Ant. Prop., *AP-14,* No. 5 (September 1966), pp. 656–8.
21. C. Dragone, "Attenuation and Radiation Characteristics of the HE_{11}-Mode," IEEE Trans. MTT, *28,* No. 7 (July 1980), pp. 704–10.
22. C. Dragone, "High-Frequency Behavior of Waveguides with Finite Surface Impedances," B.S.T.J., *60,* No. 1 (January 1981), pp. 89–116.
23. C. C. Cutler, "Parabolic-Antenna Design for Microwaves," Proc. IRE, *35,* No. 11 (November 1947), pp. 1284–94.

CONTRIBUTORS TO THIS ISSUE

Sergio M. Brecher, Diploma Engineer, 1964 (Electrical Engineering), University of Buenos Aires, Argentina, M.S., 1969, Ph.D., 1972 (Electrical Engineering), Columbia University; Bell Laboratories 1972—. Before joining Bell Laboratories, Mr. Brecher worked in Argentina in high-frequency point-to-point communications and in automatic control for nuclear reactors. At Bell Laboratories, he has been working in switching systems engineering problems for No. 1/1A ESS, No. 10A RSS, and, more recently, No. 5 ESS. Member, Sigma Xi, IEEE.

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Libera Docenza, 1968, Ministero della Pubblica Istruzione (Italy); Bell Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently concerned with problems involving electromagnetic wave propagation and microwave antennas.

Kai Y. Eng, B.S.E.E. (summa cum laude), 1974, Newark College of Engineering; M.S. (Electrical Engineering), 1976, Dr. Engr. Sc. (Electrical Engineering), 1976, Dr. Engr. Sc. (Electrical Engineering), 1979, Columbia University; RCA Astro-Electronics, 1974–1979; Bell Laboratories, 1979—. Mr. Eng has worked on various areas of microwave transmission, spacecraft antenna analysis, and communications satellites. He is presently a member of the Radio Research Laboratory, studying TV transmission through satellites. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Phi Eta Sigma.

Barry G. Haskell, B.S. (Electrical Engineering) 1964, M.S., 1965, and Ph.D., 1968, University of California, Berkeley; University of California, 1965–1968; Bell Laboratories, 1968—; Rutgers University, 1977–1979. Mr. Haskell was a Research Assistant at the University of California Electronics Research Laboratory and a part-time faculty member of the Department of Electrical Engineering at Rutgers University. At Bell Laboratories, he is presently head of the Radio Communications Research Department, where his research interests include television picture coding and transmission of digital and analog information via microwave radio. Member, IEEE, Phi Beta Kappa, Sigma Xi.

Markus S. Mueller, E.E. Diploma, 1970, Ph.D. (Electronic Engineering), 1976, Swiss Federal Institute of Technology, Zurich, Switzerland; Swiss Federal Institute of Technology, Zurich, 1971–1976; GenRad, Zurich, 1976–1978; Bell Laboratories, 1978–1981. Mr. Mueller was teaching and research assistant at the Swiss Federal Institute of Technology, where he worked in various fields, including filter theory, data transmission, and adaptive signal processing. He was a product specialist with GenRad for computer controlled automatic test systems. At Bell Laboratories, he was a member of the Data Theory Group in the Data Systems and Technology Department, and his interests were in data communication, digital signal processing, and adaptive systems.

Vasant K. Prabhu, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Laboratories, 1966—. Mr. Prabhu has been concerned with various theoretical problems in solid-state microwave devices and digital and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, Commission 6 of URSI.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests include compartmental models, the theory of digital filtering, global implicit-function theorems, and functional expansions for nonlinear systems. Former Vice Chairman IEEE Group on Circuit Theory, and Former Guest Editor IEEE Transactions on Circuit Theory Special Issue on Active and Digital Networks. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, National Academy of Engineering.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked on remote line concentrators for the electronic switching system. Since 1968 he has supervised a group engaged in theoretical studies in data communications and is currently a member of the Communications Methods Research Department. During the academic year 1967–68, he was on leave as Professor of Electrical Engineering at the University of Florida. In Spring, 1981, he was a visiting lecturer at Stanford University. Member, Sigma Xi.

Hans S. Witsenhausen, Ph.D. (Electrical Engineering), 1966 Massachusetts Institute of Technology; Bell Laboratories, 1966—. Mr. Witsenhausen has been associated with the Université Libre de Bruxelles, with the European Computation Center, Brussels, and with the Princeton Research Division of Electronics Associates. At M.I.T. he was with the Electronic Systems Laboratory as a Lincoln Laboratory Associate and a Hertz Fellow. At Bell Laboratories, he is associated with the Mathematics and Statistics Research Center. He was a Senior Fellow at the Imperial College of Science and Technology, London, in 1972, a Visiting Professor at M.I.T. in 1973, and Vinton Hayes Senior Fellow at Harvard University in 1975–76. He has worked on problems of hybrid computation, control theory, optimization, geometric inequalities, and other applied mathematical fields.

Aaron D. Wyner, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, Ph.D., 1963, Columbia University; Bell Laboratories 1963—. Mr. Wyner has been doing research in various aspects of information and communication theory and related mathematical problems. He is presently Head of the Communications Analysis Research Department. He spent the year 1969–1970 visiting the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, at Technion, Haifa, Israel, on a Guggenheim Foundation Fellowship. He has also been a full- and part-time faculty member at Columbia University and the Polytechnic Institute of Brooklyn. He was chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, associate editor of the Group's *Transactions*, and co-chairperson of two international symposia. In 1976, he was president of the IEEE Information Theory Group. Fellow, IEEE; member, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Machine-Readable Output From Online Searches. D. T. Hawkins, J Amer Soc Info Sci, *32,* No. 4 (July 1981), pp 253-6.
Maximum Likelihood Ridge Regression and the Shrinkage Pattern Hypotheses. R. L. Obenchain, Inst Math Statist Bull, *10,* No. 1 (January 1981), pp 37.

ENGINEERING

Analysis and Prediction of Cross Polarization on Earth Space Links. T. S. Chu, Annales des Télécommun, France, *36,* No. 1-2 (January-February 1981), pp 140-7.
Automatic Component Placement in an Interactive Minicomputer Environment. F. Shupe, ACM IEEE Eighteenth Design Automation Conf Proc (June 1981), pp 145-52.
On the Effect of Uplink Noise on a Nonlinear Digital Satellite Channel. M. L. Steinberger, P. Balaban, and K. S. Shanmugam, 1981 Int Commun Conf Record, *1* (June 1981), pp. 20.2.1-5.
Evaluation of Advanced Switching Technologies for Integrated Voice/Data Applications. A. M. Rudrapatna, IEEE Commun Magazine, *19,* No. 2 (March 1981), pp. 38-44.
Laser Processing in Silicon Microelectronics Technology. H. J. Leamy, 1981 Proc Univ/Gov Ind Microelectron Symp, Catalogue No. 81CH1620-4, New York: IEEE, 1981, pp VI-I-II.
LSI Product Quality and Fault Coverage. V. D. Agrawal, S. C. Weth, and P. Agrawal, ACM IEEE Eighteenth Design Automation Conf Proc (June 1981), pp 196-203.
A Moisture Protection Screening Test For Hybrid Circuit Encapsulants. R. G. Mancke, Proc 1981—31st Electron Components Conf (1981), pp 119-25.
Reliability Considerations for Metallized Plastic Film Capacitors Under High-Stress AC Waveforms. R. A. Frantz, 1981 Power Electronics Specialists Conf Record (June 1981), pp 81-90.
A Photoconductive Detector for High-Speed Fiber Communication. J. C. Gammel, G. M. Metze, and J. M. Ballantyne, IEEE Trans Electron Devices, *ED-28,* No. 7 (July 1981), pp 841-9.
The Pseudoelastic Force Balance and its Application to β Fe-Be Alloys. M. L. Green, M. Cohen, and G. B. Olson, Matls Sci and Eng, *50,* No. 1 (1981), pp 109-16.
Scattered Light Speckle Metrology. T. D. Dudderar and P. G. Simpkins, Proc OSA Topical Meeting Hologram Interferometry and Speckle Metrology (June 1980), pp 71-5.
Technologies and Future Trends in Rectifiers for Customer Premises Telecommunications Systems. C. O. Riddleberger, IEE 1981 Proc Third Int Telecommun Energy Conf, No. 196 (May 1981), pp 313-8.
Transmission Cathodoluminescence as a Screening Technique for Rake Lines in (Al, Ga) as DH Laser Material. C. A. Gaw and C. L. Reynolds, Jr., Electron Lett, *17,* No. 8 (April 16, 1981), pp 285-6.

PHYSICAL SCIENCES

Aging Behavior of Piezoelectric Poly(vinylidene Fluoride) Films Irradiated by γ Rays. T. T. Wang, J Poly Sci, Polymer Lett Ed, *19* (June 1981), pp 289-93.
The b-Component of the Transition Moment for the ν_2 Band of Nitric Acid Vapor. L. A. Farrow and R. E. Richton, J Chem Phys, *74,* No. 10 (May 1981), pp 5474-8.
Efficient Solar to Chemical Conversion: 12% Efficient Photoassisted Electrolysis in the p-Type InP(Ru) HCl-KCl/Pt(Rh) Cell. A. Heller and R. G. Vadimsky, Phys Rev Lett, *46,* No. 17 (April 27, 1981), pp 1153-60.

The First General Index of Molecular Complexity. S. H. Bertz, J Am Chem Soc, *103* (July 1981), pp 3599–601.
The Hydrolytic Stability of Some Commercially Available Polycarbonates. C. A. Pryde, P. G. Kelleher, M. Y. Hellman, and R. P. Wentz, Soc Plastics Engineers 39th Annual Technical Conf, Technical Papers, *28* (May 1981), pp 98–100.
Pulsed Opto-Acoustic Spectroscopy of Condensed Matter. C. K. N. Patel and A. C. Tam, Rev Mod Phys, *53,* No. 3 (July 1981), pp 517–50.
Surface Reaction Studies on Roughened Silver Using Enhanced Roman Scattering. T. H. Wood and D. A. Zwemor, J Vacuum Sci and Tech, *18,* No. 2 (March 1981), pp 649–50.
Multiple-Image Dark-Field Electron Microscopy of Beam-Sensitive Materials. A. J. Lovinger and H. D. Keith, J Poly Sci, Phys Ed, *19,* No. 7 (July 1981), pp 1163–6.
Photoelectrochemical Etching of p-GaAs. F. W. Ostermayer, Jr. and P. A. Kohl, Appl Phys Lett, *39,* No. 1 (July 1981), pp 76–8.

CONTENTS, JANUARY 1982

ERRATUM

Articles in the September 1981 (Part 2) Bell System Technical Journal mistakenly specify *UNIX* as a registered trademark of Bell Laboratories. *UNIX* is an unregistered trademark.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION

(Act of August 12, 1970: Section 8685. Title 39, U.S. Code)

1. Title of pubication: Bell System Technical Journal.
2. Date of filing: November, 1981
3. Frequency of issue: Monthly, except combined May–June and July–August.
4. Publication office: Bell Laboratories, 185 Monmouth Pky., West Long Branch, N. J. 07764.
5. Headquarters: American Telephone and Telegraph Co., 195 Broadway, New York, N. Y.
6. Editor and publisher: B. G. King, Editor, Bell Laboratories, 185 Monmouth Pky., West Long Branch, N. J. 07764. Publisher, American Telephone and Telegraph Co., 195 Broadway, New York, N. Y.
7. Owner: American Telephone and Telegraph Co., 195 Broadway, New York, N. Y.
8. Bondholders, mortgages and other security holders: none.
9. Extent and nature of circulation:

	Average number of copies of each issue during preceding 12 months	Actual number of copies of single issue published nearest to filing date
A. Total no. copies printed (Net press run)	11,937	12,429
B. Paid circulation		
1. Sales through dealers and carriers, street vendors and counter sales	1,953	2,973
2. Mail subsciptions	4,550	3,619
C. Total paid circulation	6,503	6,592
D. Free distribution by mail, carrier or other means, samples, complementary, and other free copies	4,651	5,047
E. Total distribution (sum of C and D)	11,154	11,639
F. Copies not distributed		
1. Office use, left over, unaccounted, spoiled after printing	783	790
2. Returns from news agents	None	None
G. Total (sum of E and F—should equal net press run shown in A)	11,937	12,429

I certify that the statements made by me above are correct and complete.

Bernard G. King, Editor

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering*, *Applied Mechanics Review*, *Applied Science & Technology Index*, *Chemical Abstracts*, *Computer Abstracts*, *Current Contents/ Engineering, Technology & Applied Sciences*, *Current Index to Statistics*, *Current Papers in Electrical & Electronic Engineering*, *Current Papers on Computers & Control*, *Electronics & Communications Abstracts Journal*, *The Engineering Index*, *International Aerospace Abstracts*, *Journal of Current Laser Abstracts*, *Language and Language Behavior Abstracts*, *Mathematical Reviews*, *Science Abstracts* (*Series A*, *Physics Abstracts*; *Series B*, *Electrical and Electronic Abstracts*; and *Series C*, *Computer & Control Abstracts*), *Science Citation Index*, *Sociological Abstracts*, *Social Welfare, Social Planning and Social Development*, and *Solid State Abstracts Journal*. Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System