THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 55 April 1976 Number 4

Step Response of an Adaptive Delta Modulator

By W. M. BOYCE

(Manuscript received May 6, 1974)

N. S. Jayant has proposed a simple but effective form of adaptive delta modulation which uses two positive parameters, P and Q, to adjust the step size. The values $P = Q = 1$ describe linear delta modulation (LDM), and Jayant has recommended using $Q = 1/P$ and $1 < P < 2$. In this paper, we study the step response of this scheme for arbitrary P and Q. For each P and Q, we are able to define an integer n, the stability exponent for P and Q, such that the step response is unstable when $P^nQ > 1$, it converges to the new level when $P^nQ < 1$, and when $P^nQ = 1$, it eventually settles into a periodic $(2n + 2)$-step cycle, for almost all initial conditions. For $P \geqq 2$, and for some combinations of P and Q with P between 1.6 and 2, it is possible to have both $PQ < 1$ and $P^nQ \geqq 1$, so that $PQ < 1$ is not sufficient for convergence. When a system is convergent, but a minimum step size δ is imposed, the eventual periodic hunting will not necessarily resemble that of LDM, but will be bounded by δP^n.

I. INTRODUCTION

The basic concepts of delta modulation (DM) have been thoroughly discussed in several recent publications.[1,2] In its simpler forms, delta modulation is a method of digitally encoding an input signal $\mathbf{X} = \{x_i\}$ into binary pulses $\mathbf{C} = \{c_i\}$ (where each $c_i = \pm 1$) so that an approximation $\mathbf{Y} = \{y_i\}$ of $\mathbf{X}$ may be reconstructed from the pulses $\mathbf{C}$ by a simple decoding scheme. The signal $\mathbf{X}$, although presented to the encoder as a discrete-time sequence, will normally be a sampled (and

perhaps digitized) version of a continuous-time analog signal. The encoder works by comparing each x_i with y_{i-1} through a feedback circuit to determine the sign of the subsequent pulse c_i, according to the equations

$$\begin{aligned} c_i &= \text{sign}\ (x_i - y_{i-1}) \\ m_i &= c_i M_i, \qquad \text{where} \qquad M_i = |m_i| > 0 \\ y_i &= y_{i-1} + m_i. \end{aligned}$$

Various forms of delta modulators differ primarily in the manner of determining the step-size M_i; of course, since only the pulses **C** are to be transmitted to the decoder, what is required is a rule for determining M_i from **C**. In conventional *linear* delta modulation (LDM), the step-size M_i is taken to be a constant δ, independent of the pulses **C** (and the signal **X**), so that each step $m_i = \pm\delta$, resulting in the familiar "staircase" appearance of **Y** under LDM. Since in this simplest form of DM, **Y** can change by only δ per step, no matter how far x_i is from y_{i-1}, **Y** has a very limited ability to keep up with **X** when **X** has a steep slope, which results in the condition known as slope overload. In contrast to LDM, *adaptive* delta modulation (ADM) permits M_i to be modified depending on **X**, especially as the slope of the signal **X** changes. Since this relieves the slope-overload problem, such adaptation can result in better encoding, and several types of adaptive delta modulators have been described in the literature (for a survey, see Ref. 2).

In this paper, we are concerned with the particular ADM scheme devised by N. S. Jayant,[3] and with certain generalizations of this scheme which arise naturally in the course of the investigation. Jayant's one-bit-memory scheme has been characterized by Steele[2] as "instantaneously companded" (that is, having an "instantaneous" adjustment of the step-size M_i), and Steele refers to Jayant's ADM as "first order constant factor delta modulation." The method is "first order," since Jayant computes M_i using only c_{i-1} in addition to M_{i-1} and c_i; the "one-bit memory" is used to save c_{i-1}. When c_i and c_{i-1} are equal, so that **Y** has not yet crossed **X**, there is a possibility of slope overload, so that M_i should be *increased*, and Jayant uses a "constant factor" $P \geqq 1$ so that $M_i = PM_{i-1}$ (and $m_i = Pm_{i-1}$) when $c_i = c_{i-1}$. To keep the step size from growing continuously with time, a second positive constant factor $Q \leqq 1$ is chosen, so that when c_i and c_{i-1} have *different* signs, indicating that **Y** has crossed **X**, the step size is reduced: $M_i = QM_{i-1}$, so $m_i = -Qm_{i-1}$. (Jayant concluded that values of P and Q with $PQ = 1$ gave the best performance on segments of speech, and he especially recommended $P = \frac{3}{2} = 1.5$, $Q = \frac{2}{3}$.) We note that when $P = Q = 1$, we recover LDM, with $M_i = \delta$ and $m_i = \pm\delta$ for all i.

As even basic LDM has proved to be quite difficult to analyze (see Refs. 5 and 6 for some recent successful efforts), it is hardly surprising that there are few definite analytical conclusions concerning the behavior of Jayant's ADM. This is confirmed by Steele's comment that "An interesting feature of instantaneously adaptive [delta modulators] is their resistance to mathematical analysis $\cdots$." Thus, in this paper, we restrict our attention to the comparatively simple problem of the step response of the approximating signal **Y** for Jayant's ADM, where by step response we mean the ultimate behavior of **Y** when **X** assumes a constant value $\bar{x}$, $x_j = \bar{x}$ for all $j \geqq i$.

For LDM, if **X** becomes constant, $x_j = \bar{x}$ for $j \geqq i$, then **Y** will eventually enter a "hunting" phase having a two-step period in which adjacent values of **Y** bracket $\bar{x}$ (see Fig. 1); for some k and all $j \geqq 0$,

$$y_{k+2j} = y_k \leqq \bar{x},$$
$$y_{k+2j+1} = y_k + \delta \geqq \bar{x}.$$

Thus, for LDM, **Y** will eventually get and remain no more than δ away from a constant signal **X**, which is a very desirable characteristic. This approximation error, which occurs because **Y** is discrete and cannot exactly match a constant or slowly varying signal **X**, is called "granular error" (or "quantization error"), in contrast to the "slope-overload" error which results from the inability of **Y** to keep up with a steeply climbing **X**. For LDM, a one-time compromise between these two types of error must be made in the choice of the sampling rate and step-size δ; then the granular error is known to be bounded by δ, but the slope-overload error can be severe for unexpectedly steep slopes. For ADM the step size can be varied with the signal, thus reducing the slope-overload error, but nature and magnitude of the granular error is less understood than for the LDM case, a situation which it is hoped that this paper will help resolve.

The question of the nature of the step response of Jayant's ADM was briefly discussed by Jayant in Section 2.3 of Ref. 3, but his conclusions were limited to the finding that in contrast to LDM, the characteristics of the "hunting" phase of the ADM, particularly the minimum step size and maximum error, were very dependent on the mag-

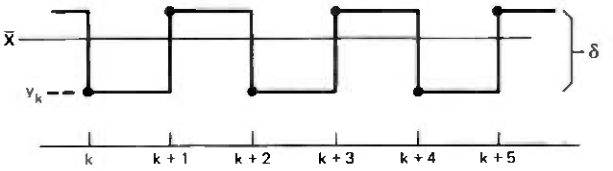


Fig. 1—Period-two (LDM) hunting.

nitude of the constant value $\bar{x}$ (with y_0 and m_0 held fixed). Figure 2, taken from Ref. 3, shows the behavior for $P = \frac{3}{2}$, $Q = \frac{2}{3}$, $y_0 = 0$, $m_0 = 1$, and $\bar{x} = 9, 10, 12$. Steele's analysis[2] showed that the *four*-step cycle exhibited in all three cases of Fig. 2 is exact and sustainable; as shown in Fig. 3, for some k and all $j \geqq 0$, the cycle is given by

$$\begin{aligned} y_{k+4j} &= y_k < \bar{x} \\ y_{k+4j+1} &= y_k + m < \bar{x} \\ y_{k+4j+2} &= y_k + m(P+1) > \bar{x} \\ y_{k+4j+3} &= y_k + mP > \bar{x}, \end{aligned}$$

where $m = m_{k+1} > 0$. Steele further indicated that this four-step periodic behavior is the typical ultimate step response of Jayant's ADM when $PQ = 1$. He also concluded that $PQ < 1$ was necessary for **Y** to converge to **X** for a step input, but he did not provide a complete proof, and he did not claim that $PQ < 1$ was sufficient for the decay of **Y** to a constant $\bar{x}$. (We note that when **Y** is in this four-step cycle, which is a "pure hunting" phase, the signal **X** is crossed only on alternate steps, and the signal value is typically not in the middle of the crossing step, calling into question assumptions used in Section IV of Ref. 3 and in Ref. 4.)

Even before the appearance of Steele's work, experimental results and preliminary analysis had given rise to the general supposition that

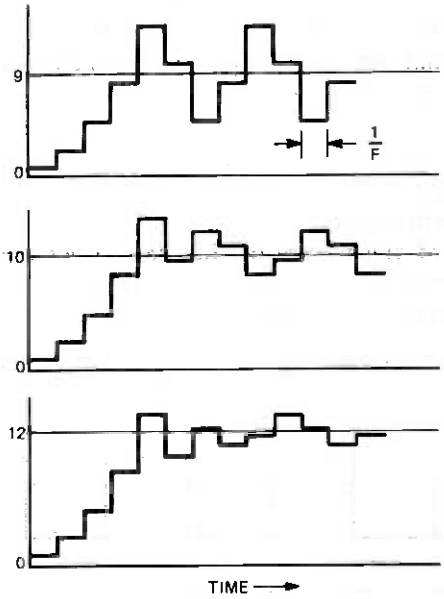


Fig. 2—$PQ = 1$ step responses (from Jayant[3]).

for a step input, (*i*) **Y** would be unstable when $PQ > 1$ (as it was for Jayant's speech data[3]), (*ii*) that when $PQ = 1$, **Y** would ultimately fall into the periodic four-step cycle, but with very large hunting amplitudes possible, and (*iii*) that for $PQ < 1$, **Y** would converge to the constant $\bar{x}$, with both step size and maximum hunting amplitude approaching zero. (Although having the step size get too small is considered undesirable in case **X** should begin to vary, it was generally thought that enforcing a well-chosen minimum step-size δ, as Jayant did in Ref. 3, would avoid this problem.) The question of convergence of **Y** for $PQ < 1$ is the most important of these, since as Steele and others have observed, using a value of PQ slightly less than 1, together with a minimum step size, would eliminate the problem of large-amplitude hunting cycles in **Y** during times when **X** was carrying no signal, while Jayant's results[3] indicate that for $PQ < 1$ but close to 1, the resulting penalty in signal-to-noise ratio during speech segments is negligible.

II. SUMMARY

Our findings on the step response of a P, Q delta modulator confirm that for almost all initial conditions, **Y** will be unstable when $PQ > 1$, and will eventually fall into the four-step cycle shown in Fig. 3 when $PQ = 1$. (We say "almost all" because for each P and Q with $PQ \geqq 1$, there is a set W of initial conditions, negligible in the sense of Lebesgue measure, for which **Y** converges to **X**. In Fig. 2, there would be convergence for $\bar{x} = 11.1625$, so that $y_0 = 0$, $m_0 = 1$, and $\bar{x} = 11.1625$ is a point of W.) More importantly, we find that $PQ < 1$ is *not* sufficient to insure that **Y** will converge to a step input **X**. Rather, in addition to those values of P and Q with $PQ < 1$ for which **Y** converges to **X**, there are values of P and Q with $PQ < 1$ for which **Y** is unstable, and also some combinations for which **Y** is eventually periodic, with a period even and greater than four. However, our results establish that

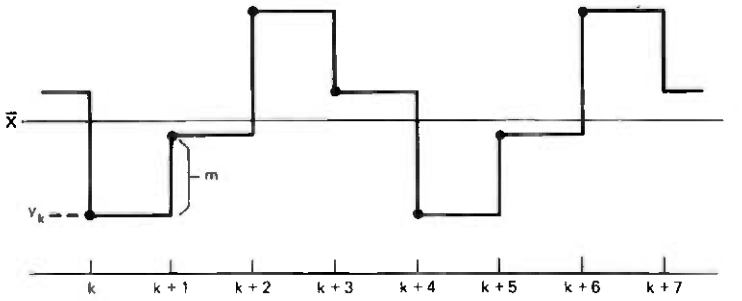


Fig. 3—Period-four ADM hunting.

when $PQ < 1$ and *either* $P \leqq 1.6$ *or* $PQ \geqq 1 - Q$, which are the cases of most practical interest at present, then the $PQ < 1$ conjecture is true, and **Y** converges to a step-input **X** for all initial conditions.

Our basic result is that for each P and Q, we can define an integer n, which we call the *stability exponent for P and Q*, such that the stability of the step response of **Y** depends not on the product PQ, as had been supposed, but on the product P^nQ. Thus, for almost all initial conditions, **Y** is unstable if $P^nQ > 1$, and is eventually periodic with period $2n + 2$ if $P^nQ = 1$; while for $P^nQ < 1$ (or whenever the initial conditions fall in W), **Y** converges to **X**. The generally expected findings for $PQ \geqq 1$ result from the fact that $n = 1$ when $PQ \geqq 1$.

It is useful to describe the stability exponent n in terms of P and PQ. If we define $F_k(P) = P(P - 1)/(P^k - 1)$, then n is the stability exponent for P and Q if and only if $F_{n+1}(P) \leqq PQ < F_n(P)$. Figure 4 shows the graphs of $F_k(P)$ for $k = 1, 2, 3, 4$. We see that $F_{k+1}(P) < F_k(P)$ for $P > 1$, so that n is well defined, and that $F_{k+1}(P)$ approaches zero with increasing k. Thus, n becomes unbounded as Q approaches zero.

The cases of most interest are those for which $PQ < 1$ and **Y** is not convergent, that is, when $F_{n+1}(P) \leqq PQ < F_n(P)$ *and* $P^nQ \geqq 1$. Since $F_{n+1}(P) \leqq P^{-n+1}$, $P^nQ \geqq 1$ implies $PQ \geqq F_{n+1}(P)$, so the bind-

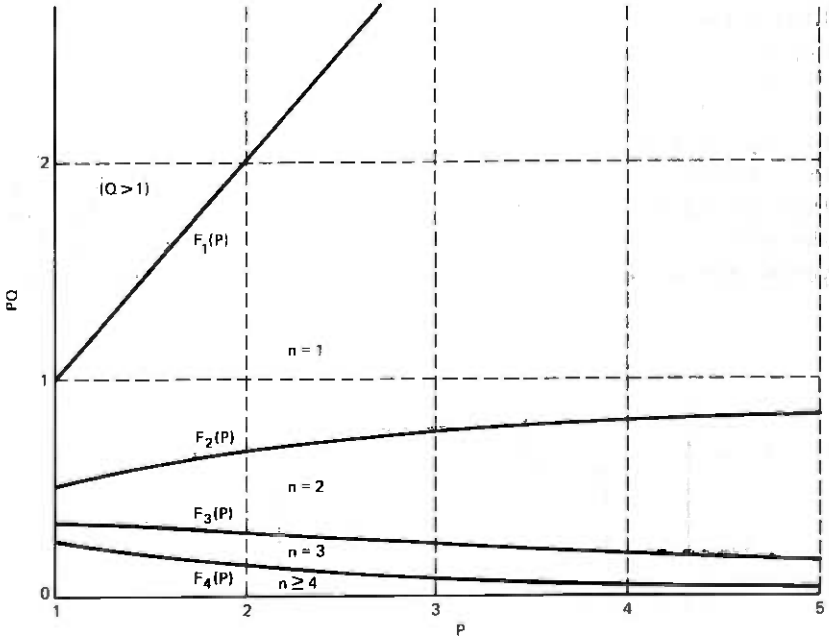


Fig. 4—Domains of the stability exponent n.

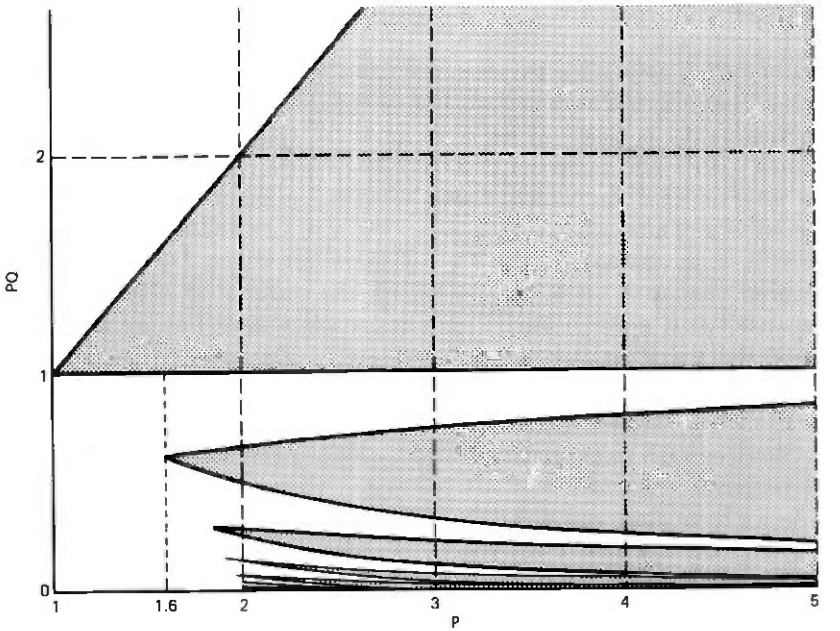


Fig. 5—Domains of unstable step response.

ing constraints are that $PQ < F_n(P)$ and $PQ \geqq P^{-n+1}$. In Fig. 5, those areas for which $P^{-n+1} < PQ < F_n(P)$ are shaded; they represent those values of P and PQ for which **Y** is unstable for almost all initial conditions. Looking particularly at the cases with $PQ < 1$, we see that when $P \leqq 1.6$, **Y** is never unstable, but even such seemingly safe cases as $P = 2$, $Q = 0.3$ fall in the shaded region. As P is made larger, which might be useful in some applications, the combinations for which **Y** is unstable become dominant, so that for $P = 4$, not only those values of Q above $\frac{1}{4}$ cause instability, but also all those between $\frac{1}{16}$ and $\frac{1}{5}$, as well as most values below $\frac{1}{16}$. The basic point of these examples is, of course, that it is not PQ which determines the stability of **Y**, but P^nQ.

The combinations for which $P^nQ = 1$ are interesting in that their step response is a straightforward generalization of that of Jayant's $PQ = 1$ ADM. Specifically, if we first decide on the stability exponent n, choose a $P \geqq 1$ which satisfies

$$P^{n+1} - 2P^n + 1 > 0,$$

and then set $Q = P^{-n}$ so that $P^nQ = 1$, we find that for almost all initial conditions, **Y** will eventually settle into a cycle of period $2n + 2$ steps. The $PQ = 1$ ADM thus appears as the $n = 1$ member of this

family, while LDM may be viewed as the $n = 0$ case: $P^0Q = Q = 1$, with a $2 \cdot 0 + 2 = 2$ step period. For each n, the set of P which satisfies the inequality consists of an open interval $(p_n, +\infty)$, where p_n increases with n and approaches 2 from below; $(p_n, +\infty)$ is also exactly the interval of P for which **Y** can be unstable when the stability exponent is n. Thus, when $n > 1$, the $P^nQ = 1$ ADM is feasible primarily for $P \geqq 2$, in contrast to the $PQ = 1$ ADM, for which Jayant has conjectured that 2 is an upper bound on the optimal P. These "high-response" ADM may be useful for some applications, but we have not tested them against any data. They seem to offer yet another method of trading off granular error against slope overload. Of course, as for the $PQ = 1$ case, one would actually set P^nQ slightly less than 1, but large enough to preserve n as the stability exponent and thus insure convergence.

As we have observed, the primary current interest is in combinations of P and Q for which **Y** converges to a step input **X**, so any practical system will provide for a minimum step-size δ. Thus, for a step input, the theoretically convergent **Y** will eventually encounter the minimum step size and become periodic, hunting about the constant $\bar{x}$. We have considered the step response of a P, Q delta modulator with minimum step size and stability exponent n, and we find that the eventual periodic behavior is exactly that of a P, $Q' = P^{-k}$ delta modulator with stability exponent k, where $0 \leqq k \leqq n$ and $P > p_k$, and where the value of k depends on the initial conditions. Thus, the hunting amplitude is bounded by $\delta P^k \leqq \delta P^n$. Moreover, all those k for which $0 \leqq k \leqq n$ and $P > p_k$ occur for initial conditions having positive Lebesgue measure. In particular, when $1 > PQ \geqq 1 - Q$, so that the stability exponent is $n = 1$, the four-step hunting cycle with range $\delta(1 + P)$ cannot be rejected. Thus, Steele's conclusion that the $k = 0$, LDM-type hunting is the only type that can occur when a minimum step size is imposed does not appear to be justified.

Our investigations also shed some light on the question of recognizing when the slope-overload condition is occurring. Since in the limit for $P^nQ = 1$, the sequence is n "forwards," one "reverse," etc., with only the nth forward crossing the signal, a sequence of n or fewer forwards should not be considered indicative of slope overload. But for $n + k$ forwards in a row, even if we decide to label k of them as slope overload, it is not clear which k of them: first? last? middle? Perhaps the magnitude of the error must be considered as well as crossings. On the other hand, for $P^nQ = 1$, distance *alone* cannot be used as the definition since the amplitude of the hunting can be quite large, depending on the initial conditions. For $P^nQ < 1$ with a minimum step size, much

the same considerations apply, except that in this case, the error magnitude would be very useful in recognizing hunting.

III. ANALYSIS

We assume that $i \geqq 0$ for all i, and that "initial conditions" x_0, y_0, and m_0 are given. Since there are no bounds on **X** or **Y**, we may assume that $\bar{x} = 0$, and that the "step" in **X** occurs at $i = 1$, that is, that $x_i = \bar{x} = 0$ for $i \geqq 1$. The effects of the previous history of **Y** and **X** can be summarized in the selection of y_0 and m_0. The step response of **Y** for a P, Q delta modulator is then characterized by how well **Y** can approximate $\bar{x} = 0$ as a function of the parameters P and Q and the initial conditions y_0 and m_0.

Jayant's ADM calculates **Y** from **X** by the following equations:

$$c_i = \text{sign } (x_i - y_{i-1})$$

$$m_i = \begin{cases} Pm_{i-1} & \text{if} \quad c_i = c_{i-1} \\ -Qm_{i-1} & \text{if} \quad c_i = -c_{i-1} \end{cases}$$

$$y_i = y_{i-1} + m_i.$$

Since $(x_i - y_{i-1})$, c_i, and m_i will always have the same sign, we may summarize the first two equations as

$$m_i = \begin{cases} Pm_{i-1} & \text{if } (x_i - y_{i-1}) \text{ and } m_{i-1} \text{ have the same sign} \\ -Qm_{i-1} & \text{if they have different signs.} \end{cases}$$

There is ambiguity in this definition, as the sign of zero is not defined; that is, what value of c_i is chosen when $x_i = y_{i-1}$? Our later analysis indicates that the proper choice is $c_i = -c_{i-1}$ when $x_i = y_{i-1}$, so that equality is considered to be a "crossing." After making this convention, and after observing that $x_i = \bar{x} = 0$ implies sign $(x_i - y_{i-1}) = -\text{sign } (y_{i-1})$ for $i \geqq 1$, we obtain the equations

$$m_{i+1} = \begin{cases} Pm_i & \text{if } y_i \text{ and } m_i \text{ have different signs} \\ -Qm_i & \text{if they have the same sign (or if } y_i = 0) \end{cases}$$

$$y_{i+1} = y_i + m_{i+1}.$$

This is a two-state system whose state equations have a discontinuity at $y_i = 0$, but we can transform it into a single-state continuous system if we note that the conditions on the comparative signs of y_i and m_i may be expressed as a condition on the sign of their *ratio*, which is always defined since m_i is never zero.

We define the *error-step-size ratio* r_i by $r_i = y_i/m_i$. Then we have

$$\begin{aligned} r_{i+1} &= y_{i+1}/m_{i+1} = 1 + y_i/m_{i+1} \\ &= 1 + (y_i/m_i)(m_i/m_{i+1}) = 1 + r_i(m_i/m_{i+1}), \end{aligned}$$

where

$$\frac{m_{i+1}}{m_i} = \begin{cases} P & \text{if } y_i \text{ and } m_i \text{ have different signs} \\ -Q & \text{if they have the same sign (or } y_i = 0) \end{cases}$$
$$= \begin{cases} P & \text{if } y_i/m_i = r_i < 0 \\ -Q & \text{if } y_i/m_i = r_i \geqq 0; \end{cases}$$

so the state equation for the ratio may be written simply as

$$r_{i+1} = \begin{cases} 1 + r_i/P & \text{if } r_i < 0 \\ 1 - r_i/Q & \text{if } r_i \geqq 0. \end{cases}$$

Thus, the sequence of ratios r_i arises from repeated applications, beginning with $r_0 = y_0/m_0$, of the function $f(\cdot)$ given by

$$f(r) = \begin{cases} 1 + r/P & \text{if } r < 0 \\ 1 - r/Q & \text{if } r \geqq 0. \end{cases}$$

This function is graphed in Fig. 6 for $P = \frac{3}{2}$, $Q = \frac{2}{3}$. Note that $f(\cdot)$ is continuous at $r = 1$, and the continuity is not dependent on our choice of c_i when $x_i = y_{i-1}$, since $f(0) = 1$ simply says that $y_i = x_i + m_i$ when $x_i = y_{i-1}$, which is true no matter how one computes m_i from m_{i-1}. But an important observation is that a particular sequence of r_i's computed from $r_{i+1} = f(r_i)$, together with an initial step m_0, gives the complete sequence of m_i's, since a negative r_i indicates $m_i = Pm_{i-1}$, while an r_i which is positive *or zero* indicates $m_i = -Qm_{i-1}$. Thus, the convention on the sign of zero affects not the sequence of r_i's but the sequence of m_i's derived from it.

We shall henceforth restrict ourselves to combinations of P and Q for which $P > 1$ and $Q < 1$, since this is the only case (other than $P = Q = 1$) that is suitable for practical applications.

In our subsequent analysis we are primarily concerned with the function $f(\cdot)$, which describes how the ratio $r_i = y_i/m_i$ changes from one step to another. Since $f(r) \leqq 1$ for all r, except for r_0 no r_i can

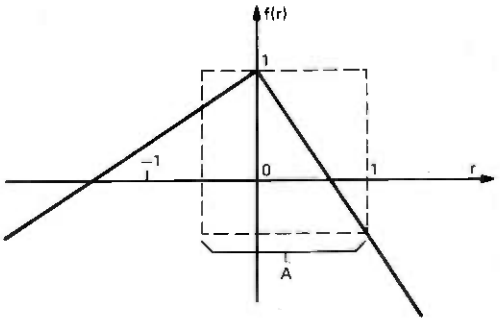


Fig. 6—The graph of $f(r)$ and A for $P = \frac{3}{2}$, $Q = \frac{2}{3}$.

exceed 1. Thus, after the first step we are not concerned with the behavior of $f(r)$ for $r > 1$.

We are not only interested in the change in the error-step-size ratio r_i during one step, which is given by $r_{i+1} = f(r_i)$, but also in the change over two steps, three steps, etc. The change in the ratio over j steps may be determined by applying the function j times, e.g., $r_{i+2} = f(r_{i+1}) = f(f(r_i))$, $r_{i+3} = f(f(f(r_i)))$, etc. The function obtained by applying $f(\cdot)$ j times we call the jth *iterate* of $f(\cdot)$, which we write $f^j(\cdot)$. Thus, we have $r_{i+j} = f^j(r_i)$, and by convention $f^1(r) = f(r)$ and $f^0(r) = r$.

Since $f(r) = 1 + r/P \geqq 1 + r$ when $P > 1$ and $r < 0$, when r is negative, the successive values of $f^j(r)$ will increase by at least 1 per step until finally one of the values $f^j(r)$ is nonnegative, that is, $0 \leqq f^j(r) \leqq 1$. This is just another way of saying that the signal **Y** will eventually cross zero on step y_{i+j} beginning at $r = r_i < 0$. But once $f^j(r)$ is in the interval $[0, 1]$, the next value of the ratio, namely $f^{j+1}(r)$, can be no smaller than $f(1) = 1 - 1/Q$, which we denote by q. If $f^{j+1}(r) < 0$, then the subsequent ratios will increase again until they reach $[0, 1]$, etc. Thus, the ratios can never escape the interval $[q, 1] = [1 - 1/Q, 1] = A$ once they enter, and we have proven:

Theorem 1: *If* $q = f(1) = 1 - 1/Q < 0$ *and* $A = [q, 1]$, *then for each* r *there is a* j *such that* $f^j(r) \in A$, *and* $r_i \in A$ *implies* $r_k \in A$ *for all* $k \geqq i$.

So the ultimate behavior of the ratios is determined by the function $f(\cdot)$ and its iterates on the interval $A = [q, 1]$, and thus by the graph of $f(\cdot)$ on the square $A \times A$, denoted by the dotted lines in Fig. 6.

We shall need more precise information on how many steps are necessary to go from a given ratio r to a zero crossing, or a nonnegative value of $f^j(r)$. We define $a_1 = 0$, $a_2 = -P$ and, in general, $a_{i+1} = a_i - P^i = -\sum_{j=1}^{i} P^j$. We further define $A_0 = [0, 1]$, and $A_i = [a_{i+1}, a_i)$ for $i \geqq 1$. Since $P > 1$, this set of intervals forms a disjoint cover of the range $(-\infty, 1]$ of $f(\cdot)$.

Theorem 2: *If* $r \in A_j$, *then* j *is the least integer such that* $f^j(r)$ *is nonnegative, so that* $r \in A_j$ *if and only if* $r \leqq 1$ *and exactly* j *steps produce a zero crossing of* **Y**. *Also, the sequence* $f^i(r)$ *is increasing for* $0 \leqq i \leqq j$.

Proof: Since $f(a_{i+1}) = a_i$ for $i \geqq 1$, it follows that $f(A_{i+1}) = A_i$ for $i \geqq 1$. Thus, if $r \in A_j$, after $j - 1$ steps, $f^{j-1}(r) \in A_1$. Then, $f(A_1) = [0, 1) \subset [0, 1] = A_0$, so $f^j(r) \in [0, 1]$.

Corollary: *For every* $r_0 = y_0/m_0$, *the ratios eventually enter and remain in* A.

Proof: For $r \leqq 1$, we have $f^j(r) \in [0, 1] \subset A$, while for $r > 1$, $f(r) \leqq 1$ so that $f(r) \in A_j$ for some j, so that $f^{j+1}(r) \in [0, 1]$.

We can now define n, the *stability exponent for P and Q*, as the largest value of j such that A_j intersects A; that is, it is the maximum number of steps from a ratio r *in* A to a zero crossing. Clearly, n is determined by the fact that $q < 0$, so that $q \in A_n$ for some $n > 0$, and this n is the stability exponent. More explicitly, P and Q must satisfy

$$a_{n+1} \leqq q < a_n$$

or

$$-\sum_{j=1}^{n} P^j \leqq 1 - 1/Q < -\sum_{j=1}^{n-1} P^j$$

or

$$\sum_{j=0}^{n} P^j \geqq 1/Q > \sum_{j=0}^{n-1} P^j.$$

To obtain the conditions cited in the summary, we invert and multiply by P to obtain

$$F_{n+1}(P) \leqq PQ < F_n(P),$$

where

$$F_k(P) = P \Big/ \sum_{j=0}^{k-1} P^j = \frac{P(P-1)}{P^k - 1}.$$

Another way of expressing this condition is

$$\frac{P^n - 1}{P - 1} < \frac{1}{Q} \leqq \frac{P^{n+1} - 1}{P - 1},$$

so multiplying by $(P - 1)Q$ and adding Q gives

$$P^nQ < P + Q - 1 \leqq P^{n+1}Q.$$

Thus, the stability exponent n is the largest n such that P^nQ is strictly less than the quantity $P + Q - 1$. Note that $Q < 1$ implies $P^nQ < P + Q - 1 < P$, so that $P^{n-1}Q < 1$ whenever n is the stability exponent for P and Q.

Theorem 3: *If n is the stability exponent for P and Q, and $P^nQ < 1$, then Y converges for all initial conditions, that is, both m_i and y_i tend to zero with increasing i.*

Proof: Once the ratios enter A, no more than n negative ratios can occur without an intervening nonnegative ratio. Thus, as m_i evolves by multiplication of P's and $(-Q)$'s, each $-Q$ can be grouped with at most n P's with no P's left over. Since $P^nQ < 1$, the absolute value of m_i will be decreasing by a factor bounded away from 1 at least every

$(n + 1)$ steps, and hence going to zero. Each time a ratio is nonnegative, which occurs at least once every $(n + 1)$ steps, **Y** has just crossed zero, so y_i must go to zero along with m_i.

The next theorem is the basic result of the theory of Jayant's adaptive delta modulation. It states that not only is the stability exponent n the *maximum* number of successive negative ratios that can occur once the ratios enter A, but that for *almost all* initial conditions (initial ratios r_0), a sequence of n negative ratios all in A will eventually occur. (Here by "almost all" we mean that the set of initial conditions for which this is false has Lebesgue measure zero—it can be covered by a family of open intervals of arbitrarily small total length.) This result is the key to the analysis for $P^nQ \geqq 1$.

Theorem 4: *Let* $B_n = A \cap A_n = [q, a_n)$ *and let* B *be the set of* $r \in A$ *such that* $f^j(r) \in B_n$ *for some* j *(so that* n *successive negative ratios eventually occur). Then* B *is open (as a subset of* A*) and has Lebesgue measure* $\mu(B) = 1/Q = 1 - q$*, the length (and Lebesgue measure) of* A*. Thus,* $A \backslash B$ *(the points of* A *not in* B*) is a measurable set of Lebesgue measure zero.*

Proof: B_n is open in A, and B can be written as

$$B = \bigcup_{i=0}^{\infty} \{r \mid f^i(r) \in B_n\}.$$

Since each $f^i(\cdot)$ is a continuous function from A into A, each set in the union is open, so B itself is open. Thus, B and its complement $A \backslash B$ are measurable. Clearly, if S is a subset of B, and S' is a subset of A such that $f(S') = S$, then S' is a subset of B also. In addition, if $f(\cdot)$ is linear with slope $1/s$ on S', $f(S') = S$, and S and S' are measurable, then $\mu(S') = |s| \cdot \mu(S)$. For each i, $0 \leqq i \leqq n$, let $B_i = A_i \cap B$, so that each B_i is measurable with measure $\mu(B_i)$. Now $f(\cdot)$ maps A_0 linearly onto A, with slope $-1/Q$, so $f(\cdot)$ must map B_0 linearly onto B, and $\mu(B_0) = Q \cdot \mu(B)$. Similarly, for each i such that $n - 1 > i > 0$, $f(\cdot)$ maps A_{i+1} linearly onto A_i with slope $1/P$, so $f(\cdot)$ must map B_{i+1} linearly onto B_i, and $\mu(B_{i+1}) = P \cdot \mu(B_i)$. When $i = 0$, $f(\cdot)$ maps A_1 linearly onto $A_0 \backslash \{1\}$, so $\mu(B_1) = P \cdot \mu(B_0 \backslash \{1\})$; but since $\{1\}$ has measure zero, $\mu(B_1) = P \cdot \mu(B_0)$ also. Thus, for $0 \leqq i < n$, we have $\mu(B_i) = P^i \cdot \mu(B_0)$. But since B is the disjoint union of the B_i, we have

$$\mu(B) = \sum_{i=0}^{n} \mu(B_i) = \mu(B_n) + \sum_{i=0}^{n-1} P^i \cdot \mu(B_0)$$

$$= (a_n - q) + \mu(B_0) \cdot \sum_{i=0}^{n-1} P^i.$$

Since $\mu(B_0) = Q \cdot \mu(B)$, and $\sum_{i=0}^{n-1} P^i = 1 - a_n$,

$$\mu(B_0)/Q = (a_n - q) + (1 - a_n)\mu(B_0)$$

or

$$\mu(B_0)(1/Q - 1 + a_n) = \mu(B_0)(a_n - q) = a_n - q.$$

Since $q < a_n$ (this relies on the convention that $m_{i+1} = -Q \cdot m_i$ when $y_i = 0$), we have $\mu(B_0) = 1$, so $\mu(B) = 1/Q = 1 - q = \mu(A)$. Thus $\mu(A \backslash B) = 0$.

Corollary: *Let W be the set of real numbers r such that $f^i(r) \notin B_n$ for all i, i.e., once $f^j(r)$ is in A, no sequence of n successive negative ratios ever occurs. Then W has Lebesgue measure zero.*

Proof: Let $W_0 = A \backslash B$ and for all i, let W_i be the set of r for which $f^i(r) \in W_0$. Since each $f^i(\cdot)$ is piecewise linear, each W_i has measure zero, so $W = \bigcup_{i=0}^{\infty} W_i = \{r \mid f^i(r) \in W_0 \text{ for some } i\}$ has measure zero. But since W_0 is the set of $r \in A$ such that $f^i(r) \notin B_n$ for all i, W is the set of (unrestricted) r such that $f^i(r) \notin B_n$ for all i.

We note that W is nonempty for all $P > 1$ and $Q < 1$, since $f(\cdot)$ has a fixed-point $w = Q/(Q + 1) \in (0, 1)$, and w and all its preimages (r such that $f^i(r) = w$ for some i) will be in W. In addition, for all $i \geqq 2$, $f^i(\cdot)$ will have fixed points in addition to w, and many of these fixed points and their preimages will be in W also.

Theorem 5: *With n the stability exponent for P and Q, on $A_n = [a_{n+1}, a_n)$ the function $f^{n+1}(\cdot)$ is linear with slope $-(P^nQ)^{-1}$ and has a fixed point $z \in [q, a_n) = B_n$.*

Proof: Since $f(A_{i+1}) \subset A_i$ and $f(\cdot)$ is linear on each A_i, $f^j(\cdot)$ is linear on each A_i for $j \leqq i + 1$. Clearly, $f^n(a_{n+1}) = 0$ and $f^n(a_n) = 1$, so $f^{n+1}(a_{n+1}) = 1$ and $f^{n+1}(a_n) = f(1) = q = 1 - 1/Q$. The slope of $f^{n+1}(\cdot)$ on A_n is thus $(q - 1)/(a_n - a_{n+1}) = (-1/Q)/P^n = -(P^nQ)^{-1}$. Since $q \in A_n$ by definition, $q = f^{n+1}(a_n) < a_n$, but since $f^{n+1}(\cdot)$ has negative slope, $f^{n+1}(q) > f^{n+1}(a_n) = q$. Thus, $f^{n+1}(q) > q$, $f^{n+1}(a_n) < a_n$, and so $f^{n+1}(\cdot)$ has a fixed point z between q and a_n.

Theorem 6: *If $P^nQ > 1$, then $f^{n+1}(B_n) \subset B_n$, that is, if $r_i \in B_n = [q, a_n)$ then $r_{i+(n+1)k} \in B_n$ for all $k \geqq 0$. Thus, except for $r_0 \in W$, the ratios eventually enter B_n and return to B_n every $n + 1$ steps thereafter. Moreover, the ratios falling in B_n converge to the fixed point z of $f^{n+1}(\cdot)$.*

Proof: $f^{n+1}(a_n) = q$, and the absolute value of the slope of $f^{n+1}(\cdot)$ on A_n is $(P^nQ)^{-1} < 1$ so $|f^{n+1}(q) - f^{n+1}(a_n)| < |q - a_n|$ and so $f^{n+1}(B_n) = (q, f^{n+1}(q)] \subset (q, a_n) \subset B_n$. Each $f^{(n+1)k}(B_n)$ is an in-

terval containing z, and each increase in k (each $n + 1$ steps) reduces the length of the interval by a factor $(P^nQ)^{-1} < 1$, so for each $r \in B_n$ we have $f^{(n+1)k}(r)$ approaching z with increasing k. Thus, except for initial conditions in W, the ratios not only eventually enter B_n (by the corollary to Theorem 4) but return there every $n + 1$ steps, each time coming closer to z.

Corollary: *If n is the stability exponent for P and Q and $P^nQ > 1$, then for all initial conditions which are not in W, the signal $\mathbf{Y}$ is unstable. Also, if $r_i \in B_n$, then $M_{i+j} > M_i$ for all $j > 0$, where $M_i = |m_i|$.*

Proof: Once r_i is in B_n, every $n + 1$ steps M_i increases by a factor of $P^nQ > 1$; hence the step size increases without bound.

The next theorem and its corollary establish the nature of the stable, periodic step response which is characteristic of the Jayant family of delta modulators.

Theorem 7: *If n is the stability exponent for P and Q and $P^nQ = 1$, then $f^{2n+2}(\cdot)$ is the identity on B_n, and if y_i and m_i are such that $r_i = y_i/m_i \in B_n$, then whenever $j \geqq i$, $k \geqq 0$, and $l = (2n + 2)k$, we have $y_{j+l} = y_j$ and $m_{j+l} = m_j$, so that $\mathbf{Y}$ becomes periodic with period $2n + 2$ steps. Thus for all initial conditions which are not in W, $\mathbf{Y}$ eventually settles into a periodic $(2n + 2)$-step cycle.*

Proof: If $P^nQ = 1$, then the slope of $f^{n+1}(\cdot)$ is -1, so that $f^{n+1}(q) = a_n$ in addition to $f^{n+1}(a_n) = q$. Thus, $f^{2n+2}(a_n) = a_n$, $f^{2n+2}(q) = q$, so $f^{2n+2}(\cdot)$ is the identity on $[q, a_n]$ and hence on $B_n = [q, a_n)$ itself. Thus, when $r_j \in B_n$, $r_{j+2n+2} = r_j$. But by Theorem 2 we know that among the $2n + 2$ successive values of r_{j+i} there are $2n$ negative ones and 2 nonnegative ones, so that $m_{j+2n+2} = P^{2n}(-Q)^2 m_j = (-P^nQ)^2 m_j = m_j$. Thus, $y_{j+2n+2} = y_j$ as well. The connection with W is made as in previous theorems.

Theorem 8: *If $P^nQ \geqq 1$ and $r_0 \in W$, then y_i and m_i both converge to 0, i.e., for initial conditions in W, $\mathbf{Y}$ is neither unstable nor periodic but converges to $\mathbf{X}$.*

Proof: For all initial conditions, the ratios eventually enter and remain in A, but if $r_0 \in W$, then all ratios in A fall in the A_i with $i < n$. Thus, at most, $n - 1$ successive negative ratios can occur; hence, each $-Q$ can be grouped with no more than $n - 1$ P's with no P's left over. But $P^iQ < 1$ for $i < n$ even if $P^nQ > 1$, so at intervals of no more than n steps m_i will be reduced in absolute value by a factor bounded away from 1; hence m_i will converge to zero, and with it Y, since a zero crossing will occur at least every n steps.

We can now relate our findings to the general supposition on the stability of Jayant's delta modulator: that is, that the system is unstable, periodic, or convergent according to whether PQ exceeds, equals, or is less than 1. We see that the general supposition is in fact correct when $PQ \geqq 1 - Q$ and $r_0 \notin W$.

Theorem 9: *If $PQ \geqq 1 - Q$, then the stability exponent for P and Q is 1. Thus, $\mathbf{Y}$ converges to $\mathbf{X}$ when $1 - Q \leqq PQ < 1$ (or when $PQ \geqq 1$ and $r_0 \in W$), settles into a four-step cycle when $PQ = 1$ and $r_0 \notin W$, and is unstable when $PQ > 1$ and $r_0 \notin W$.*

Proof: All we must show is that $q = 1 - 1/Q \geqq a_2 = -P$, so that $q \in A_1$. But dividing $1 - Q \leqq PQ$ by $-Q$ yields $q \geqq -P$ as required. The rest follows from our earlier theorems, taking $n = 1$.

The most unexpected results of our analysis are the existence of both unstable combinations of P and Q with $PQ < 1$ and Jayant-type delta modulators that satisfy $P^nQ = 1$ and are eventually periodic with a $2n + 2$ step period when $n > 1$ (and $r_0 \notin W$). The next three theorems establish that since n depends on P and Q, in order to attain $P^nQ \geqq 1$ we must have $P > p_n$, where $p_1 = 1$, $p_2 \approx 1.62$, $p_i < p_{i+1}$, and $\lim_{i\to\infty} p_i = 2$. Thus, for $P \geqq 2$, all values of n are realizable, while for $P \leqq p_2 \approx 1.62$, only the $n = 1$ value will allow $P^nQ \geqq 1$. (The sequence $\{p_i\}$ that we define here comes up again in our subsequent analysis of a P, Q delta modulator with a minimum step size.)

Theorem 10: *If $P^kQ \geqq 1$, then $q \geqq a_{k+1}$, so the stability exponent for P and $Q \geqq P^{-k}$ cannot exceed k.*

Proof: Since $q = 1 - 1/Q \geqq 1 - P^k$, all we need show is that $1 - P^k \geqq a_{k+1} = -\sum_{i=1}^{k} P^i$, or $P^k \leqq \sum_{i=0}^{k} P^i$, which always holds. Thus, if $q \in A_n = [a_{n+1}, a_n)$, then $a_n > q \geqq a_{k+1}$ so $n \leqq k$.

Theorem 11: *We can choose a Q such that $P^nQ \geqq 1$, where n is the stability exponent for P and Q, if and only if P satisfies $P^{n+1} - 2P^n + 1 > 0$. Equivalently, n is the stability exponent for P and $Q = P^{-n}$ ($P^nQ = 1$) if and only if $P^{n+1} - 2P^n + 1 > 0$.*

Proof: If $P^nQ \geqq 1$, then $q = 1 - 1/Q \geqq 1 - P^n$. By the definition of n, $1 - P^n \leqq q < a_n = -\sum_{j=1}^{n-1} P^j$, so $P^n > \sum_{j=1}^{n-1} P^j = (P^n - 1)/(P - 1)$. But then $P^n(P - 1) = P^{n+1} - P^n > P^n - 1$, and $P^{n+1} - 2P^n + 1 > 0$. Since each of these steps can be reversed, if $P^{k+1} - 2P^k + 1 > 0$, then setting $Q = P^{-k}$, we have $q < a_k$, so $n \geqq k$. Since q is strictly less than a_k and $\partial q/\partial Q > 0$, there is an open interval of values of $Q \geqq P^{-k}$ for which $n \geqq k$. But by Theorem 10, $n \leqq k$

when $P^kQ \geq 1$, so for these values of Q we have $n = k$ and $P^nQ = 1$ or $P^nQ > 1$, respectively.

Theorem 12: *For each $k \geq 1$, let $\mathcal{P}_k$ be the set of $P > 1$ which satisfy $P^{k+1} - 2P^k + 1 > 0$. Then, each $\mathcal{P}_k$ is an open half-line $(p_k, +\infty)$, where $p_k < p_{k+1} < 2$ and $\lim_{k\to\infty} p_k = 2$.*

Proof: For $k = 1$, the requirement is simply that $(P - 1)^2 > 0$, so $p_1 = 1$. For $k \geq 2$, differentiating $g(P) = P^{k+1} - 2P^k + 1$ gives $g'(P) = (k + 1)P^k - 2kP^{k-1}$, whose only zero besides $P = 0$ is $P = 2k/(k + 1)$, which lies between 1 and 2 and approaches 2 with increasing k. Since $g(1) = 0$, $g'(1) = 1 - k < 0$, and $g(2) = 1$, $g(P)$ has a zero p_k between $2k/(k + 1)$ and 2, and $g(P) > 0$ for $P \geq 2$. Thus, $P > p_k$ implies $g(P) > 0$, and $1 < P < p_k$ implies $g(P) < 0$. Since $2k/(k + 1) < p_k < 2$, p_k approaches 2 with increasing k. Since $g(p_{k+1}) = p_k - 1 > 0$, $p_{k+1} > p_k$, so the sequence $\{p_k\}$ converges monotonically to 2.

In fact, since $g(2) = 1$ and $g'(2) = 2^k$, a good approximation for p_k is $2 - 2^{-k}$. For $k = 2, 3, 4$, the approximations are 1.75, 1.875, 1.9375 and the actual values 1.6180, 1.8393, 1.9275.

We have previously observed that the periodicity that occurs when $P^nQ = 1$ is undesirable in practical systems, since it may result in **Y** having significant power when **X** is zero or close to it. This problem is aggravated by the fact that the amplitude of the periodic hunting is unpredictable and can be quite large. To overcome this problem, Steele and others have suggested setting P^nQ slightly less than 1, so as to make the **Y** converge to **X**, and using a minimum step size, which we call δ, to prevent the step size from getting so close to zero during long stretches of zero (or constant) signal **X** that **Y** cannot quickly respond when **X** begins to vary. Indeed even when studying the case $PQ = 1$, Jayant used a minimum step size, although it was seldom binding (see Fig. 3 of Ref. 3).

In our final three theorems we treat the case of a P, Q delta modulator with a minimum step-size δ, so that when $M_i < \delta/Q$ and a zero crossing occurs, instead of the next step having magnitude $M_{i+1} = QM_i < \delta$, we set $M_{i+1} = \delta$. Thus, $M_i \geq \delta$ for all i. We note that if **Y** would be unstable or periodic in the absence of a minimum step size, then the step sizes may never be reduced to the point that the minimum becomes binding. If a step of size δ does occur, however, with $M_i = \delta$ and $r_{i-1} \in [0, 1] = A_0$, we show that **Y** eventually becomes periodic with a $2J + 2$ step cycle, where $0 \leq J \leq n$ (the stability exponent for P and Q) and $P > p_J$; with the exception of the case $P^nQ > 1$, $r_i \in B_n \subset A_n$, for which **Y** is unstable and a step of

size δ never reoccurs. Thus, in contradiction to Steele's conclusion, the step response of a P, Q delta modulator with minimum step size does not reduce to the LDM case, but is fully as complex as the $P^nQ = 1$ case with no minimum. However, it is true that if $P^nQ < 1$, or $P^nQ \geqq 1$ and $r_0 \in W$, the minimum step size will eventually occur and the hunting amplitudes be thereafter bounded by $P^n\delta$.

Theorem 13: *If* $r_i < 0$, *then* $r_{i+1} = f(r_i)$; *if* $r_i \geqq 0$ *and* $M_i \geqq \delta/Q$, *then* $r_{i+1} = f(r_i)$; *and if* $r_i \geqq 0$ *and* $\delta \leqq M_i < \delta/Q$, *then* $f(r_i) < r_{i+1} \leqq 1$. *Thus, for all initial conditions,* $r_i \in A$ *for some* i, *and if* $r_i \in A$ *then* $r_j \in A$ *for all* $j \geqq i$.

Proof: When $r_i \leqq 0$, we have $m_{i+1} = Pm_i$, so the minimum is not relevant, and when $r_i \geqq 0$ and $M_i \geqq \delta/Q$, we have $m_{i+1} = -Qm_i$, so the minimum is not binding. Thus, for these cases, $r_{i+1} = f(r_i)$. But when $r_i \geqq 0$ and $M_i < \delta/Q$, we have $f(r_i) = 1 - r_i/Q$ but $r_{i+1} = y_{i+1}/m_{i+1} = 1 + (m_i/m_{i+1})(y_i/m_i) = 1 + (m_i/m_{i+1})r_i$. Since $m_i/m_{i+1} < 0$, we can write this $r_{i+1} = 1 - (M_i/M_{i+1})r_i$. But $M_{i+1} = \delta$, $M_i < \delta/Q$ so $M_i/M_{i+1} < (\delta/Q)/\delta = 1/Q$, $0 \leqq 1 - r_{i+1} = r_i(M_i/M_{i+1}) < r_i/Q$, and so $1 \geqq r_{i+1} > 1 - r_i/Q = f(r_i)$. Thus, the evolution of r_i for $r_i < 0$ is given by $f(\cdot)$, so $r_i \in A$ and $r_i < 0$ implies $r_{i+1} \in A$; while if $r_i \in [0, 1]$, $q \leqq f(r_i) \leqq r_{i+1} \leqq 1$ so $r_{i+1} \in A$ in this case also.

For the next two theorems, we assume that a minimum step size has occurred, with $M_i = \delta$, and that $r_{i-1} \in A_0$ so that $r_i \in A$. Since $r_i \in A$, we must have $r_i \in A_J$ for some J, $0 \leqq J \leqq n$, where n is the stability exponent for P and Q. For almost all cases of interest, steps of size δ will continue to occur at least every J steps, and $\mathbf{Y}$ will be periodic; the sole exception, which we dispose of first, is when $P^nQ > 1$ and $J = n$, in which case $\mathbf{Y}$ is unstable and a step of size δ *never* reoccurs.

Theorem 14: *If* $M_i \geqq \delta$, $r_i \in A \cap A_n = B_n$, *and* $P^nQ > 1$, *then* $r_{i+j} = f^j(r_i)$ *and* $M_{i+j} > \delta$ *for all* $j > 0$, *so that Theorem 6 and its corollary apply and* $\mathbf{Y}$ *is unstable.*

Proof: If $M_i \geqq \delta$ and $r_i \in B_n$, then by Theorem 13, $r_{i+n} = f^n(r_i) \in A_0$, and $M_{i+n} = P^nM_i$. But $P^nQ > 1$, so $M_{i+n} > M_i/Q \geqq \delta/Q$, and $M_{i+n+1} = QM_{i+n} = P^nQM_i \geqq \delta P^nQ > \delta$. Thus, $r_{i+n+1} = f^{n+1}(r_i) \in B_n$, $M_{i+n+1} \geqq \delta P^nQ$, and so $r_{i+(n+1)k} \in B_n$ and $M_{i+(n+1)k} \geqq \delta(P^nQ)^k$ for all $k \geqq 0$.

The next theorem characterizes the ultimate behavior of the P, Q delta modulator with minimum step size for the more interesting cases —those not covered by Theorem 14. Thus, we assume that $M_i = \delta$ and $r_i \in A$, with $r_i \in A_J$, where $P^JQ \leqq 1$. Without loss of generality,

we choose signs so that $m_i = M_i = \delta$ and $y_i - \delta = y_{i-1} \leqq 0$ (we continue to assume $\bar{x} = 0$, i.e., **X** is identically zero). Since $r_i \in A_J$, we have $r_{i+J} \in A_0$ so $r_{i+J+1} \in A_K$ for some K, $0 \leqq K \leqq n$. To simplify the notation, we set $l = 2J + 2$.

Theorem 15: *If* $m_i = \delta$, $r_i \in A \cap A_J$, $P^JQ \leqq 1$, *and* $r_{i+J+1} \in A_K$, *then* $K \leqq J$. *If* $K = J$, *then* $P > p_J$, *and* $y_{i+l} = y_i$, $m_{i+l} = m_i$, *and* **Y** *is periodic with period* $2J + 2$ *and maximum amplitude* $\delta P^J \leqq \delta P^n$. *Moreover, for each* j *such that* $0 \leqq j \leqq n$ *and* $P > p_j$, *the set of initial conditions which produce a* $(2j + 2)$*-step period has positive Lebesgue measure.*

Proof: When $J \geqq 2$, we have $y_{i+1} = y_i + \delta P < 0$, $m_{i+1} = \delta P$; $y_{i+2} = y_i + \delta(P + P^2)$, $m_{i+2} = \delta P^2$; and, in general (even for $J = 0, 1$), we have $y_{i+J} = y_i + \delta \sum_{j=1}^{J} P^j \geqq 0$, $m_{i+J} = \delta P^J$. Since $P^JQ \leqq 1$, $\delta P^J \leqq \delta/Q$ so that $m_{i+J+1} = -\delta$. If $K \geqq J$, then

$$y_{i-1+l} = y_{i+J} - \delta \sum_{j=0}^{J} P^j = y_i - \delta = y_{i-1} \leqq 0,$$

so that $K \leqq J$; thus, $K \geqq J$ implies $K = J$, so we have proven that $K \leqq J$. If $K = J$, then we have seen that $y_{i-1+l} = y_{i-1}$; also, $m_{i+l} = \delta$ since $m_{i-1+l} = -\delta P^J$ and $P^J \leqq 1/Q$. Thus, $y_{i+l} = y_{i-1+l} + \delta = y_{i-1} + \delta = y_i$, and $m_{i+l} = \delta = m_i$, so **Y** is periodic with period $l = 2J + 2$. To show that $P > p_J$ when $K = J$ and **Y** has period $l = 2J + 2$, we observe that by the definition of J and K $(=J)$ we have (see Fig. 7,

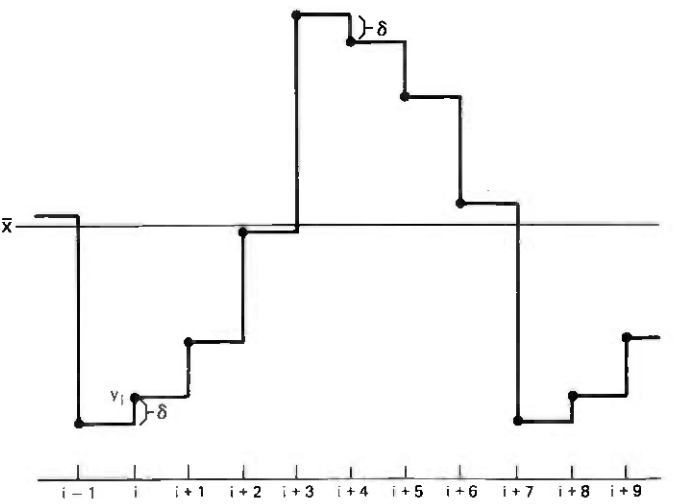


Fig. 7—Period-eight ADM hunting with minimum step-size δ.

with $J = 3$)

$$y_{i+J-1} < 0, \qquad y_{i+J} \geqq 0,$$
$$y_{i-2+l} > 0, \qquad y_{i-1+l} \leqq 0,$$

so that

$$y_{i+J-1} = y_{i-1} + \delta \sum_{j=0}^{J-1} P^j < 0$$

$$y_{i-2+l} = y_{i-1+l} + \delta P^J = y_{i-1} + \delta P^J > 0,$$

so that

$$y_{i-1} + \delta \sum_{j=0}^{J-1} p^j < y_{i-1} + \delta P^J,$$

so

$$\frac{P^J - 1}{P - 1} < P^J$$

from which

$$P^{J+1} - 2P^J + 1 > 0.$$

But this is the defining condition for $P > p_J$. To show that each j satisfying $0 \leqq j \leqq n$ and $P > p_j$ comes up with positive measure, it is only necessary to observe that choosing y_0, m_0 such that $\delta \leqq -m_0 \leqq \delta/Q$ and

$$-\delta P^J < y_0 < -\delta \sum_{j=1}^{J-1} P^j$$

will realize the $2J + 2$ step period analyzed above with $i = 1$.

We note that once a minimum step occurs, the series of "reversal numbers" (of which the J and K are two adjacent elements) is monotone decreasing ($K < J$) until it repeats itself ($K = J$), after which it is constant, and Y is periodic. This monotonicity holds only *after* δ occurs; when there is no minimum step size, there is no monotonicity, except that when $P^nQ \geqq 1$ an occurrence of $J = n$ will result in nothing but n's thereafter. What we have shown is:

Corollary: *If δ is the minimum step size and $M_i = \delta$ where $r_i \in A \cap A_j$, then unless $P^nQ > 1$ and $j = n$, within $(j + 1)^2$ steps Y will become periodic with period $2J + 2$, where $0 \leqq J \leqq j$.*

Proof: Until the reversal numbers become constant, at least every $j + 1$ steps a new, lower reversal number occurs, and there are only $j + 1$ possible such numbers; thus, within $(j + 1)^2$ steps the minimum number J is obtained and Y is periodic.

IV. ACKNOWLEDGMENTS

Special thanks are due to S. J. Brolin for a critical reading of the original manuscript and valuable suggestions on the most useful interpretation of the analytical results. The encouragement and helpful comments of N. S. Jayant and D. J. Goodman are also gratefully acknowledged.

REFERENCES

1. H. R. Schindler, "Delta Modulation," IEEE Spectrum, October 1970, pp. 69–78.
2. R. Steele, *Delta Modulation Systems*, New York: John Wiley, 1975. (Especially pp. 243–251 of Chapter 8.)
3. N. S. Jayant, "Adaptive Delta Modulation With a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
4. N. S. Jayant and A. E. Rosenberg, "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech," B.S.T.J., *50*, No. 10 (December 1971), pp. 3117–3125.
5. D. Slepian, "On Delta Modulation," B.S.T.J., *51*, No. 10 (December 1972), pp. 2101–2137.
6. A. Gersho, "Stochastic Stability of Delta Modulation," B.S.T.J., *51*, No. 4 (April 1972), pp. 821–841.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1976
Printed in U.S.A.

On the Design of All-Pass Signals With Peak Amplitude Constraints

By L. R. RABINER and R. E. CROCHIERE

(Manuscript received November 24, 1975)

In this paper, the problem is discussed of designing a signal other than the standard impulse function to be used to test a digital system of limited dynamic range. The constraints on such a signal are that it must be all-pass, of limited duration (approximately), and peak-amplitude-limited so as to utilize the limited dynamic range of the system as far as possible. Stated another way, the goal is to spread out the energy in the signal as much as possible to reduce its peak amplitude and therefore to be able to pass higher energy signals through the system without clipping them. The class of all-pass signals (obtained as the impulse response of a variable order all-pass filter) was investigated for use as the test signal. The parameters of the all-pass filter of a given order were optimized to give an all-pass signal whose peak amplitude was the smallest possible. Filter orders from first to eighth order were designed and investigated. It was found that reductions in the peak signal level of up to 11.2 dB (relative to the signal level of an equivalent energy impulse) could be obtained for an eighth-order all-pass signal. Interpolated versions of these all-pass signals showed that the peak value of the interpolated waveform was only on the order of 6 dB. Thus, the use of an all-pass signal, rather than the standard impulse, for testing a digital system can result in about 1 bit extra dynamic range.

I. INTRODUCTION

The problem of designing digital signals for testing (e.g., evaluating the impulse response) digital systems is one which has received very little attention in the digital signal-processing literature. This is because the impulse function is used as the standard test signal for most systems. Although the impulse function is suitable for this purpose in a wide variety of digital systems, there are cases in which the use of the impulse function leads to problems. Generally, such systems are those that have limited dynamic range—e.g., digital hardware implementations of a system, or fixed-point, finite, precision, software implementation of a digital system. In this paper, the problem is con-

sidered of designing signals other than the standard impulse function to be used to test digital systems of limited dynamic range.

The desirable features of a test signal for digital systems are

(*i*) It must be an all-pass signal in that it must be capable of testing the system (i.e., determining the frequency response of the system) for any admissible frequency.
(*ii*) It should be of limited duration.
(*iii*) It should be peak-amplitude-limited, to give the maximum utilization of the limited dynamic range of the system.

The above features define a desirable test signal as one whose energy is spread out as much as possible to reduce the peak signal amplitude and therefore be able to pass higher energy signals through the system without clipping.

If we let $x(n)$ denote the test signal, then the requirements described above can be related to $x(n)$ and $X(e^{j\omega})$, the Fourier transform of $x(n)$, in the following manner. For the signal to be all-pass implies

$$|X(e^{j\omega})| = C, \qquad \text{all } \omega, \tag{1}$$

where C is an arbitrary constant value. If we let $C = 1$, then by Parseval's theorem we have

$$\frac{1}{2\pi}\int_0^{2\pi} |X(e^{j\omega})|^2 d\omega = 1 = \sum_{n=0}^{\infty} x^2(n), \tag{2}$$

i.e., the overall energy of the test signal is unity. For the signal to be of limited duration (at least approximately) requires

$$\sum_{n=0}^{N_1-1} x^2(n) = \gamma, \tag{3}$$

where $\gamma \approx 1$ and N_1 is the signal duration in samples. (The constraint of (3) has not been used directly in the work presented here, since it was found that it was satisfied by all the signals that were designed.) Finally, the constraint that the peak signal amplitude be as small as possible requires that $\max_n |x(n)|$ be minimized over the design parameters of the signal.

Besides the standard impulse function, the only other class of signals that is appropriate for a test function (i.e., that has the set of features described above) is the set of all-pass filter impulse responses. Such signals can be optimized to meet the design requirements by varying the parameters of the all-pass network to minimize the peak signal amplitude.

The purpose of this paper is to discuss the issues in the design of all-pass signals to be used to test a digital system. In Section II, the design

methods used to optimize these all-pass signals are discussed. In Section III, considerations dealing with the interpolation of the resulting all-pass test signals are given. Finally, in Section IV a brief discussion of the effects of filtering these all-pass signals is given.

II. DESIGN TECHNIQUES FOR ALL-PASS SIGNALS

The signal design problem is one of choosing the parameters (the filter coefficients) in the implementation of an Nth-order all-pass filter to minimize the peak amplitude of the resulting impulse response. For the actual implementation of most all-pass filters, it is generally convenient to consider the cascade realization which is of the form

$$X(z) = \prod_{i=1}^{N_s} H_i(z), \tag{4}$$

where N_s is the number of sections in the cascade and $H_i(z)$ are the individual sections, which generally are either first-order or second-order sections. A first-order all-pass section has the system function

$$H_i(z) = \frac{-a + z^{-1}}{1 - az^{-1}}, \tag{5}$$

whereas a second-order all-pass section has the system function

$$H_i(z) = \frac{b_i - c_i z^{-1} + z^{-2}}{1 - c_i z^{-1} + b_i z^{-2}}. \tag{6}$$

The design problem is thus to choose the all-pass parameters (a, b_i, c_i) to minimize the peak signal amplitude in the impulse response of the filter.

For the first-order case, the parameter a can be analytically determined. In this case, the difference equation is

$$x(n) = u_0(n-1) - au_0(n) + ax(n-1), \tag{7}$$

where

$$u_0(n) = \begin{cases} 1 & n = 0 \\ 0 & \text{otherwise,} \end{cases}$$

or

$$\begin{aligned} x(n) &= 0 && n < 0 \\ x(0) &= -a \\ x(1) &= (1 - a^2) \\ x(n) &= (1 - a^2)a^{n-1}, && n \geqq 2. \end{aligned} \tag{8}$$

Since $|a| < 1$ for stability, it is seen from (8) that the largest possible

samples are $x(0)$ and $x(1)$. Thus, to minimize the larger of $|x(0)|$ and $|x(1)|$ requires a choice of a such that

$$|x(0)| = |x(1)| \tag{9}$$

or

$$|a_{\min}| = |1 - a_{\min}^2|, \tag{10}$$

The solution to (10) gives $a_{\min} = 0.618$.

For optimization of higher-order all-pass filters, no analytical solution could be found. Thus, an optimization method was used to obtain the desired solutions. In particular, a nonlinear unconstrained optimization method developed by Powell[1] was used in which the evaluation of derivatives was not required. The maximum peak amplitude of the all-pass signal can be minimized by minimizing the function

$$G = \lim_{p \to \infty} \left[\sum_{n=0}^{\infty} |x(n)|^p \right]^{1/p}. \tag{11}$$

In practice, however, the function of (11) is not unimodal or smooth, and thus it is not practical to find the optimum choice of parameters without a good starting point (initial choice of parameters) for the optimization routine. To obtain such starting points, (11) was used as the objective function for a value of $p = 4$. A variety of randomly chosen starting points was used to obtain the best solutions for $p = 4$. The $p = 4$ solutions were then used as starting points to determine the optimum $p = \infty$ solutions.

The parameters that were varied within the optimization program were the b_i's and c_i's of the second-order sections within the cascade and the a for a first-order section (used whenever the order of the all-pass filter was odd). The advantage of using the cascade realization is that it is simple to ensure stability of the resulting filter. Additionally, instabilities occurring during the optimization program because of poles drifting outside the unit circle were easily detected and corrected with minimal computational effort.

Using the Powell optimization method, the optimum all-pass signals of order 1 to 8 were designed. Table I gives values of the optimum all-pass filter parameters and the resulting peak signal level for each of these cases. It is seen in this table that the peak signal level falls from 0.618 to 0.275 as the all-pass filter order varies from first to eighth order. Further, it can be seen that progressive increases in the order of the all-pass filter result only in very modest reductions of the peak signal level beyond a second-order filter. Figures 1 and 2 show the positions of the poles and zeros of the optimum all-pass filters and their group delay responses for each of the filters of Table I.

Table I — Filter coefficients for optimum all-pass filters with peak amplitude constraints

Filter Order	Maximum Signal Level	a	b_1	c_1	b_2	c_2	b_3	c_3	b_4	c_4
1	0.618	0.6180	—	—	—	—	—	—	—	—
2	0.500	—	0.5	1.0	—	—	—	—	—	—
3	0.428	0.8698	0.4915	0.6961	—	—	—	—	—	—
4	0.386	—	0.8008	0.5137	−0.4823	0.3491	—	—	—	—
5	0.3380	0.6734	0.6081	0.4462	0.7996	−0.5253	—	—	—	—
6	0.3183	—	0.6151	1.4640	0.8228	0.3748	−0.6290	0.0197	—	—
7	0.2895	−0.5183	0.8339	−0.4254	0.7668	1.5407	0.8735	0.4700	—	—
8	0.2748	—	0.8149	1.2308	−0.4970	−0.1060	0.8621	−0.2135	0.7870	1.5727

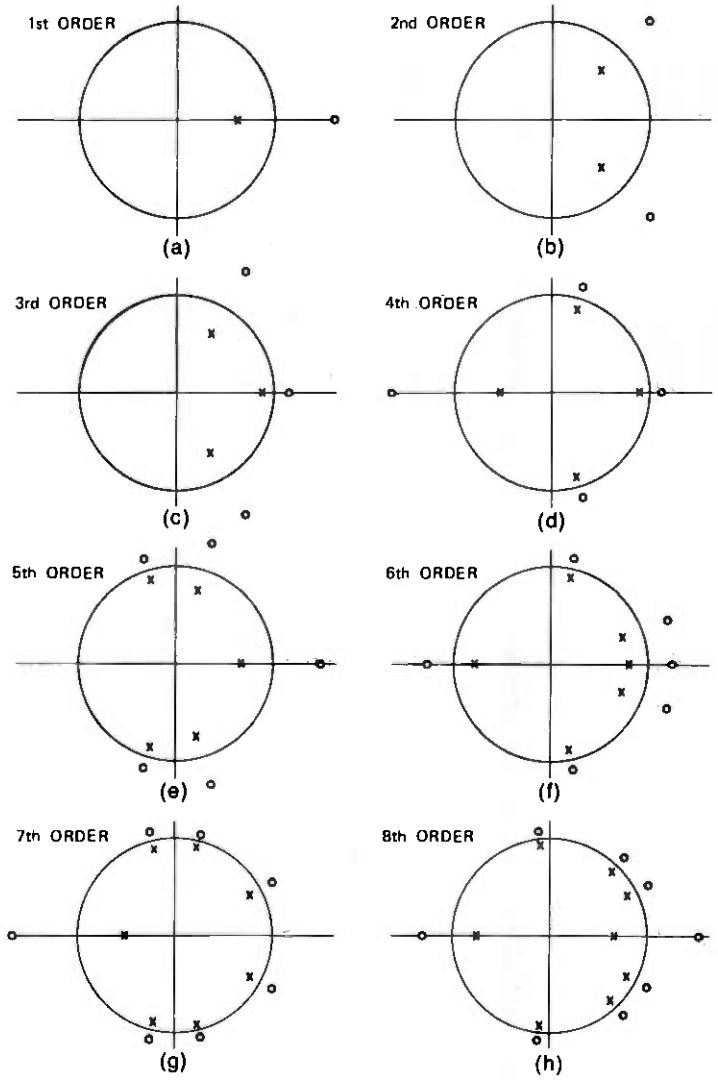


Fig. 1—Positions of the poles and zeros of the optimized all-pass signals of order 1 to 8.

An interesting property of this class of signals is that the optimum all-pass filter is not unique. This result is readily seen since the simple replacement of z by z^{-1} in the z transform leads to a multiplication of the signal by $(-1)^n$, which does not affect the signal magnitude at all. Thus, each pole and zero of Fig. 1 could equally be shown reflected about the imaginary z axis and still be a valid optimum solution.

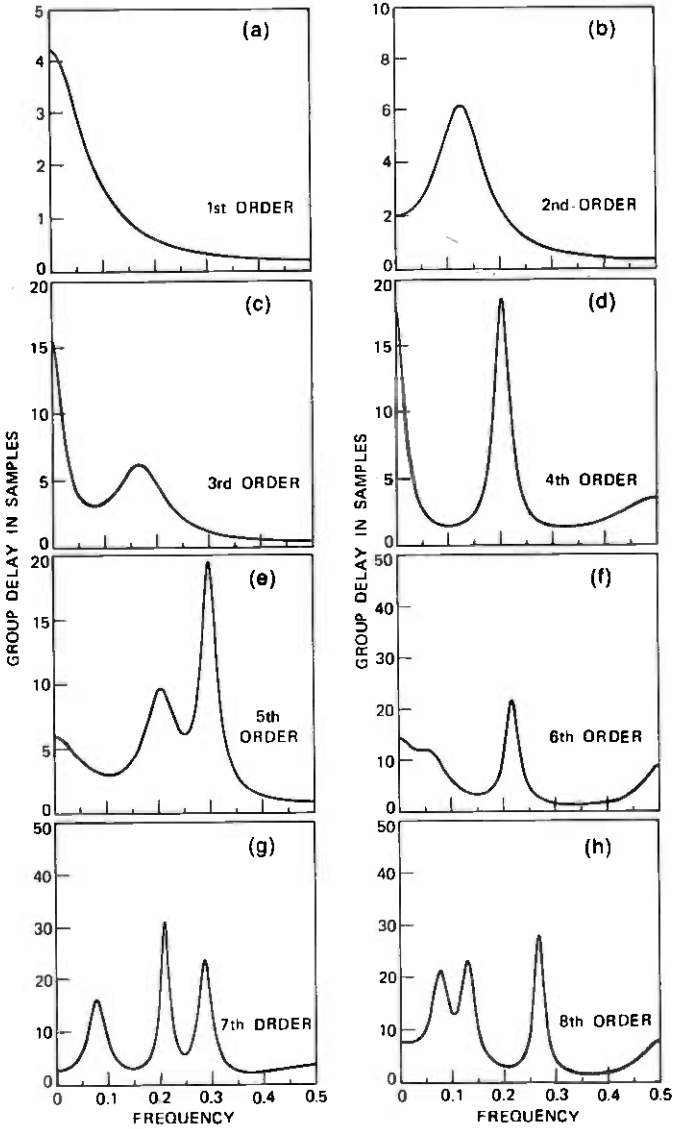


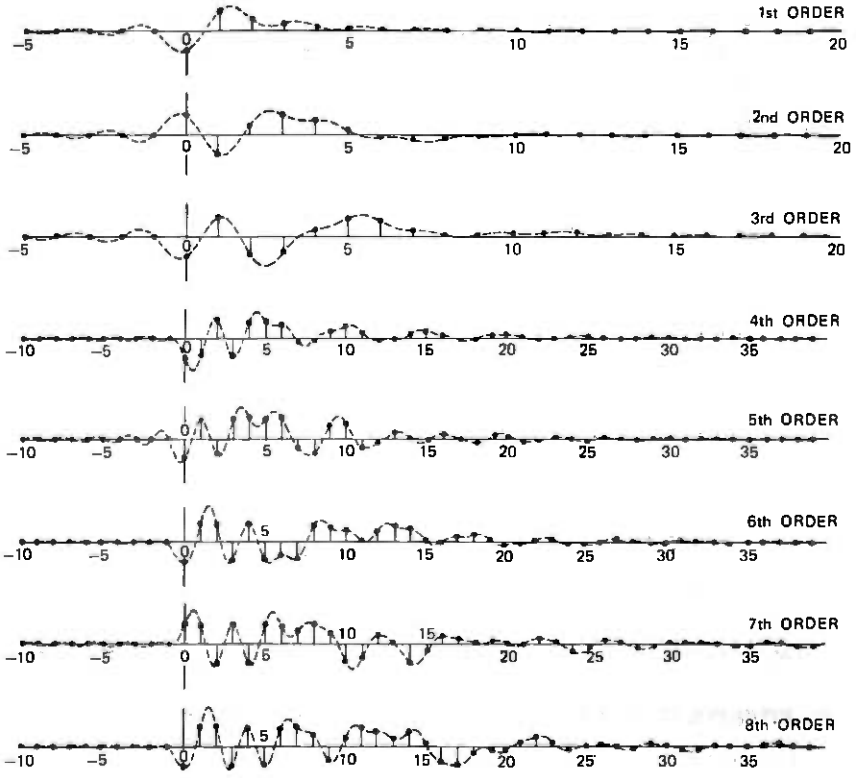
Fig. 2—Group delay responses of the optimized all-pass signals of order 1 to 8.

III. INTERPOLATION OF THE OPTIMUM ALL-PASS SIGNALS

The results of the preceding section indicate that reductions in the peak level of the optimized all-pass signal on the order of 4 to 1 can be obtained with an eighth-order filter. This result can be somewhat misleading, however, since the continuous waveform (from which the

signal samples could be derived) could peak up between samples—i.e., the actual reduction in signal level could be a fortuitous result obtained by sampling the waveform at the most opportune sampling intervals. If this were the case, and the test signal was used as input to a network which approximated a noninteger delay, the output signal could be of higher amplitude than the input signal simply because of the interpolative properties of the network.

To investigate the true peak amplitude of the continuous waveform associated with the test signal, each of the eight test signals of Table I were interpolated using a 20-to-1 interpolator implemented using the methods described by Crochiere and Rabiner.[2,3] Figure 3 and Table II show the results of interpolating the test signals. Figure 3a shows both the test signal samples as well as the interpolated waveforms (dotted



(a)

Fig. 3—Samples and interpolated waveforms of (a) the all-pass signals for orders 1 to 8 and (b) the all-pass signals modulated by $(-1)^n$.

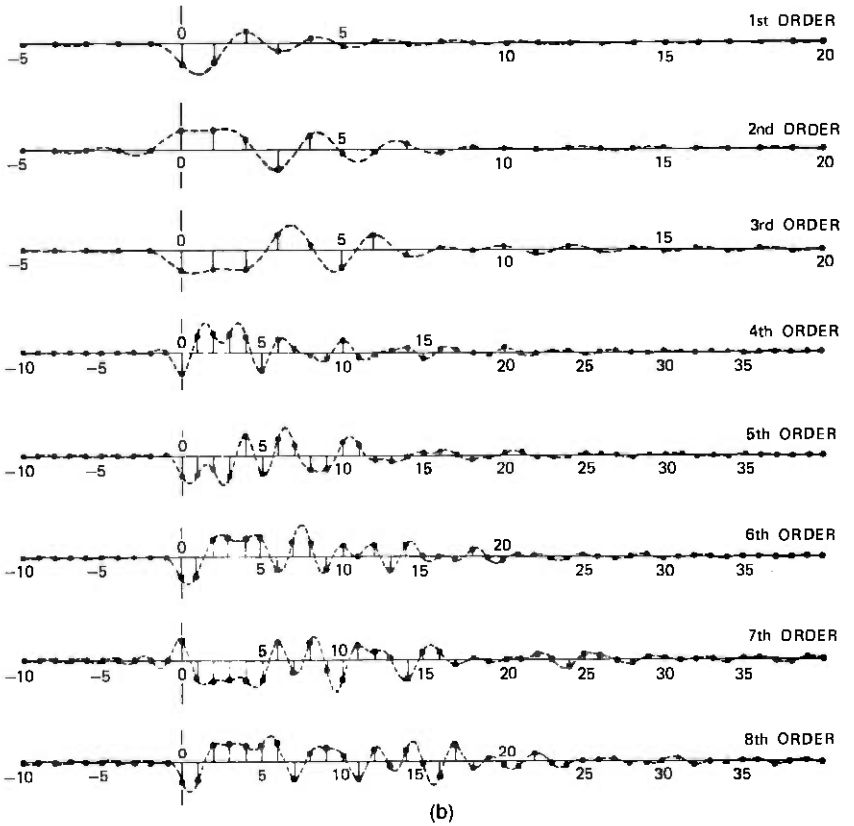


Fig. 3 (continued).

lines) associated with the signals. Figure 3b shows the alternate set of peak-limited waveforms formed by multiplication of the signals in Fig. 3a by $(-1)^n$. Although each test signal attains its peak amplitude at a number of different sampling instants, its interpolated waveform generally shows a distinct maximum amplitude. Table II also shows that the peak interpolated waveform amplitude ranged from 0.766 for the first-order signal to 0.421 for the seventh-order signal. Thus, in terms of the interpolated waveform, on the order of a 2-to-1 reduction in peak signal level was obtained for these test signals.

One more observation can be obtained from Fig. 3 and that is that the test signals, although generated as the output of a recursive structure, damp out in level extremely rapidly and could be considered finite duration signals. It was found that 128 samples of the test signal were sufficient for obtaining 16-bit test signals to full 16-bit accuracy.

Table II — Comparison of signal level of peak-limited signals to a unit sample, all with unit signal energy

Filter Order	Peak-Limited Signal				Waveform (Interpolated) of Peak-Limited Signal				Waveform (Interpolated) of Peak-Limited Signal with $(-1)^n$ Modulation			
	Max.	Min.	Ratio	Ratio dB	Max.	Min.	Ratio	Ratio dB	Max.	Min.	Ratio	Ratio dB
1	0.618	−0.618	0.618	−4.18	0.766	−0.633	0.766	−2.32	0.377	−0.924	0.924	−0.68
2	0.500	−0.500	0.500	−6.02	0.595	−0.522	0.595	−4.52	0.533	−0.501	0.533	−5.46
3	0.428	−0.428	0.428	−7.37	0.465	−0.632	0.632	−3.98	0.545	−0.502	0.545	−5.27
4	0.375	−0.386	0.386	−8.27	0.505	−0.627	0.627	−4.06	0.586	−0.391	0.586	−4.65
5	0.338	−0.334	0.338	−9.42	0.526	−0.391	0.526	−5.57	0.474	−0.503	0.503	−5.97
6	0.301	−0.318	0.318	−9.95	0.585	−0.345	0.585	−4.65	0.525	−0.442	0.525	−5.60
7	0.290	−0.290	0.290	−10.75	0.497	−0.367	0.497	−6.07	0.338	−0.457	0.457	−6.79
8	0.273	−0.275	0.275	−11.21	0.544	−0.313	0.544	−5.29	0.387	−0.421	0.421	−7.51

IV. APPLICATION OF PEAK-LIMITED SIGNALS AS TEST SIGNALS

One application of the above class of peak-limited signals is for use as test signals for systems of limited dynamic range. By spreading the signal energy among many samples, a test signal of greater total energy than an impulse can be used without exceeding the dynamic range of the system. This then enhances the signal-to-noise ratio (s/n) of the measurement.

For a system that has approximately a linear-phase response, s/n improvements of the orders shown in Table II can be expected. If the system has considerable phase distortion, the amount of s/n enhancement may be less. In an extreme case, a system could act as a "matched filter" to a particular test signal and compress all the signal energy back into a single sample. In this case, no s/n improvement would be possible with that test signal, although other peak-limited test signals in this class might be useful.

To investigate the use of the peak-limited signals as test signals, we chose a system that consists of a complex modulator, a decimator, an interpolator, and another complex modulator. The system was implemented on a 16-bit computer, and the decimator and interpolator were designed as discussed in Refs. 2 and 3. The net function of the above system is that of a bandpass filtering operation. It represents a useful type of system for speech-processing applications (e.g., vocoders).

The frequency response of the system is shown in Fig. 4a. It was measured by exciting the system with the peak-limited signal for $N = 7$ and taking the Fourier transform of the output. The largest peak amplitude signal which could be used without overflow was 16384, or 2^{14}. Similarly, the largest impulse that could be used as a test signal was 2^{14}. The frequency response measurement in this case was essentially equivalent to that using the peak-limited signal (see Fig. 4a). The reason for this is apparent. The 16-bit system has a large dynamic range (about 90 dB) compared to the frequency response of the filter (about 45 dB). Obviously, the use of peak-limited signals is not warranted.

We next considered a 12-bit implementation of the same system.* This would very likely be the available word length of a practical hardware implementation or small minicomputer implementation. In this case, the dynamic range of the system is about 66 dB, and we can expect that roundoff noise will affect the frequency response measurement. The largest magnitude impulse that could be used to test this

* This was simulated on the 16-bit system by not allowing the use of the four most significant bits.

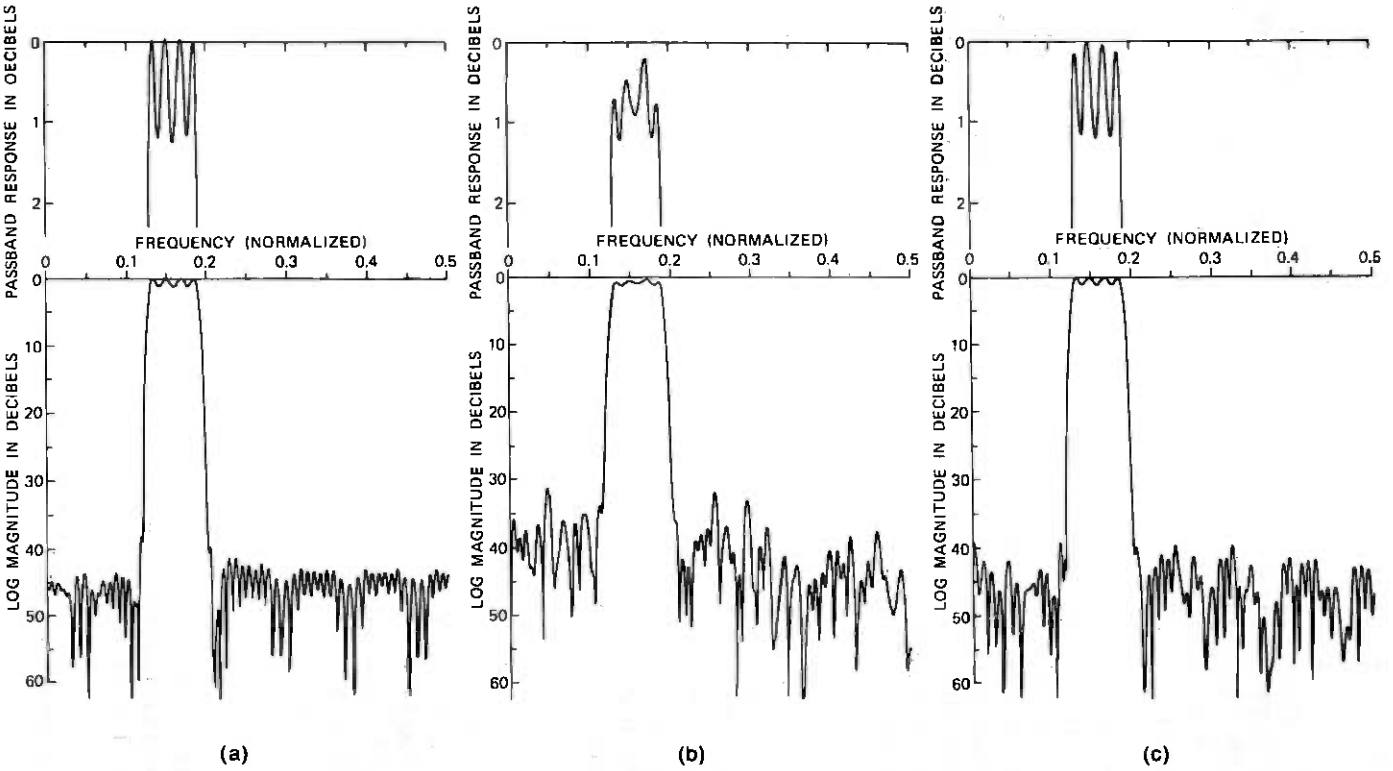


Fig. 4—Frequency response measurement using (a) an $N = 7$ all-pass test signal in a 16-bit system, (b) an impulse test signal in an equivalent 12-bit system, and (c) an $N = 7$ all-pass test signal in the 12-bit system.

system without overflow was 1024, or 2^{10}. A measurement of the frequency response based on this impulse response is shown in Fig. 4b. It is apparent that the roundoff noise has degraded the measurement considerably. The passband response has been distorted, and the peak stopband signal rejection measures only 31 dB compared to 41 dB in Fig. 4a.

Figure 4c shows the frequency response measurement of the same 12-bit system based on the peak-limited signal for $N = 7$. The maximum amplitude that could be used for this signal was 2^{10} and, as can be seen from Table II, it contains 10.75 dB more signal energy than an impulse of the same amplitude. In comparing Figs. 4a, b, and c, it is clear that the use of the peak-limited signal has improved the frequency response measurement of the 12-bit system. The measurement of the stopband rejection is on the order of 40 dB, or 9 dB better than in Fig. 4b. The passband response looks more like the essentially noiseless measurement in Fig. 4a.

V. CONCLUSIONS

It has been shown that a class of peak-limited and essentially finite duration signals can be generated by optimizing the $\rho = \infty$ norm of the impulse responses of the class of all-pass networks. Signals were generated for all-pass filter orders from $N = 1$ to $N = 8$. It was demonstrated that this class of signals is useful as test signals for systems of limited dynamic range. Improvements of up to 11 dB in s/n enhancement were found to be possible.

REFERENCES

1. M. J. Powell, "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," Computer J., *7*, 1964, pp. 155–162. (Computer code available in J. L. Kuester and J. H. Mize, *Optimization Techniques with Fortran*, New York: McGraw-Hill, 1973.)
2. R. E. Crochiere and L. R. Rabiner, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation, and Narrow Band Filtering," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 5 (October 1975), pp. 444–456.
3. R. E. Crochiere and L. R. Rabiner, "Further Considerations in the Design of Decimators and Interpolators," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-24*, No. 4 (August 1976).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1976
Printed in U.S.A.

Analysis of a Gradient Algorithm for Simultaneous Passband Equalization and Carrier Phase Recovery

By D. D. FALCONER

(Manuscript received December 11, 1975)

A two-dimensional receiver structure has been proposed, incorporating two innovations: passband equalization, which mitigates intersymbol interference, and data-directed carrier recovery and demodulation following equalization, which enables compensation of carrier frequency offset and phase jitter, but does not require transmission of a separate pilot tone with the data signal. The receiver is fully adaptive; the adjustment of the equalizer tap coefficients and of the estimate of the current channel phase shift is based on a gradient algorithm for jointly minimizing the mean squared error with respect to those parameters.

In this paper, we analyze the dynamic behavior of the deterministic gradient algorithm (where channel parameters entering into the gradient expression are assumed known in advance). The corresponding estimated gradient algorithm (where these parameters are initially unknown) has previously been studied experimentally, but is not treated here.

The first part of the present study concerns system start-up (or transient) response when the channel's phase shift is fixed. Examination of the analytical solution leads to the qualitative conclusion that, if the equalizer tap adaptation coefficient β is small relative to the phase-tracking coefficient α, the added phase estimation feature does not strongly affect the start-up behavior of the passband equalizer under typical operating conditions. Indeed, if the equalizer tap coefficients all start at zero, their evolution in the deterministic gradient algorithm is completely unaffected by the phase-tracking loop.

The second situation analyzed is the steady-state response of the system to a constant carrier frequency offset. In this case, the phase-tracking loop is found to reduce the resulting rate of rotation of the equalizer taps to about $\beta/(\alpha + \beta)$ of the original frequency offset. As a result, the degradation in system mean squared error due to frequency offset is typically quite small.

The final analysis is of the response of a linearized version of the receiver structure to sinusoidal phase jitter. When the channel's linear

distortion is not too severe and the coefficient β is small, the system mean squared error owing to phase tracking error is found to approximate that of a simple, first-order, phase-locked loop.

I. INTRODUCTION

The combination of adaptive equalization and decision-directed estimation of a fixed carrier phase offset in suppressed-carrier PAM modems by means of a gradient algorithm has been suggested by Chang[1] and by Kobayashi,[2] the latter also including adaptive timing recovery. The receivers contemplated in those papers demodulated the received data signal prior to equalization and carrier phase estimation.

Reference 3 describes an alternative receiver configuration for two-dimensional modulated data transmission systems, combining equalization and carrier recovery. This receiver's distinction is that it employs a passband equalizer[4] whose reference signal consists of receiver decisions amplitude-modulating a carrier whose phase shift is the receiver's estimate of the channel phase shift. Following the passband equalizer is a demodulator which compensates for the channel's phase shift (which may be time-varying as a result of frequency offset or phase jitter).

The receiver's estimation of the carrier phase shift is based on a decision-directed gradient algorithm for estimating a fixed phase shift, as proposed in Refs. 1, 2, 5, and 6. An advantage of the demodulator following the equalizer is that the demodulator's phase reference is delayed relative to the actual channel phase shift by only one symbol interval instead of by the entire equalizer delay as in the traditional "baseband" receiver configuration. This fact, plus the provision of a sufficiently large gain coefficient in the phase-tracking gradient algorithm, makes possible tracking and compensation of typical conditions of frequency offset and phase jitter that may occur on voiceband telephone channels. Computer simulations, reported in Refs. 3 and 7, have confirmed this capability.

In this paper, we study the dynamic behavior of the gradient algorithm for jointly adjusting the equalizer tap coefficients and the phase estimate in each of the following situations: (*i*) start-up (transient response) for a fixed carrier phase shift; (*ii*) steady-state response to a frequency offset; (*iii*) steady-state response to sinusoidal phase jitter. Throughout, we consider only the deterministic gradient algorithm; that is, receiver decisions are assumed perfect, and the gradient of the mean squared error as a function of equalizer tap coefficients and carrier reference is assumed known. A stochastic gradient algorithm, which would be used in practice, has been simulated,[3,7] but is not treated in this paper.

II. SYSTEM EQUATIONS

The transmitted two-dimensional modulated data signal is assumed to be of the form

$$s(t) = \text{Re}\,\{\sum_n A_n g(t - nT)e^{j2\pi f_c t}\},$$

where A_n is a two-dimensional (complex-valued) data symbol transmitted in the nth symbol interval, $g(t)$ is a band-limited baseband pulse waveform, T is the duration of a symbol interval, and f_c is the carrier frequency. The set of possible discrete complex values that each A_n can assume constitutes the signal *constellation*. Quadrature amplitude modulation (QAM) and digital phase modulation (PM) systems are familiar examples of two-dimensional modulation systems. We shall assume that successive data symbols are uncorrelated; i.e.,

$$\begin{aligned} \langle A_n A_m^* \rangle &= 1 \qquad \text{for } n = m \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Figure 1 shows the receiver structure. The received signal, after transmission through a noisy, dispersive channel which may introduce a slowly time-varying phase shift, is passed through a phase splitter to produce parallel in-phase and quadrature components. These parallel waveforms can be represented as a single complex waveform that is sampled and passed on to a passband transversal equalizer with, say, $2N + 1$ complex-valued tap coefficients. In the nth symbol interval, when a decision is to be made on the nth data symbol, the latest $(2N + 1)$ complex-valued samples stored in the $(2N + 1)$-tap passband equalizer can be represented by the complex $(2N + 1)$-dimensional vector $\mathbf{R}_n e^{j\theta_n}$, where θ_n is the channel phase shift (assumed quasi-stationary in the nth symbol interval). A sequence $\{\theta_n\}$ changing at a constant rate with time is an example of frequency offset, while $\{\theta_n\}$ varying randomly or quasi-periodically constitutes phase jitter. Typically, the change in θ_n in one or two symbol intervals is so small as to allow us to neglect the phase-to-amplitude modulation conversion effected by filtering the sequence of incidental frequency-modulated components $\{e^{j\theta_n}\}$.

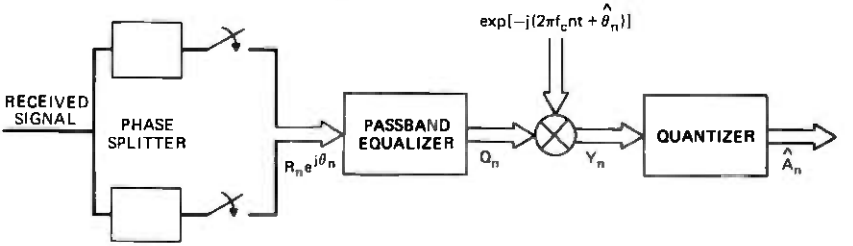


Fig. 1—Two-dimensional receiver.

The $(2N + 1)$ complex equalizer tap coefficients in the nth symbol interval are denoted by the complex $(2N + 1)$-dimensional vector $\mathbf{C}_n = (c_{-N}, \cdots, c_0, \cdots, c_N)$.* The symbol * will denote transposed complex conjugate throughout. Then the nth complex equalizer output is

$$Q_n = \mathbf{C}_n^* \mathbf{R}_n e^{j\theta_n}, \tag{1}$$

the real part being interpreted as the in-phase component and the imaginary part as the quadrature component.

The receiver's estimate of θ_n is a real quantity denoted by $\hat{\theta}_n$, and the demodulator output is written

$$Y_n = Q_n e^{-j(2\pi f_c nT + \hat{\theta}_n)}. \tag{2}$$

This quantity is passed into a simple quantizer to produce $\hat{A}_n$, which is the receiver's decision on A_n. Based on this decision, the complex reference signal used for updating the equalizer taps and the phase estimate is

$$\hat{Q}_n = \hat{A}_n e^{j(2\pi f_c nT + \hat{\theta}_n)}. \tag{3}$$

We define the properties of the channel in terms of expectations (denoted by $\langle\ \rangle$) with respect to the ensembles of information symbol sequences and additive noise samples. The complex impulse response $\mathbf{X}$ is defined by

$$\mathbf{X} = \frac{1}{\langle |A_n|^2 \rangle} \langle A_n^* \mathbf{R}_n \rangle e^{-j2\pi f_c nT}. \tag{4}$$

The positive definite Hermitian $\mathcal{A}$ matrix of the channel is defined by

$$\mathcal{A} = \frac{1}{\langle |A_n|^2 \rangle} \langle \mathbf{R}_n \mathbf{R}_n^* \rangle. \tag{5}$$

The normalized mean squared error in the nth symbol interval is defined to be

$$\epsilon_n = \frac{1}{\langle |A_n|^2 \rangle} \langle |Q_n e^{-j(2\pi f_c nT + \hat{\theta}_n)} - A_n|^2 \rangle, \tag{6a}$$

which, by virtue of (1), (4), and (5), can be rewritten as

$$\epsilon_n = 1 - \mathbf{X}^* \mathcal{A}^{-1} \mathbf{X} + \gamma_n, \tag{6b}$$

where $\gamma_n \equiv \mathbf{E}_n^* \mathcal{A} \mathbf{E}_n \geq 0$ is the excess mean squared error and $\mathbf{E}_n$ is a tap-error vector,

$$\mathbf{E}_n \equiv \mathbf{C}_n e^{j(\hat{\theta}_n - \theta_n)} - \mathcal{A}^{-1} \mathbf{X}. \tag{7}$$

Since $\mathcal{A}$ is positive definite, the value of ϵ_n is a positive minimum,† $1 - \mathbf{X}^* \mathcal{A}^{-1} \mathbf{X}$, when the equalizer taps $\mathbf{C}_n$ and phase shift estimate $\hat{\theta}_n$

† The positive quantity $\mathbf{X}^* \mathcal{A}^{-1} \mathbf{X}$ is therefore less than unity, a fact which is exploited in the appendix.

are adjusted so that $\mathbf{E}_n = 0$, or

$$\mathbf{C}_n e^{j\hat{\theta}_n} = \mathfrak{A}^{-1}\mathbf{X}e^{j\theta_n}. \tag{8}$$

This equation is also the condition for the gradients of ϵ_n with respect to $\mathbf{C}_n$ and $\hat{\theta}_n$ to be jointly zero; it is satisfied by an infinitude of points $(\mathbf{C}_n, \hat{\theta}_n)$.

Thus, a gradient algorithm can be used to adjust the tap coefficients $\mathbf{C}_n$ and phase estimate $\hat{\theta}_n$ recursively toward optimal values. The equations governing the evolution of $\{\mathbf{C}_n\}$ and $\{\hat{\theta}_n\}$ are[3]

$$\mathbf{C}_{n+1} = (I - \beta\mathfrak{A})\mathbf{C}_n + \beta\mathbf{X}e^{-j(\hat{\theta}_n - \theta_n)} \tag{9}$$

and

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \alpha \operatorname{Im}\left[\mathbf{C}_n^*\mathbf{X}e^{-j(\hat{\theta}_n - \theta_n)}\right], \tag{10}$$

where I is the identity matrix and β and α are positive gain coefficients. These equations [or the equivalent equations (13) and (14)] form the basis for the results in this paper.

In practice, $\mathbf{X}$ and $\mathfrak{A}$ would generally not be known in advance, and the following *stochastic gradient* algorithm,[3] involving the equalizer inputs $\mathbf{R}_n e^{j\theta_n}$, outputs Q_n, and modulated decisions $\hat{Q}_n$, would replace the deterministic gradient algorithm described by eqs. (9) and (10).

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \beta\mathbf{R}_n e^{j\theta_n}(Q_n^* - \hat{Q}_n^*). \tag{11}$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{\alpha}{|A_n|^2} \operatorname{Im}\,(Q_n\hat{Q}_n^*). \tag{12}$$

These are coupled stochastic difference equations, since successive vectors $\{\mathbf{R}_n\}$ are correlated random variables. Simple stochastic gradient algorithms have been studied by Widrow.[8] The application to equalizer adaptation, where no phase recovery is involved and under the assumption that the $\{\mathbf{R}_n\}$ are uncorrelated, has been studied by Ungerboeck,[9] by Gersho,[10] and by Gitlin, Mazo, and Taylor.[11] The extension to correlated vectors $\{\mathbf{R}_n\}$ has been introduced by Daniell.[12]

That the algorithm specified by (11) and (12) converges and can perform satisfactorily is confirmed by the computer simulations reported in Refs. 3 and 7. Analysis of the stochastic gradient algorithm is complicated by the possibility of a cycle-slipping phenomenon as in phase-lock loop systems. References 5 and 6 deal with continuous-time, decision-directed, phase-locked loops in the absence of adaptive equalization.

However, insight can be gained by studying instead the deterministic gradient algorithm of (9) and (10), since the estimated gradient algorithm can be interpreted as implicitly performing the averaging involved in determining $\mathbf{X}$ and $\mathfrak{A}$, provided the signal-to-noise ratio is high and the gain coefficients α and β are sufficiently small.

Using definition (7), we can rewrite the coupled difference equations as

$$\mathbf{E}_{n+1} = (I - \beta\mathcal{A})\mathbf{E}_n e^{j(\hat{\Delta}_{n+1}-\Delta_{n+1})} + \mathcal{A}^{-1}\mathbf{X}(e^{j(\hat{\Delta}_{n+1}-\Delta_{n+1})} - 1) \quad (13)$$

and

$$\hat{\Delta}_{n+1} = \alpha \operatorname{Im}(\mathbf{E}_n^*\mathbf{X}), \quad (14)$$

where

$$\Delta_{n+1} = \theta_{n+1} - \theta_n \qquad \text{and} \qquad \hat{\Delta}_{n+1} = \hat{\theta}_{n+1} - \hat{\theta}_n.$$

III. SYSTEM START-UP WITH FIXED CHANNEL PHASE SHIFT

In this section, we study the behavior of the deterministic gradient algorithm during start-up, assuming the channel's phase shift is fixed: $\theta_n = 0$.† General theorems tell us that, if the initial error and the coefficient of the gradient algorithm are small enough, convergence is guaranteed.[13] However, we are interested in sharper results for the specific problem at hand.

The solution of (13) and (14) will depend on the initial choice of $\mathbf{E}_0$ (or $\mathbf{C}_0$) and $\hat{\theta}_0$. It is interesting to consider first the special case $\mathbf{C}_0 = \mathbf{0}$, the all-zero vector; i.e., $\mathbf{E}_0 = -\mathcal{A}^{-1}X$. In this case,

$$\hat{\Delta}_1 = -\alpha \operatorname{Im}[\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X}] = 0,$$

since $\mathcal{A}$ is Hermitian, and

$$\mathbf{E}_1 = -(I - \beta\mathcal{A})\mathcal{A}^{-1}\mathbf{X}.$$

Continuing, it is easy to show that

$$\Delta_n = 0 \qquad \text{for all } n$$

and that

$$\mathbf{E}_n = -(I - \beta\mathcal{A})^n\mathcal{A}^{-1}\mathbf{X}. \quad (15)$$

Thus, at least for this special all-zero starting condition, the estimated carrier phase shift $\hat{\theta}_n$ does not change at all and the start-up behavior of the deterministic algorithm is exactly the same as that of the passband equalizer alone.[4]

Let us now consider the more general case, when $\mathbf{E}_0$ is not necessarily equal to the right-hand side of (15) for some $n \geq 0$. We remark that the mathematical formulation of this start-up situation will be basically the same as that of a system transient caused by an abrupt change in the channel's carrier phase shift.

Expression (6b) for the normalized mean squared error involves the positive definite quadratic form $\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n \equiv \gamma_n$. We can bound this term

† There is no loss of generality in assuming a fixed phase shift of zero, since any nonzero fixed phase-shift factor $e^{j\theta}$ can be incorporated in the complex channel impulse response $\mathbf{X}$.

and study its evolution by writing down a recursive expression for it and upper-bounding the right-hand side of that expression. Using (13) with $\Delta_{n+1} = 0$ for all n, we can write

$$\gamma_{n+1} = \mathbf{E}_n^*(I - \beta\mathfrak{A})\mathfrak{A}(I - \beta\mathfrak{A})\mathbf{E}_n + \mathbf{X}^*\mathfrak{A}^{-1}\mathbf{X}|e^{j\hat{\Delta}_{n+1}} - 1|^2 + 2\,\mathrm{Re}\,\{\mathbf{E}_n^*(I - \beta\mathfrak{A})\mathbf{X}(1 - e^{-j\hat{\Delta}_{n+1}})\}. \quad (16)$$

The right-hand side of expression (16) is upper-bounded in the appendix. The derivation of the bound requires the following assumptions about the channel and algorithm parameters.

Assumption (1): The initial value $\gamma_0 \equiv \mathbf{E}_0^*\mathfrak{A}\mathbf{E}_0$ is less than unity. This condition is fulfilled, for example, if $\mathbf{C}_0 = \mathbf{0}$; i.e., $\mathbf{E}_0 = -\mathfrak{A}^{-1}\mathbf{X}$, for then $\gamma_0 = \mathbf{X}^*\mathfrak{A}^{-1}\mathbf{X} \leqq 1$, since the positive quadratic form $\mathbf{X}^*\mathfrak{A}^{-1}\mathbf{X}$, which is one minus the minimum mean squared error, must be less than unity.

Assumption (2): $\alpha < \alpha_0$, where α_0 is the solution of

$$\alpha_0(1 + \sqrt{\gamma_0}) = 2\,\mathrm{sinc}\,(\alpha_0\sqrt{\gamma_0}),$$

where

$$\mathrm{sinc}\,\theta = \frac{\sin\theta}{\theta}.$$

Assumption (3): Let the maximum and minimum eigenvalues of the positive definite Hermitian matrix $\mathfrak{A}^*$ be denoted respectively by $\lambda_{\max}$ and $\lambda_{\min}$. Then the gain coefficient β must satisfy

$$0 < \beta < \frac{2\lambda_{\min}}{\lambda_{\max}^2(1 + \epsilon_0^2)},$$

where ϵ_0^2 is defined in terms of α by

$$\alpha\left(1 + \sqrt{\gamma_0} + \frac{1}{\epsilon_0^2}\right) = 2\,\mathrm{sinc}\,(\alpha\sqrt{\gamma_0}), \qquad \alpha < \alpha_0.$$

Figure 2 illustrates the solution of the equations defining ϵ_0^2 and α_0. For example, if we assume $\alpha = 0.5$ and $\gamma_0 = 1$, then α_0 is 0.88 and ϵ_0^2 is 0.543.

The upper bound obtained in the appendix is

$$\gamma_{n+1} \equiv \mathbf{E}_{n+1}^*\mathfrak{A}\mathbf{E}_{n+1} \leqq \mathbf{E}_n^*\mathfrak{A}\mathbf{E}_n - 2\beta\mathbf{E}_n^*\mathfrak{A}^2\mathbf{E}_n + \beta^2(1 + \epsilon_0^2)\mathbf{E}_n^*\mathfrak{A}^3\mathbf{E}_n. \quad (17a)$$

An explicit bound on γ_{n+1} is obtained by first weakening (17a) using (41) of the appendix to obtain

$$\gamma_{n+1} \leqq (1 - 2\beta\lambda_{\min} + \beta^2(1 + \epsilon_0^2)\lambda_{\max}^2)\gamma_n, \quad (17b)$$

so that

$$\gamma_{n+1} \leqq (1 - 2\beta\lambda_{\min} + \beta^2(1 + \epsilon_0^2)\lambda_{\max}^2)^n\gamma_0. \quad (17c)$$

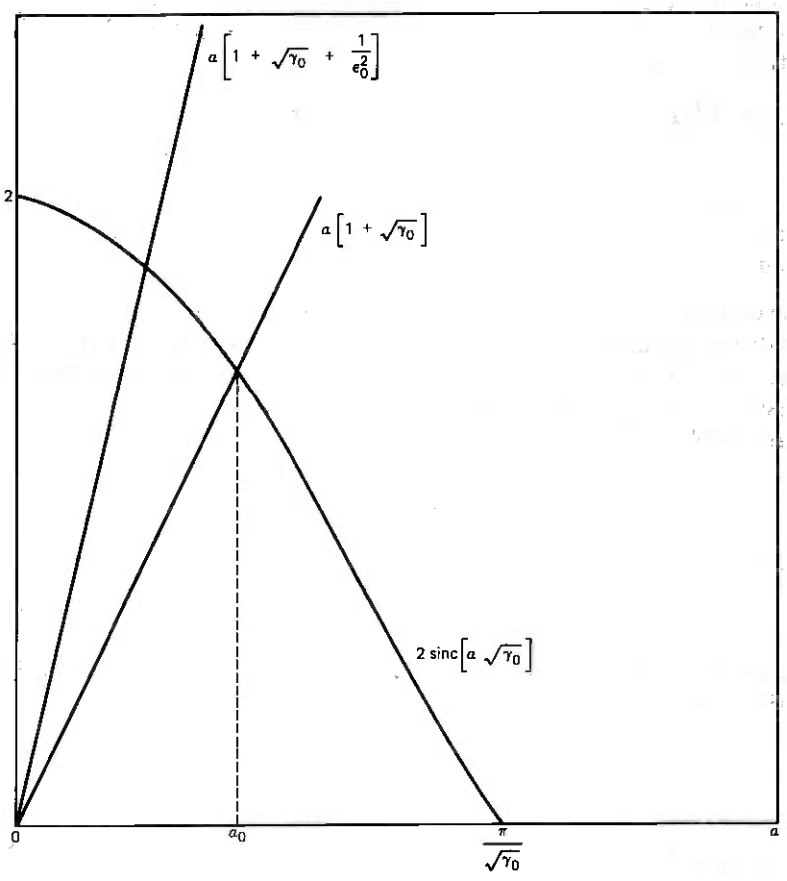


Fig. 2—Illustration of the definitions of ϵ_0^2 and α_0.

In the absence of phase tracking, $\alpha = \hat{\Delta}_n = 0$, and the mean squared error at step $n + 1$ of the deterministic gradient algorithm is obtained directly from expression (15)[9,10,14] as

$$\gamma_{n+1} = \sum_i \lambda_i (1 - \beta\lambda_i)^{2n} |\mathcal{E}_{0i}|^2, \tag{18}$$

where the summation is over all the eigenvalues of the matrix $\mathfrak{A}$, the $\{\lambda_i\}$ are the set of eigenvalues, and $\mathcal{E}_{0i}$ is the inner product of $\mathbf{E}_0$ with the normalized ith eigenvector.

Comparison of the upper bound (17c) for the joint equalizing and phase-tracking receiver and the exact expression (18) for the equalizer alone yields some insight into the penalty in convergence rate imposed by the additional phase-tracking algorithm. Consider an example where

all $\{\lambda_i\}$ (and therefore λ_{max} and λ_{min}) are equal to a common value λ. This would represent the case of a channel with delay distortion but not amplitude distortion (flat Nyquist equivalent frequency characteristic). Then inequality (17c) becomes

$$\gamma_{n+1} \leqq [(1 - \beta\lambda)^2 + \beta^2\lambda^2\epsilon_0^2]^n\gamma_0, \tag{19}$$

and, recognizing that

$$\gamma_0 \equiv \sum_i \lambda_i |\mathcal{E}_{0i}|^2,$$

we can write equality (18) for the case of no-phase tracking as

$$\gamma_{n+1} = (1 - \beta\lambda)^{2n}\gamma_0. \tag{20}$$

In practice, the equalizer adaptation coefficient β is small ($\beta \ll 1/\lambda$), to minimize the mean squared error resulting from a practical stochastic gradient algorithm.[9] Thus the right-hand sides of (19) and (20) should be nearly equal, and we conclude that an ideal gradient algorithm for joint phase tracking and equalization should not converge appreciably slower than the equalizer adjustment algorithm alone. An exact analytical evaluation of the effect of phase tracking on the convergence of a practical stochastic gradient algorithm for a severely distorted ($\lambda_{max} \gg \lambda_{min}$) channel remains elusive. However, the results of this section suggest that the influence of the phase-tracking parameter α in the convergence is relatively small. This conjecture is bolstered by the experimental results summarized in Figs. 3a and 3b. A 9600-b/s two-dimensional data transmission system was simulated, employing the stochastic gradient algorithm described by eqs. (11) and (12). The transmission channel, whose frequency characteristics are shown in Fig. 3a, was regarded as severely distorted (it violates the minimum standard for private line voiceband channel data transmission). The plots of measured mean squared error versus time for $\alpha = 0$ and for $\alpha = 0.2$ shown in Fig. 3b are very similar, indicating that little penalty in convergence rate is to be ascribed to the use of joint decision-directed phase tracking.

IV. CASE OF FREQUENCY OFFSET

In this section, we study the behavior of the system in the presence of frequency offset by obtaining steady-state solutions to eqs. (13) and (14) when the channel phase shift increases linearly with time; i.e., $\Delta_n = 2\pi\Delta T$, where Δ is the frequency offset. In this case, eq. (13) becomes

$$\mathbf{E}_{n+1} = (I - \beta\mathcal{A})\mathbf{E}_n e^{j(\hat{\Delta}_{n+1} - 2\pi\Delta T)} + \mathcal{A}^{-1}\mathbf{X}(e^{j(\hat{\Delta}_{n+1} - 2\pi\Delta T)} - 1). \tag{21}$$

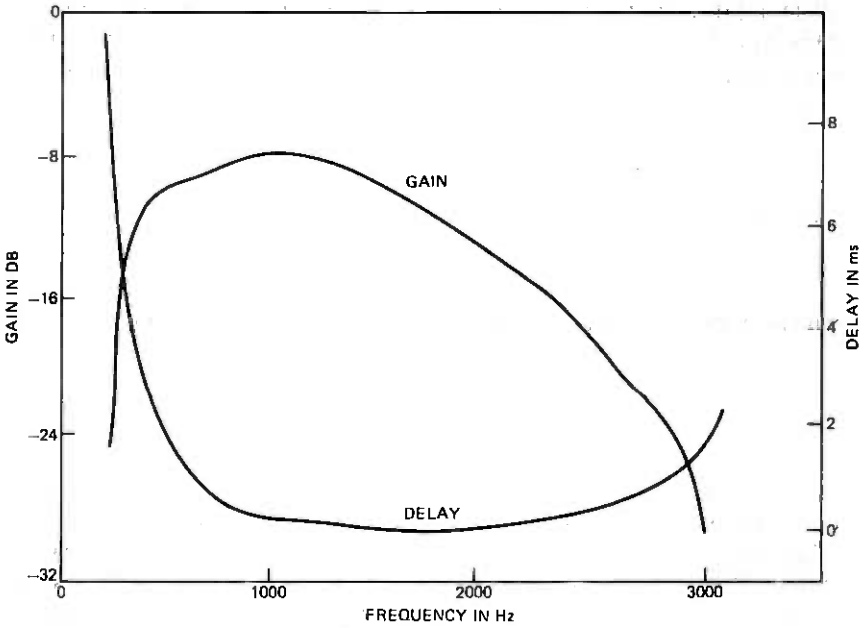


Fig. 3a—Frequency characteristics of the simulated channel.

A steady-state solution to (21) and (14) is obtained by substituting the trial solution,

$$\mathbf{E}_n = \mathbf{E}$$
$$\hat{\Delta}_n = 2\pi(\Delta + \delta)T,$$

and then solving for the fixed quantities $\mathbf{E}$ and δ. The substitution results in

$$\mathbf{E} = (e^{j2\pi\delta T} - 1)M^{-1}\mathfrak{A}^{-1}\mathbf{X}, \tag{22}$$

where M is the matrix

$$M = I - e^{j2\pi\delta T}(I - \beta\mathfrak{A})$$

and

$$\begin{aligned} 2\pi(\Delta + \delta)T &= \alpha \operatorname{Im}(\mathbf{E}^*\mathbf{X}). \\ &= \alpha \operatorname{Im}[(e^{-j2\pi\delta T} - 1)\mathbf{X}^*\mathfrak{A}^{-1}M^{*-1}\mathbf{X}]. \end{aligned} \tag{23}$$

It is clear from the definition of M that the eigenvectors $\{\mathfrak{y}_i\}_{-N}^{N}$ of $\mathfrak{A}$, which form a complete orthonormal set, are also those of M. Thus, expressing the vector $\mathbf{X}$ as a linear combination of $\mathfrak{y}_i$, we write

$$\mathbf{X} = \sum_{i=-N}^{N} G_i\mathfrak{y}_i,$$

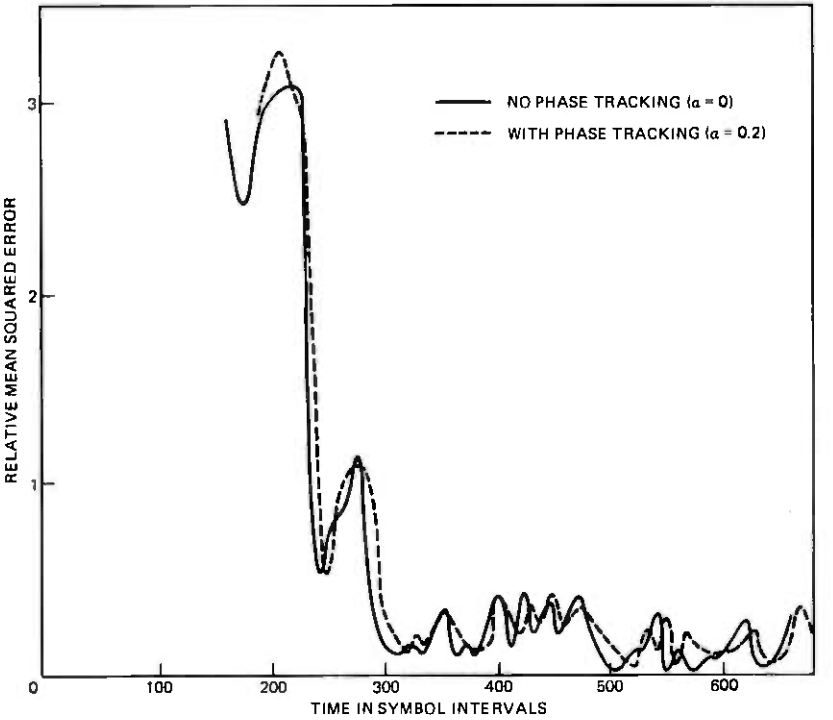


Fig. 3b—Convergence with and without phase tracking (ideal reference; all equalizer top coefficients start at zero).

we can rewrite (23) after a little algebra as

$$2\pi(\Delta + \delta)T = \alpha \operatorname{Im}\left[(e^{-j2\pi\delta T} - 1)\right] \sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i[1 - e^{-j2\pi\delta T}(1 - \beta\lambda_i)]},$$

$$= -\alpha\beta \sin 2\pi\delta T \sum_{i=-N}^{N} \frac{|G_i|^2}{1 - 2(1 - \beta\lambda_i)\cos 2\pi\delta T + (1 - \beta\lambda_i)^2}, \quad (24)$$

where $\{\lambda_i\}_{i=-N}^{N}$ are the eigenvalues of $\mathfrak{A}$ and are positive and real.

The excess mean squared error is similarly given by

$$\gamma_n = \mathbf{E}_n^* \mathfrak{A} \mathbf{E}_n = |e^{j2\pi\delta T} - 1|^2 \mathbf{X}^* \mathfrak{A}^{-1} M^* \mathfrak{A} M^{-1} \mathfrak{A}^{-1} \mathbf{X}$$

$$= 2(1 - \cos 2\pi\delta T) \sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i[1 - 2(1 - \beta\lambda_i)\cos 2\pi\delta T + (1 - \beta\lambda_i)^2]}. \quad (25)$$

Equation (24) is a transcendental equation whose solution δ is clearly not zero in general. The quantity δ may be interpreted as a bias in the receiver's estimate of the frequency offset. This "residual"

frequency offset then must be compensated for by a rotation of the equalizer complex tap coefficients at rate δ Hz.

For purposes of illustration, we again consider only a special case of a "good" channel, for which all $\lambda_i = 1$ and $\sum_i |G_i|^2 = 1$. Then (24) becomes

$$2\pi(\Delta + \delta)T = \frac{-\alpha\beta \sin 2\pi\delta T}{\beta^2 + 2(1 - \beta)(1 - \cos 2\pi\delta T)}. \tag{26}$$

Typically, $\beta \ll \alpha < 1$; for example, $\beta = 0.001$ and $\alpha = 0.2$. The left- and right-hand sides of (26) as functions of $2\pi\delta T$ are sketched in Fig. 4. Apparently in the region of intersection, $2\pi\delta T \ll \beta$ and $\sin 2\pi\delta T \approx 2\pi\delta T$. Solving (26) with this approximation yields

$$2\pi(\Delta + \delta)T \approx \frac{-2\pi\alpha}{\beta}\,\delta T.$$

Thus

$$\delta \approx \frac{-\beta\Delta}{\alpha + \beta}, \tag{27}$$

and the necessary rate of rotation of the equalizer taps has been reduced by a factor of $\beta/(\alpha + \beta)$, which is about 1/200 for a typical case, $\alpha = 0.2, \beta = 0.001$. The corresponding normalized excess mean squared error is

$$\mathbf{E}_n^* \mathfrak{A} \mathbf{E}_n \approx \frac{(2\pi\delta T)^2}{\beta^2 + (2\pi\delta T)^2} \approx \frac{(2\pi\Delta T)^2}{(\alpha + \beta)^2 + (2\pi\Delta T)^2}. \tag{28}$$

If $\Delta = 1$ Hz, $\alpha = 0.2$, $\beta = 0.001$, $T = 1/2400$ s. This amounts to about 10^{-4}.

V. STEADY-STATE SINUSOIDAL RESPONSE

The phase jitter process $\{\theta_n\}$ that occurs in telephone channels is typically quasi-periodic. It is thus of interest to determine the steady-state solution of the coupled difference equations (13) and (14) when the driving term $\{\theta_n\}$ is sinusoidal.

It is convenient at this point to rewrite eqs. (13) and (14) further in terms of eigenvalues and eigenvectors of the matrix $\mathfrak{A}$. Since $\mathfrak{A}$ is Hermitian, its eigenvalues $\{\lambda_i\}_{i=-N}^{N}$ are positive real and its eigenvectors $\{\boldsymbol{\mu}_i\}_{i=-N}^{N}$ form an orthonormal set which is a basis in $2N + 1$-dimensional space. Using these properties and expressing the vectors $\mathbf{E}_n$ and $\mathbf{X}$ as linear combinations of the $\{\boldsymbol{\mu}_i\}$,

$$\mathbf{E}_n = \sum_{i=-N}^{N} \mathcal{E}_{n,i}\boldsymbol{\mu}_i$$

$$\mathbf{X} = \sum_{i=-N}^{N} G_i\boldsymbol{\mu}_i,$$

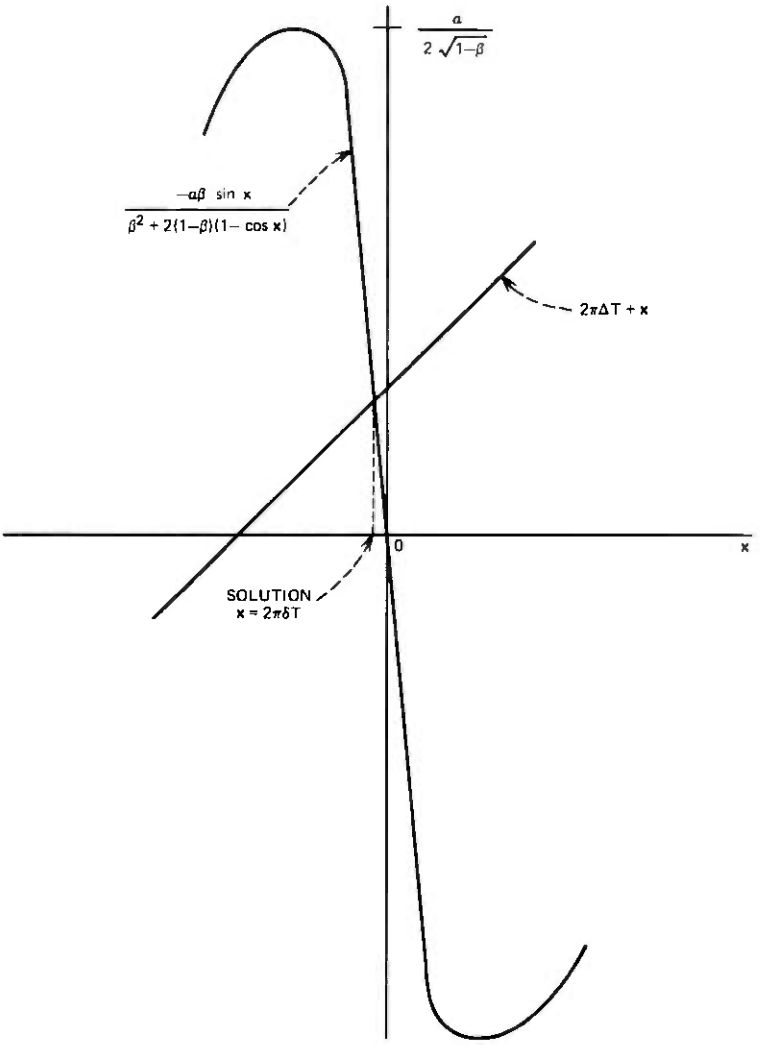


Fig. 4—Illustration of the solution of

$$2\pi(\Delta + \delta)T = \frac{-\alpha\beta \sin 2\pi\delta T}{\beta^2 + 2(1-\beta)(1-\cos 2\pi\delta T)}.$$

we can write (13) and (14) as

$$\mathcal{E}_{(n+1)i} = (1 - \beta\lambda_i)\,\mathcal{E}_{ni} e^{j(\hat{\Delta}_{n+1} - \Delta_{n+1})} + \frac{G_i}{\lambda_i}\,(e^{j(\hat{\Delta}_{n+1} - \Delta_{n+1})} - 1) \qquad -N \leqq i \leqq N \quad (29)$$

and

$$\hat{\Delta}_{n+1} = \alpha \sum_{i=-N}^{N} \operatorname{Im}\,(\mathcal{E}_{ni}^{*} G_i). \quad (30)$$

We now make the following change of variable in (29) and (30). Define

$$\mathcal{E}_{ni}^{*} G_i e^{j(\hat{\theta}_n - \theta_n)} = u_{ni} + jv_{ni}. \tag{31}$$

Then we can write the real and imaginary parts of (29) as

$$u_{(n+1)i} = (1 - \beta\lambda_i)u_{ni} + \frac{|G_i|^2}{\lambda_i}[\cos(\hat{\theta}_n - \theta_n) - \cos(\hat{\theta}_{n+1} - \theta_{n+1})], \quad -N \leqq i \leqq N \tag{32}$$

and

$$v_{(n+1)i} = (1 - \beta\lambda_i)v_{ni} + \frac{|G_i|^2}{\lambda_i}[\sin(\hat{\theta}_n - \theta_n) - \sin(\hat{\theta}_{n+1} - \theta_{n+1})] \quad -N \leqq i \leqq N, \tag{33}$$

and we can write (30) in the form

$$\hat{\theta}_{n+1} - \hat{\theta}_n = \alpha \sum_{i=-N}^{N} [v_{ni}\cos(\hat{\theta}_n - \theta_n) - u_{ni}\sin(\hat{\theta}_n - \theta_n)]. \tag{34}$$

Equations (32), (33), and (34) are a set of nonlinear coupled difference equations. In particular, eq. (34) is reminiscent of the equation governing a discrete-time, first-order, phase-locked loop. We shall solve linearized versions of (32), (33), and (34). Assuming the steady-state error angle $(\hat{\theta}_n - \theta_n)$ for $n \gg 1$ is very small, we replace $\cos(\hat{\theta}_n - \theta_n)$ by 1 and $\sin(\hat{\theta}_n - \theta_n)$ by $(\hat{\theta}_n - \theta_n)$. Then (32) becomes

$$\begin{aligned} u_{(n+1)i} &= (1 - \beta\lambda_i)u_{ni}, \\ &= (1 - \beta\lambda_i)^{n+1}u_{0i}, \qquad -N \leqq i \leqq N, \end{aligned}$$

which approaches zero in the steady state (assuming $\beta < 1/\lambda_i$ for all i). Thus in the steady state we are left with the linearized versions of (33) and (34):

$$v_{(n+1)i} = (1 - \beta\lambda_i)v_{ni} + \frac{|G_i|^2}{\lambda_i}(\theta_{n+1} - \theta_n - \hat{\theta}_{n+1} + \hat{\theta}_n) \quad -N \leqq i \leqq N \tag{35}$$

and

$$\hat{\theta}_{n+1} - \hat{\theta}_n = \alpha \sum_{i=-N}^{N} v_{ni}. \tag{36}$$

Equations (35) and (36) are linear and can be solved for a given sequence of channel phase shifts $\{\theta_n\}$. We consider the case where the phase jitter is sinusoidal with frequency ω rad/s; i.e.,

$$\theta_n = \text{Re}\,(Je^{j\omega nT}),$$

where J is a complex constant. The solution for $\{v_{ni}\}$ is also sinusoidal:

$$v_{ni} = \text{Re}\,(V_i e^{j\omega nT}). \qquad -N \leqq i \leqq N. \tag{37}$$

Substitution of this trial solution in (35) and (36) yields a value of V_i after some algebraic manipulations.

$$V_i = \frac{J(1 - e^{j\omega T})|G_i|^2}{\lambda_i(1 - \beta\lambda_i - e^{j\omega T})\left(1 - \alpha \sum_{k=-N}^{N} |G_k|^2/[(1 - \beta\lambda_k - e^{j\omega T})\lambda_k]\right)}. \tag{38}$$

It follows from the sinusoidal variation of $\{v_{ni}\}_{i=-N}^{N}$ that the error angle $\{\hat{\theta}_n - \theta_n\}$ and the equalizer tap coefficient vector $\mathbf{C}_n$ also vary sinusoidally with frequency ω in the steady state.

The excess time-averaged mean squared error can be calculated from expression (31), (37), and (38).

$$\begin{aligned}\gamma = \langle \gamma_n \rangle &= \langle \mathbf{E}_n^* \mathfrak{A} \mathbf{E}_n \rangle \\ &= \sum_{i=-N}^{N} \lambda_i \langle |\mathcal{E}_{ni}|^2 \rangle \\ &= \frac{|J|^2|1 - e^{j\omega T}|^2 S_1}{2|1 - e^{j\omega T} - \alpha S_2|^2},\end{aligned} \tag{39}$$

where

$$S_1 = \sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i|1 - \beta\lambda_i/(1 - e^{j\omega T})|^2}$$

and

$$S_2 = \sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i[1 - \beta\lambda_i/(1 - e^{j\omega T})]}.$$

The total mean squared error is, from (6b),

$$\begin{aligned}\langle \epsilon_n \rangle &= 1 - \mathbf{X}^* \mathfrak{A}^{-1} \mathbf{X} + \gamma \\ &= 1 \sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i} + \frac{|J|^2|1 - e^{j\omega T}|^2 S_1}{2|1 - e^{j\omega T} - \alpha S_2|^2}.\end{aligned} \tag{40}$$

Typically, if the overall mean squared error is close to zero,

$$\sum_{i=-N}^{N} \frac{|G_i|^2}{\lambda_i} \approx 1 \qquad \text{and} \qquad \beta\lambda_i \ll |1 - e^{j\omega T}|.$$

Then the excess mean squared error in (40) is approximately

$$\frac{|J|^2|1 - e^{j\omega T}|^2}{2|1 - e^{j\omega T} - \alpha|^2}.$$

This expression corresponds to a previously derived, approximate, mean squared error due to sinusoidal jitter in the absence of noise [see eq. (39) of Ref. 3]. That equation, valid for a first-order, phase-locked loop,

was derived ignoring the coupling between eqs. (13) and (14) and assuming perfect equalization. Calculated curves of mean squared error versus α are found in Ref. 3.

VI. CONCLUSIONS

Previous studies have shown that the functions of joint passband equalization and data-directed carrier recovery in a QAM receiver can be formulated as a gradient search algorithm. If the channel parameters entering into the expression for the gradient of the mean squared error are known, it is termed a deterministic gradient algorithm. In this paper we have analyzed the start-up behavior of the deterministic gradient algorithm and also the steady-state response to frequency offset and to sinusoidal phase jitter. The more practically motivated stochastic or estimated gradient algorithm, in which the channel parameters are initially unknown, has been studied experimentally and awaits further analytical study.

It was shown that, under typical channel conditions, when the carrier phase offset is fixed, phase tracking does not greatly slow down the start-up behavior of the deterministic gradient algorithm, at least provided the equalizer adaptation coefficient β is much less than that of the phase estimator α.

The phase estimator was first proposed as an adjunct to the passband equalizer, to mitigate the effects of too-rapid tap-coefficient rotation in the presence of channel frequency offset. It has been shown that frequency offset still causes tap rotation in the equalizer-plus-phase estimator system, but that the rate of rotation is tolerable, being on the order of $1/[1 + (\alpha/\beta)]$ times the amount of frequency offset.

The steady-state response of the linearized system to sinusoidal phase jitter was obtained. When linear distortion in the channel is not severe and the coefficient β is small, the system mean squared error due to tracking error approximates that of a first-order, phase-locked loop, as was assumed in an earlier paper.

APPENDIX

We wish to upper-bound the right-hand side of (16), given assumptions (1), (2), and (3) of Section III.

$$\mathbf{E}_{n+1}^{*}\mathcal{A}\mathbf{E}_{n+1} = \mathbf{E}_{n}^{*}(I - \beta\mathcal{A})\mathcal{A}(I - \beta\mathcal{A})\mathbf{E}_{n} + \mathbf{X}^{*}\mathcal{A}^{-1}\mathbf{X}\left|e^{j\tilde{\Delta}_{n+1}} - 1\right|^{2} + 2\,\mathrm{Re}\,\{\mathbf{E}_{n}^{*}(I - \beta\mathcal{A})\mathbf{X}(1 - e^{-j\hat{\Delta}_{n+1}})\}, \quad (16)$$

where $\hat{\Delta}_{n+1}$ was given by (14).

The first term on the right-hand side can be written

$$\mathbf{E}_{n}^{*}\mathcal{A}\mathbf{E}_{n} - 2\beta\mathbf{E}_{n}^{*}\mathcal{A}^{2}\mathbf{E}_{n} + \beta^{2}\mathbf{E}_{n}^{*}\mathcal{A}^{3}\mathbf{E}_{n}.$$

The matrix $\mathcal{A}$ is positive definite and Hermitian; hence,

$$-\mathbf{E}_n^*\mathcal{A}^2\mathbf{E}_n \equiv -(\mathcal{A}^{\frac{1}{2}}\mathbf{E}_n)^*\mathcal{A}(\mathcal{A}^{\frac{1}{2}}\mathbf{E}_n) \leqq -\lambda_{\min}\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n,$$

where $\lambda_{\min}$ is the minimum eigenvalue of $\mathcal{A}$. Similarly, $\mathbf{E}_n^*\mathcal{A}^3\mathbf{E}_n \leqq \lambda_{\max}^2\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n$, where $\lambda_{\max}$ is the maximum eigenvalue. Thus we note for future reference that the first term in (16) is bounded as

$$E_n^*(I - \beta\mathcal{A})\mathcal{A}(I - \beta\mathcal{A})\mathbf{E}_n \leqq (1 - 2\beta\lambda_{\min} + \beta^2\lambda_{\max})\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n. \quad (41)$$

The second term in (16) is

$$\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X}|e^{j\hat{\Delta}_{n+1}} - 1|^2 \leqq \sin^2\frac{\hat{\Delta}_{n+1}}{2},$$

since $\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X} \leqq 1$. Upper-bounding $\sin^2(\hat{\Delta}_{n+1}/2)$ by $(\hat{\Delta}_{n+1}/2)^2$ and substituting expression (14) for $\hat{\Delta}_{n+1}$, we have

$$\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X}|e^{j\hat{\Delta}_{n+1}} - 1|^2 \leqq \alpha^2[\text{Im }(\mathbf{E}_n^*\mathbf{X})]^2. \quad (42)$$

The third term in (16) can be written as the sum of three terms.

$$2\text{ Re }\{\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}(1 - e^{-j\hat{\Delta}_{n+1}})\} = 4\text{ Re }[\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}]\sin^2\frac{\hat{\Delta}_{n+1}}{2}$$
$$- 2\text{ Im }(\mathbf{E}_n^*X)\sin\hat{\Delta}_{n+1} + 2\beta\text{ Im }(\mathbf{E}_n^*\mathcal{A}\mathbf{X})\sin\hat{\Delta}_{n+1}. \quad (43)$$

As in the inequality (42), the first term in (43) is upper-bounded by

$$\alpha^2|\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}|[\text{Im }(\mathbf{E}_n^*\mathbf{X})]^2.$$

The matrix $I - \beta\mathcal{A}$ is Hermitian; its eigenvalues are $\{1 - \beta\lambda_i\}$, where the $\{\lambda_i\}$ are the eigenvalues of $\mathcal{A}$. Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximum and minimum eigenvalues, respectively. By assumption (3), $1 - \beta\lambda_{\max} > 0$ and thus $I - \beta\mathcal{A}$ is positive definite. Therefore,

$$|\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}| = |\mathbf{E}_n^*(I - \beta\mathcal{A})^{\frac{1}{2}}\mathcal{A}^{\frac{1}{2}}\mathcal{A}^{-\frac{1}{2}}(I - \beta\mathcal{A})^{\frac{1}{2}}\mathbf{X}|$$
$$\leqq [\mathbf{E}_n^*\mathcal{A}^{\frac{1}{2}}(I - \beta\mathcal{A})\mathcal{A}^{\frac{1}{2}}\mathbf{E}_n]^{\frac{1}{2}}[\mathbf{X}^*\mathcal{A}^{-\frac{1}{2}}(I - \beta\mathcal{A})\mathcal{A}^{\frac{1}{2}}\mathbf{X}]^{\frac{1}{2}}, \quad (44)$$

where we have used Schwartz's inequality. Using the positive definiteness of $I - \beta\mathcal{A}$ and $\mathcal{A}$, we can further upper-bound the right-hand side of (44) by

$$|\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}| \leqq (1 - \beta\lambda_{\min})^2(\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n)^{\frac{1}{2}}(\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X})^{\frac{1}{2}}$$
$$\leqq (\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n)^{\frac{1}{2}}, \quad (45)$$

since the quantities $1 - \beta\lambda_{\min}$ and $\mathbf{X}^*\mathcal{A}^{-1}\mathbf{X}$ are less than unity. Thus we have upper-bounded the first term in (43) by

$$4\text{ Re }\{\mathbf{E}_n^*(I - \beta\mathcal{A})\mathbf{X}\}\sin^2\frac{\hat{\Delta}_{n+1}}{2} \leqq \alpha^2(\mathbf{E}_n^*\mathcal{A}\mathbf{E}_n)^{\frac{1}{2}}[\text{Im }(\mathbf{E}_n^*\mathbf{X})]^2. \quad (46)$$

After substituting for $\hat{\Delta}_{n+1}$ using eq. (14), we can express the second term in (43) as

$$-2 \text{ Im } (\mathbf{E}_n^* \mathbf{X}) \sin \hat{\Delta}_{n+1} = -2\alpha[\text{Im } (\mathbf{E}_n^* \mathbf{X})]^2 \text{ sinc } [\alpha \text{ Im } (\mathbf{E}_n^* \mathbf{X})], \quad (47)$$

where

$$\text{sinc } \theta = \frac{\sin \theta}{\theta}.$$

The third term in (43) is

$$2\beta \text{ Im } (\mathbf{E}_n^* \mathcal{A} \mathbf{X}) \sin \hat{\Delta}_{n+1},$$

which can be upper-bounded, using (14) and the inequality $|\sin \hat{\Delta}| \leqq |\hat{\Delta}|$, by

$$2\alpha\beta(|\mathbf{E}_n^* \mathcal{A} \mathbf{X}|)[|\text{Im } (\mathbf{E}_n^* \mathbf{X})|] \leqq \epsilon^2 \beta^2 |\mathbf{E}_n^* \mathcal{A} \mathbf{X}|^2 + \frac{\alpha^2}{\epsilon^2} [\text{Im } (\mathbf{E}_n^* \mathbf{X})]^2 \quad \text{for any arbitrary } \epsilon,$$

where we have used the simple inequality

$$2\alpha\beta AB \leqq \epsilon^2 \beta^2 A^2 + \frac{\alpha^2}{\epsilon^2} B^2.$$

But

$$\begin{aligned} |\mathbf{E}_n^* \mathcal{A} \mathbf{X}|^2 &= |(\mathbf{E}_n^* \mathcal{A}^{\frac{1}{2}})(\mathcal{A}^{-\frac{1}{2}} \mathbf{X})|^2 \\ &\leqq (\mathbf{E}_n^* \mathcal{A}^3 \mathbf{E}_n)[\mathbf{X}^* \mathcal{A}^{-1} \mathbf{X}] \\ &\leqq \mathbf{E}_n^* \mathcal{A}^3 \mathbf{E}_n, \end{aligned}$$

by Schwartz's inequality and the fact that $\mathbf{X}^* \mathcal{A}^{-1} \mathbf{X} \leqq 1$.

Thus the third term in (43) is upper-bounded by

$$\epsilon^2 \beta^2 (\mathbf{E}_n^* \mathcal{A}^3 \mathbf{E}_n) + \frac{\alpha^2}{\epsilon^2} [\text{Im } (\mathbf{E}_n^* \mathbf{X})]^2. \quad (48)$$

Finally, substituting (42), (46), (47), and (48) into the right-hand side of (16), we have

$$\gamma_{n+1} \equiv \mathbf{E}_{n+1}^* \mathcal{A} \mathbf{E}_{n+1} \leqq \mathbf{E}_n^* \mathcal{A} \mathbf{E}_n - 2\beta \mathbf{E}_n^* \mathcal{A}^2 \mathbf{E}_n + \beta^2(1 + \epsilon^2) \mathbf{E}_n^* \mathcal{A}^3 \mathbf{E}_n + \alpha \mathcal{R}_n [\text{Im } (\mathbf{E}_n^* \mathbf{X})]^2, \quad (49)$$

where ϵ is arbitrary and

$$\mathcal{R}_n = \alpha \left[1 + (\mathbf{E}_n^* \mathcal{A} \mathbf{E}_n)^{\frac{1}{2}} + \frac{1}{\epsilon^2} \right] - 2 \text{ sinc } [\alpha \text{ Im } (\mathbf{E}_n^* \mathbf{X})]. \quad (50)$$

We make the following choice of ϵ: $\epsilon = \epsilon_0$, where ϵ_0 is defined by (with $\gamma_0 = \mathbf{E}_0^* \mathcal{A} \mathbf{E}_0$)

$$\alpha \left(1 + \sqrt{\gamma_0} + \frac{1}{\epsilon_0^2} \right) = 2 \text{ sinc } (\alpha \sqrt{\gamma_0}). \quad (51)$$

Figure 2 is a sketch of the left- and right-hand sides of eq. (51) as functions of α for various values of ϵ_0. Equation (51) has a unique solution with $0 \leqq \epsilon_0^2 < \infty$ as long as $0 \leqq \alpha < \alpha_0$, where α_0 is defined by

$$\alpha_0(1 + \sqrt{\gamma_0}) = 2 \operatorname{sinc} (\alpha_0\sqrt{\gamma_0}).$$

Note also that, by assumption (3), the coefficient $1 - 2\beta\lambda_{\min} + \beta^2\lambda_{\max}$ of $\mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n$ in the bound (41) is less than 1 and hence (49) can be weakened to

$$\mathbf{E}_{n+1}^*\mathcal{Q}\mathbf{E}_{n+1} \leqq \mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n + \mathcal{R}_n[\operatorname{Im} (\mathbf{E}_n^*\mathbf{X})]^2. \tag{52}$$

Lemma: $\mathcal{R}_n$ *is negative, and hence the sequence* $\{\gamma_n \equiv E_n^*\mathcal{Q}E_n\}$ *is monotone decreasing.*

Proof: We first observe that the sinc function in (50) defining $\mathcal{R}_n$ is even, positive, and monotone decreasing provided its argument's absolute value is less than π. But its argument is

$$\alpha \operatorname{Im} (\mathbf{E}_n^*\mathbf{X}) \leqq \alpha|\mathbf{E}_n^*\mathcal{Q}^{\frac{1}{2}}\mathcal{Q}^{-\frac{1}{2}}\mathbf{X}|.$$

This can be bounded, using Schwartz's inequality, by

$$\alpha(\mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n\mathbf{X}^*\mathcal{Q}^{-1}\mathbf{X})^{\frac{1}{2}} \leqq \alpha(\mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n)^{\frac{1}{2}}$$

and so

$$-\operatorname{sinc} [\alpha \operatorname{Im} (\mathbf{E}_n^*\mathbf{X})] \leqq -\operatorname{sinc} [\alpha(\mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n)] \qquad \text{for } \alpha(\mathbf{E}_n^*\mathcal{Q}\mathbf{E}_n)^{\frac{1}{2}} < \pi. \tag{53}$$

In particular,

$$\alpha \operatorname{Im} (\mathbf{E}_0^*\mathbf{X}) \leqq \alpha\sqrt{\gamma_0} < \pi$$

by assumption (1), and hence we can upper-bound $\mathcal{R}_0$ by

$$\mathcal{R}_0 \leqq \alpha\left(1 + \sqrt{\gamma_0} + \frac{1}{\epsilon_0^2}\right) - 2 \operatorname{sinc} (\alpha\sqrt{\gamma_0}). \tag{54}$$

According to our choice of $\epsilon = \epsilon_0$, defined by (51), the right-hand side of (54) is zero, and so $\mathcal{R}_0 \leqq 0$. It follows from (52) that $\sqrt{\gamma_1} \leqq \sqrt{\gamma_0}$, which is less than π by hypothesis. Thus $\mathcal{R}_1$ is bounded, using (53) and $\epsilon = \epsilon_0$, by $\mathcal{R}_1 \leqq \hat{\mathcal{R}}_1$, where $\hat{\mathcal{R}}_n$ is defined by

$$\hat{\mathcal{R}}_n = \alpha\left[1 + \gamma_n^{\frac{1}{2}} + \frac{1}{\epsilon_0^2}\right] - 2 \operatorname{sinc} (\alpha\gamma_n^{\frac{1}{2}}), \tag{55}$$

and $\hat{\mathcal{R}}_0 = 0$ by the definition of ϵ_0^2. Now since $\sqrt{\gamma_1} \leqq \sqrt{\gamma_0} \leqq \pi$,

$$-2 \operatorname{sinc} (\alpha\gamma_1^{\frac{1}{2}}) < -2 \operatorname{sinc} (\alpha\gamma_0^{\frac{1}{2}})$$

and so

$$R_1 \leqq \hat{R}_1 \leqq \hat{R}_0 = 0.$$

Similarly, from (52), $\gamma_2 < \gamma_1$ and by induction

$$\gamma_n \leqq \gamma_{n-1} \leqq \cdots \leqq \gamma_0$$

and all $\mathcal{R}_n \leqq 0$.

Q.E.D.

Finally, since $\mathcal{R}_n$ is negative, we obtain the following recursive upper bound from (49):

$$\gamma_{n+1} \equiv \mathbf{E}_{n+1}^* \mathcal{A} \mathbf{E}_{n+1} \leqq \mathbf{E}_n^* \mathcal{A} \mathbf{E}_n - 2\beta \mathbf{E}_n^* \mathcal{A}^2 \mathbf{E}_n + \beta^2 (1 + \epsilon_0^2) \mathbf{E}_n^* \mathcal{A}^3 \mathbf{E}_n. \quad (56)$$

REFERENCES

1. R. W. Chang, "Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1069–1104.
2. H. Kobayashi, "Simultaneous Adaptive Estimation and Decision Algorithm for Carrier-Modulated Data Transmission Systems," IEEE Trans. Commun. Technol., *COM-19*, No. 3 (June 1971), pp. 268–280.
3. D. D. Falconer, "Jointly Adaptive Equalization and Carrier Recovery in Two-Dimensional Digital Communication Systems," B.S.T.J., *55*, No. 3 (March 1976), pp. 317–334.
4. R. D. Gitlin, E. Y. Ho, and J. E. Mazo, "Passband Equalization for Differentially Phase-Modulated Data Signals," B.S.T.J., *52*, No. 2 (February 1973), pp. 219–238.
5. W. C. Lindsey and M. K. Simon, "Carrier Synchronization and Detection of Polyphase Signals," IEEE Trans. Commun., *COM-20*, No. 6 (June 1972), pp. 441–454.
6. M. K. Simon and J. G. Smith, "Carrier Synchronization and Detection of QASK Signal Sets," IEEE Trans. Commun., *COM-22*, No. 2 (February 1974), pp. 98–106.
7. R. R. Anderson and D. D. Falconer, "Modem Evaluation on Real Channels Using Computer Simulation," National Telecommunications Conference Record, San Diego, December 1974, pp. 877–883.
8. B. Widrow, *Adaptive Filters I: Fundamentals*, TR6764-6, System Theory Laboratory, Stanford Electronics Laboratories, Stanford University, December 1966.
9. G. Ungerboeck, "Theory on the Speed of Convergence in Adaptive Equalizers for Digital Communication," IBM J. Research and Development, November 1972, pp. 546–555.
10. A. Gersho, "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48*, No. 1 (January 1969), pp. 55–70.
11. R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally-Implemented Adaptive Filters," IEEE Trans. on Circuit Theory, March 1973.
12. T. P. Daniell, "Adaptive Estimation with Mutually Correlated Training Sequences," IEEE Trans. on Systems Science and Cybernetics, *SSC-6*, No. 1 (January 1970).
13. A. A. Goldstein, *Constructive Real Analysis*, New York: Harper and Row, 1967.
14. 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.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1976
Printed in U.S.A.

Spectral Occupancy of Digital Angle-Modulation Signals

By V. K. PRABHU

(Manuscript received December 2, 1975)

The spectral or band occupancy of an RF signal is often defined as the bandwidth that contains a specified fraction (usually 99 percent) of the modulated RF power. The band occupancy of binary and quaternary PSK signals with and without RF filtering and with modulation pulses of several shapes has been evaluated and the results presented in graphical and tabular form. For a binary FSK signal with phase deviation of $\pm\pi/2$, sometimes called an FM-PSK signal, numerical values of the spectral occupancy with rectangular and raised-cosine signaling have been obtained and the results given in graphical form. For a binary PSK signal with signaling rate $1/T$ and with arbitrary baseband pulse shaping, we have derived a lower bound on the fraction of the continuous power contained outside any given band, but have not been able to get a bound on the total band occupancy. However, for an FM-PSK signal, a lower bound on the total band occupancy has been derived, and it is shown that the value of this lower bound for 99-percent power occupancy is $1.117/T$. The 99-percent power occupancy bandwidth of an FM-PSK signal is $1.170/T$ with rectangular signaling and $2.20/T$ with raised-cosine signaling.

I. INTRODUCTION AND SUMMARY

Efficiency of use of the radio spectrum has recently become the subject of increased attention since terrestrial and satellite communication needs have placed an increasing burden on the available RF bands.[1,2] For spectrum conservation, the band occupancy of the chosen modulation scheme must be small so that as many channels as possible can be accommodated in a given band. Since the band occupancy of analog signals has been extensively discussed in the literature,[3–5] we shall deal here only with digital signals.

For radio systems, the "occupied bandwidth" is often specified by the spectral band which contains a certain fraction of the total RF power.* The Federal Communications Commission (FCC) presently

* For analog FM systems, an alternate way of specifying bandwidth is discussed in Ref. 5.

specifies this power to be 99 percent and requires that not more than 1 percent of the power be contained outside the assigned band.[6] For radio transmission using digital modulation techniques, the additional requirements presently specified by the FCC are in terms of the spectral density of out-of-band emission rather than just total out-of-band power.* For operating frequencies below (above) 15 GHz, the attenuation A, expressed in dB and equal to the mean output power divided by the power measured in any 4-kHz (1-MHz) band, the center frequency of which differs from the assigned frequency by 50 percent or more of the authorized bandwidth, shall not be less than 50 dB (11 dB) and shall satisfy the relation $A \geqq 35 + 0.8\ (P - 50) + 10 \log_{10} B$ for operating frequencies below 15 GHz and the relation $A \geqq 11 + 0.4\ (P - 50) + 10 \log_{10} B$ for operating frequencies above 15 GHz where P is the percent difference from the carrier frequency and B is the authorized bandwidth in megahertz. For operating frequencies below (above) 15 GHz, attenuation greater than 80 dB (56 dB) is not required for any value of P. While this is the "necessary bandwidth" specified by the FCC, the quantity "occupied bandwidth" still remains as one of the parameters used to specify the assigned band.†

The spectral occupancy of binary and quaternary PSK signals with nonoverlapping pulses of several shapes has been determined and the results presented in graphical form. The 99-percent power occupancy band of a PSK signal with rectangular signaling is extremely large; hence, for this case we also give the band occupancy when different RF filters are used to confine the spectrum.

By using the classical work of Slepian, Landau, and Pollak,[8,9] we derive a lower bound on the fractional power, contained outside any given band, of the *continuous part* of the binary PSK spectrum.[10] It is shown that the lower bound can be achieved if the baseband pulse is the inverse sine function of a certain prolate spheroidal wave function. It is also shown that the smaller the value of the lower bound, the smaller the amount of total power that can be contained in the continuous part (the total RF power has been normalized to unity). We have not been able to get a bound on the total fractional power that may be contained outside the assigned band of a binary PSK signal or find an optimum pulse shape if the total power contained in the continuous part is assumed to be a specified fraction of the total RF power.

For a binary PSK signal with phase deviation of $\pm\pi/2$, sometimes called an FM-PSK signal, numerical values of the spectral occupancy

---

* For details, see FCC Docket 19311, FCC 71-940, adopted September 8, 1971, released September 15, 1971; FCC 73-445, adopted May 3, 1973, released May 8, 1973; FCC 74-985, adopted September 19, 1974, released September 27, 1974.

† Another method of determining "sufficient bandwidth" for PSK systems is discussed in Ref. 7.

with rectangular and raised-cosine signaling have been obtained and the results are given in graphical form. For such a binary FSK signal with arbitrary baseband pulse shaping, a lower bound on the *total* band occupancy has been derived, and it is shown that the value of this lower bound for 99-percent power occupancy is $1.117/T$, where T is the signal interval. The 99-percent power occupancy bandwidth of an FM-PSK signal is $1.170/T$ with rectangular signaling and $2.20/T$ with raised-cosine signaling. The good spectral properties of an FM-PSK signal with rectangular signaling are well known,[11] and it may be detected as a PSK signal with the same bit error rate performance as that of BPSK.[12]

II. SPECTRAL OCCUPANCY OF DIGITAL SIGNALS

In our analysis for PSK and FM-PSK systems, we assume that the baseband signaling pulses have a common shape and that all signaling pulses are equally likely. We also assume that symbols transmitted during different time slots are statistically independent and identically distributed.

If the digital angle-modulated (PSK or FSK) wave is represented as

$$x(t) = \text{Re} \exp \{j[2\pi f_c t + \Phi(t) + \theta]\}, \tag{1}$$

it is shown in Refs. 10 and 13 that the power spectral density $\mathbf{P}_x(f)$ of $x(t)$ can be expressed as

$$\mathbf{P}_x(f) = \tfrac{1}{4}\mathbf{P}_v(f - f_c) + \tfrac{1}{4}\mathbf{P}_v(-f - f_c), \tag{2}$$

where $\mathbf{P}_v(f)$ is the power spectral density of

$$v(t) = e^{j\Phi(t)} \tag{3}$$

and f_c is the carrier frequency. In (1), θ is assumed to be a random variable uniformly distributed over $[0, 2\pi)$.

The fractional power Δ^2 contained outside the band $[f_c - W, f_c + W]$ can be shown to be

$$\Delta^2 = 2\int_W^\infty \mathbf{P}_v(f)df - \int_{2f_c - W}^{2f_c + W} \mathbf{P}_v(f)df. \tag{4}$$

In most cases of practical interest, $\mathbf{P}_v(f)$ is a rapidly decreasing function of f, $f_c/W \gg 1$, and*

$$\Delta^2 = 2\int_W^\infty \mathbf{P}_v(f)df. \tag{5}$$

* Since $\mathbf{P}_v(f) \geq 0$, $\Delta^2 \leqq 2\int_W^\infty \mathbf{P}_v(f)df$ for any f_c/W.

2.1 Spectral density of an M-ary PSK signal

For an M-ary PSK signal (we assume $M = 2^N$, N an integer) with signaling rate $1/T$,

$$\Phi(t) = \sum_{k=-\infty}^{\infty} \underline{\mathbf{a}_k} \cdot \underline{g(t - kT)}, \tag{6}$$

where $\mathbf{a}_k$ is a vector-valued stationary random process and $\underline{\mathbf{g}(t)}$ are the pulse shapes corresponding to the M symbols.

If the signaling pulses in different time slots never overlap, it is shown in Ref. 10 that $\mathbf{P}_s(f)$ consists of a line component part $\mathbf{P}_{s_l}(f)$ and a continuous part $\mathbf{P}_{s_c}(f)$, $\mathbf{P}_s(f) = \mathbf{P}_{s_l} + \mathbf{P}_{s_c}(f)$,

$$\mathbf{P}_{s_l}(f) = \frac{1}{T^2} |\underline{\mathbf{w}} \cdot \mathbf{R}(f)|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right), \tag{7}$$

$$\mathbf{P}_{s_c}(f) = \frac{1}{2T} \sum_{i=1}^{M} \sum_{j=1}^{M} w_i w_j |R_i(f) - R_j(f)|^2, \tag{8}$$

where $\underline{\mathbf{w}} = [w_1, w_2, \cdots, w_M]$, w_i is the probability that the ith signaling waveform $g_i(t)$ is transmitted in any time slot and $R_i(f)$ is the Fourier transform of $r_i(t)$,

$$r_i(t) = \begin{cases} \exp[jg_i(t)], & 0 < t \leqq T \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Since we assume that the M signaling pulses have a common shape,

$$\underline{\mathbf{g}(t)} = [\alpha_1, \alpha_2, \cdots, \alpha_M] g(t), \tag{10}$$

where α_i is the peak phase value of the ith symbol and the maximum value of $g(t)$ has been normalized to unity.

From (5),

$$\Delta^2 = \Delta_l^2 + \Delta_c^2, \tag{11}$$

where

$$\Delta_l^2 = 2 \int_W^{\infty} \mathbf{P}_{s_l}(\mu) d\mu = \text{the fractional part of line power contained outside the band} \tag{12}$$

and

$$\Delta_c^2 = 2 \int_W^{\infty} \mathbf{P}_{s_c}(\mu) d\mu = \text{the fractional part of continuous power contained outside the band.} \tag{13}$$

2.2 Spectral density of an M-ary FSK signal

For an M-ary FSK signal,[13]

$$\Phi(t) = \int^t f_d(\mu) d\mu, \tag{14}$$

$$f_d(t) = \sum_{k=-\infty}^{\infty} \underline{a_k} \cdot \mathbf{h}(t - kT)\underline{]}, \qquad \underline{\mathbf{h}(t)} = \underline{0}, \; t \leqq 0, \; t > T, \quad (15)$$

$$\mathbf{P}_v(f) = \frac{1}{T}\underline{\mathbf{R}(f)} \cdot (\mathbf{A} + \mathbf{A}^\dagger) \cdot \mathbf{R}^*(f)\underline{]}, \quad (16)$$

where

$$\mathbf{A} = \frac{1}{2}\mathbf{w}_d + \frac{e^{-j2\pi fT}\mathbf{w}\underline{]} \cdot \underline{\mathbf{r}(T)} \cdot \mathbf{w}_d}{1 - e^{-j2\pi fT}\underline{\mathbf{w}} \cdot \mathbf{r}(T)\underline{]}}, \qquad |\underline{\mathbf{w}} \cdot \mathbf{r}(T)\underline{]}| < 1, \quad (17)$$

$$\mathbf{w}_d = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_M \end{bmatrix}, \quad (18)$$

$R_i(f)$ is the Fourier transform of $r_i(t)$, and

$$r_i(t) = \begin{cases} \exp\left[j \int_0^t h_i(\mu)d\mu\right], & 0 < t \leqq T \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

We make the same assumptions for FSK as for PSK. However, note that $\mathbf{P}_v(f)$ does not contain any lines if $\mathbf{w}\underline{]}$ and $\mathbf{r}(t)\underline{]}$ satisfy the inequality in (17). Since spectral lines do not often contain any useful information (except for carrier recovery), their presence indicates nonoptimum pulse shaping. In this paper, we shall not consider FSK with spectral lines. For FSK, $\Delta^2 = \Delta_c^2$ from (5).

III. BAND OCCUPANCY OF A BINARY PSK SIGNAL

For binary PSK, we assume that $\alpha_1 = -\alpha_2 = \pi/2$ and that both symbols are equally likely. From (8),

$$\mathbf{P}_{v_c}(f) = \frac{1}{4T}\,|R_1(f) - R_2(f)|^2, \quad (20)$$

where

$$\begin{aligned} R_1(f) - R_2(f) &= \int_0^T \left[e^{j(\pi/2)g(t)} - e^{-j(\pi/2)g(t)}\right]e^{-j2\pi ft}dt \\ &= 2j \int_0^T \sin\left\{\frac{\pi}{2}g(t)\right\} e^{-j2\pi ft}dt. \end{aligned} \quad (21)$$

For rectangular, cosinusoidal, raised-cosinusoidal, trapezoidal, and triangular $g(t)$, we have calculated $\mathbf{P}_v(f)$ from (7) and (8) and Δ^2 from (5). For these cases, the total out-of-band power ratio Δ^2 for binary PSK is plotted in Figs. 1 and 2. The 99-percent (or any other fractional) power bandwidth occupancy for binary PSK may be determined from

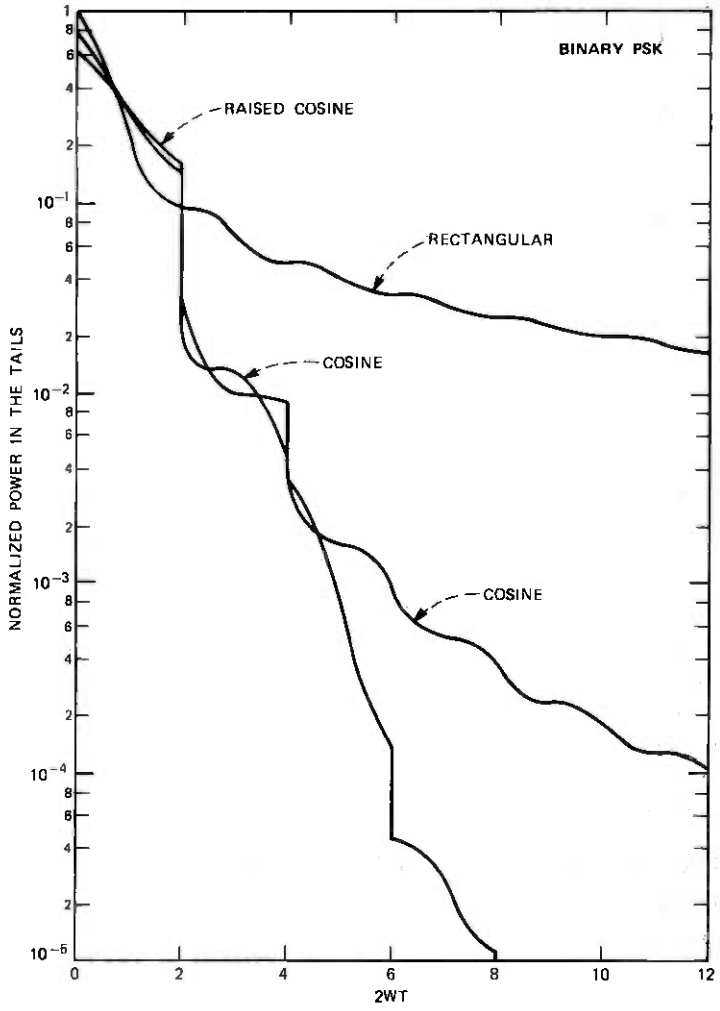


Fig. 1—Normalized power contained outside the band $[-W, W]$ for binary PSK with different baseband signaling waveforms.

these figures. Since the 99-percent power occupancy of binary PSK with rectangular signaling is very large, we show the bandwidth occupancy with RF filtering in Figs. 3, 4, and 5.

IV. BAND OCCUPANCY OF A QPSK SIGNAL

For QPSK modulation and for equally likely symbols,

$$\mathbf{P}_{v_c}(f) = \frac{1}{32T} \sum_{i=1}^{4} \sum_{j=1}^{4} |R_i(f) - R_j(f)|^2, \tag{22}$$

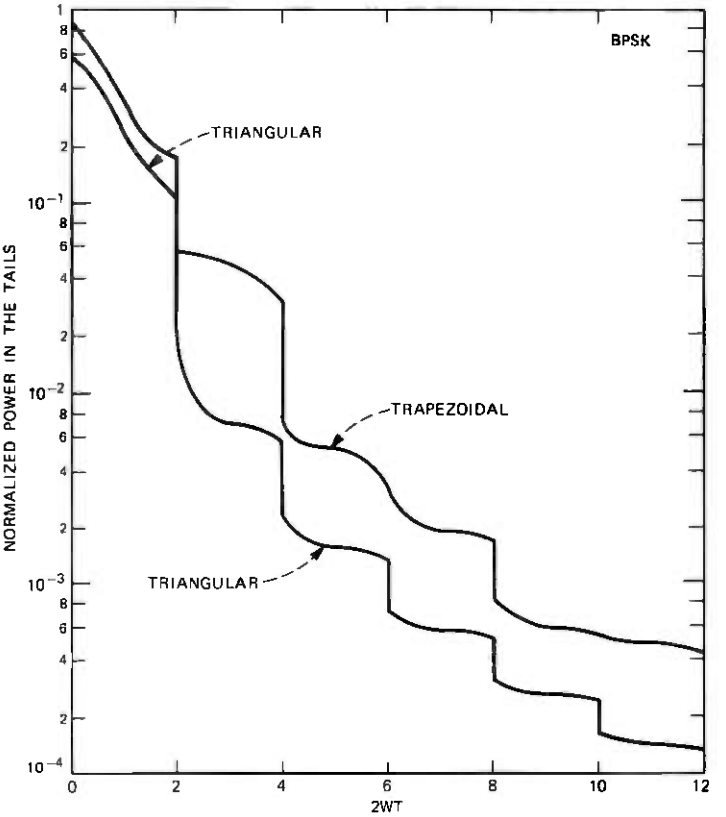


Fig. 2—Normalized power contained outside the band $[-W, W]$ for binary PSK with different baseband signaling waveforms.

where $R_l(f)$ is the Fourier transform of $r_l(t)$,

$$r_l(t) = \begin{cases} \exp[j\alpha_l g(t)], & 0 < t \leqq T \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

and

$$\alpha_l = (2l - 5)\frac{\pi}{4}, \qquad l = 1, 2, 3, 4. \tag{24}$$

$\mathbf{P}_{v_l}(f)$ is given by (7).

For rectangular, cosinusoidal, raised-cosinusoidal, trapezoidal, and triangular $g(t)$, we have calculated $\mathbf{P}_v(f)$ from (7) and (8) and Δ^2 from (5). For these cases, the total out-of-band power ratio Δ^2 is plotted in Figs. 6 and 7. The 99-percent (or any other fractional) power bandwidth occupancy for quaternary PSK may be determined from these figures. Since the spectral density of QPSK with rectangular signaling

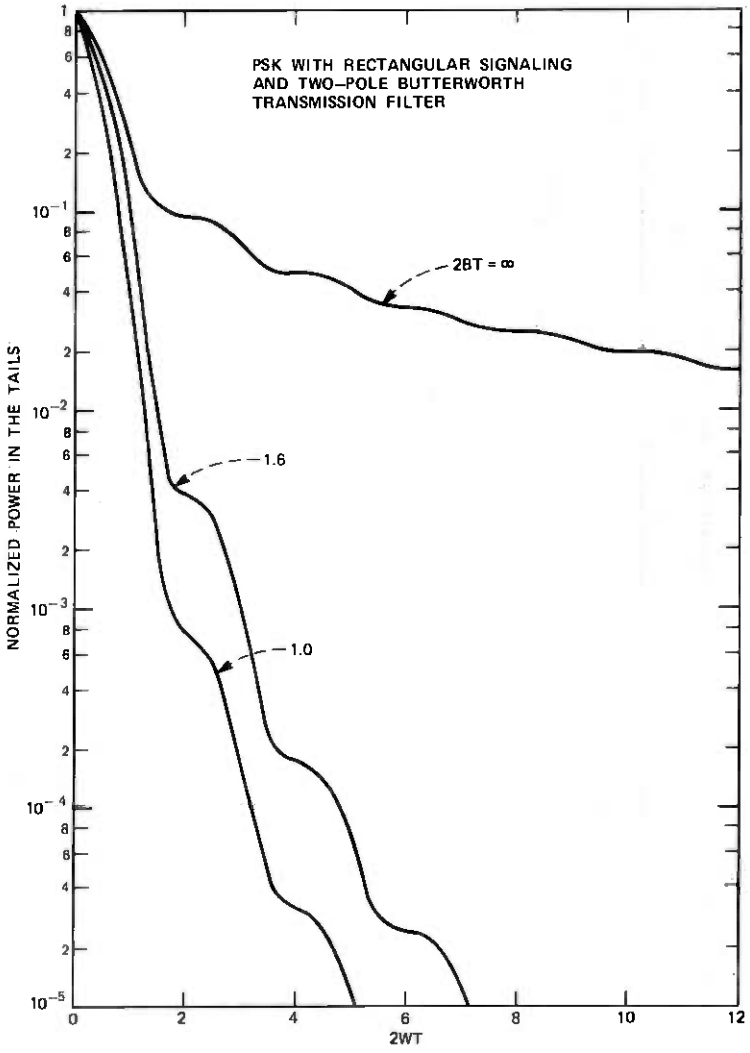


Fig. 3—Normalized power contained outside the band $[-W, W]$ for M-ary PSK ($M = 2^N$, $N \geq 1$) with rectangular signaling and a two-pole Butterworth transmission filter. The squared amplitude characteristic of the equivalent low-pass filter is assumed to be given by $|H_T(f)|^2 = 1/[1 + (f/A)^4]$, where $2B = 2A(\pi/4)/\sin \pi/4$ is the noise bandwidth of the filter.

is the same as that of BPSK, the bandwidth occupancy of QPSK with RF filtering is also given by Figs. 3, 4, and 5.

V. BAND OCCUPANCY OF AN FM-PSK SIGNAL

A binary FM-PSK signal is a special case of the binary continuous-phase FSK modulation where the phase deviation in one signaling

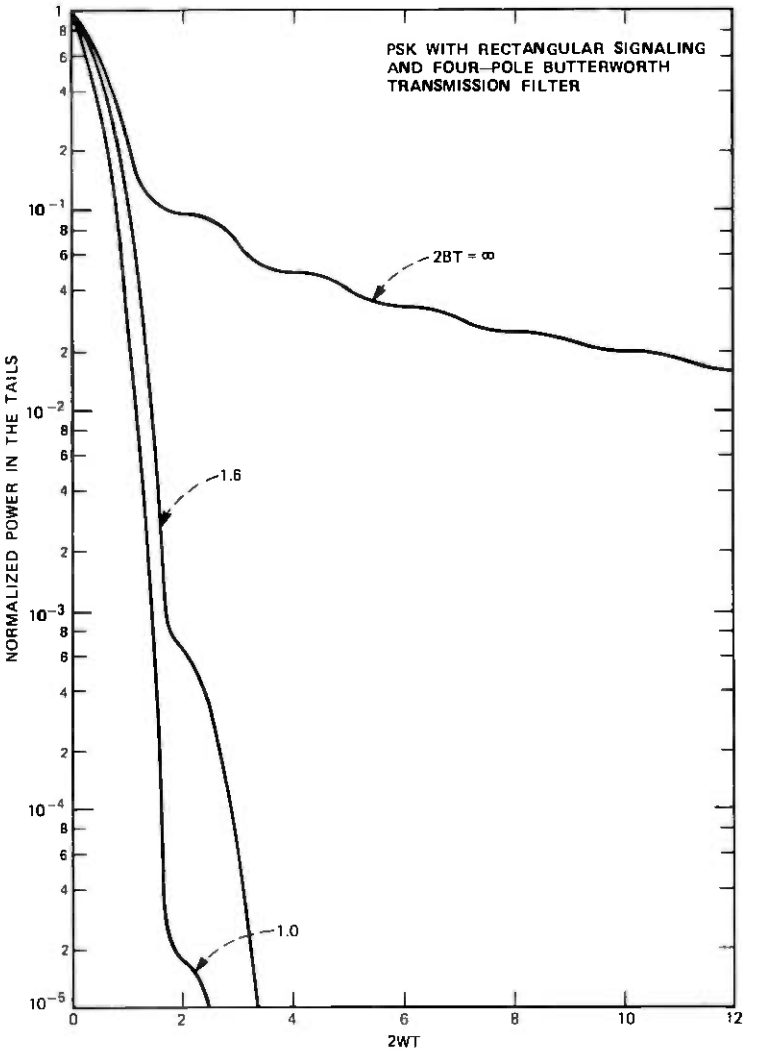


Fig. 4—Normalized power contained outside the band $[-W, W]$ for M-ary PSK ($M = 2^N$, $N \geq 1$) with rectangular signaling and a four-pole Butterworth transmission filter. The squared amplitude characteristic of the equivalent low-pass filter is assumed to be given by $|H_T(f)|^2 = 1/[1 + (f/A)^8]$, where $2B = 2A(\pi/8)/\sin \pi/8$ is the noise bandwidth of the filter.

interval is $\pm\pi/2$ and which can be detected as a PSK signal. Note that one may use a four-phase demodulator to detect a binary FM-PSK signal* to have the same bit error rate performance as that of BPSK.[14–16]

* A form of binary FM-PSK can be shown to be equal to the sum of two offset quadrature-phase binary PSK signals. A form of it is, therefore, sometimes referred to as offset QPSK (Ref. 2). An FM-PSK with rectangular frequency modulation signaling is called fast FSK in Ref. 12 and MSK in Ref. 14.

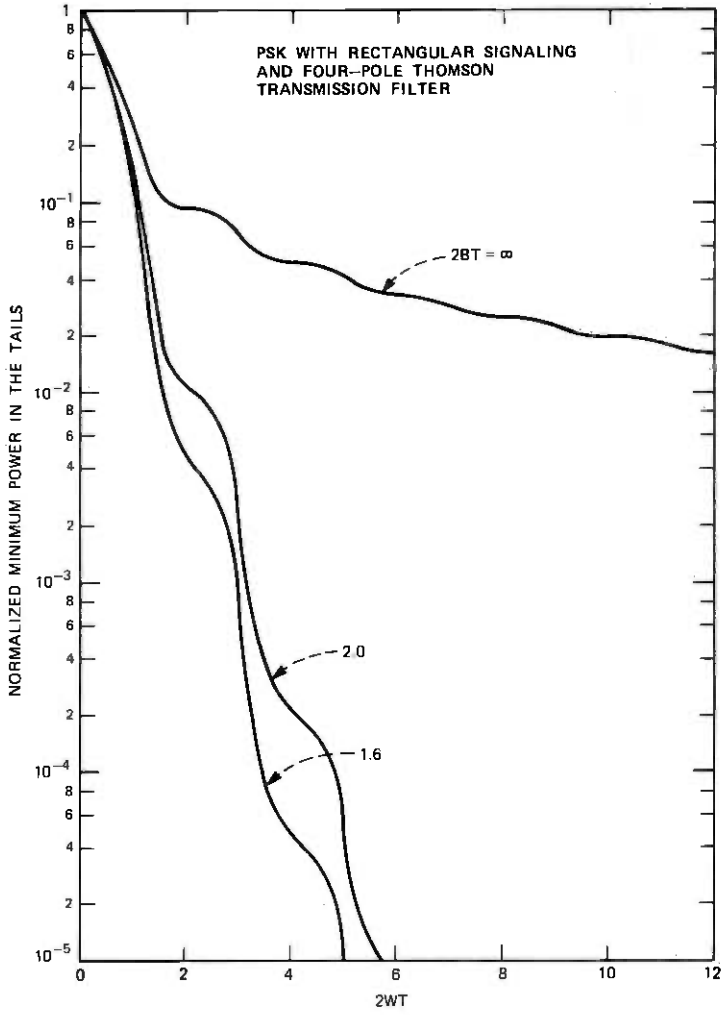


Fig. 5—Normalized power contained outside the band $[-W, W]$ for M-ary PSK ($M = 2^N$, $N \geq 1$) with rectangular signaling and a four-pole Thomson transmission filter. The squared amplitude characteristic of the equivalent low-pass filter is assumed to be given by $|H_T(f)|^2 = 11025/(z^8 + 10z^6 + 135z^4 + 1575z^2 + 11025)$, $z = f/A$ and $2B = 4.4238A$ is the noise bandwidth of the filter.

There are no discrete lines in the FM-PSK spectrum, but standard techniques (such as the Costas loop) can be used to recover the coherent carrier (it is necessary to use differential encoding or prior knowledge of framing polarity, etc., to resolve the ambiguity present in the phase of the recovered carrier).[12]

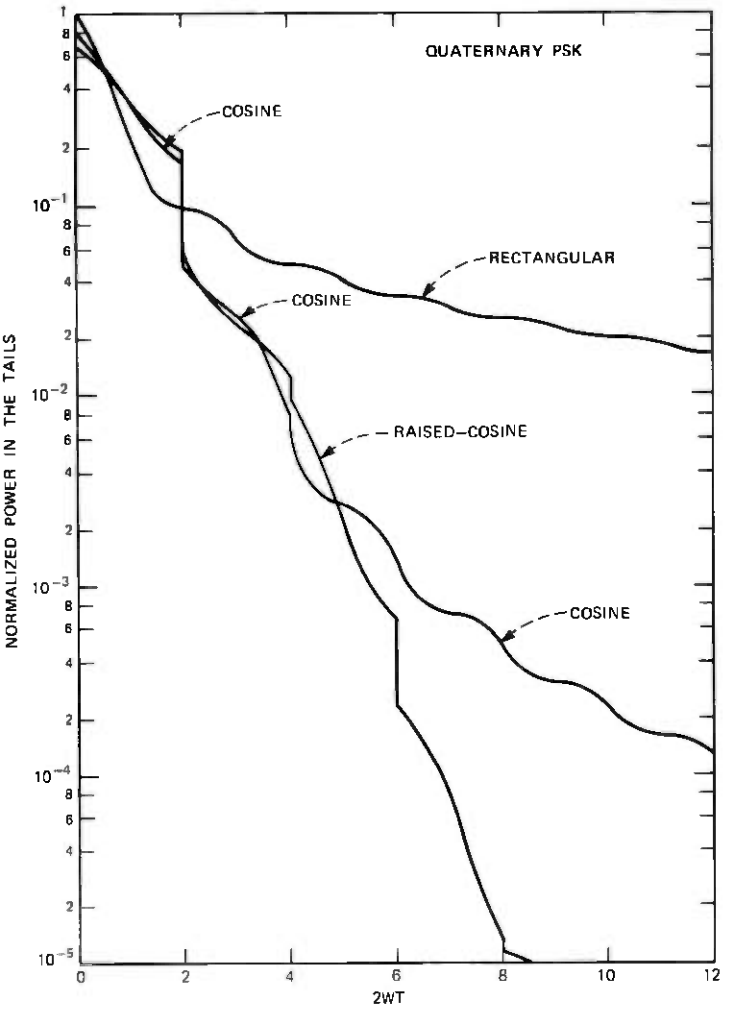


Fig. 6—Normalized power contained outside the band $[-W, W]$ for quaternary PSK with different baseband signaling waveforms.

To get the spectral density of binary FSK, we put

$$r(T)] = \begin{matrix} e^{j(\pi/2)} \\ e^{-j(\pi/2)} \end{matrix} \Bigg] \tag{25}$$

in (16), (17), and (19) for any baseband signaling waveform $h(t)$. We assume that we transmit $+1$ by shifting the carrier frequency by $+f_d g(t)$, $0 < t \leqq T$, and -1 by shifting the carrier frequency by

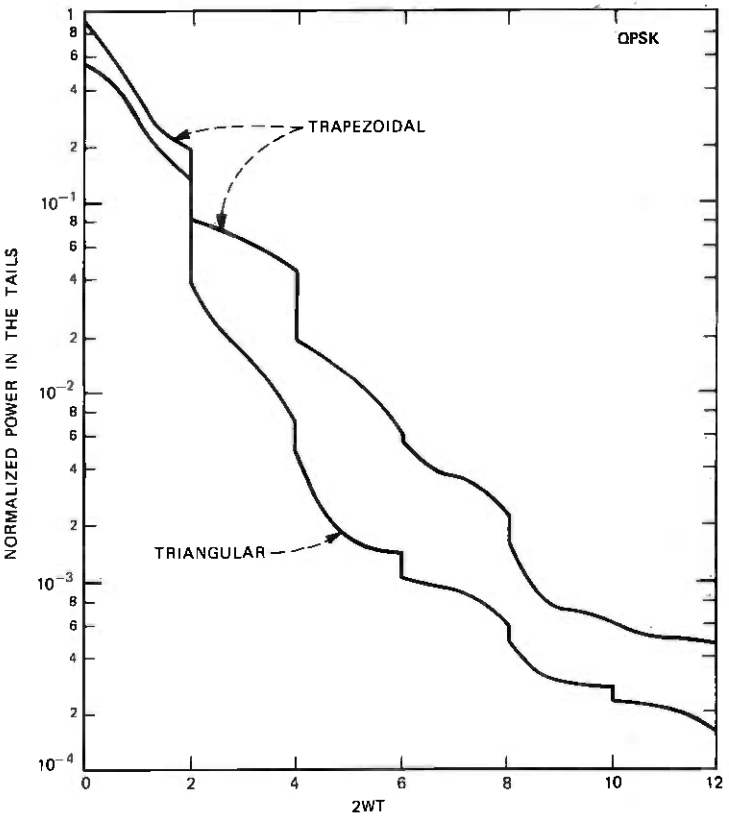


Fig. 7—Normalized power contained outside the band $[-W, W]$ for quaternary PSK with different baseband signaling waveforms.

$-f_d g(t)$, $0 < t \leqq T$. For rectangular signaling,

$$f_d = \frac{1}{4}\frac{1}{T}, \tag{26}$$

and for raised-cosine signaling,

$$f_d = \frac{1}{2}\frac{1}{T}, \tag{27}$$

so that the peak frequency deviation with raised-cosine signaling is larger than that with rectangular signaling.

From (16), (17), and (25) one can show that the spectral density $\mathbf{P}_v(f)$ of binary FM-PSK is

$$\mathbf{P}_v(f) = \mathbf{P}_{v_c}(f) = \frac{1}{T} R_1 R_2 \cdot \begin{bmatrix} \frac{1}{2}\{1 + \sin(2\pi fT)\} & -\frac{j}{2}\cos(2\pi fT) \\ \frac{j}{2}\cos(2\pi fT) & \frac{1}{2}\{1 - \sin(2\pi fT)\} \end{bmatrix} \cdot \begin{bmatrix} R_1^* \\ R_2^* \end{bmatrix}, \quad (28)$$

where R_1, R_2 are the Fourier transforms of $r_1(t)$, $r_2(t)$

$$r_1(t) = \begin{cases} \exp\left[j2\pi f_d \int_0^t g(t)dt\right], & 0 < t \leqq T \\ 0, & \text{otherwise,} \end{cases} \quad (29)$$

$$r_2(t) = \begin{cases} \exp\left[-j2\pi f_d \int_0^t g(t)dt\right], & 0 < t \leqq T \\ 0, & \text{otherwise.} \end{cases} \quad (30)$$

For rectangular and raised-cosine signaling, we plot for binary FM-PSK the out-of-band power ratio Δ^2 in Fig. 8. The 99-percent (or any other fractional) power bandwidth occupancy may be determined from results given in this figure.

VI. TIME-LIMITED AND BAND-LIMITED SIGNALS

We shall derive the lower bound on the band occupancy of binary PSK and FSK signals by using the results obtained for time-limited and band-limited functions.

In their classical papers, Slepian, Landau, and Pollak have derived[8,9] the pulse waveform of given duration that has a maximum of its energy concentrated below a certain frequency band. These optimum pulse waveforms are the well-known prolate spheroidal wave functions. A widespread opinion is that pulses with minimum energy at high frequencies should have a rounded form with many continuous derivatives. Since the optimum pulses (the prolate spheroidal wave functions) are usually not continuous at the limits of their truncation interval, this opinion does not seem to be justified. In fact, Hilberg and Rothe[17] have shown recently that constraints of continuous derivatives tend to increase the total out-of-band energy. We shall now state the bounds given by Slepian, Landau, and Pollak.

If we define

$$\alpha^2 = \frac{\int_{t_0 - T'/2}^{t_0 + T'/2} |f(t)|^2 dt}{\int_{-\infty}^{\infty} |f(t)|^2 dt}, \quad (31)$$

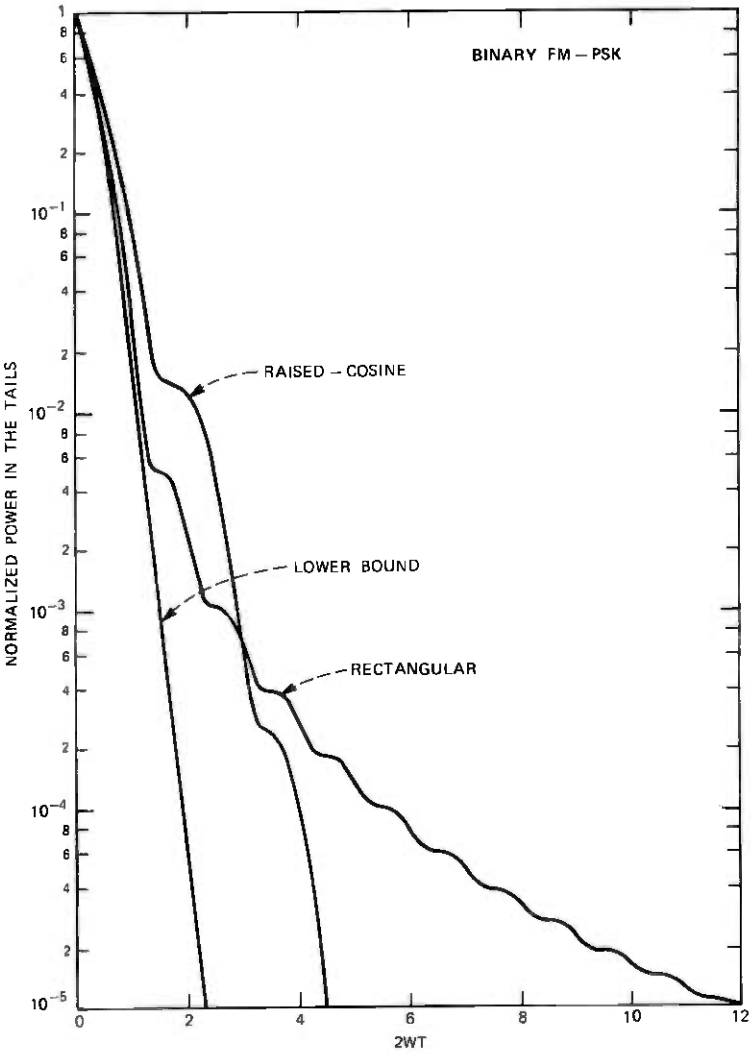


Fig. 8—Normalized power Δ^2 contained outside the band $[-W, W]$ for binary FM-PSK with different baseband signaling waveforms and also the lower bound on Δ^2 for any baseband signaling waveform.

and

$$\beta^2 = \frac{\int_{-W}^{W} |F(f)|^2 df}{\int_{-\infty}^{\infty} |F(f)|^2 df}, \qquad F(f) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi ft} dt, \tag{32}$$

it is shown in Ref. 9 that

$$\cos^{-1}(\alpha) + \cos^{-1}(\beta) \geqq \cos^{-1}\sqrt{\lambda_0}, \tag{33}$$

where λ_0 is the largest eigenvalue of the integral equation

$$\lambda f(t) = \frac{1}{\pi} \int_{-(T'/2)}^{T'/2} f(s) \frac{\sin \{2\pi W(t-s)\}}{(t-s)} ds. \tag{34}$$

In (31), we assume that $f(t) \in \mathfrak{L}_\infty^2$ where $\mathfrak{L}_\infty^2$ is the set of all complex-valued functions defined on the real line and integrable in absolute square [$f(t)$ has finite energy].

In binary PSK and certain binary FSK, we shall show that $\mathbf{P}_v(f)$ or $\mathbf{P}_{v_c}(f)$ can be expressed as the energy density spectrum [$|X(f)|^2$] of a certain $x(t)$, time-limited to a duration T_{eq}.* From (31), if $x(t)$ is of duration T_{eq},

$$\alpha^2 = 1, \qquad T' = T_{eq}, \qquad \beta^2 \leqq \lambda_0 \tag{35}$$

and the maximum value of β^2 is attained when $x(t)$ is a prolate spheroidal wave function $\psi_0(t, d)$ given in Refs. 8 and 9, $d = \pi W T_{eq}$. The fractional energy Λ^2 contained outside the band $[-W, W]$ is, therefore, lower-bounded by

$$\Lambda^2 = 1 - \beta^2 \geqq 1 - \lambda_0 = \Lambda_{\min}^2. \tag{36}$$

The values of $\Lambda_{\min}^2$ computed from the relations given in Refs. 8, 9, and 18 are shown in Fig. 9. It therefore follows that it is impossible to find an $\mathfrak{L}^2$-integrable pulse waveform $x(t)$ which has a duration T_{eq} and which has a fractional energy less than $\Lambda_{\min}^2(WT_{eq})$ outside the band $[-W, W]$.

VII. LOWER BOUND ON THE BAND OCCUPANCY OF PSK AND FM-PSK SIGNALS

Let us first consider the band occupancy of the continuous part $\mathbf{P}_{v_c}(f)$ of a BPSK spectrum.

From (20) and (21),

$$\mathbf{P}_{v_c}(f) = |X(f)|^2, \tag{37}$$

where $X(f)$ is the Fourier transform of

$$x(t) = \begin{cases} \dfrac{\sin\left\{\dfrac{\pi}{2} g(t)\right\}}{\sqrt{T}}, & 0 < t \leqq T \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

In (37) we have expressed the continuous part of the spectral density of a binary PSK signal in terms of the energy density spectrum

* In FM-PSK, it will turn out that $T_{eq} = 2T$, where T is the duration of the signaling waveform $g(t)$. Hence, we use the symbol T_{eq} to denote the duration of $x(t)$.

of an arbitrary pulse waveform $x(t) \in \Omega \subset \mathfrak{L}_w^2$.* $x(t)$ can be nonzero only for $0 < t \leqq T$. From Section VI, it therefore follows that the out-of-band power ratio Λ_c^2 of a binary PSK signal is lower-bounded by

$$\Lambda_c^2 \geqq \Lambda_{\min}^2(WT), \tag{39}$$

where

$$\Lambda_c^2 = \frac{\text{Continuous power contained outside the band } (-W, W)}{\text{Total power contained in the continuous part}}. \tag{40}$$

Note that $\Lambda_c^2 \neq \Delta^2$ or Δ_c^2, but

$$\frac{\Lambda_c^2}{\Delta_c^2} = \frac{\text{Total power in } \mathbf{P}_v(f)}{\text{Total power in } \mathbf{P}_{v_c}(f)} \geqq 1. \tag{41}$$

Now Λ_c^2 can be made equal to $\Lambda_{\min}^2(WT)$ by choosing

$$x(t) = k\psi_0(t - T/2, d), \qquad d = \pi TW, \qquad T' = T, \tag{42}$$

where $\psi_0(t, d)$ is a prolate spheroidal wave function and k is a normalizing constant.† We choose k so that the total power E contained in the information-bearing part $\mathbf{P}_{v_c}(f)$ [equivalently, the total energy contained in $x(t)$] is maximum. Since $\psi_0(t, d)$ is maximum at $t = 0$, E is maximized by choosing

$$g(t) = \begin{cases} \frac{2}{\pi} \operatorname{Sin}^{-1}\left\{\dfrac{\psi_0(t - T/2, d)}{\psi_0(0, d)}\right\}, & 0 < t \leqq T,‡ \\ 0, & \text{otherwise.} \end{cases} \tag{43}$$

For this value of $g(t)$,

$$E = \frac{\lambda_0}{T\psi_0^2(0, d)}. \tag{44}$$

For $x(t)$ in (42) and $g(t)$ in (43), the minimum out-of-band power ratio $\Lambda_{\min}^2(WT)$ can be attained, and

$$\Lambda_{\min}^2(WT) = 1 - \lambda_0. \tag{45}$$

For some values of d, the minimum out-of-band power ratio $\Lambda_{\min}^2(WT)$ and the maximum power contained in the continuous part are listed in Table I.§ The rest of the power in the PSK signal is contained in $\mathbf{P}_{v_l}(f)$ or the discrete lines. For binary PSK, it follows from Sec. VI and eqs. (39) and (42) that $[\Lambda_c^2]_{\min}$ is given in Fig. 9.

---

* Since $|x(t)| \leqq 1$, note that Ω is a proper subset of $\mathfrak{L}_w^2$.

† Our letter d in $\psi_0(t, d)$ corresponds to the letter c used in Refs. 8, 9, and 18.

‡ $\theta = \operatorname{Sin}^{-1}(x)$ denotes the principal value of the inverse sine, $-\pi/2 \leqq \theta \leqq \pi/2$.

§ We chose the values of d given in Table I so that we can make use of the results given in Refs. 8, 9, and 18.

Table I — Minimum out-of-band power ratio of binary PSK

$d = \pi TW$	WT	Minimum Out-of-Band Power Ratio $\Delta^2_{\min}(WT)$	Maximum Normalized Power Contained in the Continuous Part of $\mathbf{P}_v(f)$
0.5	0.1592	0.6903	0.9730
1.0	0.3183	0.4274	0.9015
2.0	0.6366	0.1194	0.7122
4.0	1.2732	0.00411	0.4736

For $d = 0.5$, 1.0, 2.0, and 4.0, we plot the optimum $g(t)$ from (43) in Fig. 10. For $g(t)$ in (43) and Fig. 10, we plot the spectral density $\mathbf{P}_{v_c}(f)$ of binary PSK in Fig. 11.

From (11), $\Delta^2 = \Delta_i^2 + \Delta_c^2$, and since one usually specifies the total out-of-band power ratio, we list in Table II Δ_i^2, Δ_c^2, and Δ^2 for $g(t)$ in (43) and WT in Table I. Also for $g(t)$ in (43), we plot the *total* out-of-band power in Fig. 12. Comparing Figs. 1, 2, and 12, note that the total out-of-band power for the optimum pulse is very close to that for the rectangular pulse for $0 \leqq 2WT \leqq 1$. In the neighborhood of

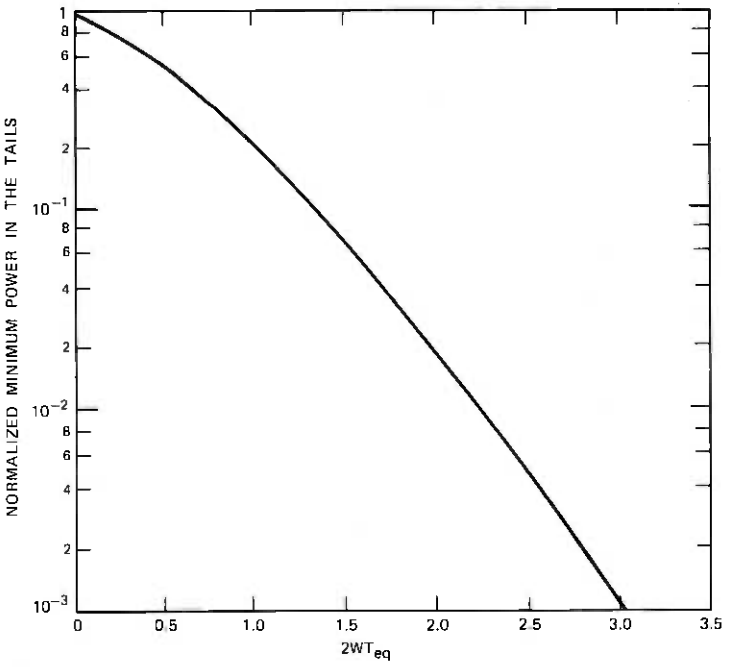


Fig. 9—Lower bound on the fractional energy contained outside the band $[-W, W]$ when the pulse $f(t)$ is of duration T_{eq}.

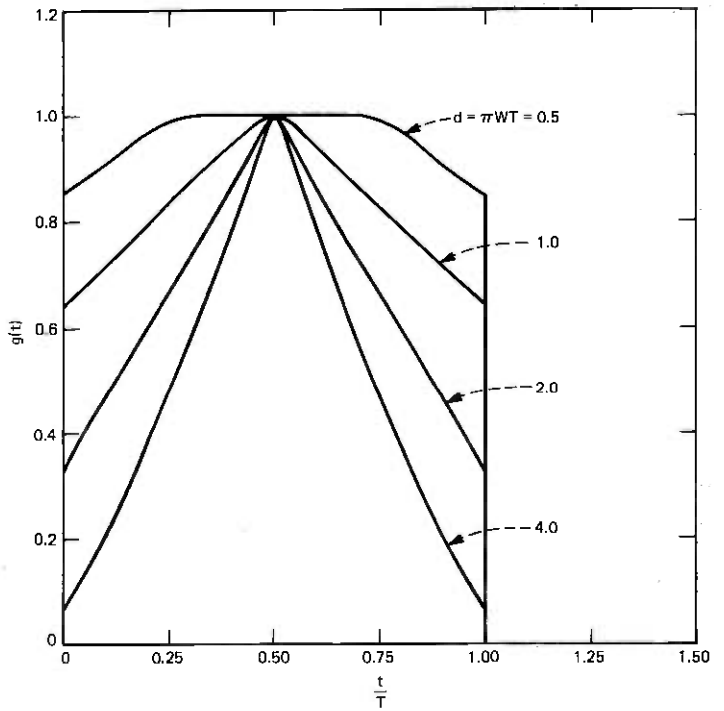


Fig. 10—Phase modulation pulse $g(t)$ for binary PSK for optimum continuous spectral occupancy.

$2WT = 0$, the total out-of-band power with the optimum pulse is greater than that with cosine, raised-cosine, triangular, or trapezoidal pulse. This is because the optimum pulse minimizes the fractional out-of-band *continuous* power and not the total power. For $g(t)$ in (43), it must be noted that the smaller the out-of-band continuous power ratio, the smaller the maximum amount of power contained in the continuous part. The rest of the power is contained in the discrete lines.*

One must note that, in general,

$$\Delta^2 \neq \Lambda^2_{\min}(WT) \tag{46}$$

the total out-of-band power ratio (the total out-of-band power divided

* The total out-of-band power with optimum pulse increases as a function of $2WT$ if we use the pulse in (43) and if $2WT > 2.5$. This is because an increasingly large amount of power is contained in the discrete lines and the total out-of-band discrete power very much dominates the out-of-band continuous power. By choosing the pulse which is optimum for $2WT \leqq 2.5$, we can make the total out-of-band power a monotone-decreasing function of $2WT$.

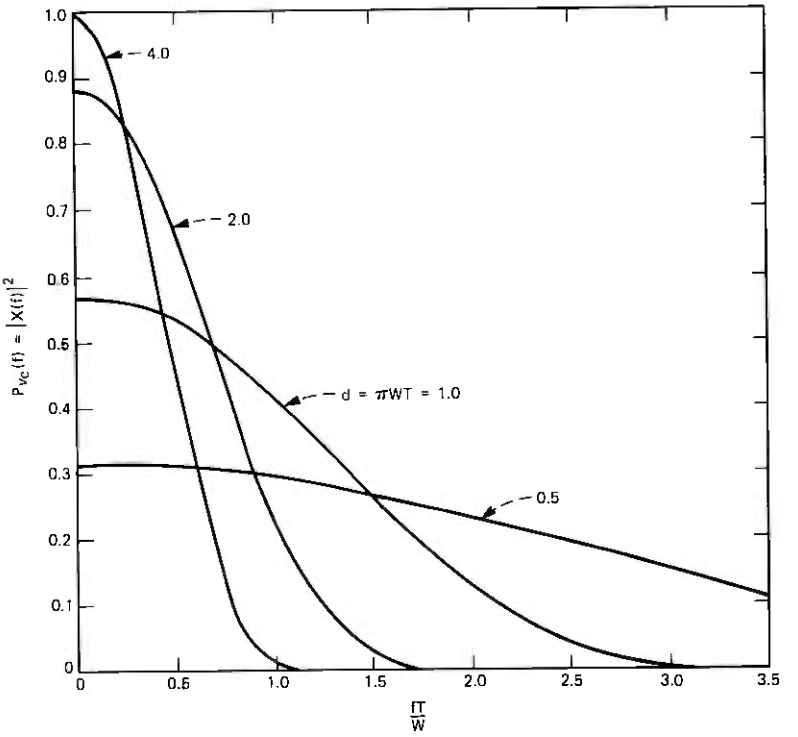


Fig. 11—Continuous spectral density $\mathbf{P}_{v_c}(f)$ of binary PSK for optimum continuous spectral occupancy.

by total power) is not equal to the out-of-band continuous power (the out-of-band continuous power divided by power contained in the continuous part). Also note that we have obtained a lower bound on $\Lambda^2_{\min}(WT)$ and not on Δ^2. Since any time function $y(t)$ containing discrete lines does not belong to $\mathcal{L}^2_\infty$, analysis given in Refs. 8 and 9 does not enable the optimization of Δ^2.

Our efforts to find a lower bound on the *total* band occupancy of a BPSK signal have not been successful so far, and it is suggested as an interesting problem for the reader.

So that we may compare the spectral occupancy of binary PSK with several different modulation pulses for $\Delta^2 = 0.1$, 0.01, and 0.001, we list in Table III the values of $2WT$.

Let us now consider a QPSK signal. From (22) one can show that no single function $x(t)$ can be found such that its energy density spectrum $|X(f)|^2$ is equal to $\mathbf{P}_{v_c}(f)$. If $2WT$ is large so that a small amount of total power is contained in the tails, we feel that the total out-of-band

Table II — Total out-of-band power ratio for binary PSK with $g(t)$ given by (43)

d	WT	Normalized Power Contained in the Continuous Part $\mathbf{P}_{v_c}(f)$	Normalized Power Contained in the Discrete Part $\mathbf{P}_{v_l}(f)$	Minimum Out-of-Band Continuous Power Δ_c^2	Out-of-Band Discrete Power Δ_l^2	Total Out-of-Band Power Ratio Δ^2	Normalized Power in the Lines at Frequency n/T from the Carrier	
0.5	0.1592	0.9730	0.0269	0.6717	0.00668	0.6784	$n = 0$	0.02024
1.0	0.3183	0.9015	0.0985	0.3853	0.02368	0.4090	$n = 0$	0.07483
2.0	0.6366	0.7122	0.2878	0.08506	0.06258	0.1476	$n = 0$	0.2252
4.0	1.2732	0.4736	0.5264	0.001943	0.00721	0.009153	$n = 0$	0.4364
							$n = \pm 1$	0.04139

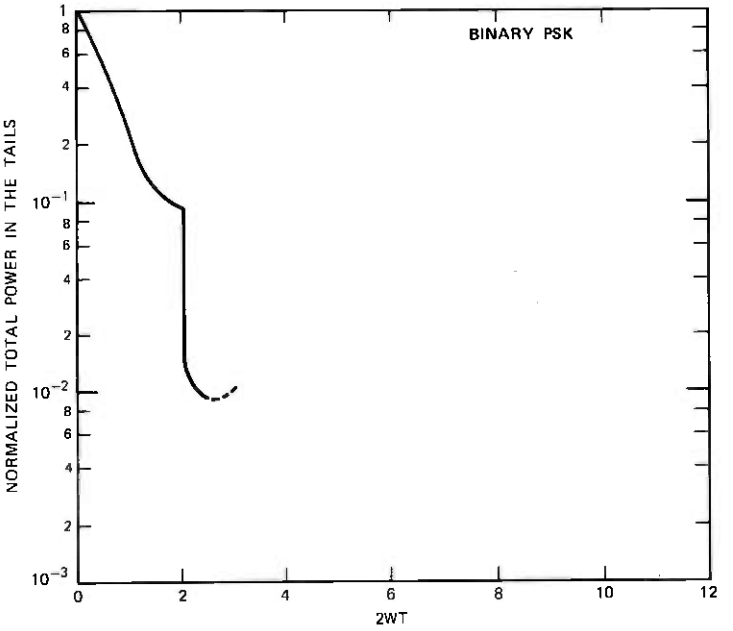


Fig. 12—Normalized total power contained outside the band $[-W, W]$ for $g(t)$ in (43). Observe that the total out-of-band power for $g(t)$ in (43) increases (see the dashed portion of the figure) as a function of $2WT$ for $2WT > 2.5$. This is because the total out-of-band discrete power, which is not optimized, very much dominates the out-of-band continuous power. Note that $g(t)$ in (43) only minimizes the fraction of the continuous power contained outside the band $[-W, W]$. For $2WT > 2.5$, by choosing $g(t)$ which is optimum for $2WT \leqq 2.5$, we can make the total out-of-band power decrease as a function of $2WT$.

power for a QPSK signal is lower-bounded by the results given for a BPSK signal. The band occupancy of QPSK for $\Delta^2 = 0.1$, 0.01, and 0.001 for different modulation pulses is listed in Table IV.

We now derive a lower bound on the *total* band occupancy of an FM-PSK signal. In (29) and (30), $g(t) \in \mathfrak{L}_\infty^2$ is assumed to be completely arbitrary.

By defining

$$R_{t_1} = j \frac{e^{j(\pi fT - \pi/4)} + e^{-j(\pi fT - \pi/4)}}{2} R_1, \tag{47}$$

$$R_{t_2} = j \frac{e^{j(\pi fT - \pi/4)} - e^{-j(\pi fT - \pi/4)}}{2} R_2, \tag{48}$$

we can show from (28) that

$$\mathbf{P}_v(f) = \frac{1}{T} |R_{t_1} - R_{t_2}|^2 = |X(f)|^2, \tag{49}$$

Table III — Values of out-of-band power ratio Δ^2 for binary PSK with different baseband modulation pulses

Pulse $g(t)$ $g(t) = 0, t \leqq 0, t > T$	Normalized Power Contained in $\mathbf{P}_{v_c}(f)$	Total Out-of-Band Power Ratio Δ^2		
		0.1	0.01	0.001
		$2WT$	$2WT$	$2WT$
Rectangular $g(t) = 1, 0 < t \leqq T$	1.000	1.807	19.295	
Trapezoidal $g\left(t + \frac{T}{2}\right) = \begin{cases} 1, 0 < \lvert t \rvert \leqq \frac{T}{4} \\ 2\left(1 - \frac{2\lvert t \rvert}{T}\right), \frac{T}{4} \leqq \lvert t \rvert \leqq \frac{T}{2} \end{cases}$	0.750	2.000	4.000	8.000
Triangular $g\left(t + \frac{T}{2}\right) = 1 - \frac{2\lvert t \rvert}{T}$ $0 < \lvert t \rvert \leqq \frac{T}{2}$	0.500	2.000	3.283	6.000
Cosinusoidal $g\left(t + \frac{T}{2}\right) = \cos \frac{\pi t}{T}$, $0 < \lvert t \rvert \leqq \frac{T}{2}$	0.652	2.000	3.744	6.246
Raised-Cosinusoidal $g(t) = \frac{1}{2}\left(1 - \cos \frac{2\pi t}{T}\right)$, $0 < t \leqq T$	0.500	2.000	2.958	4.904

Table IV — Values of out-of-band power ratio Δ^2 for quaternary PSK with different baseband modulation pulses. Expressions for $g(t)$ are given in Table III

Pulse $g(t)$	Normalized Power Contained in $\mathbf{P}_{v_c}(f)$	Total Out-of-Band Power Ratio Δ^2		
		0.1	0.01	0.001
		$2WT$	$2WT$	$2WT$
Rectangular	1.000	1.807	19.295	
Trapezoidal	0.769	2.000	5.389	8,672
Triangular	0.538	2.000	3.651	6.274
Cosinusoidal	0.682	2.000	3.839	6.270
Raised-Cosinusoidal	0.526	2.000	4.000	5.491

where $X(f)$ is the Fourier transform of $x(t)$ and

$$x(t) = \begin{cases} \frac{1}{\sqrt{T}} \sin \left\{ 2\pi f_d \int_0^t g(\mu) d\mu \right\}, & 0 < t \leqq T, \\ \frac{1}{\sqrt{T}} \cos \left\{ 2\pi f_d \int_0^{t-T} g(\mu) d\mu \right\}, & T \leqq t \leqq 2T, \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

Since $x(t)$ may be nonzero only over an interval $(0, 2T)$ it follows that the minimum out-of-band power ratio Δ^2 of a binary FM-PSK is lower-bounded by*

$$\Delta^2 \geqq \Lambda^2_{\min}(WT_{eq}), \qquad T_{eq} = 2T, \tag{51}$$

where $\Lambda^2_{\min}$ is defined by (45). For $2WT \gg 1$, one can show[19] that

$$\Delta^2 \geqq \Lambda^2_{\min}(2WT) \sim 4\pi\sqrt{2WT}\left(1 - \frac{3}{64\pi WT}\right) \exp(-4\pi WT). \tag{52}$$

Note that $x(t)$ is not completely arbitrary over the interval $(0, 2T)$. From (50) one can show that if

$$x(t) = x_T(t), \qquad 0 < t \leqq T, \tag{53}$$

then

$$x(t) = \sqrt{\frac{1}{T} - x_T^2(t - T)}, \qquad T \leqq t \leqq 2T. \tag{54}$$

Equations (25) and (50) also yield

$$x(0) = 0 \tag{55}$$

and

$$x(T) = \frac{1}{\sqrt{T}}. \tag{56}$$

When $x(t) \in \mathfrak{L}^2_\infty$ is completely arbitrary, the lower bound in (51) is attained when

$$x(t) = k\Psi_0(t - T, d), \qquad d = 2\pi TW, \qquad T' = 2T, \tag{57}$$

where $\Psi_0(t, d)$ is defined in Section VI. Any function other than (57) has a larger out-of-band power ratio. Since $x(t)$ in (57) does not satisfy (53) to (56), it follows that the bound in (51) is strictly a lower bound and is not attainable.†

---

* Note that there is no discrete power contained in an FM-PSK signal.

† The derivation of an attainable lower bound is extremely complicated and will not be attempted here. Also, Table V shows that rectangular signaling gives a bandwidth occupancy which is very close to the lower bound when $\Delta^2 \approx 0.01$, the region of interest.

Table V — Values of the lower bound on $2WT$ and of bandwidth occupancy for binary FM-PSK with rectangular and raised-cosine signaling

Pulse $g(t)$ $g(t) = 0, t \leqq 0, t > T$	Rectangular $g(t) = 1, 0 < t \leqq T$	Raised-Cosinusoidal $g(t) = \frac{1}{2}\left(1 - \cos\frac{2\pi t}{T}\right), 0 < t \leqq T$	Lower Bound on $2WT$
(Peak-to-Peak Frequency Deviation) $\times T$	0.5	1.0	
Out-of-Band Power Ratio Δ^2	(Bandwidth Occupancy $2W$) $\times T$		
0.1	0.773	0.930	0.675
0.01	1.170	2.200	1.117
0.001	2.578	2.874	1.517

The values of the lower bound on $2WT$ and of band occupancy of binary FM-PSK for $\Delta^2 = 0.1$, 0.01, and 0.001 with rectangular and raised-cosine signaling are listed in Table V. The lower bound on Δ^2 given by (51) is also plotted in Fig. 8. Note that the lower bound is very close to Δ^2 with rectangular signaling for $1 \leqq \Delta^2 \leqq 0.01$.

Note that the bandwidth occupancy of binary FM-PSK with rectangular signaling is smaller than that with raised-cosine signaling if $\Delta^2 \geqq 0.001$.* Note also that the peak-to-peak frequency deviation with raised-cosine signaling is larger than that with rectangular signaling. The phase deviation in one signaling interval is always $\pm\pi/2$.

VIII. CONCLUSIONS

For binary and quaternary PSK systems, the band occupancy results presented here can be combined with the results given in Ref. 7 so that channel bandwidth and channel spacing can be chosen to produce minimum distortion transmission and to satisfy any specified power occupancy criterion. The band occupancy of PSK with overlapping baseband pulses is known to be narrower,[10] but we have not considered such signals in this paper.

The 99-percent power occupancy bandwidth of an FM-PSK signal with rectangular signaling is shown to be only 4.7 percent higher than the lower bound. The channel spacing requirements of FM-PSK, from

* The tails of the FM-PSK spectra with raised-cosine signaling go as $\sim 1/f^8$, with rectangular signaling as $\sim 1/f^4$. Hence, the bandwidth occupancy with raised-cosine signaling becomes smaller than that with rectangular signaling for small enough $\Delta^2 (\Delta^2 < 7.5 \times 10^{-4})$.

the point of view of distortion produced by adjacent channel interference, will be treated in subsequent work.

An attempt is also being made to derive a lower bound on the band occupancy if the total power in the continuous part of a BPSK signal is a specified fraction of the total RF power.

IX. ACKNOWLEDGMENTS

Discussions with Larry J. Greenstein, John J. Kenny, and Harrison E. Rowe are gratefully acknowledged.

REFERENCES

1. L. C. Tillotson, C. L. Ruthroff, and V. K. Prabhu, "Efficient Use of the Radio Spectrum and Bandwidth Expansion," Proc. IEEE, *61* (April 1973), pp. 445–452.
2. C. K. H. Tsao and E. M. Perdue, "RF Wideband Data Terminal," Raytheon Company Technical Report, Wayland, Mass., February 1974.
3. H. E. Rowe, *Signals and Noise in Communication Systems*, New York: Van Nostrand, 1965, pp. 57–202.
4. L. Lundquist, "Channel Spacing and Necessary Bandwidth in FDM-FM Systems," B.S.T.J., *50*, No. 3 (March 1971), pp. 869–880.
5. A. Anuff and M. L. Liou, "A Note on Necessary Bandwidth in FM Systems," Proc. IEEE, *59*, No. 10 (October 1971), pp. 1522–1523.
6. V. M. Ray, *Interpreting FCC Broadcast Rules and Regulations*, Blue Ridge Summit, Pa.: TAB Books, 1966.
7. V. K. Prabhu, "Bandwidth Occupancy in PSK Systems," IEEE Trans. Comm., *COM-25* (April 1976), pp. 456–462.
8. D. Slepian and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—I," B.S.T.J., *40*, No. 1 (January 1961), pp. 43–63.
9. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—II," B.S.T.J., *40*, No. 1 (January 1961), pp. 65–84.
10. V. K. Prabhu and H. E. Rowe, "Spectra of Digital Phase Modulation by Matrix Methods," B.S.T.J., *53*, No. 5 (May–June 1974), pp. 899–935.
11. T. T. Tjhung, "Band Occupancy of Digital FM Signals," IEEE Trans. Comm., *COM-12* (December 1964), p. 211.
12. R. deBuda, "Coherent Demodulation of Frequency-Shift Keying with Low Deviation Ratio," IEEE Trans. Comm., *COM-20* (June 1972), pp. 429–435.
13. H. E. Rowe and V. K. Prabhu, "Power Spectrum of a Digital FM Signal," B.S.T.J., *54*, No. 6 (July–August 1975), pp. 1095–1125.
14. W. A. Sullivan, "High-Capacity Microwave System for Digital Data Transmission," IEEE Trans. Comm., *COM-20* (June 1972), pp. 466–470.
15. D. M. Brady, "FM-CPSK: Narrowband Digital FM with Coherent Phase Detection," Proc. Int. Conf. on Communications, Philadelphia, Pa., June 1972, pp. 44.12–44.16.
16. D. M. Brady, "Spectra for FM-DCPSK Modulation," unpublished work.
17. W. Hilberg and P. G. Rothe, "The General Uncertainty Relations for Real Signals in Communication Theory," Information and Control, *18* (1971), pp. 103–125.
18. C. Flammer, *Spheroidal Wave Functions*, Stanford, Calif.: Stanford University Press, 1957.
19. D. Slepian and Mrs. E. Sonnenblick, "Eigenvalues Associated with Prolate Spheroidal Wave Functions of Zero Order," B.S.T.J., *44*, No. 8 (October 1965), pp. 1745–1759.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1976
Printed in U.S.A.

A *Touch-Tone*® Receiver-Generator With Digital Channel Filters

By B. GOPINATH and R. P. KURSHAN

(Manuscript received December 5, 1975)

A Touch-Tone® *receiver with cyclotomic digital channel filters introduced in a companion paper is presented in this paper. A comparison with standard digital channel filters reveals that the number of additions per second needed to implement the channel filters is significantly reduced using cyclotomic filters. The performance of cyclotomic filters as a function of their period is presented in graphic form. The results presented here simulating the filter with random inputs indicates that the filters can effectively reject non-*Touch-Tone *signals. Sensitivity of some important criteria as a function of the accuracy of the clock used to control the digital filters is summarized. The results show that the filters are not particularly sensitive to nonaccurate clocks.*

I. INTRODUCTION

In Ref. 1 we describe a family of filters with several advantages over existing filters, which can be used to generate and detect single tones. Here, we describe how such filters can be used in the construction of a *Touch-Tone*® receiver.

The standard *Touch-Tone* receiver is described in Ref. 2; many other receivers have been proposed in the literature; one which is completely digital is described in Refs. 3 and 4, and an analog receiver with a digitally controlled center frequency is described in Ref. 5. The basic *Touch-Tone* telephone must generate tones to identify the ten basic possible inputs (1, 2, $\cdots$, 9, 0) or, in the case of augmented telephones, 12 to 16 possible inputs (including, for example, * and #). This is done by arranging the input buttons in a grid of four rows and three or four columns. Associated with each row is one of four "low" frequencies (697, 770, 852, or 941 Hz), and associated with each column is one of three or four high frequencies (1209, 1336, 1477, or 1633 Hz). When a button is pushed, one low and one high frequency are simultaneously generated, corresponding to the row and column in which the button is situated. In the central office, a detector decodes the incoming pair of tones to determine which button was pushed.

An incoming signal first passes through a series of tuned filters that filter out dial tones, ring tones, busy tones, and power harmonics (which have amplitude too large to be accommodated by the subsequent channel filters). Next, the signal passes through two parallel bandpass filters (BPF) (see Fig. 1), one to reject the four high-frequency tones (low BPF) and one to reject the four low-frequency tones (high BPF). The output of each BPF passes through a limiter, and the limited signal passes through four parallel channel filters. Each channel filter is connected to a threshold detector which, in 40 ms, makes a determination of whether the tone was present or absent.

In analog receivers, the most critical section consists of the channel filters. Hence, these have to be made with precision components to meet the specifications for station sets. Use of a completely digital receiver requires analog-to-digital (A-to-D) conversion, and special care has to be taken to avoid problems caused by roundoff errors in the BPFs. Furthermore, use of the receiver to generate *Touch-Tone* signals leads to unwanted limit cycles, impairing performance (see Ref. 6).

We propose here a hybrid receiver based on the cyclotomic filters presented in Ref. 1. In the hybrid receiver, the filters that attenuate the dial tone, etc., are the standard analog filters which, using RC active circuitry, can be integrated.[6] The digital part of the receiver follows the limiting circuits (see Fig. 1), which in this case are hard-clippers, thus eliminating the need for separate A-to-D conversion, and at the same time replacing a significant portion of the receiver by digital circuitry. The analog part need not be made with precision components, since variation in the gains of the bandpass filters does not affect the output of the hard limiter significantly. Only the sign of the outputs of the BPFs are used in the digital part of the receiver. The digital filters in the receiver are all operated with perfect arithmetic. All channels have identical filters operating on samples of the output of the hard limiters. However, for each channel, the sampling frequency is proportional to the channel frequency.

Some important features of the system can be summarized as follows:

(*i*) Compared to the channel filters in the all-digital receiver presented in Ref. 3, the number of additions needed to detect tones is relatively small. Hence, fewer adders are needed.

(*ii*) All digital channel filters are mechanized with perfect arithmetic, thus avoiding problems of roundoff.

(*iii*) Since we use perfect arithmetic, we can also generate *Touch-Tone* receiver frequencies using the same channel filters in the receiver.

(*iv*) Without using any A-to-D conversion, we use digital channel filters with analog BPFs.

(*v*) By resetting the filters periodically, we lessen the chance that noise during inter-digit silences, or residual ring tone signals before the first digit, will affect performance.

(*vi*) Since the channel filters have infinite Q, it is possible to increase the signaling rate.

(*vii*) Although the filters have infinite Qs, the peak-to-threshold rejection is kept below 3 dB, thus still preserving the guard action of the hard limiters.

We assume that the reader is familiar with Refs. 1 and 2. Section II gives a description of the hybrid receiver. Section III deals with the performance of the channel filters. Some remarks concerning the factors that enter into choosing the period of cyclotomic filters and interval of operation are contained in Section IV.

II. DESCRIPTION OF THE HYBRID RECEIVER

Figure 1 is a block diagram of the general receiver. The structure of the hybrid receiver is very similar to the standard receiver which is described in Ref. 2. The analog part of the receiver includes both the BPFs and the filters which attenuate power harmonics, ring tones, etc. The outputs of each BPF go into hard-limiters, which convert the analog output of the BPFs into a signal which is either $+1$ or -1, depending

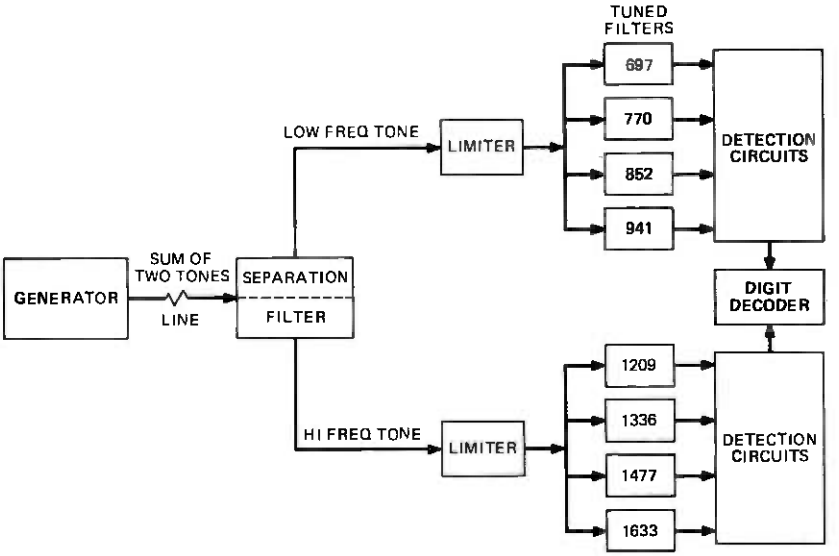


Fig. 1—General receiver.

on whether the analog signal is nonnegative or negative, respectively. This entire analog part can be integrated using active RC circuitry (see Ref. 4). The channel filters which follow the hard-limiters (see Fig. 1) are identical cyclotomic filters (see Ref. 1 and Fig. 2). The cyclotomic filter for each channel has as its input the output of the hard-limiter sampled at a rate p times the channel frequency, where p is the period of the cyclotomic filter used. This requires clock pulses of different frequencies for the different channels.

The channel filters are run periodically for an interval of time inversely proportional to the channel frequency, called the interval of operation. At the beginning of each such interval, the filters are set to zero. The magnitude of the output of each of the filters is compared with a fixed threshold; when the magnitude exceeds this level, a tone corresponding to this frequency is assumed to be present (during the entire interval of operation). The length of the interval of operation is dependent on the permissible error. An interval of operation corresponding to seven cycles of the channel frequency was found to be sufficient (see Section 3.2). This corresponds to 10 ms for the channel corresponding to the lowest *Touch-Tone* frequency, 697 Hz. Hence, if the 697-Hz channel tone is present for the required 40 ms (Ref. 2, p. 11), then in at least three consecutive intervals the tone will produce a signal above the threshold. For higher frequencies, the interval of

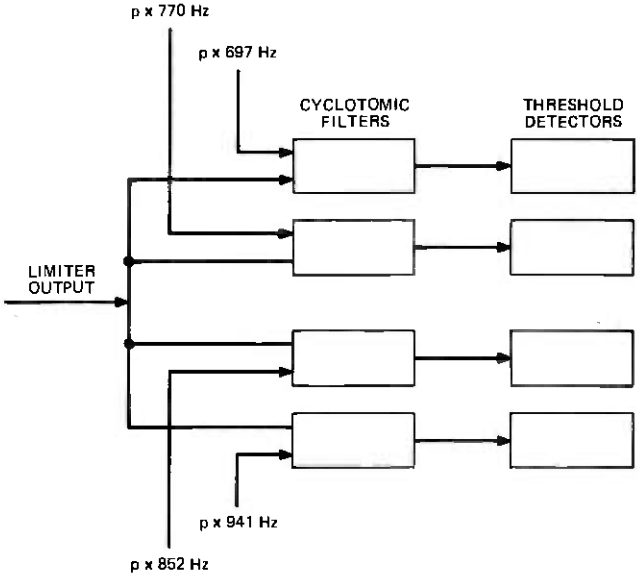


Fig. 2—Channel filters of the low group.

operation is shorter. By synchronizing the intervals of operation of all channels, testing is made for the simultaneous presence of a high tone and a low tone. When a high tone and a low tone are each present for three consecutive intervals, a valid *Touch-Tone* signal is assumed to be present. The digit corresponding to a pair of tones is decoded in the standard way, as described in Section I. Modification of the elementary decision process could be made to increase the signal rate, since the interchannel rejection achieved in a single operating interval is sufficient (see below).

We will not be concerned here with details of hardware in the mechanization of the receiver, but will describe some ways in which the computations in the channel filters can be performed in a multiplexed system.

Two basic modes of implementation will be discussed. One involves individual channel filters dedicated to a fixed frequency. These could be multiplexed to receive inputs from many sources (Fig. 3). This may be more useful in central office applications, where a substantial number of *Touch-Tone* receivers have to be operating at the same time. In this case, the channels controlled by the same clock can be multiplexed in the usual way using serial arithmetic as described in Ref. 1. A system of 20 receivers would require eight clocks (or clock pulses derived from a simple high-frequency clock). For a system using, for example, six times the channel frequency as sampling rate, one adder per channel seems adequate. From Table II, Ref. 1, computations show that the cyclotomic polynomial of period 6, F_6, needs 84 adds per period. The channel corresponding to the highest frequency, 1633 Hz, will need ($1633 \times 84 \times 20$) adds/s. This implies that an add must not take more than about 0.36 μs. So, with 0.36-μs adders, eight adders would be needed for the whole system. This is, of course, excluding the logic involved in the decision process. If in the system we allow for buffers in the higher frequency channels, then a slower adder could be used, since we wait 10 ms before a decision is made. In this case, the speed of the adder is determined by the channel corresponding to the

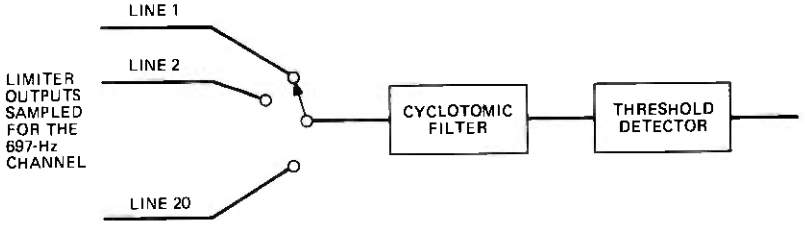


Fig. 3—System amenable to serial multiplexing.

lowest frequency. The lowest frequency channel requires (697 × 84 × 20) adds/s, corresponding to an add in ~ 0.85 μs.

Another system involves buffering the input in such a way that a single filter can be used for two or more channels (Fig. 4). This might prove useful when an adder is mutliplexed between channels corresponding to the same receiver. In this case, buffers for each channel store the output from the limiter in segments corresponding to the seven-cycle interval of operation. For the filter based upon F_6, this would be 42 bits long. Since the buffer corresponding to a higher frequency would fill up faster than one corresponding to a lower frequency, the channel corresponding to the highest frequency, i.e., 1633 Hz, is fed into the filter first, say, after 5 ms (the buffer of this channel fills up in less than 5 ms). After completing the operation on all the 42 bits of input of this channel, the filter is multiplexed to operate on the next highest frequency channel, and so on. This requires that the adder be fast enough to do 7 × 84 adds in less than $\frac{5}{8}$ ms, i.e., 940,800 adds/s so a 1-μs adder would suffice. Since this adder is idle for every 5 ms of the 10-ms cycle, it can be used for another receiver. Hence, a 1-μs adder could do all the additions for the channel filters of two receivers. Modification of this elementary decision process could be made depending on the statistics of noise in the channel and sensitivity of the limiter. When a high and a low tone

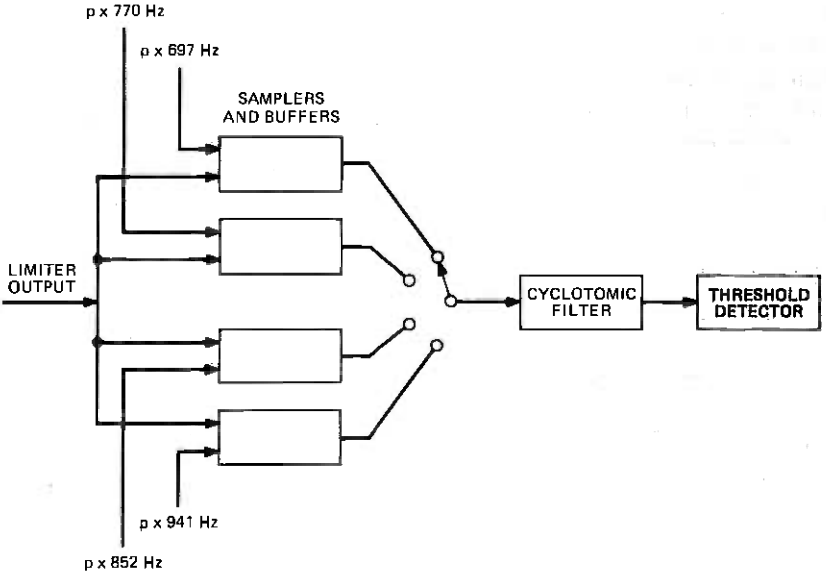


Fig. 4—Multiplexing using buffers.

have been simultaneously detected for three consecutive 10-ms periods, then a decision is made that a *Touch-Tone* signal has been received and the digit is identified in the usual way. The regular second-order filter used in the all-digital receiver of Ref. 3 requires a minimum of 2400 adds/ms and a total of 96,000 adds and achieves an interchannel rejection of ~ 7 dB. Using a cyclotomic filter of period 6 (based on F_6) would require 840 adds to give the same interchannel rejection. This corresponds to a rate up to 60 adds/ms for the 697-Hz channel. If the period were raised to 30 and no use of read-only memory were made, it would still only require a maximum of 56,700 adds to achieve the same rejection; this corresponds to approximately 4010 adds/ms. Of course, intermediate periods would give intermediate statistics, which can be readily computed for systems based on F_p ($p = 8, 9, 12, 15, 16, 18, 24$; see Ref. 1).

III. PERFORMANCE OF THE CHANNEL FILTERS

To discuss the performance of the channel filters, we need to define certain terms. Let f_i, $i = 1, 2, \cdots, 8$ be the eight channel frequencies. As described earlier, each channel filter is a cyclotomic filter of some period p, based on the cyclotomic polynomial F_p. The order of the filter is denoted by k (the degree of F_p). The fundamental resonance frequency of each filter is determined by τ_i, the sampling interval in seconds of the output of the hard-limiter. In order that the fundamental resonance of the filter be at frequency f_i, τ_i should satisfy

$$p\tau_i = \frac{1}{f_i}.$$

From Ref. 1 we see that the operation of any channel filter can be modeled by

$$x_n = \sum_{i=1}^{k} a_i x_{n-i} + u_n$$

$$y_n = \sum_{i=1}^{k} c_i x_{n-i} \qquad n = 0, 1, \cdots, N$$

$$x_j = 0 \qquad \text{for} \quad j < 0,$$

where x_{n-i}, $i = 1, \cdots, k$ are the numbers stored in the shift register implementing the particular channel filter, y_n the output of the filter, and u_n the sampled output of the hard-limiter, which is, of course, the input to the filter. Hence, if the output of the BPF is a sinusoid of frequency f,

$$\begin{aligned} u_n &= 1 \qquad \text{if} \quad \sin 2\pi n f\tau \geqq 0 \\ &= -1 \qquad \text{if} \quad \sin 2\pi n f\tau < 0, \end{aligned}$$

where τ is the sampling interval associated with the channel. So that we may use the same threshold for all channels, we normalize the interval of operation by the fundamental resonant frequency. Hence, if each filter is operated for N steps, this corresponds to operating the filter for $N\tau_i$ s. N/p describes the same interval in units corresponding to a period of the fundamental resonant frequency, hence, an interval of operation of seven periods of the fundamental, i.e., $7 \cdot 1/f_i$ s. We will compare performance of cyclotomic filters of different periods operating for the same number of periods of the fundamental.

Let $M(f)$ denote the maximum absolute value of y_n in the interval of operation when the input square wave is of frequency f. Detection of the fundamental frequency is based on $M(f)$ exceeding a preassigned threshold. A plot of $M(f)$ vs frequency for various cylotomic filters when operated for seven periods of the fundamental is given in Ref. 1. The curves serve to indicate how well the filter performs in distinguishing between tones. The model of a typical curve is shown in Fig. 5. Following standard terminology, we use the term power gain or gain at f to mean $20 \log_{10} M(f)$. Difference between power gains at two frequencies is related in the obvious way to the ratio of $M^2(f)$ at these two frequencies. By scaling the frequency axis linearly, the fundamental resonant frequency can be shifted arbitrarily. The specifica-

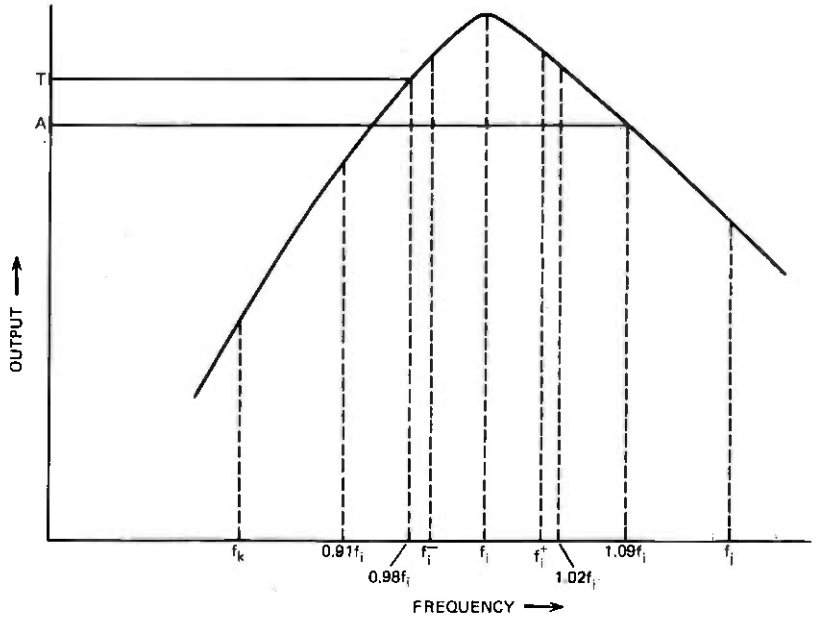


Fig. 5—Specifications for a typical channel.

tions (see Ref. 2, p. 11) for the *Touch-Tone* receiver require that any tone of frequency f lying in the interval I_i defined by $f_i^- \equiv 0.987 f_i - 4 \leqq f \leqq 1.013 f_i + 4 \equiv f_i^+$ be accepted as a tone corresponding to frequency f_i. This band of frequencies is referred to as the accept band of channel i. The threshold T_i has to be set such that $M(f) > T_i$ for all f in the accept band. Therefore, $T_i \leqq \text{Min}_{f_i^- \leqq f \leqq f_i^+} M(f)$. We call $20 \log_{10} T_i$ the "maximum threshold" for channel i. On the other hand, T_i has to be greater than $M(f)$ for $f \in I_j$, $j \neq i$. We call $A_i \equiv [\text{Max}_{j \neq i} M(f_j)]$ the "maximum gain at reject channels." If the gain at any other channel j exceeds T_i, then a tone corresponding to channel j could be mistaken as one corresponding to i. The threshold with 3-dB rejection is merely $20 \log_{10} A_i + 3$. Use of this threshold assures that if the input to channel i is a signal corresponding to some other channel, then the signal level in the filter is at least 3 dB below threshold. Finally, the "rejection at edge" is the measure of the maximum drop in signal level at the edge frequencies f_i^- and f_i^+ from the center frequency f_i.

Evidently, these parameters are different for different channels. However, by setting certain standards for a typical threshold and maximum reject channel gain, a worst-case standard set for the whole receiver can be found to compare the performance of cyclotomic filters of different periods. It is easily seen that I_i is contained in the interval $[0.98 f_i, 1.02 f_i]$; on the other hand, this interval is not significantly bigger than I_j for any j. For each channel frequency f_i, every $f_j j \neq i$ lies outside the interval $[0.91 f_i, 1.09 f_i]$. The rejection of every alien channel is greater than the rejection of frequencies at ends of this interval because of the bell-shaped nature of the curve in the intervals of interest.

Now that the ends of the intervals of interest have been scaled with respect to the resonant frequency, we can define

$$T = \text{Min}\,[M(0.98 f_i), M(1.02 f_i)]$$
$$A = \text{Max}\,[M(0.91 f_i), M(1.09 f_i)].$$

Then $20 \log_{10} T$ and $20 \log_{10} A$ serve as standards for threshold and maximum reject channel gain for all channels. Figure 6* is a plot of $M(f_i)$, T, and A for cyclotomic filters of periods 6 through 30, run for seven periods of the fundamental resonant frequency. Although the T and A as a percentage of $M(f_i)$ do not change appreciably as the period of the filter increases, the effect of increasing the period of the cyclotomic filter is not equivalent to scaling the input to the filter.

* In Fig. 6, ○, +, and □ correspond to $M(f_i)$, T, and A adjusted for phase shift of input signal as described in Section 3.2.

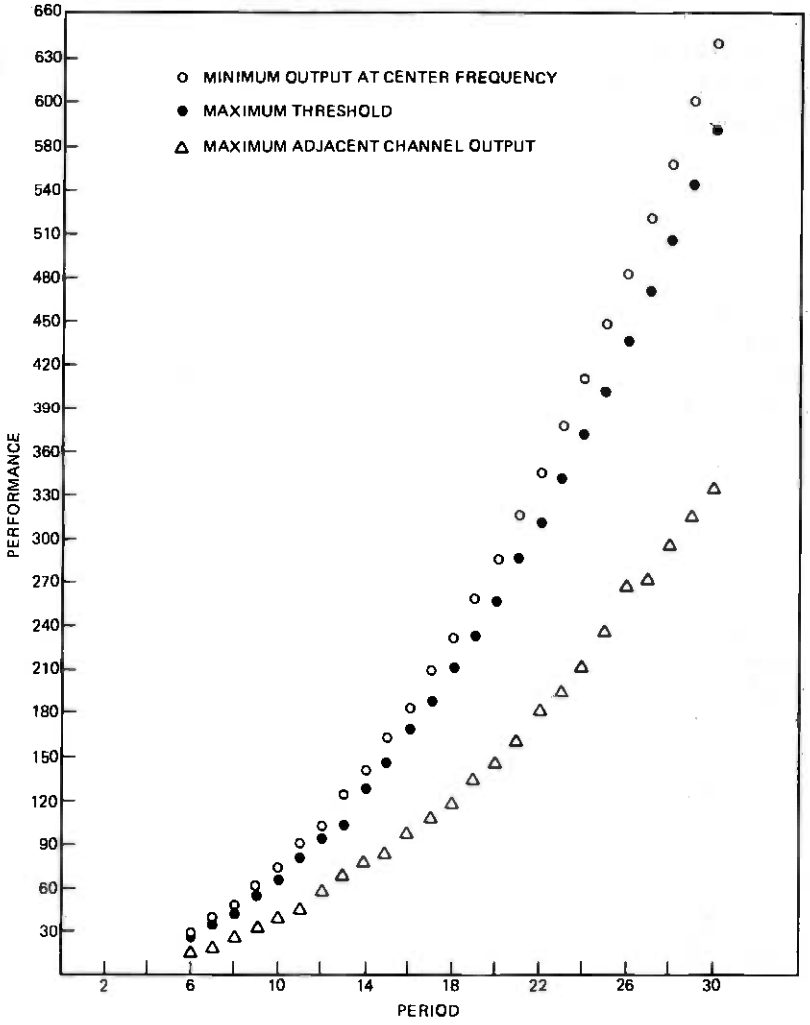


Fig. 6—Performance vs period.

This is because filters of larger periods assume a larger number of distinct levels. Furthermore, increasing the period of the filter may be a way of reducing the effect of noise at the limiter as described in Ref. 1. Although rejection in decibels is a conventional method of describing performance of the tuned filter, the actual level of the signal may be more pertinent to digital applications; hence, the plot is a linear scale. We can now discuss some specific aspects of the performance of these filters.

3.1 Higher-order resonances of filters

Since the channel filters are discrete-time filters, spurious resonances could affect performance, especially since the inputs are hard-clipped, and hence have all odd harmonics (see Ref. 1, Section III). The higher harmonics introduced resulting from clipping could interfere with the fundamental. However, for a cyclotomic filter with transversal weighting function (see Ref. 1) of period p, the spurious resonance closest to the fundamental resonant frequency is $(p - 1)$ times the fundamental. Hence, for example, for $p = 6$ (the lowest period considered here) the closest spurious resonance is five times the fundamental. Therefore, for the channel corresponding to the lowest *Touch-Tone* frequency (697 Hz), the first spurious resonance occurs at 3485 Hz, well outside the *Touch-Tone* band. The higher the period of the cylotomic filter, the further away this resonance will move.

3.2 Interchannel rejection

It was observed above that the ratio $T_i/A_j i \neq j$ was greater than T/A for all channels. Hence, the minimum interchannel rejection is greater than $20 \log_{10} (T/A)$. We will use $20 \log_{10} (T/A)$ as a measure of interchannel rejection. The interchannel rejection for all filters of periods between 6 and 30 varies between 4.2 and 4.9 dB. This is predicated on the assumption that the tone was synchronized with the switching on of the filter. This, of course, need not be the case in practice. Hence, this figure was adjusted for the worst-case phase difference between switching on of the receiver and zero of the time signal. Calculations showed that in all cases the rejection was not lowered by more than 0.5 dB for all filters. The values shown in Fig. 5 are corrected for worst-case phase difference. By increasing the interval of operation to 10 periods of the fundamental, the minimum interchannel rejection for all channels can be increased to about 7 dB. If the interval of operation is of the form $(m + \frac{1}{2})$ periods of the fundamental for any integer m, no correction for phase shift seems to be necessary.

3.3 Sensitivity to clock rate

Some important parameters of the filters corresponding to each channel as a function of percentage variation in sampling rate was calculated. The results when cyclotomic filters of period 6 are used for seven cycles of channel frequency show that with a threshold set at 28 dB above the unit signal level of the hard-clipper, a ± 2-percent change in sampling rate can be tolerated. Hence, even though we have to use eight different clock pulses, these clock pulses do not have to be controlled especially accurately. For cyclotomic filters of period 30,

the largest period considered in Ref. 1, similar observations can be made based on computational results. It can be deduced then that the performance of the channel filters are not especially sensitive to clock rate. This allows for the use of cheaper clocks, when each channel is clocked separately.

IV. REJECTION OF PSEUDO *TOUCH-TONE* SIGNALS

Whenever the input to the hard-limiter is a sinusoid, $M(f)$ gives an indication of the signal level in the filter. However, when no *Touch-Tone* signal is present, the output of the BPFs are not sinusoids. Owing to the nonlinear nature of hard-limiting, the curve on Fig. 5 does not lend significant insight into the signal level for complex signals. To simulate a family of non-*Touch-Tone* receiver inputs to the filter, we modeled the output of the hard-limiters as a two-state symmetric Markov chain such that the average number of changes of sign in the interval of operation was equal to the number of changes of sign of a tone corresponding to the channel frequency. Then a simulation of the filter operating on such inputs was made. The noise level was about 12 dB below the level in the accept band for all cyclotomic filters of periods 6 through 30.

V. SOME REMARKS ON THE CHOICE OF INTERVAL OF OPERATION AND PERIOD OF CYCLOTOMIC FILTERS USED

As mentioned earlier, an interval of operation corresponding to seven periods is sufficient to provide adequate interchannel rejection. Hence, for signaling it is possible that a 20-ms on-time requirement for tones might be sufficient. In this case, one can eliminate the need for bandpass filters by altering the signaling process somewhat. Instead of transmitting two tones simultaneously for 40 ms, the tones can be sent one after the other, each being 20 ms at present. However, it would be necessary to determine whether this scheme can provide adequate speech immunity. This would reduce the number of channel filters to four, since only one frequency from the two groups of frequencies is present at a time. Because of simplifications effected in the receiver, this method of signaling might prove more useful for transmitting information using *Touch-Tone* signaling.

As for the period of the cyclotomic filter used, it is clear from Table II, Ref. 1, that the number of adds/s increases as the period increases. However, depending on the signal-to-noise ratio at the input to the hard limiter, the use of a period high enough to make the frequency of errors in detection small might be necessary.

VI. ACKNOWLEDGMENTS

The authors wish to acknowledge many useful conversations with H. Breece, J. Condon, and D. Haglebarger.

REFERENCES

1. B. Gopinath and R. P. Kurshan, "Digital Single-Tone Generator-Detectors," B.S.T.J., this issue, pp. 469–496.
2. R. N. Battista, C. G. Morrison, and D. H. Nash, "Signalling System and Receiver for TONE-TONE® Calling," IEEE Trans. Comm. and Elect., *82* (March 1963), pp. 9–16.
3. L. B. Jackson and R. T. Piotrowski, "A Preliminary Study of a Digital *Touch-Tone*® Receiver," unpublished document.
4. L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-16* (September 1968), pp. 413–421.
5. M. Awipi and D. S. Levinstone, "*Touch-Tone*® Channel Filters Using Bucket Brigade Delay Lines," unpublished document.
6. J. J. Friend, C. A. Harris, and D. Hilberman, "STAR: An Active Biquadratic Filter Section," unpublished document.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1976
Printed in U.S.A.

Digital Single-Tone Generator-Detectors

By R. P. KURSHAN and B. GOPINATH

(Manuscript received December 5, 1975)

A class of digital, linear generator-detectors, based upon cyclotomic polynomials, which have simple implementation and operate without roundoff errors, is proposed. It is shown how these filters are optimal among all linear generator-detectors which have no roundoff required in the feedback loop. The complexity of various cyclotomic filters are compared. These filters in general require far fewer binary adds/s than conventional second-order filters used for the same purpose.

I. INTRODUCTION

Devices for pure tone generation and detection have widespread applications. The most notable examples are *Touch-Tone*® signaling, frequency shift keying (FSK), and multifrequency (MF) signaling. Associated with such devices are problems of stability and predictability, which in practice are dealt with on an individual basis, using techniques peculiar to the particular application. When these devices are realized digitally, the above problems are manifest from errors due to operational roundoff.

Generally, tones for signaling are analog signals of the form $A \sin \omega t$ (A is the amplitude, $2\pi/\omega$ is the period, and $\omega/2\pi$ is the frequency). Devices that generate these tones are usually oscillators of various kinds. Because of the requirement of structural stability, in practice these devices are limit cycle oscillators. These are simulations and realizations in hardware of nonlinear differential equations that have limit cycles. Because of the complexity of these equations, the amplitude and frequency are not easily predicted from given values of resistors and capacitors in the network.

For detection of these tones, linear analog filters are frequently used. These are also used as generators, when the duration of the signal is not too long compared to the period. However, passive linear analog oscillators require inductors which are bulky, and the frequency and amplitude of these oscillators can vary with changes in value of the inductors and capacitors due to environmental conditions. Active

linear oscillators using RC elements are used in many applications. However, they also generally need some form of limiting and end up being nonlinear devices, thus usually preventing them from being used as receivers.

Digital oscillators, on the other hand, are almost insensitive to changing parameter values and produce stable repeatable waveforms. However, in the mechanization of these oscillators (which are usually based upon second-order linear equations), roundoff in multiplication and addition produce errors in the feedback that lead to limit cycles and can significantly impair the signal quality. Also, when such linear digital devices are used as receivers, the precision required for satisfactory performance goes up quite rapidly with increasing Q. Although the effects of this can be satisfactorily controlled in certain specific applications (see, for example, Ref. 1), the difficulties, in general, cannot be ameliorated except by increasing the accuracy of computations.[2]

In this paper, we present a class of digital filters that operate without arithmetic roundoff. These filters are linear, and can be used both as oscillators for signal generation and also as receivers for signal detection. The feedback loop of each filter is constructed in such a way as to eliminate the possibility of roundoff or truncation errors, thus insuring perfect arithmetic. This entirely eliminates the problem of limit cycles. The filters presented, when used as generators, produce quantized values of $A \sin \omega t$ of arbitrary accuracy. Implementation of these filters as receivers involves first sampling an analog input signal to produce a digital input into the filter. The filter is designed to resonate for a particular input frequency, thus enabling detection.

The means by which arithmetic errors are eliminated in the feedback loop involves constraining all feedback coefficients to be integers (a constraint which turns out to be necessary to guarantee perfect arithmetic in any digital filter). Thus multiplication by these coefficients can be performed as additions, simplifying implementation.

The behavior of the feedback loop of this filter is modeled by a linear recursion whose characteristic polynomial is a cyclotomic polynomial. In recognition of this, we call the filter consisting of the feedback loop alone a "cyclotomic filter." It will be demonstrated that the only way to ensure perfect arithmetic with no limit on the period of operation (and thus avoid limit cycles) in a filter modeled by a linear recursion (i.e., a linear digital filter) is to constrain the feedback coefficients to be integers. Furthermore, it will be shown that, with this constraint on the feedback coefficients and also subject to minimizing memory and eliminating as many resonant harmonics as possible, the cyclotomic filter is uniquely optimal among all digital linear filters, both for the purpose of tone generation and the purpose of tone detection.

In subsequent sections of this paper, it is demonstrated how a weighting function can be applied externally to the cyclotomic filter to drastically reduce the impact of those higher-order resonances that remain. This is applied also to determine those impulse responses which have a small number of integer levels and lack higher-order harmonics. All the cyclotomic filters of practical significance, along with their associated weighting functions and impulse responses, are examined.

In Ref. 3, a specific proposal is described for the *Touch-Tone* receiver (and tone generator), utilizing eight cyclotomic filters.

II. CYCLOTOMIC FILTERS

The purpose of this filter, as discussed in Section I, is to generate or detect a single pure tone $u(t) = A \sin (2\pi ft + \varphi)$ of frequency f. Digital implementation involves realizing a discrete time filter with k stages of memory (see Fig. 1), which is described recursively in terms of an input sequence u_n as

$$x_n = \sum_{i=1}^{k} a_i x_{n-i} + u_n. \tag{1}$$

The numbers $a_i (i = 1, \cdots, k)$ are the feedback coefficients of the filter. The filter is driven by a clock with the time interval τ between pulses. In tone generation, the filter must satisfy

$$x_n = u(n\tau), \tag{2}$$

at least for some initial conditions $x_0, \cdots, x_{k-1}$. When used as a receiver, the analog input $u(t)$ is sampled, producing a discrete input $u_n = A \sin (2\pi fn\tau + \varphi)$; the filter (1) must distinguish between the desired frequency f_0 and all other frequencies in a band containing f_0.

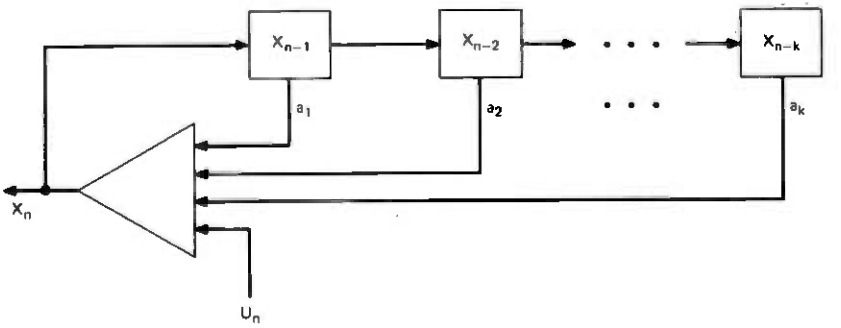


Fig. 1—Recursive filter in k stages of memory.

Specifically, it must satisfy the resonance property

$$\limsup |x_n| = \infty, \tag{3}$$

when $f = f_0$, and in a sufficiently large band B including f_0 there must be no other such resonances. Then $|x_n|$ will be uniformly bounded in B in the complement of any small interval δ about f_0, say, $|x_n| \leqq m(\delta)$ for all $f \in B$, $f \notin \delta$, for all n. A threshold detector can thus detect in a finite amount of time $N\tau$, the presence (or absence) within B of an input frequency f_0 (with error $\pm\frac{1}{2}|\delta|$). It does this by comparing the gain $\sup_{n \leqq N} |x_n|$ with the bound $m(\delta)$; if $\sup |x_n| > m(\delta)$, then $f \in \delta$; otherwise it is not. Of course, the smaller the allowable error δ, the larger N must be.

To know precisely when an input u_n will resonate with respect to this filter, we first observe that the general solution to (1) is

$$x_n = \sum_{j=0}^{n} \sum_{i=1}^{k} b_i \rho_i^{n-j} u_j, \tag{4}$$

where $\rho_1, \cdots, \rho_k$ are the roots (assumed to be distinct) of the characteristic equation

$$\lambda^k - \sum_{i=1}^{k} a_i \lambda^{k-i} = 0 \tag{5}$$

and $b_1, \cdots, b_k$ are complex functions of the roots. [This is derived in (17) below.] If the magnitude of a root of (5) is greater than 1, the filter will be unstable. However, if all roots are inside the unit circle, then (1) will not have any resonance as defined in (3). Hence, in general we will assume that all roots of (5) lie inside or on the unit circle.

Hence, the resonance (3) will occur if and only if the frequency f is such that with $\theta(i) = \arg \rho_i$ either

$$2\pi f\tau \equiv \theta(i) \text{ (modulo } 2\pi) \quad \text{or} \quad 2\pi f\tau \equiv -\theta(i) \text{ (modulo } 2\pi) \tag{6}$$

for some $i = 1, \cdots, k$ with the property that $|\rho_i| = 1$. That is, the detector (see Fig. 2) will give a "yes" response iff (6) is satisfied. As we are trying to detect the presence of the frequency $f = f_0$, let us suppose by way of example that $\theta(1) = 2\pi f_0 \tau$ ($|\rho_1| = 1$). Then an input $A \sin (2\pi f_0 t + \varphi)$ would elicit a "yes" response from our receiver. (Any phase shift of $A \sin 2\pi f_0 t$ will not affect the resonance of this signal, as $A \sin (2\pi f_0 t + \varphi) = (A \cos \varphi) \sin 2\pi f_0 t + (A \sin \varphi) \cos 2\pi f_0 t$, and $\cos \varphi$ and $\sin \varphi$ never simultaneously vanish.) However, let us now suppose that also $\theta(2) = 2\pi f_1 \tau$ ($|\rho_2| = 1$). Then the receiver would also detect an input frequency f_1 (and would not differentiate between f_0 and f_1). Hence, one would know only whether or not *either* f_0 or f_1 is among the inputs. To positively identify the presence of f_0, one

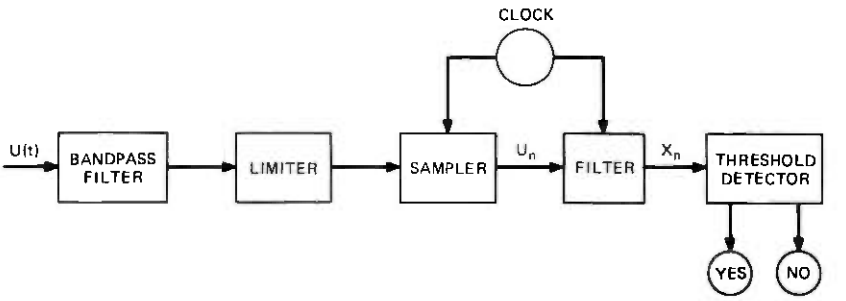


Fig. 2—Structure of a tone detector.

must either insure that f_1 is out of band or use some other means to differentiate between f_0 and f_1.

Similarly, because of (6), the filter cannot distinguish between the frequency f and the frequency $\tau^{-1} - f$, since $2\pi(\tau^{-1} - f)\tau = 2\pi - 2\pi f\tau \equiv -2\pi f\tau$ (modulo 2π). In fact, $2\pi f\tau$ and $-2\pi f\tau$ are the respective arguments of complex conjugates, and thus we see from (6) that no new resonances can occur if the characteristic polynomial (5) is altered to include among its roots any complex conjugates of $\rho_1, \cdots, \rho_k$. We shall use this fact in our determination of a good structure for the recursion (4). When the filter is such that an input of frequency f will resonate, we shall say that the filter resonates (or has a resonance) at f.

Recapitulating, because of (6), whenever the filter has a resonance at a frequency f, it will also necessarily and unavoidably resonate at the frequency $\tau^{-1} - f$. To counter the effect of this in practice, τ must be made sufficiently small so that $\tau^{-1} - f$ is out of band. In keeping with (6), we refer to resonance at the frequency f as "resonance at the root $e^{i2\pi f\tau}$," and resonance at $\tau^{-1} - f$ as "resonance at the conjugate root $e^{-i2\pi f\tau}$" [the roots in question being, of course, roots of (5)].

The remaining resonances described by (6) are those due to aliasing. These also are intrinsic to the system—a consequence of using discrete (rather than continuous) input samples u_n. Indeed, if resonance occurs at a frequency f (or, equivalently, at the root $e^{i2\pi f\tau}$), it will also occur at all the frequencies $f + m\tau^{-1}$ for any integer m, as $2\pi f\tau \equiv 2\pi(f + m\tau^{-1})\tau$ (modulo 2π) or, equivalently,

$$e^{i2\pi f\tau} = \exp[i2\pi(f + m\tau^{-1})\tau].$$

In practice, if conjugate resonances are out of band, resonances due to aliasing will also necessarily be out of band.

Hence, if $\rho_1, \cdots, \rho_m$ are those roots of (5) of modulus 1, the filter will have resonances within the band $[0, \tau^{-1}]$ at the frequencies $\theta(1)/2\pi\tau$, $[2\pi - \theta(1)]/2\pi\tau$, $\theta(2)/2\pi\tau$, $\cdots$, $[2\pi - \theta(m)]/2\pi\tau$. The number of distinct resonances in the interval $[0, (2\tau)^{-1}]$ is m, less the number of roots among $\rho_1, \cdots, \rho_m$ which appear along with their conjugates. This picture is repeated in each successive interval $[n\tau^{-1}, (n+1)\tau^{-1}]$ $(n = \pm 1, \pm 2, \cdots)$.

It should be clear that, in choosing the recursion (1), one desires to have the number of resonances as small as possible—for the purpose of generation, to minimize the number of harmonics that can be produced by perturbations of the initial conditions; and for the purpose of detection, to maximize the band in which the filter can detect a unique signal. Also, of course, one desires to have the memory k (a measure of the complexity of implementation) as small as possible.

Ideally, one would like to have only one resonance, namely at the frequency one is trying to detect or generate. This is possible within the band $[0, \tau^{-1}]$, by using the recursion $x_n = -x_{n-1} + u_n$. However, this resonates at a frequency equal to half the clock frequency τ^{-1} and thus also resonates at the third harmonic $(2\tau)^{-1} + \tau^{-1}$ due to aliasing. As the third harmonic is frequently in band, this recursion is generally not satisfactory.

On the other hand, for some complex number ρ of unit modulus, one could use the recursion $x_n = \rho x_{n-1} + u_n$ which also has memory one. By adjusting τ, one could make the argument of $\rho = \exp(i2\pi f_0\tau)$ small, thus avoiding any resonance up to as high a frequency as desired. However, there are problems with this recursion. First of all, the memory (in implementation) is not really one but two, as the real and imaginary parts of ρ must be handled separately. In fact, as seen before, no new resonances would be introduced by including the complex conjugate $\bar{\rho}$ of ρ to form a recursion of order two. Hence, one does just as well by replacing the characteristic equation $\lambda - \rho = 0$ with $0 = (\lambda - \rho)(\lambda - \bar{\rho}) = \lambda^2 - a\lambda + 1$ (where the real number $a = \rho + \bar{\rho}$). The corresponding recursion replacing $x_n = \rho x_{n-1} + u_n$, also (but now explicitly) of memory two, is $x_n = ax_{n-1} - x_{n-2} + u_n$. This is the recursion after which digital linear filters are customarily modeled. However, as a (ρ) is, in general, not a rational number (gaussian rational*), it must in general be truncated, leading to slight frequency shifts, and multiplication round-off error in the feedback loop of these filters (Fig. 1); this could lead to unwanted limit cycles.[2] To avoid this, a (ρ) is restricted to be rational (gaussian rational). Even for rational numbers, however, truncation error would occur if

* Has rational real and imaginary parts.

the number of bits necessary to represent the number x_n exceeded the word length allowed. In Section V we show that this can be controlled only if a is an integer.

Hence, in the case of the real recursion, we restrict a to be an integer, and the only possibilities are $a = 0, \pm 1, \pm 2$. We have already ruled out $a = -2$ (this gives the square of the characteristic equation of $x_{n+1} = -x_n + u_n$). If $a = 2$, this gives the square of the characteristic equation of $x_{n+1} = x_n + u_n$, which is even worse, as it produces resonance at the second harmonic. The remaining three possibilities for a correspond to cyclotomic polynomials of orders 3, 4, and 6 (as defined subsequently in this section). It will be shown that, by taking a cyclotomic polynomial for the characteristic equation (5), one always obtains the best possible recursion (1) for the given amount of memory.

In general, to have perfect arithmetic (the only means by which to uniformly avoid unwanted limit cycles), it is necessary to constrain the feedback coefficients a_i, $i = 1, \cdots, k$ [see (1)] to be gaussian integers (see Section V). In fact, it will be shown that one can take each $a_i = 0, \pm 1$ so that each tap in the feedback loop involves at most changing the sign. Hence, from here on we restrict ourselves to three cases: the a_i's are gaussian integers, are integers, or are $0, \pm 1$. In what follows, we will show that the three are, for practical purposes, equivalent.

For the first case, in the recursion corresponding to $\lambda - \rho = 0$ (no complex conjugate), the restriction to integer real and imaginary parts requires $\rho = \pm 1, \pm i$ resulting in less generality than possible, as this corresponds to the recursions of the previous example with $a = \pm 2$ only. In fact, we can generalize this, and say it is always better to include among the roots of (5) all the complex conjugates, and thus to have a recursion (1), all of whose coefficients are real (and hence integers). We will make this explicit in a moment, but let us first indicate the reasoning. First of all, by including the conjugates, no new resonances are introduced (as has already been demonstrated). Second, if among the roots of (5) even one conjugate were missing, the coefficients of (1) would not all be real. In this case, the real and imaginary parts of x_n would have to be considered separately, and one would thus need an effective memory of $2k$. On the other hand, if one multiplies (5) by factors of the form $(\lambda - \bar{\rho})$, one for each root ρ of (5) whose complex conjugate is not also a root of (5), then the resulting polynomial and the corresponding recursion will have real coefficients. The respective degree and memory will thus be raised to no more than $2k$ (the effective memory of the complex recursion). Furthermore, as will be shown in Theorem 1 below, the new polynomial (and recursion) obtained from multiplication by the factors $(\lambda - \bar{\rho})$ will also be

guaranteed to have integer coefficients. Thus, we will do at least as well (and, as we have seen above, even better) by restricting all the recursions (1) to have real (and hence integer) coefficients.

Let us now make this explicit. Suppose one has the recursion

$$x_n = \sum_{j=1}^{k} \alpha_j x_{n-j} + u_n,$$

where $\alpha_j (j = 1, \cdots, k)$ are gaussian integers: $\alpha_j = a_j + b_j i$ (a_j and b_j integers, $i = \sqrt{-1}$). Let y_n and z_n be, respectively, the real and imaginary parts of x_n. Then

$$y_n = \sum_{j=1}^{k} (a_j y_{n-j} - b_j z_{n-j}) + u_n,$$

$$z_n = \sum_{j=1}^{k} (b_j y_{n-j} + a_j z_{n-j}).$$

The only feature possibly mitigating in favor of the complex recursion is this: We are constrained to have a_j and b_j be integers. If the new recursion with added roots did not have integer coefficients, then in spite of the other considerations above, one would choose the complex recursion. However, in the following theorem we show this is not possible.

Theorem 1: *Suppose $F(\lambda)$ is a polynomial with gaussian integer coefficients, and suppose $\rho_1, \cdots, \rho_m$ are those roots of $F(\lambda)$ whose complex conjugates are not also roots of F. Then $F(\lambda) \prod_{i=1}^{m} (\lambda - \bar{\rho}_i)$ has integer coefficients. Furthermore, if $F(\lambda)$ has no polynomial with integer coefficients as a factor, then deg $F = m$.*

Proof: Write $F(\lambda) = g(\lambda)h(\lambda)$, where $h(\lambda) = \prod (\lambda - \rho_i)$. Then g has real coefficients. Let $p(\lambda)$ be any irreducible factor of $F(\lambda)$ (considered as a polynomial over the gaussian integers). Suppose p has the root r in common with g and the root s in common with h. Then $\bar{p}$ (the polynomial in λ whose coefficients are the complex conjugates of the coefficients of p) has $\bar{r}$ as a root, and hence $\bar{p}$ must also be a factor of F. But $\bar{s}$ is also a root of $\bar{p}$, whereas $\bar{s}$ is expressly not a root of F. Hence, any irreducible factor of F must be a factor of either g or h. It follows that g has integer coefficients, and h (and thus $\bar{h}$) have gaussian integer coefficients. As $h(\lambda)\bar{h}(\lambda)$ has real, and hence integer, coefficients the theorem follows.

Thus, it is best to take the coefficients of the recursion (1) to be integers. The theorem which follows completely characterizes those recursions.

First, however, a short description of cyclotomic polynomials must be given. The Euler φ-function is a function on the positive integers, defined as follows: $\varphi(m)$ is the number of positive integers less than or equal to m and having no integer factor in common with m, other than 1 (such integers are said to be *relatively prime* to m). For example, $\varphi(1) = \varphi(2) = 1$, $\varphi(3) = \varphi(4) = 2$, $\varphi(9) = 6$. The cyclotomic ("circle-dividing") polynomial of order m, denoted $F_m(\lambda)$, is that monic polynomial (coefficient of the term of highest degree is 1) with integer coefficients all of whose roots are primitive mth roots of unity (that is, $r^m = 1$, and $r^n \neq 1$ for $0 < n < m$). Over the integers, $F_m(\lambda)$ is irreducible (not a nontrivial product of polynomials with integer coefficients).[4] From the definition, one can explicitly determine that $F^m(\lambda) = \prod_d (\lambda - \exp[2\pi i(d/m)])$, where the product is taken over all d, $1 \leq d < m$ such that d and m are relatively prime. Thus, the degree of F_m is $\varphi(m)$.

The next theorem shows that, whatever constraints there are on available memory and acceptable resonant harmonics, the characteristic polynomial of the optimal recursion will be a cyclotomic polynomial.

Theorem 2: Let $F(\lambda) = \lambda^k - \sum_{i=1}^{k} a_i \lambda^{k-i}$, *where* $a_k \neq 0$, a_i $(i = 1, \cdots, k)$ *are integers. Suppose every root* ρ *of* $F(\lambda) = 0$ *satisfies* $|\rho| \leq 1$. *Then* F *is a product of cyclotomic polynomials.*

This is proved in Section V. Recall from our prior discussion that all the roots of F must be chosen to satisfy $|\rho| \leq 1$ to have stable detection. As it is, of course, better to have fewer resonances, one would hence choose for (4) a single cyclotomic polynomial. The cyclotomic polynomials make very desirable characteristic polynomials because of their extremely simple structure. For example, for $m < 105$ or for m a product of two primes, the coefficients of F_m are all 0, ± 1! For m a power of a single prime, the coefficients are all 0, 1 and for $m < 385$, the coefficients do not exceed 2 in absolute value. If m is a product of three distinct odd primes, all the coefficients are less than the smallest of those primes. These assertions are cited in Ref. 5.

This means that implementation of the recursion (1) in the filter shown in Fig. 1 is very simple indeed. For all cases of practical interest, the feedback coefficients a_i will be 0, ± 1. Of course, when $a_i = 0$, one simply does not put a tap on the ith stage. Because of the relation

$$F_{p_1^{\alpha} \cdots p_n^{\alpha_n}}(\lambda) = F_{p_1 \cdots p_n}(\lambda^{p_1^{\alpha_1 - 1} \cdots p_n^{\alpha_n - 1}})$$

(p_i distinct primes—see Ref. 4), most of the coefficients of F_m will usually be zero, and hence the taps-to-memory ratio is generally low (see Table I).

In the preceding discussion, the principal emphasis has been on the use of the filter as a receiver. However, considerations relating to its

use as a generator lead to the same conclusion: that the characteristic polynomial (5) of the recursion (1) should be a cyclotomic polynomial. Indeed, for a generator, the problem of unwanted limit cycles is more critical. There is again the requirement that all the roots ρ_i of (5) satisfy $|\rho_i| \leqq 1$, as small perturbations in the initial conditions $x_0, \cdots, x_{k-1}$ from the (ideal) values 0, $\sin 2\pi f_0\tau, \cdots, \sin 2\pi f_0(k-1)\tau$ (to generate $\sin 2\pi f_0 n\tau$) are inevitable; if such a perturbation occurs along an eigenvector corresponding to a root ρ_i, where $|\rho_i| > 1$, it produces a nonzero coefficient b_i for that root in the general solution $x_n = \sum_{i=1}^{k} b_i\rho_i^n$ (where $b_1, \cdots, b_k$ are functions of the initial conditions $x_0, \cdots, x_{k-1}$; see Section III). This component would attain an arbitrarily large amplitude (with time) and overwhelm the desired tone.

Hence, one again requires a filter that can perform perfect arithmetic and whose characteristic equation has all its roots on the unit disc. From Theorem 2 we thus deduce that (5) should be a product of cyclotomic polynomials for the generator as well. As tone generation is impeded by the presence of harmonic resonances at other roots (due, again, to perturbation of initial conditions), one takes for (5) a single cyclotomic polynomial.

Thus we have shown that, for both generating and receiving, the best linear recursion is one whose characteristic polynomial is cyclotomic. As the roots in this case are all of the form $\exp[2\pi i(d/m)]$, the resonant frequencies can be expressed as

$$2\pi f\tau \equiv 2\pi \frac{d}{m} \text{ (modulo } 2\pi) \tag{7}$$

for all positive integers $d < m$ such that d is relatively prime to m. Resonance at the fundamental is described by $2\pi f\tau = 2\pi(1/m)$, that is, the fundamental of the filter is $f = \tau^{-1}/m$. Hence, if one requires a fundamental frequency of f_0 (i.e., if f_0 is the frequency of the tone to be generated or detected) and one intends to use a filter with memory $k = \varphi(m)$, the clock rate τ^{-1} is set at $\tau^{-1} = f_0 m$. All other resonances occur at various harmonics (multiples of f_0) as follows: the resonant harmonics in the band $0 \leqq f \leqq \tau^{-1}$ occur when $f\tau = d/m$, that is, at $f = df_0$ for all those integers d as above. For example, if $m = 30$ then $k = 8$ and d assumes the values 1, 7, 11, 13, 17, 19, 23, 29. Hence, this filter has no resonances between the fundamental f_0 and the seventh harmonic. It resonates at the seventh harmonic $7f_0$, and thereafter at $11f_0$, $13f_0$, and so on. The resonances are at all the prime harmonics greater than 5, since in general those integers less than and relatively prime to the product m of the first p primes, are those primes lying between the pth prime and m. Furthermore, note that $30 = 1 + 29 = 7 + 23 = 11 + 19 = 13 + 17$. The first resonance due to aliasing

will always be at $f = f_0 + \tau^{-1} = f_0 + f_0 m = (m + 1) f_0$. In the case of the previous example, this is the thirty-first harmonic.

Factors pertinent to the choice of which cyclotomic polynomial to use are relegated to Section VI. Suffice it to say at this point that the more memory available, the farther away from the fundamental can the first resonance be made due to aliasing. However, except for the cases $m = 1$ and $m = 2$, the first resonance after the fundamental will be below the clock frequency τ^{-1}. In these cases, for a given amount of memory k, if the interest is to have the first higher-order resonance as far from the fundamental as possible, one would find the largest integer r such that the product m of the first r primes satisfies $\varphi(m) \leqq k$. Then the first higher-order resonance would occur at the qth harmonic, where q is the $(r + 1)$st prime.

III. ELIMINATING IN-BAND HIGHER-ORDER RESONANCES

The preceding analysis has indicated that, within the constraints established, various higher-order resonances are unavoidable. This could lead to difficulties. In practice, many higher-order harmonics are introduced in the process of limiting the input signal. The limiter (see Fig. 2) limits the amplitude of the input signal $u(t)$. For example, a common limiter is a "hard-clipper." This has output ± 1, depending upon whether $u(t) \geqq 0$ or $u(t) < 0$. The effect of hard-clipping on an input is to produce all the odd harmonics: $\sin 2\pi ft \rightarrow 2/\pi \sin 2\pi ft + 2/3\pi \sin 6\pi ft + 2/5\pi \sin 10\pi ft + \cdots$. Hence, a filter with more resonances frequently must be run for a longer period of time to attain a threshold sufficiently high to reject spurious signals. Also, when used as a generator, perturbations of the initial conditions of the filter could lead to unwanted harmonics at all the resonances of the filter. As such perturbations are inevitable, it is usually necessary to make allowance for eliminating these harmonics.

While resonances due to aliasing are inherent to the discrete-time nature of the system and are hence unavoidable, resonances below the clock frequency τ^{-1} can be handled outside the feedback loop. In particular, it is possible (in theory) to eliminate (in practice, to reduce the Fourier coefficients of) any or all resonances at a frequency f, $0 < f < (2\tau)^{-1}$, along with the conjugate resonance at $\tau^{-1} - f$. This is effected through operations *outside* the feedback loop. Specifically, this is accomplished either through alteration of the input before it enters the filter: $u_n \rightarrow v_n = \sum_{i=1}^{d} c_i u_{n-i}$, or equivalently through alteration of the filter output before it enters the threshold detector: $x_n \rightarrow y_n = \sum_{i=1}^{d} c_i x_{n-i}$ (see Figs. 3 and 8). Although these two options are mathematically equivalent, considerations with respect to minimizing the word length necessary for perfect arithmetic would mitigate in

favor of one or the other. This will be discussed in Section V. Here, we will describe the latter option only.

Let $X(\lambda)$ be the generating function for the sequence

$$x_n = \sum_{i=1}^{k} a_i x_{n-i} + u_n,$$

and let $U(\lambda)$ be the generating function for the input u_n. That is,

$$X(\lambda) = \sum_{n=0}^{\infty} x_n \lambda^n, \qquad U(\lambda) = \sum_{n=0}^{\infty} u_n \lambda^n. \tag{8}$$

Then

$$X(\lambda) = \sum_{i=1}^{k} a_i \lambda^i X(\lambda) + U(\lambda), \qquad \text{or} \qquad X(\lambda) = \frac{1}{1 - \sum_{i=1}^{k} a_i \lambda^i} U(\lambda).$$

Notice that defining $F(\lambda) \equiv \lambda^k - \sum a_i \lambda^{k-i}$, the characteristic polynomial of the filter, we obtain

$$X(\lambda) = \frac{1}{\lambda^k F(\lambda^{-1})} U(\lambda). \tag{9}$$

Since $F(\lambda)$ is assumed to be a cyclotomic polynomial, it is real and all its roots are of unit modulus. Hence ρ is a root if and only if $\bar{\rho} = \rho^{-1}$ is a root. It follows that $\lambda^k F(\lambda^{-1}) = F(\lambda)$. Thus (9) may be rewritten as

$$X(\lambda) = \frac{1}{F(\lambda)} U(\lambda). \tag{10}$$

We define a weighting function $W(\lambda)$ with the property that the resulting output function

$$Y(\lambda) \equiv W(\lambda) X(\lambda) \tag{11}$$

has poles only at those roots of $F(\lambda)$ corresponding to those resonances actually desired. Specifically, $W(\lambda)$ will be a real polynomial of degree $k - 2r$, where r is the number of resonances desired in the band $[0, (2\tau)^{-1}]$; the roots of W shall be those roots of F corresponding to the unwanted resonances. Typically, one desires to eliminate all resonances but the fundamental, in which case $r = 1$ and $W(\lambda)/F(\lambda) = 1/(\lambda^2 - a\lambda + 1)$ for an appropriate real number a. Then, from (10) and (11), one obtains $Y(\lambda) = W(\lambda)X(\lambda) = (1/(\lambda^2 - a\lambda + 1))U(\lambda)$ so $Y(\lambda) = -\lambda^2 Y(\lambda) + a\lambda Y(\lambda) + U(\lambda)$, and

$$y_n = a y_{n-1} - y_{n-2} + u_n. \tag{12}$$

This corresponds to a second-order filter with only one resonance in the band $[0, (2\tau)^{-1}]$ as shown in Fig. 3. Although there will be trunca-

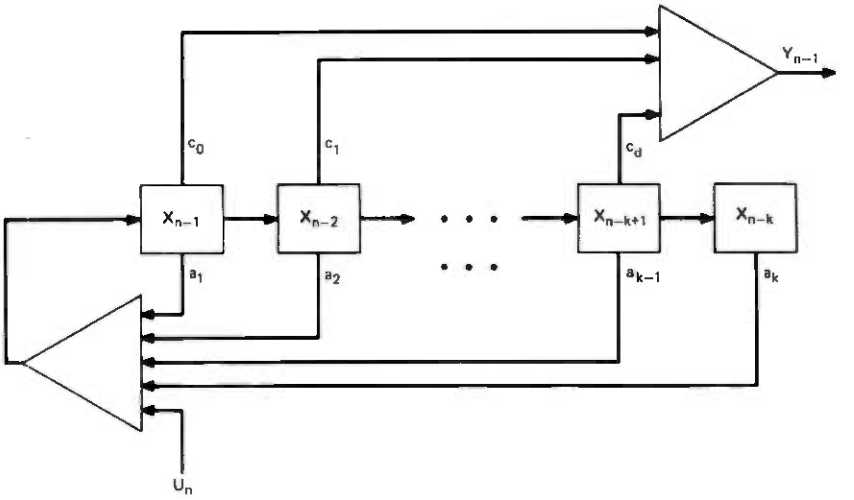


Fig. 3—Implementation of the weighting function.

tion error in (12), this will not lead to limit cycles, as there is no feedback from this to the filter [although (12) represents the performance of the filter in terms of resonances, the filter, of course, is not realized in this way]. Specifically, the weighting function is implemented as in Fig. 3. This is derived from definition (11): if $W(\lambda) = \sum_{i=0}^{d} c_i \lambda^i$, then equating terms in (11) yields

$$y_n = \sum_{i=0}^{d} c_i x_{n-i}, \tag{13}$$

where, typically, $d = k - 2$.

As mentioned earlier, the arithmetic of the weighting function is only approximate; since there is truncation error in the computation of the coefficients c_i, the roots of W will not precisely cancel out the roots of F. Rather, the roots of W will be slightly perturbed from the corresponding roots of F. The effect of this, as will be shown, is that all the resonances due to the roots of F (i.e., all the resonant harmonics of the original feedback loop) will be present in the output y_n—however, they will have reduced energy (but for the fundamental). That is, the less the error in the implementation of W, the smaller the Fourier coefficients of the higher resonant harmonics of the filter. This is demonstrated below.

Suppose F is the cyclotomic polynomial of order m (or any polynomial whose roots $\rho_i, \cdots, \rho_k$ are distinct mth roots of unity, so that each $\rho_j = e^{i2\pi q/m}$ for some integer q, $0 \leqq q < m$). A continuous-time

extension $x(t)$ of the discrete-time function x_n, satisfying $x(n\tau) = x_n$ can be defined as

$$x(t) = \sum_{n=0}^{m-1} x_n v(t - n\tau), \tag{14}$$

where $v(t)$ describes a continuous-time extension of x_n. Specifically, $v(t)$ is a periodic input pulse satisfying $v(t + m\tau) = v(t)$ for all t [typically, $v(t) = 1$ for $0 \leq t < \tau$]. In (4), set $u_n = v(n\tau)$ and normalize $v(0) = 1$. Then $x_n = \sum_{j=1}^{k} b_j \rho_j^n$ for $n < m$. Let $\hat{x}(q)$ [$\hat{v}(q)$] denote the qth Fourier coefficient of $x(t)$ [$v(t)$]. It follows that

$$\begin{aligned}
\hat{x}(q) &\equiv \frac{1}{m\tau} \int_0^{m\tau} x(t) \exp\left(-i2\pi \frac{q}{m} t\right) dt \\
&= \hat{v}(q) \sum_{n=0}^{m-1} x_n \exp\left(-i2\pi \frac{q}{m} n\right) \\
&= \hat{v}(q) \sum_{j=1}^{k} b_j \sum_{n=0}^{m-1} \rho_j^n \exp\left(-i2\pi \frac{q}{m} n\right) \\
&= \hat{v}(q) b_j,
\end{aligned} \tag{15}$$

where j is that index such that $\rho_j = \exp[i2\pi(q/m)]$; if no such index exists, then $\hat{x}(q) = 0$. To simplify matters, we will use the expression "the Fourier coefficient at (the root) ρ_j" to indicate what in the case of (15) is the qth Fourier coefficient $\hat{x}(q)$.

These Fourier coefficients can be computed explicitly from (9). Indeed, factoring $\lambda^k F(\lambda^{-1}) = \prod_{j=1}^{k} (1 - \rho_j \lambda)$ obtains

$$\begin{aligned}
X(\lambda) &= \prod_{j=1}^{k} \frac{1}{1 - \rho_j \lambda} U(\lambda) \\
&= \sum_{j=1}^{k} B_j \frac{1}{1 - \rho_j \lambda} U(\lambda),
\end{aligned} \tag{16}$$

where the B_j's are the coefficients of the partial fraction decomposition, derived explicitly in Lemma 3 below (it is assumed that all the roots ρ_j are distinct; in the case of multiple roots, however, similar results obtain). From (16) one obtains

$$\begin{aligned}
X(\lambda) &= \sum_{j=1}^{k} B_j \sum_{n=0}^{\infty} (\rho_j \lambda)^n \sum_{i=0}^{\infty} u_i \lambda^i \\
&= \sum_{j} B_j \sum_{i,n} \rho_j^{n-i} u_i \lambda^n,
\end{aligned} \tag{17}$$

so $x_n = \sum_{j=1}^{k} B_j \sum_{i=0}^{n} \rho_j^{n-i} u_i$ [which is (4) above]. Hence, $B_j = b_j$ ($j = 1, \cdots, k$) and their explicit form is given in the following lemma.

Lemma 3: *Suppose* $\rho_1, \cdots, \rho_k$ *are distinct numbers. Then*

$$\prod_{i=1}^{k} \frac{1}{1 - \rho_i \lambda} = \sum_{i=1}^{k} \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{\rho_i}{(\rho_i - \rho_j)} \right) \frac{1}{1 - \rho_i \lambda}.$$

Proof: The residue of the left-hand side at the ith pole is the coefficient of that term in the sum above. The decomposition follows from the Cauchy residue theorem.

Notice that, as the roots of F occur in conjugate pairs, a direct consequence of (17) is that, if ρ_i and ρ_j are conjugate roots, then the corresponding Fourier coefficients are also conjugate: $b_i = \bar{b}_j$.

The Fourier coefficients for the sequence y_n can be determined as in (15). For $x_n = \sum b_j \rho_j^n$ as before, we obtain from (13)

$$\begin{aligned} y_n &= \sum_{i=0}^{d} c_i \sum_{j=1}^{k} b_j \rho_j^{n-i} \\ &= \sum_{j=1}^{k} W(\bar{\rho}_j) b_j \rho_j^n. \end{aligned} \tag{18}$$

Thus, the Fourier coefficient of the sequence y_n at the root ρ_j is $W(\bar{\rho}_j) b_j$ (as could be expected, since Fourier transformations are multiplicative). Again, the conjugate coefficient $W(\rho_j) = \overline{W(\bar{\rho}_j)}$. Observe that, if ρ_j is a root of W, then the Fourier coefficients of y_n vanish at the roots ρ_j and $\bar{\rho}_j$ (W was chosen to be real). If W' is the result of perturbing the coefficients of W to correspond to truncation error, then $W'(\bar{\rho}_j)$ is (by continuity) close to zero. Hence, as errors in the weighting functions are reduced, so is the power at each of the resonant harmonics above the fundamental (running the system for finite time, of course). Surprisingly, W is very stable; if the coefficients of W' are simply those of W rounded to the nearest integer (!), the results are frequently virtually as good as if W itself were used. This is exhibited in Table I and illustrated in Figs. 4, 5, and 6. These figures correspond to a filter using the cyclotomic polynomial F_{30}. The input is a hard-clipped sine wave for each given frequency up to 15 times the fundamental. The input frequencies are normalized to units of the fundamental frequency for each filter. For each input frequency, the filter is run for an amount of time equal to seven cycles of the fundamental. If this time corresponds to N steps of the filter, the output is $\max_{n \leq N} |x_n|$, as measured at each input frequency (1500 samples). Using a W' with integer coefficients (or any W' with uniformly truncated coefficients) enables one to perform all the multiplications as additions, simplifying implementation and eliminating any further errors. As one expects, upon

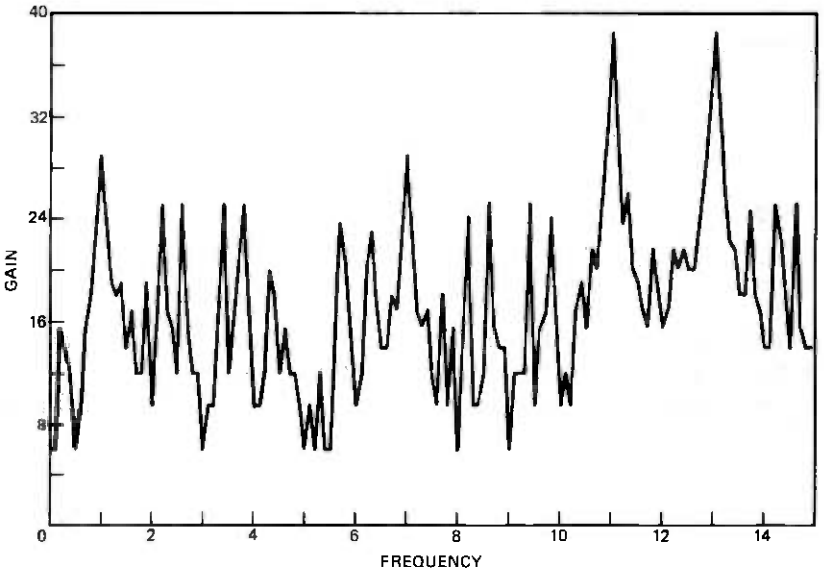


Fig. 4—Hard-clipped/no weighting.

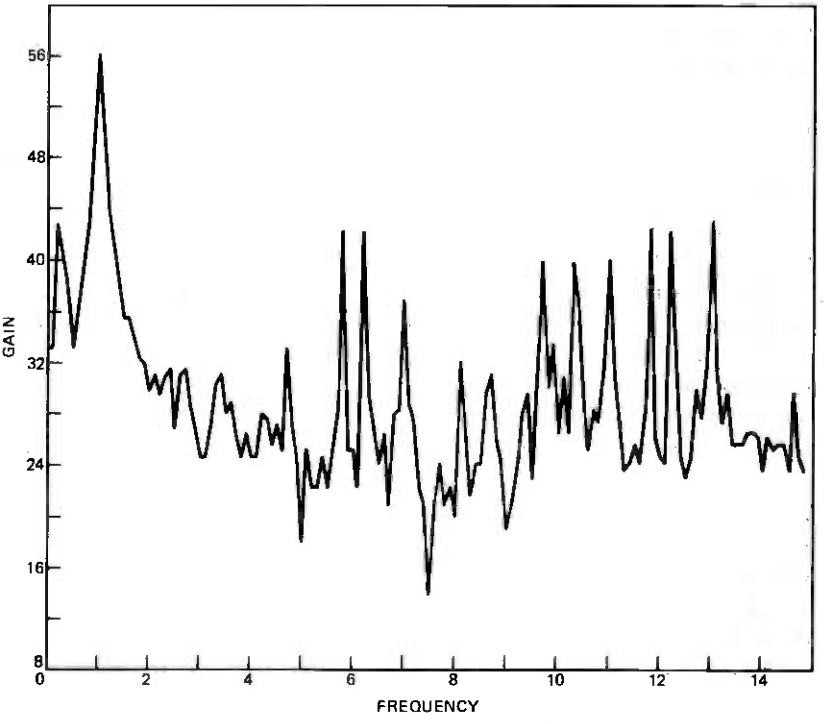


Fig. 5—Hard-clipped/rounded weighting.

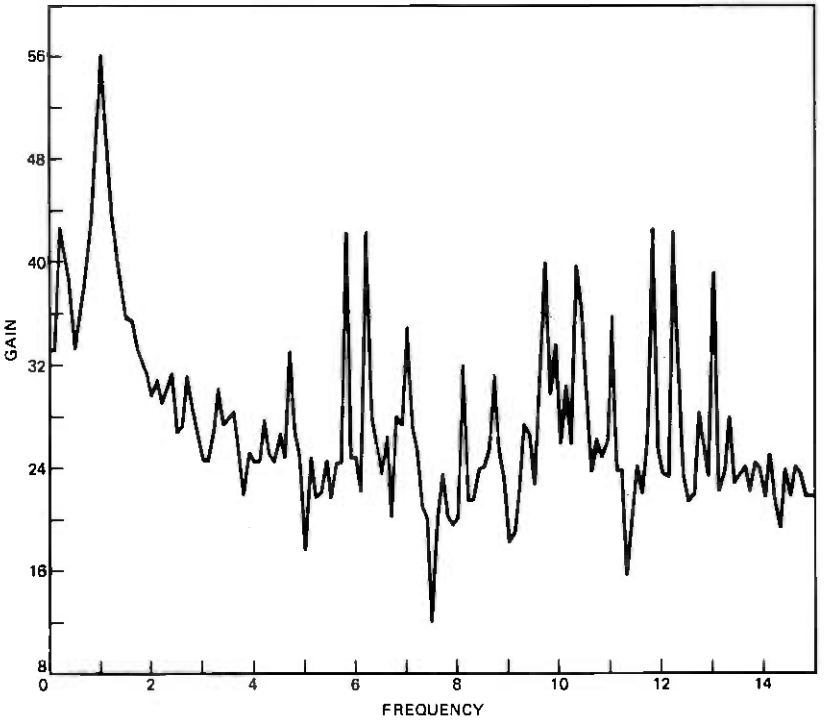


Fig. 6—Hard-clipped/exact weighting.

setting $\rho_2 = \bar{\rho}_1$, the Fourier coefficient of y_n at the fundamental

$$W(\bar{\rho}_1) \prod_{j=2}^{k} \frac{\rho_1}{\rho_1 - \rho_j} = \frac{\rho_1}{\rho_1 - \rho_2} \prod_{j=3}^{k} \frac{\rho_1(1 - \rho_j\bar{\rho}_1)}{\rho_1 - \rho_j} = \frac{\rho_1}{\rho_1 - \rho_2},$$

is the Fourier coefficient of (12) at the fundamental.

IV. IMPULSE RESPONSE

The impulse response is the output resulting from an input of a single pulse: $u_0 = 1$, $u_{n>0} = 0$. Since this output can also be produced by appropriately setting initial conditions, we will refer to it as a pulse train. From (4) we see that if the input u_n is a single pulse, then the output x_n reduces to

$$x_n = \sum_{i=1}^{k} b_i\rho_i^n. \tag{19}$$

In the context of the previous sections, it is assumed that the characteristic polynomial of the sequence x_n is cyclotomic. Since each ρ_i is then an mth root of unity, the sequence x_n is periodic: $x_{n+m} = x_n$ for all n. As before, the resonant harmonics present in the pulse trains x_n correspond to the mth roots of unity which are roots ρ_i $(i = 1, \cdots, k)$

of $F = F_m$; the Fourier coefficient of the pulse train at the root ρ_j is b_j (see Section III).

In particular, using the notations of Section III, $U(\lambda) = 1$ and thus (10) reduces to

$$X(\lambda) = \frac{1}{F(\lambda)}. \tag{20}$$

But $X(\lambda) = \sum_{n=0}^{\infty} x_n\lambda^n = (\sum_{n=0}^{m-1} x_n\lambda^n)(\sum_{n=0}^{\infty} \lambda^{mn})$ since $x_{n+m} = x_n$. Defining $f(\lambda) = \sum_{n=0}^{m-1} x_n\lambda^n$, one obtains

$$f(\lambda) = \frac{1 - \lambda^m}{F(\lambda)} \tag{21}$$

from (20). Notice that f has integer coefficients (the input u_n is integer, as are the coefficients a_i). Indeed, $1 - \lambda^m$ is a product of cyclotomic polynomials, one of which is $F(\lambda)$. Specifically,

$$1 - \lambda^m = \pm \prod_{n|m} F_n(\lambda)$$

[the product is taken over all n which divide m; hence, for example,

$$1 - \lambda^6 = -F_1(\lambda)F_2(\lambda)F_3(\lambda)F_6(\lambda)$$
$$= (\lambda - 1)(\lambda + 1)(\lambda^2 + \lambda + 1)(\lambda^2 - \lambda + 1)];$$

and, from (21),

$$f(\lambda) = \pm \prod_{\substack{n|m \\ n \neq m}} F_n(\lambda)$$

obtains. Consequently, $f(\rho) = 0$ for all mth roots of unity ρ, *except* for the primitive roots of unity [the roots of $F_m(\lambda)$]. This was anticipated by E. N. Gilbert in Ref. 6, where he showed that a pulse train x_n of period m has resonances at those harmonics corresponding to the mth roots of unity which are not roots of $\sum_{n=0}^{m-1} x_n\lambda^n = 0$. Equation (21) covers the general situation where $f(\lambda)$ [and consequently $F(\lambda)$] are arbitrary products of cyclotomic factors of $1 - \lambda^m$.

In the same paper, Gilbert was concerned about the problem of increasing the power of the pulse train at the fundamental (relative to the power at the other resonances). This could be done by shaping the input u_n for one period, but it is usually undesirable to do this. As explained in Section III, however, the same effect is obtained by utilizing a weighting function W. If utilized directly, this will introduce noninteger levels into the pulse train. Nonetheless, it is possible to avoid this by replacing W with W^I where the latter is obtained through rounding off to the nearest integer the coefficients of the former. The pulse train resulting from W^I will have integer levels, but the trunca-

tion error will again introduce higher-order resonances. However, Table I shows that these are very small indeed, leaving typically about 98 percent of the power at the fundamental. This compares with 25 percent or less (for F_{16}, F_{24}, F_{30}) without W^I. Note, for example, F_8. The pulse train 1, 0, 0, 0, −1, 0, 0, 0 has resonances at the third, fifth, and seventh harmonics. However, by simply altering this to 1, 1, 1, 0, −1, −1, −1, 0, the first appreciable resonance does not come until the seventh harmonic. In this case, use of W^I does not introduce any new levels in the pulse train.

The worst case in Table I is F_9 where 92 percent of the power is at the fundamental. E. N. Gilbert has pointed out that if one wished to increase the proportion of the power at the fundamental of this train (or any other), one could multiply the output $Y(\lambda)$ by some constant $c > 1$, chosen so that the roundoff error of $cW \to (cW)^I$ is smaller than that for W^I alone [recall (11)]. This, however, would introduce more levels into the pulse train (although no more than c times as many).

Table I gives an indication of the possibilities for various filters. Included are the filters with memory less than 12 which provide the greatest separation between the fundamental and the first resonant harmonic, either with or without the weighting function. The asterisks and daggers indicate those which, for the amount of memory, have the largest possible separation without or with the weighting function. For utilization with a "hard-clipper" (which has all odd harmonics), F_8, F_9, and F_{15} are included. Although these resonate at all even harmonics, they have the same response to a hard-clipped input at the fundamental as the respective cyclotomic filters of twice the sampling rate. To have the first resonant harmonic higher than the seventh (without W) would require a memory of 48 (and F to have a coefficient of 2). The next interesting entry with respect to W is F_{36} with memory 12. The columns to the right of the double line all deal with the integer-rounded transfer function W^I. Columns A and B give $|b_1|^2/\sum_{i=1}^{k/2} |b_i|^2$ and $|b_1 W^I(\bar{p}_1)|^2/\sum_{i=1}^{k/2} |b_i W^I(\bar{p}_i)|^2$ as a percent, respectively, where b_i is the Fourier coefficient of the sequence x_n at the root ρ_i [see (6) and Section III)]. Column C gives $(\max_{2 \le i \le k/2} |b_i W^I(\bar{p}_i)|^2)/|b_1 W^I(\bar{p}_1)|^2$ as a percent. The roots ρ_i $(i = 1, \cdots, k/2)$ are assumed to be in order of ascending argument $< \pi$ (so ρ_1 is the fundamental). Columns D and E give the moduli of the Fourier coefficients b_i and $b_i W^I(\bar{p}_i)$ of the sequences x_n and y_n, respectively. Columns F and G give the pulse trains of x_n and y_n, respectively, with initial pulse $u_0 = 1$, $u_{n>0} = 0$. The exponent denotes repeated digit; the arrow indicates that the preceding train is followed by another identical train, but that each digit is the negative of what it was.

Memory	Characteristic Polynomial F	Resonant Harmonics in the Band $[0, \tau^{-1}]$, Aside From Fundamental		First Resonant Harmonic Due to Aliasing	Number of Taps on Filter (Without Weighting Function)	Integer-Rounded Weighting Function W^I
		Without Weighting Function	With Weighting Function			
*1	$F_2=\lambda+1$	none	—	3	1	—
2	$F_3=\lambda^2+\lambda+1$	2	—	4	2	—
2	$F_4=\lambda^2+1$	3	—	5	1	—
*2	$F_6=\lambda^2-\lambda+1$	5	—	7	2	—
4	$F_5=\lambda^4+1$	3, 5, 7	7	9	1	$1+\lambda+\lambda^2$
†4	$F_{12}=\lambda^4-\lambda^2+1$	5, 7, 11	11	13	2	$1+2\lambda+\lambda^2$
6	$F_9=\lambda^6+\lambda^3+1$	2, 4, 5, 7, 8	8	10	2	$1+2\lambda+\lambda^2+2\lambda^3+\lambda^4$
†6	$F_{18}=\lambda^6-\lambda^3+1$	5, 7, 11, 13, 17	17	19	2	$1+2\lambda+3\lambda^2+2\lambda^3+\lambda^4$
8	$F_{15}=\lambda^8-\lambda^7+\lambda^5-\lambda^4+\lambda^3-\lambda+1$	2, 4, 7, 8, 11, 13, 14	14	16	6	$1+\lambda+\lambda^2+\lambda^3+\lambda^4+\lambda^5+\lambda^6$
8	$F_{18}=\lambda^6+1$	3, 5, 7, 9, 11, 13, 15	15	17	1	$1+2\lambda+2\lambda^2+3\lambda^3+2\lambda^4+2\lambda^5+\lambda^6$
8	$F_{24}=\lambda^8-\lambda^4+1$	5, 7, 11, 13, 17, 19, 23	23	25	2	$1+2\lambda+3\lambda^2+3\lambda^3+3\lambda^4+2\lambda^5+\lambda^6$
†*8	$F_{30}=\lambda^8+\lambda^7-\lambda^6-\lambda^4-\lambda^3+\lambda+1$	7, 11, 13, 17, 19, 23, 29	29	31	6	$1+3\lambda+5\lambda^2+5\lambda^3+5\lambda^4+3\lambda^5+\lambda^6$

*† See text for explanation

V. CONDITIONS FOR PERFECT ARITHMETIC

Here we indicate why cyclotomic polynomials yield optimal recursions for generating sinuosidal signals. When we use (1) to generate tones, the u_n is set to zero and some initial condition $x_0, x_1, \cdots, x_{k-1}$ is chosen to generate the required samples x_n:

$$x_n = \sum_{i=1}^{k} a_i x_{n-i}. \tag{22}$$

If we use the usual second-order recursion, then (22) is of the form

$$x_n = ax_{n-1} - x_{n-2}, \tag{23}$$

where $|a| < 2$, so we have complex roots. In this case, we show below that the number of distinct values that x_n, $n = 0, 1, \cdots, N$ can take grows at least as fast as $N/2$, with N. So, to simulate (22) with perfect arithmetic, the number of "words" needed grows at least as fast as N, the number of samples needed.

Proposition 4: Suppose $|a| < 2$, and rational but not an integer. Then for any initial conditions x_0, x_1 (not both zero) and any positive integer N, the number of distinct values among $x_0, \cdots, x_N$, where $x_n = ax_{n-1} - x_{n-2}$, for $2 \leqq n \leqq N$, is at least $N/2$.

some cyclotomic filters

A	B		D	E	F	G
% of Total Power in $[0, (2\tau)^{-1}]$ at Fundamental		Highest Power at a Rejected Resonance (% of Fundamental)	Modulus of Fourier Coefficients (in Order of Arguments $< \pi$)		Pulse Trains	
Without W^t	With W^t		x_n	y_n	x_n	y_n
100	—	—	1	—	$1, -1$	—
100	—	—	0.58	—	$1, -1, 0$	—
100	—	—	0.5	—	$1, 0, -1, 0$	—
100	—	—	1.73	—	$1^2 0 - 1^2 0$	—
50	97.1	2.9	All 0.25	0.60, 0.10	$10^2 - 10^2$	$1^2 0 - 1^2 0$
50	99.5	0.5	All 0.29	1.08, 0.08	$10 10^2 - 10 - 10^2$	$12^2 10 - 1 - 2^2 - 10$
33.3	92.7	7.8	All 0.19	0.85, 0.04, 0.24	$10^2 - 10^5$	$121^2 - 1^2 - 2 - 10$
33.3	99.6	0.3	All 0.19	7.6, 0.04, 0.06	$10^2 10^5 \rightarrow$	$123^4 210 \rightarrow$
53.3	98.9	0.8	0.33, 0.27, 0.09, 0.11	4.78, 0.51, 0.55, 0.75	$1^2 0^2 - 1^2 0^7$	$123^2 21 - 1 - 2 - 3^2 - 2 - 10$
25	98.0	1.8	All 0.13	1.29, 0.02, 0.06, 0.17	$10^7 - 10^7$	$12^2 32^2 10 \rightarrow$
25	99.5	0.3	All 0.14	1.97, 0.05, 0.09, 0.11	$10^2 10^7 - 10^2 - 10^7$	$123^4 4^2 3^2 210 \rightarrow$
6	98.4	1.0	0.11, 0.09, 0.27, 0.33	2.41, 0.04, 0.19, 0.24	$1 - 110^2 1 - 110^7 \rightarrow$	$123^2 545^4 453^2 210 \rightarrow$

Proof: We can write $x_n = b_1\rho_1^n + b_2\rho_2^n$, where ρ_1, ρ_2 are the distinct roots of $\lambda^2 - a\lambda + 1$, as in (19). Since the roots are not real, let $\rho = \rho_1 (= \bar{\rho}_2)$, $b = b_1 (= \bar{b}_2)$. Then $x_n = x_m$ implies $\mathrm{Re}\,(b\rho^n) = \mathrm{Re}\,(b\rho^m)$. In this case, letting $\theta = \arg \rho$, $\varphi = \arg b$, we obtain $\cos(\varphi + n\theta) = \cos(\varphi + m\theta)$ so $\varphi + n\theta \equiv \pm(\varphi + m\theta) \pmod{2\pi}$. Since ρ is not a root of unity, the numbers $n\theta$ ($n = 0, 1, 2, \cdots$) are all distinct and hence for fixed m either $n = m$ or $n\theta \equiv -2\varphi - m\theta \pmod{2\pi}$. As this last congruence can be satisfied by at most one n, it follows that, for each m, there is at most one $n \neq m$ such that $x_n = x_m$.

The following result shows that, if one wishes to generate $\sin \pi n\theta$ with perfect accuracy using a linear recursion, $e^{i\pi\theta}$ must be a root of the corresponding polynomial (5).

Proposition 5: *If $s_n = \sin \pi n\theta$ is a solution of $x_n = \sum a_j x_{n-j}$ and θ is not an integer, then $e^{i\pi\theta}$ is a root of the polynomial $\lambda^k - \sum a\lambda^{k-j}$.*

Proof: From $\sin \pi(n+1)\theta = \sum a_j \sin \pi(n+1-j)\theta$, we expand both sides using a familiar trigonometric identity and get

$$\begin{aligned}\sin \pi n\theta \cos \pi\theta + \cos \pi n\theta \sin \pi\theta &= \sin \pi(n+1)\theta \\ &= \sum a_j \sin \pi(n+1-j)\theta = \sum a_j \sin \pi(n-j)\theta \cos \pi\theta \\ &+ \sum a_j \cos \pi(n-j)\theta \sin \pi\theta = \sin \pi n\theta \cos \pi\theta \\ &+ \sum a_j \cos \pi(n-j)\theta \sin \pi\theta.\end{aligned}$$

Since θ is not an integer, $\sin \pi\theta \neq 0$, and thus from the equality of the first and last expressions, we obtain $\cos \pi n\theta = \sum a_j \cos \pi(n - j)\theta$. Hence, $\cos \pi n\theta$ is also a solution to the recursion, and it follows that $e^{i\pi n\theta} = \cos \pi n\theta + i \sin \pi n\theta$ is a solution too. Consequently, $e^{i\pi k\theta} - \sum a_j e^{i\pi(k-j)\theta} = 0$.

The next theorem shows that every recursion which satisfies the stability criterion $|\rho| \leqq 1$ for all its roots, and for which perfect arithmetic is possible, is cyclotomic.

Theorem 6: *Suppose every root ρ of the polynomial $F(\lambda) = \lambda^k - \sum_{i=1}^{k} a_i \lambda^{k-i}$ satisfies $|\rho| \leqq 1$.*

(*i*) *If $a_1, \cdots, a_k$ are integers and $a_k \neq 0$, then $F(\lambda)$ is a product of cyclotomic polynomials.*

(*ii*) *If $a_1, \cdots, a_k$ are rational numbers and $x_n = \sum a_i x_{n-i}$ is periodic ($x_{n+p} = x_n$ for some p, all n) for some nonzero initial conditions $x_0, \cdots, x_{k-1}$, then $F(\lambda)$ has as a factor a cyclotomic polynomial.*

Proof: For case (*i*), each irreducible factor (over the integers) of $F(\lambda)$ has the same form as $F(\lambda)$ itself by' 'Gauss' Lemma".[6] Thus, it suffices to assume that $F(\lambda)$ is irreducible, in which case all its roots are distinct. In this case, we can write $x_n = \sum b_i \rho_i^n$ where the ρ_i's are the roots of $F(\lambda)$ and x_n is as in case (*ii*). But then $|x_n| \leqq \sum |b_i|$, and as for any integer initial conditions $x_0, \cdots, x_{k-1}$, x_n will be an integer for all n, x_n can in such a case assume only a finite number m of distinct values ($m = [\sum |b_i|]$). Hence for all n, the $k$tuple $(x_{n+1}, \cdots, x_{n+k})$ can assume at most m^k distinct values, and as x_n is recursively generated with memory k, x_n must be periodic, of period $p \leqq m^k$. This brings us to case (*ii*).

For case (*ii*), let L be the rational canonical form associated with the recursion x_n (see Ref. 7, Section 5.2.1), and J be the Jordan canonical form of L. Then for some initial state vector $\mathbf{x}$, $J^p\mathbf{x} = \mathbf{x}$, and it follows that some diagonal element of J, that is, some root of $F(\lambda)$, must be a pth root of unity. Hence, the irreducible factor of $F(\lambda)$ having that root must be cyclotomic.

Hence, from the above the θ of Proposition 5 must be rational when perfect accuracy is required.

In all the preceding, the basic assumption has been that all the coefficients of the recursion (22) are real. We can infer from Theorem 1 that this is no loss of generality as, if the recursion had complex coefficients (with rational real and imaginary parts) and was irreducible over the field $Q(i)$ (the field of gaussian rationals), then the roots of the characteristic polynomial would be distinct, no pair being conjugate. Indeed, Theorem 1 remains true if the word "integer" is every-

where replaced by "rational number." The arguments of Section II show that we may as well assume all the coefficients are real.

VI. COMPUTING WORD LENGTH AND ADDITIONS PER CYCLE

To realize the cyclotomic filters in hardware with perfect arithmetic, the necessary amount of memory and adder complexity must be provided. We describe here how to estimate the word length and the rate of additions required to implement a cyclotomic filter with a weighting function. It shall be assumed that all operations are performed in binary form. The number of binary bits required to store each x_n is called the word length ω of the system. For generators that produce a signal approximating a sinusoid, the word length required will depend on the accuracy of approximation needed. When the filter is used as a tone detector, the word length required will depend on the duration of operation, since the signal level tends to build up, especially at frequencies close to any resonant frequency (Fig. 7). The signal level, of course, does not uniquely specify the minimum word length. Even though for storing x_n we may need only ω bits, it is conceivable that during the computations numbers greater in magnitude than x_n, which need more bits for storage, could arise. To perform operations in a serial-multiplexed fashion, it is desirable to have uniform word length for all operations in the feedback loop of the filter. Hence, the word length will have to be increased to accommodate any number encountered during the computations. However, for the filters considered

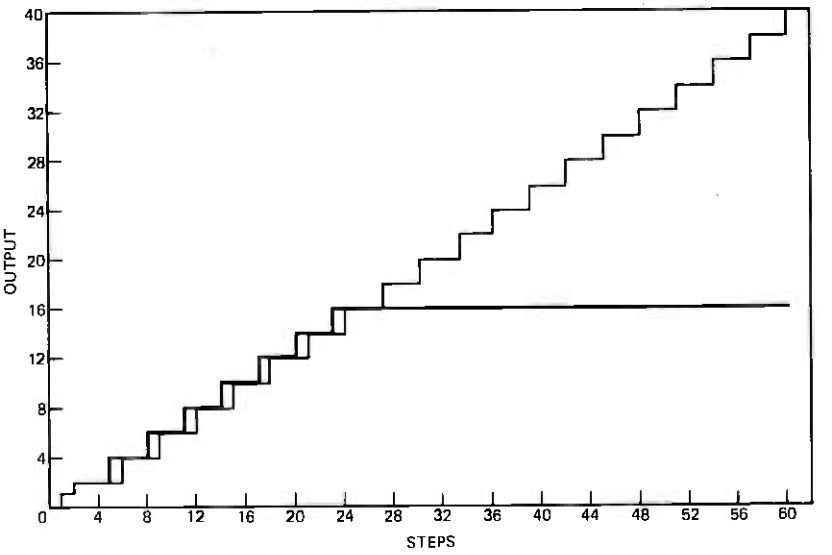


Fig. 7—Growth of output.

in Table II, it is possible to arrange the computations in such a way that the word length is determined by the maximum magnitude of x_n. In general, there are a finite number of ways in which the additions involved in the filter can be arranged. By simulation of the different arrangements, the word length required can then be determined.

There are two possible ways of implementing the cyclotomic filters as generators. The first is to generate the impulse response (19); this is generally sufficient (see Table I). In this case, the weighting function (13) shapes the effect of this impulse to simulate the initial conditions $x_0, \cdots, x_{k-1}$ of the tone being generated. As the input is zero after the initial pulse $u_0 = 1$, the weighting function need only be used during the first $d + 1$ steps of the filter. Let m be the largest number in the pulse train y_n of Table I, and let $[[x]]$ be the smallest integer larger than x. The word length necessary for perfect arithmetic is at least $\omega = [[\log_2 m]] + 1$ and, for the filters considered here, ω is also sufficient. (We add 1 for a sign bit.) This word length is shown in column B of Table II.

However, rounding off in the weighting function introduces errors in the effective initial values of the signal. If this approximation is not sufficiently good, then the initial conditions of the filter $x_0, \cdots, x_{k-1}$ can be set as accurately as needed, and then the filter is operated with the feedback loop alone. In particular, one can set the initial conditions of the filter such that $|x_n - \sin 2\pi n/p| < 2^{-m}$ $(n = 0, \cdots, k-1)$ where $\sin 2\pi n/p$ is the desired signal. One can then compute the minimum word length required by simulating the filter for one period. In all cases of interest here, the word length including sign is $(m + 1)$ for $m \leqq 12$. Hence, as an example, the cyclotomic filter of order 30 can generate a sequence (x_n) such that $|x_n - \sin 2\pi n/p| < 2^{-10}$ if the initial conditions are set such that $|x_n - \sin 2\pi n/p| < 2^{-10}$ $(n = 0, \cdots, 7)$, using a word length of 11.

To determine the number of binary additions per period of the filter (i.e., per cycle of the fundamental), one counts the number of bit additions per step. If m denotes the number of additions per step, then $pm\omega$ is the number of binary additions per cycle, where p is the period of (22) and ω the word length used in the feedback loop (see above). When the generator is implemented in the first way (using an initial pulse and the weighting function), the number of additions is shown in column C of Table II (not including those necessary in the initial $d + 1$ steps for the weighting function). When the generator is implemented in the second way (setting the initial conditions), the number of additions can be computed by multiplying the value in column C by ω/ω', where ω is the word length chosen and ω' is the corresponding word length from column B.

When the filter is used as a detector, we assume that the input to the filter is a sequence which only assumes the values $+1$ and -1. This is true, for example, when the analog signal to be detected is either hard-clipped or delta-modulated. In these cases, it is advantageous to apply the weight function to the input sequence u_n rather than to the sequence x_n; since, in general, x_n can assume many values other than $+1$ and -1, computations involving the weighting function are simplified if they are performed on the input (see Section III). In fact, applying the weighting function to the input is so simple arithmetically that it can be implemented with read-only memory. On the other hand, if read-only memory is not used and one wishes to save on computations by checking the threshold (max $\{x_n\}$) only in the last cycle of the filter (with respect to its duration of operation for detection), then the weighting function is best implemented as in Section III, on the output of the feedback loop. Then the filter can be run during all but the last cycle, without computing the weighting function.

When the weighting function is applied to the input, the filter is described by

$$v_n = \sum_{i=0}^{d} c_i u_{n-i} \tag{24}$$

$$x_n = \sum_{i=1}^{k} a_i x_{n-i} + v_n, \tag{25}$$

where u_n is the input into the filter and v_n is the result of the weighting function. Figure 8 describes this filter.

For the filters in Table I, the effect of rounding c_i to the nearest integer is slight. Hence, it is *a fortiori* suitable to round off $v_n = \sum c_i u_{n-i}$ to the nearest integer. Therefore, since the only values

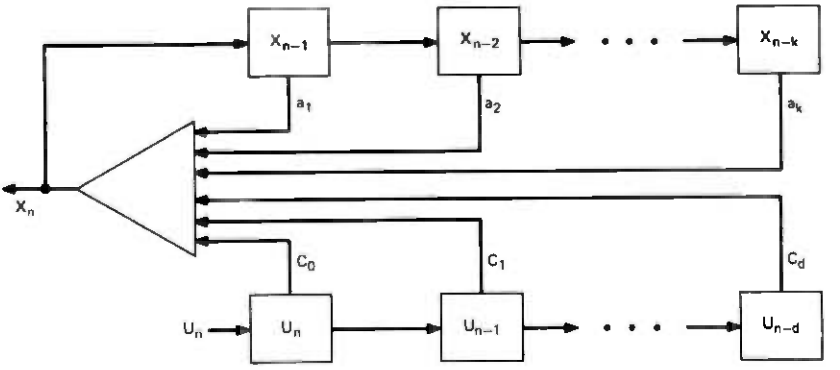


Fig. 8—Implementation of the weighting function at the input.

assumed by u_i are ± 1, it suffices to have for v_n a word length of $\omega = [[\log_2 \{\sum |c_i|\}]] + 1$ (where $\{x\}$ is the integer closest to x and $[[x]]$ is the smallest integer larger than x; 1 is added for a sign bit). The sequence v_n can then assume any value between $-\{\sum |c_i|\}$ and $\{\sum |c_i|\}$. With d as in (24) and ω as above, implementations of the weighting function with read-only memory then requires $2^{d+1}\omega$ memory bits. The respective values for this are shown in column D of Table II. When a bank of such tuned filters is used in one receiver (for example, in a *Touch-Tone*® system such as described in Ref. 3), all the filters could use one read-only memory for the weighting functions. Also, by increasing ω, we can make the round-off error as small as we wish.

To determine the word length for use in the feedback loop of the detector, the maximum signal level can be determined by using an input u_n of the same frequency as the resonant frequency. Since the impulse response [see (19)] of these filters is periodic and of the same period as the resonant frequency, the latter produces the maximum signal level $\sup_{n \leq N} x_n$, for duration of operation $N\tau$. Let this maximum be M. The word length required should then be at least $[[\log_2 M]] + 1$. For all the filters considered here, $[[\log_2 M]] + 1$ is also sufficient. The number of M, of course, is determined by N. If the cyclotomic filter is of period p (i.e., Theorem 1 is F_p), then the filter runs through N/p periods, corresponding to N/p cycles of the fundamental. Calculations have been made for two values of N/p: 7 (the number of cycles computed in Ref. 3 to be necessary for *Touch-Tone* interchannel rejection), and 10 (a more uniform point of reference).

In Table II, column E shows the word length required in the feedback loop for the indicated durations, when the weighting function is computed on the input as in (24), implemented equivalently with or without read-only memory, producing the filter response (25).

When there is no weighting function on the input, the word length required is shown in column F (of course, a weighting function may be applied to the output as in Section III).

The number of binary additions per cycle for the detector is determined in the same way as for the generator; the number is $pm\omega$ as defined above. These numbers are shown in columns G, H, and K of Table II. Column G shows the number of binary additions per cycle in the feedback loop when read-only memory is used to implement the weighting function, applied to the input as in (24). If read-only memory is not used, then the weighting function has to be computed. Since the numbers involved in the computation of the weighting function [when implemented as in (24)] are generally smaller than those in the feedback loop, the word length required for their computations are smaller. Hence, one can use two different adders, one for the weighting function

Table II — Complexity of some cyclotomic generator-detectors

Period	Generator Using Impulse Response		R.O.M.	Detector: 7 Cycles					Detector: 10 Cycles				
				Word Length for (x_n)		Adds/Cycle for (x_n)			Word Length for (x_n)		Adds/Cycle for (x_n)		
	Word Length	Adds/ Cycle		Weighted Input	Unweighted Input	Weighted Input	Unweighted Input	For (v_n)	Weighted Input	Unweighted Input	Weighted Input	Unweighted Input	For (v_n)
A	B	C	D	E	F	G	H	K	E	F	G	H	K
6	2	24	—	6	6	72	72	—	7	7	84	84	—
8	2	16	24	7	5	56	40	112	8	6	64	48	128
9	3	54	128	8	5	144	90	360	8	6	144	108	360
12	3	72	24	8	6	192	144	288	9	7	216	168	324
15	3	270	512	9	7	810	630	810	10	7	900	630	900
16	3	48	640	9	5	144	80	1276	10	6	160	96	1440
18	3	108	128	9	6	324	216	1134	10	7	360	252	1260
24	4	192	640	10	6	480	288	2160	11	7	528	336	2376
30	4	720	768	11	8	1980	1440	3630	11	8	1980	1440	3630

and one for the feedback loop. Using this arrangement, the number of additions per cycle for calculating the weighting function is shown in column K. The number of binary additions per cycle when no weighting function is used is shown in column H. This, of course, applies when the weighting function is applied to the output as in Section III (but does not include the number of additions necessary for the weighting function). To calculate the number of additions when the weighting function is applied to the input, but read-only memory is not used, add columns H and K.

Column A indicates the respective cyclotomic filters described by their periods.

One important consideration that affects the choice of the order of cyclotomic filter is the noise level at the input to limiter (together with the noise in the limiter). This affects the output of the limiter when the signal level is low. One could divide the period of the signal to be detected into regions where errors could affect the decision about the sign of the signal, and regions where no errors will occur. Those sampling instances where errors could occur lie in regions where the absolute value of the signal is small. Suppose these regions are intervals of length ϵ around the zero crossings of the signal. The worst case corresponds to a phase shift of the signal with respect to the sampling interval which maximizes the number of samples in the error regions. For $\epsilon = 1/63$ (corresponding to approximately 20 dB s/n), there are at most two samples per period that are subject to errors for all the filters we have considered here. Hence the ratio of error-susceptible

samples to error-free ones decreases in this case as the period p increases (for $p \leq 30$). This ratio indicates the perturbation of the threshold one has to make in order to compensate for errors in the limiter.

VII. APPLICATIONS

Possible uses for the systems described in this paper have been mentioned in Section I. In particular, a scheme is proposed in Ref. 3 for utilizing eight cyclotomic filters as channel detectors in a *Touch-Tone* receiver.

Another application of cyclotomic filters may be FSK. As described earlier, by selecting the initial conditions of a cyclotomic filter of period p, one can approximate uniformly sampled values of a sinusoid of period p, i.e., $\sin 2\pi n/p$. By changing the clock rate of the filter, one can shift the frequency of the sinusoid to any preassigned value. Hence, when using the filter as a generator, one can shift the clock rate to shift the frequency. This method of shifting frequencies does not introduce any "discontinuities" in the signal. If, instead of changing clock rate, one were to change the coefficient of a filter, then the filter has to be reinitialized to have constant amplitude, thus producing a discontinuity in the signal. In a similar manner, when using the filter as a detector, one can shift the resonant frequency by shifting clock rate. Hence, with the same filter, one can generate and detect both tones used in a typical FSK arrangement. Furthermore, cyclotomic filters have infinite Q, allowing for the possibility of increasing signaling rate above the presently used systems with finite Q.

REFERENCES

1. L. B. Jackson, "An Analysis of Limit Cycles due to Multiplication Roundoff in Recursive Digital Filters," Proceedings of the Seventh Allerton Conference on Circuits and Systems Theory, 1969, pp. 69–78.
2. H. Breece, private communication.
3. B. Gopinath and R. P. Kurshan, "A *Touch-Tone*® Receiver-Generator with Digital Channel Filters," B.S.T.J., this issue, pp. 455–467.
4. S. Lang, *Algebra*, New York: Addison-Wesley, 1965.
5. E. Lehmer, "On the Magnitude of the Coefficients of the Cyclotomic Polynomial," Bull. Am. Math. Soc., *42* (June 1936), p. 389.
6. E. N. Gilbert, "Pulse Trains Which Lack Prescribed Harmonics," unpublished document.
7. B. Gopinath and R. P. Kurshan, "Recursively Generated Periodic Sequences," Canadian Journal of Math., *XXVI*, No. 6, 1974, pp. 1356–1371.

Contributors to This Issue

William M. Boyce, B.A., 1959, and M.S., 1960, Florida State University; Ph.D., 1967, Tulane University; U. S. Army, 1963–65; NASA Manned Spacecraft Center, 1963 and 1965–67; Bell Laboratories, 1967—. At NASA, Mr. Boyce was head of a section working on navigation plans for the Apollo missions. At Bell Laboratories, he has worked on business data processing, financial modeling, economic theory, computational graph theory, and other topics in applied probability and management science. Since 1970, he has been head of the Mathematics Analysis Department. Member, IEEE, SIAM, ORSA, American Finance Association.

Ronald E. Crochiere, B.S., (E.E.) 1967, Milwaukee School of Engineering; M.S. (E.E.) and Ph.D. (E.E.), 1968 and 1974, Massachusetts Institute of Technology; Bell Laboratories, 1974—. Mr. Crochiere is presently engaged in research activities in speech communications and digital signal processing. Member, IEEE, Sigma Xi, IEEE Acoustics, Speech, and Signal Processing Group Ad Com; Associate Editor, IEEE Transactions on Acoustics, Speech, and Signal Processing.

David D. Falconer, B.A.Sc., 1962, University of Toronto; S.M., 1963, and Ph.D., 1967, Massachusetts Institute of Technology; post-doctoral research, Royal Institute of Technology, Stockholm, 1966–67; Bell Laboratories, 1967—. Mr. Falconer has worked on problems in coding theory, communication theory, channel characterization, and high-speed data communication. Member, Tau Beta Pi, Sigma Xi, IEEE.

B. Gopinath, M.Sc. (Mathematics), 1964, University of Bombay; Ph.D. (E.E.), 1968, Stanford University; Research Associate, Stanford University, 1967–1968; Alexander von Humbolt Research Fellow, University of Göttingen, 1971–1972; Bell Laboratories, 1968—. Mr. Gopinath is engaged in applied mathematics research in the Mathematics and Statistics Research Center.

Robert P. Kurshan, Ph.D. (Mathematics), 1968, University of Washington; Krantzberg Chair for Visiting Scientists, Technion, Haifa, Isarel, March–August 1976; Bell Laboratories, 1968—. Mr.

Kurshan is engaged in mathematics research, with an emphasis on algebra, in the Mathematics and Statistics Research Center. President, MAA-NJ, 1975—.

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, and Commission 6 of URSI.

Lawrence R. Rabiner, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. Coauthor, *Theory and Application of Digital Signal Processing*. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; former President, IEEE G-ASSP Ad Com; member, G-ASSP Technical Committee on Digital Signal Processing, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Technical Committee on Speech Communication of the Acoustical Society; former Associate Editor of the G-ASSP Transactions.

Abstracts of Bell System Papers Appearing in Other Publications

Beginning with this issue, the Journal will publish abstracts of papers written by Bell System authors for other technical and scientific publications. We hope this new section provides you, our readers, with a reference source for articles covering the broad range of research and development in the Bell System.

CHEMISTRY

Heterogeneous Removal of Free Radicals by Aerosols in the Urban Troposphere. L. A. Farrow, T. E. Graedel, and T. A. Weber, ACS Symposium Series, Removal of Trace Contaminants from the Air, ed. Victor R. Deitz, *17*, 1975, pp. 17–27. The effect of aerosols on atmospheric photochemistry has been evaluated in a computation of the gas phase chemistry of the urban troposphere for the northern New Jersey metropolitan region. It is shown that aerosol-radical interactions provide an efficient radical sink and stabilize the diurnal variation of radical concentrations.

The Influence of Aerosols on the Chemistry of the Troposphere. T. E. Graedel, L. A. Farrow, and T. A. Weber, I. J. Chem. Kinetics, Symposium No. 1, 1975; Proceedings of the Symposium on Chemical Kinetics Data for the Upper and Lower Atmosphere, pp. 581–594. Full kinetic calculations of the diurnal chemistry of the urban troposphere have been made using a formalism that includes the interactive effects of aerosols and free radicals. These effects are shown to be necessary to a unified analysis of atmospheric chemical reactions.

Liquidus-Solidus Isotherms in the In-Ga-As System. M. A. Pollack, R. E. Nahory, L. V. Deas, and D. R. Wonsidler, J. Electrochem. Soc., *122* (November 1975), pp. 1550–1552. Liquidus and solidus data are presented for the 800°, 850°, and 900°C isotherms in the In-rich corner of the In-Ga-As phase diagram. A simple solution model gives excellent agreement with the solidus data, but describes the liquidus more poorly than desired.

Ozone: Involvement in Atmospheric Chemistry and Meteorology. T. E. Graedel and L. A. Farrow, *Ozone Chemistry and Technology*, ed. J. S. Murphy and J. R. Orr, Philadelphia: Franklin Institute Press, 1975, pp. 165–175. The chemistry of ozone is closely related to virtually every gas phase chemical process that occurs in the troposphere and stratosphere of the earth. This paper reviews the current knowledge of ozone sources and sinks for the urban troposphere, the rural troposphere, the natural stratosphere, and the perturbed stratosphere.

The Synthesis and Characterization of Some Oxide Fluorides of Rhenium and Osmium. W. A. Sunder and F. A. Stevie, J. Fluorine Chem., *6* (November 1975), p. 449. Existing synthetic methods for oxide fluorides of rhenium and osmium have been reviewed. New syntheses, using static heating, have been developed for OsO_3F_2, OsO_2F_3, $OsOF_5$, $OsOF_4$, ReO_3F, ReO_2F_3, $ReOF_5$, and $ReOF_4$. The products were characterized principally by mass spectroscopy, with supporting information for X-ray powder diffraction, chemical analysis, and molecular beam deflection.

ELECTRICAL AND ELECTRONIC ENGINEERING

Using Discretionary Telecommunications. D. Gillette, IEEE Trans. Commun., *COM-23* (October 1975), pp. 1054–1058. Continuing technical effort can help reduce the cost of telecommunications and add opportunities for their use. However, the biggest task in application is organizing institutions and procedures to use existing telecommunications systems and information technologies effectively.

MATERIALS SCIENCE

Lead Alloys for High Temperature Soldering of Magnet Wire. W. G. Bader, Welding Journal, *54* (October 1975), Research Supplement, pp. 370-s to 375-s. Lead-tin solders were evaluated for use in high-temperature soldering of fine gauge, polyurethane-insulated, copper-magnetic wire. The dissolution rates of copper by molten solders were determined at temperatures to 900°F and the reduction of these rates by copper additions to the solder. Also, wetting of copper by the solders and solder joint appearance were evaluated.

GENERAL MATHEMATICS AND STATISTICS

Explicit Construction of Invariant Measures for a Class of Continuous State Markov Processes. S. Halfin, Ann. Prob., *3* (October 1975), pp. 859–864. An explicit construction of invariant measures for a certain class of continuous-state Markov processes is presented. A special version of these processes is of interest in the theory of representation of real numbers (β-expansions). Previous results of Rényi and Parry are generalized, and an open problem of Parry is resolved.

Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. R. L. Obenchain, Technometrics, *17* (November 1975), pp. 431–441 (with discussion by G. C. McDonald, pp. 443–445). Ridge analysis is a "new" form of multiple linear regression which can be helpful when the data are ill-conditioned (nearly multicollinear) and least-squares coefficients are highly intercorrelated. Utilizing the likelihood function for mean-squared-error optimality under normal distribution, a statistical test can detect situations where ridge analysis will be worthwhile.

PHYSICS

Aspects of the Band Structure of $CuGaS_2$ and $CuGaSe_2$. B. Tell and P. M. Bridenbaugh, Phys. Rev. B, *12* (October 15, 1975), pp. 3330–3335. The spin-orbit splitting has been determined in the sulfur-rich section of the system $CuGaSe_{2-2x}Se_{2x}$, which demonstrates that the spin-orbit splitting is negative in $CuGaS_2$. A model which provides adjustable coupling and separation between the p- and d-like valence band can account for the main features of the band structure of $CuGaS_2$ and $CuGaSe_2$.

Excitation of Transversely Excited CO_2 Waveguide Lasers. O. R. Wood II, P. W. Smith, C. R. Adams, and P. J. Maloney, Appl. Phys. Letters, *27* (November 15, 1975), pp. 539–541. Using a preionization scheme based on the Malter effect, small-signal gains >5%/cm at 10.6 μm have been produced in a 1-mm^2 cross-section waveguide CO_2 amplifier at total operating pressures of 0.1 to 1 atmosphere. Comparisons between this preionization scheme and those using electron beams are made.

Dynamic Spectroscopy and Subpicosecond Pulse Compression. E. P. Ippen and C. V. Shank, Appl. Phys. Letters, *27* (November 1, 1975), pp. 488–490. Picosecond pulses from a mode-locked cw dye laser have been compressed in time to produce pulses as short as a few tenths of a picosecond. Dynamic spectroscopic investigations of the laser pulses reveal temporal asymmetry and frequency chirping on a subpicosecond time scale.

Frequency Dependence of the Electron Conductivity in the Silicon Inversion Layer in the Metallic and Localized Regimes. S. J. Allen, Jr., D. C. Tsui, and F. DeRosa, Phys. Rev. Letters, *35* (November 17, 1975), pp. 1359–1362. The conductivity of electrons in the inversion layer of silicon has been measured from 0 to 40 cm^{-1} at 1.2°K in the metallic and localized regimes. The correlation between $\sigma(T)$ and $\sigma(\omega)$ in the localized regime suggests that the drop in conductivity at low electron concentrations is caused by the appearance of a gap at the Fermi level.

Elasticity Measurements in the Layered Dichalcogenides $TaSe_2$ and $NbSe_2$. M. Barmatz, L. R. Testardi, and F. J. Di Salvo, Phys. Rev. B, *12* (November 15, 1975), pp. 4367–4376. The Young's modulus and internal friction exhibit large anomalies at the commensurate charge-density wave (CDW) transition in 2H-$TaSe_2$. Hysteresis

effects (~5K) verify the first-order nature of this transition. The incommensurate CDW transitions and the superconducting transition in 2H-$NbSe_2$ show weak elastic anomalies with essentially no hysteresis effects.

Interdiffusions in Thin-Film Au on Pt On GaAs (100) Studied with Auger Spectroscopy. C. C. Chang, S. P. Murarka, V. Kumar, and G. Quintana, J. Appl. Phys., *46* (October 1975), pp. 4237–4243. Pt/GaAs heated *in vacuum* reacted initially by rapid Ga migration into Pt and formation of an As-rich layer at the Pt/GaAs interface. Ga eventually traveled entirely through even 9000 Å Pt films, while As always stopped abruptly about $\frac{2}{3}$ way into the Pt. No Au was detected (<1 atom percent) in the Pt or GaAs after extensive Pt-GaAs reaction in Au/Pt/GaAs. Pt/GaAs heated *in air* behaved similarly, but developed a Ga-O layer over the Pt and an oxygen-rich layer at the Pt/GaAs interface.

Low-Threshold Room-Temperature Double-Heterostructure $GaAs_{1-x}Sb_x/Al_yGa_{1-y}As_{1-x}Sb_x$ Injection Lasers at 1-μm Wavelengths. R. E. Nahory and M. A. Pollack, Appl. Phys. Letters, *27* (November 15, 1975), pp. 562–564. Double-heterostructure (DH) injection lasers based on the $GaAs_{1-x}Sb_x/Al_yGa_{1-y}As_{1-x}Sb_x$ system have been fabricated using liquid phase epitaxial growth techniques and operated at room temperature at wavelengths in the 1-μm region. The observed room-temperature threshold current densities, as low as 2100 A cm^{-2}, are comparable to those of GaAs/AlGaAs devices of similar geometry.

Observation of Resonance Radiation Pressure on an Atomic Vapor. J. E. Bjorkholm, A. Ashkin, and D. B. Pearson, Appl. Phys. Letters, *27* (November 15, 1975), pp. 534–537. We have used the resonance radiation pressure from 40 mW of cw dye laser light propagating axially down a tube filled with sodium vapor to increase the sodium pressure (density) up to 50 percent over a length of 20 cm. The magnitude of the effect agrees well with measurements of the absorbed power.

Optical Pumping in Nitrogen Doped GaP. R. F. Leheny and Jagdeep Shah, Phys. Rev. B, *12* (October 15, 1975), pp. 3268–3274. Absorption saturation at the A bound exciton in GaP:N is described for a pulsed pump laser tuned directly to this absorption line and for a pump laser tuned above the indirect absorption edge. The second measurement yields 10-percent capture efficiency for N impurity. These measurements are analyzed by a model three-level system for the bound exciton by states.

Physical Properties of Poly(vinylchloride)-Copolyester Thermoplastic Elastomer Mixtures. T. Nishi, T. K. Kwei, and T. T. Wang, J. Appl. Phys., *46* (October 1975), pp. 4157–4165. A study was made on the compatibility, thermal behavior, and mechanical properties of the poly(vinylchloride) blended with copolyester thermoplastic elastomer. Results from NMR, thermal expansion, tensile test, and dynamic mechanical measurements indicate extensive mixing of the segments of two polymers.

Torsional-Mode Losses at Contacts Between Homogeneous Fiber Waveguides and Supporting Structures. R. L. Rosenberg and G. D. Boyd, J. Appl. Phys., *46* (November 1975), pp. 4654–4658. The losses from an ultrasonic torsional wave in a homogeneous fiber that are caused by contacts with fiber supports are found to depend primarily on contact area for a wide range of contact forces and materials. The associated force, compliance, and frequency dependencies are used to evaluate long-waveguide potentialities.

Volume Holograms in Photochromic Materials. W. J. Tomlinson, Appl. Opt., *14* (October 1975), pp. 2456–2467. Theoretical expressions are derived describing the process of writing volume (or thick) hologram gratings in photochromic materials. The theory includes the effects of the saturation of the material response, scattering of the writing beams by the partially written hologram, and the refractive index changes that accompany the photoinduced absorption changes.