THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 50 May-June 1971 Number 5

Stability of Frequency Feedback Receivers Under Steps in Input Frequency

By V. E. BENEŠ

(Manuscript received January 19, 1971)

It is known that frequency feedback demodulators can show instability in their response to step changes (mistuning) in input frequency. This work reports on some mathematical analyses of this phenomenon as described by differential equations arising from simple IF and feedback filters in the demodulator. These equations are studied for local and global stability by geometric or phase-plane analysis, by means of Lyapunov functions, and by the topological Poincaré-Bendixson methods. A typical result is for the case of no feedback filter and one-pole baseband analog of the IF filter, and states in physical terms that if the mistuning is not too big, specifically if

$$|\ \textit{mistuning}\ | < (\textit{half-power IF bandwidth})(1 + \textit{feedback gain})$$

then solutions which are bounded away from zero amplitude approach the natural equilibrium point. Examples are given in which a sufficiently large mistuning makes the equilibrium point unstable.

I. INTRODUCTION

The frequency feedback (or frequency compression) demodulator for FM signals was proposed by J. G. Chaffee[1] in 1937. After some twenty-five years, Chaffee's idea was found to have a particularly fitting application in the satellite communications experiments Echo[2] and Telstar,[3] in which there was a high premium on detecting a low-power wide-band FM signal in noise. Nevertheless, since its invention, little progress has been made in the mathematical analysis of this circuit. Approximate methods of analysis and synthesis have been proposed, and some of them experimentally verified as useful.[4,5] However, except for unpublished works by S. O. Rice and T. R. Williams, the nonlinear character of the circuit away from equilibrium positions has not been considered.

It is the aim of this paper to formulate briefly one of the problems arising in the analysis of the FM with feedback (FMFB) receiver, namely that of stability of its response to step changes in input frequency. We shall write equations describing this response and present results about local and global stability of solutions for simple cases.

II. CIRCUIT DESCRIPTION

The FMFB receiver has been extensively discussed in recent publications,[4,5] so only a brief description of it is included here. Roughly speaking, the receiver is a conventional FM demodulator, with a local oscillator whose frequency is controlled linearly by the output of the detector. The object of this control is to reduce the index of modulation at the output of the mixer, so as to be able to use a narrower IF filter than in a conventional FM receiver, and thus to eliminate some of the noise accompanying the input signal. The action of the circuit is to follow the slowly varying frequency of an FM wave while looking at it through a moving narrow frequency "window."

The circuit is closely related to the phase-locked oscillator, but it is distinguished from that device by having amplitude effects absent in the latter. Mathematically this distinction takes the form that in FMFB there is an amplitude variable for every phase variable, while in phase-lock these variables do not appear. Their presence critically affects and complicates circuit analysis: thus the simplest FMFB equation is in two dimensions, while the simplest phase-lock equation is the pendulum equation, in one. The FMFB receiver resembles the phase-locked oscillator in that both devices work by phase-locking onto an FM wave that varies slowly over a limited range; if this range is exceeded locking fails and oscillation can set in. This phenomenon is well-known in phase-locked oscillators; in FMFB receivers a similar behavior has been described by L. H. Enloe.[4] It is to this stability problem that we address ourselves, endeavoring an analytical study of the stability of simple differential equations describing the mistuning of the incoming signal away from the normal carrier frequency.

A typical result we prove states that if the mistuning ω_d is not too big, specifically, in physical terms, if (for a one-pole baseband analog of the IF filter)

$$|\omega_d| < \text{(half-power IF bandwidth)}(1 + \text{feedback gain}),$$

then, for the simplest receiver, solutions which are bounded away from zero amplitude approach the equilibrium or critical point. This and similar results are proved by using the Poincaré-Bendixson theory, or with the help of Lyapunov functions.

III. EQUATIONS FOR RECEIVER WITH IDEAL DETECTOR

We shall write equations for the FMFB receiver (see Fig. 1) under the assumption that it contains an ideal frequency detector. That is, we assume that if the signal leaving the IF filter is $a(t) \cos (\omega t + \theta(t))$, then the detector produces the output $\dot{\theta}(t)$. Let the mixer input be

$$x_c \cos \omega_1 t - x_s \sin \omega_1 t,$$

and let the mixer multiply this input by

$$2 \cos (\omega_2 t + \beta\varphi(t))$$

where β (in practice and here > 0) is the feedback gain, and $\dot{\varphi}$ is the feedback signal. It is assumed that the IF filter is tuned to the difference frequency $\omega = \omega_1 - \omega_2$, and can be represented by an impulse response

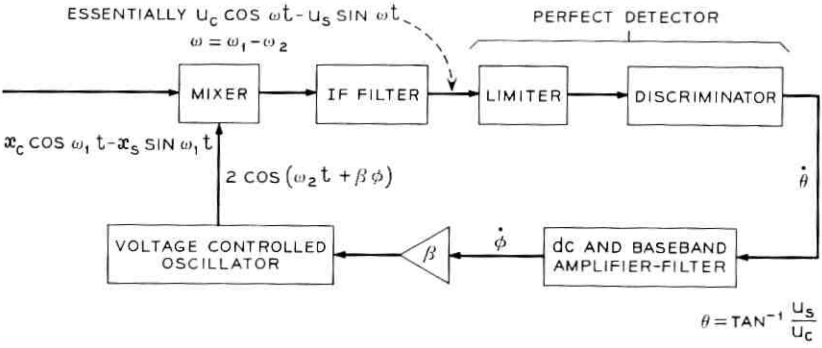
ESSENTIALLY $u_c \cos \omega t - u_s \sin \omega t$
$\omega = \omega_1 - \omega_2$
PERFECT DETECTOR
MIXER
IF FILTER
LIMITER
DISCRIMINATOR
$x_c \cos \omega_1 t - x_s \sin \omega_1 t$
$2 \cos (\omega_2 t + \beta\phi)$
$\dot{\theta}$
VOLTAGE CONTROLLED OSCILLATOR
β
$\dot{\phi}$
dc AND BASEBAND AMPLIFIER-FILTER
$\theta = \tan^{-1} \frac{u_s}{u_c}$

Fig. 1—FMFB receiver, block diagram.

of the form $2f(t) \cos \omega t$, with $f(\cdot)$ a baseband response such that $f(t) = 0$ for $t < 0$. The sum $(\omega_1 + \omega_2)$ components of the mixer are essentially removed by the IF filter, and will be ignored. The difference $(\omega_1 - \omega_2)$ components at the output of the mixer are

$$\cos \omega t\{x_c \cos \beta\varphi + x_s \sin \beta\varphi\} - \sin \omega t\{x_s \cos \beta\varphi - x_c \sin \beta\varphi\}.$$

The response of the IF filter to these components has the form

$$\cos \omega t \int_0^t f(t-u)\{x_c(u) \cos \beta\varphi(u) + x_s(u) \sin \beta\varphi(u)\}\, du$$

$$- \sin \omega t \int_0^t f(t-u)\{x_s(u) \cos \beta\varphi(u) - x_c(u) \sin \beta\varphi(u)\}\, du$$

$$+ \text{ terms around } 2\omega$$
$$+ \text{ terms representing initial conditions.}$$

We shall assume that the passband of $f(\cdot)$ is small compared to 2ω, so that the components around 2ω may be ignored as well.

To complete the loop equations we must indicate how the feedback $\varphi(\cdot)$ is determined from the output of the IF filter. We set, for $t \geq 0$

$$u_c(t) = r_c(t) + \int_0^t f(t-u)\{x_c(u) \cos \beta\varphi(u) + x_s(u) \sin \beta\varphi(u)\}\, du,$$

$$u_s(t) = r_s(t) + \int_0^t f(t-u)\{x_s(u) \cos \beta\varphi(u) - x_c(u) \sin \beta\varphi(u)\}\, du,$$

where $r_c(\cdot)$, $r_s(\cdot)$ represent the effects of initial conditions in the filter at $t = 0$. Exclusive of the carrier, the angle modulation of the IF output,

$$u_c \cos \omega t - u_s \sin \omega t,$$

is just

$$\theta = \tan^{-1} \frac{u_s}{u_c},$$

corresponding to an instantaneous frequency,

$$\dot{\theta} = \frac{u_c \dot{u}_s - u_s \dot{u}_c}{a^2},$$

where $a = (u_c^2 + u_s^2)^{1/2}$. This is the output of the ideal detector.

The feedback frequency $\dot{\varphi}(\cdot)$ controlling the voltage-controlled oscillator is obtained by filtering $\dot{\theta}(\cdot)$. Thus

$$\dot{\varphi}(t) = r(t) + \int_0^t k(t-u)\dot{\theta}(u)\,du, \qquad t \geq 0$$

where $k(\cdot)$ is the impulse response of the feedback filter, and $r(\cdot)$ represents the effect of initial conditions at $t = 0$.

IV. DIFFERENTIAL EQUATION FOR THE SIMPLEST CASE

When the baseband responses $f(\cdot)$ and $k(\cdot)$ correspond to filters with rational transfer functions, the integral equations for $u_c(\cdot)$ and $u_s(\cdot)$ can be turned into differential equations in a well-known way. In the simplest case, when there is no feedback filter and $f(\cdot)$ corresponds to a (one-pole no-zero) filter with transfer $\mu/(\lambda + s)$, we obtain the equations

$$\begin{aligned}
\dot{u}_c &= -\lambda u_c + \mu[x_c \cos \beta\theta + x_s \sin \beta\theta] \\
\dot{u}_s &= -\lambda u_s + \mu[x_s \cos \beta\theta - x_c \sin \beta\theta] \\
\dot{\theta} &= \mu \frac{u_c(x_s \cos \beta\theta - x_c \sin \beta\theta) - u_s(x_c \cos \beta\theta + x_s \sin \beta\theta)}{u_c^2 + u_s^2}.
\end{aligned}$$

The introduction of polar coordinates $u_c = a \cos \theta$, $u_s = a \sin \theta$ simplifies these equations to

$$\begin{aligned}
\dot{\theta} &= \frac{\mu}{a}(x_s \cos(\beta+1)\theta - x_c \sin(\beta+1)\theta) \\
\dot{a} &= -\lambda a + \mu(x_c \cos(\beta+1)\theta + x_s \sin(\beta+1)\theta).
\end{aligned} \tag{1}$$

We first consider the stability of equations (1) when the input to the demodulator consists of the carrier $\cos \omega_1 t$ alone, with no signal. In this case $x_c \equiv 1$, $x_s \equiv 0$, and the equations are

$$\begin{aligned}
\dot{\theta} &= -\frac{\mu}{a} \sin(\beta+1)\theta \\
\dot{a} &= -\lambda a + \mu \cos(\beta+1)\theta.
\end{aligned} \tag{2}$$

We recall that the critical points of a differential equation $\dot{x} = v(x)$ are the points x in the phase-space at which $v(x) = 0$. Those of the system (2) are then the points in the a, θ plane at which simultaneously $\dot{\theta} = \dot{a} = 0$, namely,

$$a = \frac{\mu}{\lambda}, \qquad \theta = \frac{2n\pi}{\beta+1}, \qquad n \text{ an integer}.$$

Because of the periodic dependence of the right-hand side of (2) on θ, it is possible and convenient to define $\zeta = (\beta + 1)\theta$, to write (2) as

$$\dot{\zeta} = -(\beta + 1)\frac{\mu}{a}\sin\zeta \qquad (3)$$

$$\dot{a} = -\lambda a + \mu\cos\zeta,$$

and to consider only principal values of ζ, and thus only the critical point $(\mu/\lambda, 0)$ in the plane specified by the polar coordinates (a, ζ).

Theorem 1: The equations (3) are globally asymptotically stable for all positive λ and μ; all solutions tend to the critical point μ/λ, 0 in an exponential manner; ζ is monotone, and a is either monotone or has one minimum.

Proof: We start with an heuristic direct analysis of the trajectories. Consider in Fig. 2 the circle C in the a, ζ plane defined by $\dot{a} = 0$, that is, $a = \mu/\lambda\cos\zeta$. With $x = a\cos\zeta$, $y = a\sin\zeta$ we shall examine the directions of the trajectories of (3) at points on C. The equation of C is

$$y = \left(\frac{\mu x}{\lambda} - x^2\right)^{\frac{1}{2}}.$$

Since C is the locus $\dot{a} = 0$, it is apparent that on C each trajectory (has a tangent that) is perpendicular to the radius from the origin.

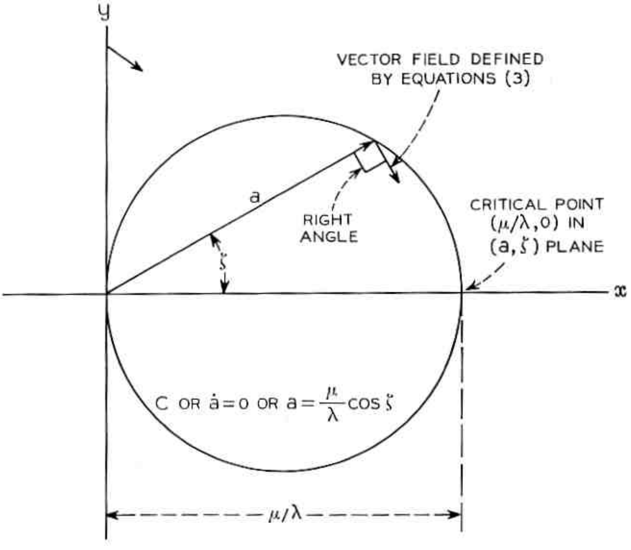
y
VECTOR FIELD DEFINED BY EQUATIONS (3)
a
RIGHT ANGLE
CRITICAL POINT (μ/λ,0) IN (a,ζ) PLANE
ζ
x
C OR ȧ=0 OR a=μ/λ cos ζ
μ/λ

Fig. 2—Phase plane for no mistuning.

By symmetry about the x-axis, we can restrict attention to $y \geqq 0$. The slope of C at x, y is

$$\frac{dy}{dx} = \frac{1}{2}\left(\frac{\mu x}{y} - x^2\right)^{-\frac{1}{2}}\left(\frac{\mu}{\lambda} - 2x\right) = \frac{1}{2y}\left(\frac{\mu}{\lambda} - 2x\right),$$

while the slope of the line through x, y perpendicular to the radius from the origin is

$$\frac{y}{x - \dfrac{\mu}{\lambda}}.$$

Since for $0 < x < \mu/\lambda$,

$$\left(\frac{\mu}{\lambda} - 2x\right)\left(x - \frac{\mu}{\lambda}\right) < 2\left(\frac{\mu x}{\lambda} - x^2\right),$$

we find

$$\frac{\mu}{\lambda} - 2x > \frac{2y^2}{x - \dfrac{\mu}{\lambda}},$$

or since $y > 0$ and $0 < x < \mu/\lambda$

$$\frac{dy}{dx} = \frac{1}{2y}\left(\frac{\mu}{\lambda} - 2x\right) > \frac{y}{x - \dfrac{\mu}{\lambda}}.$$

Thus every trajectory is entering C on $\dot{a} = 0$ except at the origin and at the critical point. ζ is decreasing in $\zeta > 0$. If a trajectory ever crosses C it can never again recross it and must approach the critical point; in this case a has a single minimum. If a trajectory never crosses C, it must simply slip into the critical point, because then both a and ζ decrease and are bounded below.

These preliminaries lead us to define the Lyapunov function

$$V = \frac{1}{2}\left(\frac{\mu}{\lambda} - a \cos \zeta\right)^2 + \frac{1}{2}(a \sin \zeta)^2$$

$$= \frac{1}{2}\,(\text{distance from } a,\, \zeta \text{ to critical point})^2$$

Evidently $V \geqq 0$, and $V = 0$ only at μ/λ, 0. The rate of change of V along trajectories of (3) is

$$\dot{V} = a\dot{a} - \dot{a}\,\frac{\mu}{\lambda}\cos \zeta + a\,\frac{\mu}{\lambda}\,\dot{\zeta}\sin \zeta$$

$$= -\lambda\left(a - \frac{\mu}{\lambda}\cos\zeta\right)^2 - \mu^2\frac{\beta+1}{\lambda}\sin^2\zeta$$

$$= -2\lambda V - \frac{\mu^2\beta}{\lambda}\sin^2\zeta < 0$$

except at the critical point, where $V = \dot{V} = 0$. It follows from Theorem II, p. 37 of Ref. 6, that the system (3) is globally asymptotically stable: all solutions tend exponentially to the critical point with reciprocal time constant 2λ. When $\lambda = \mu$, 2λ has the physical interpretation

$$2\lambda = 2 \times \text{(half-power IF bandwidth)}.$$

V. MISTUNING IN THE SIMPLEST CASE

Let us assume that in equation (1) we have

$$x_s = \sin\omega_d t, \qquad x_c = \cos\omega_d t,$$

corresponding to the "mistuned" carrier input $\cos(\omega_1 + \omega_d)t$, or to the constant modulating signal ω_d. The equation (1) assumes the form

$$\dot{\theta} = \frac{\mu}{a}\sin(\omega_d t - (\beta+1)\theta)$$

$$\dot{a} = -\lambda a + \mu\cos(\omega_d t - (\beta+1)\theta),$$

or with $\zeta = \omega_d t - (\beta + 1)\theta$,

$$\dot{\zeta} = \omega_d - \frac{\mu(\beta+1)}{a}\sin\zeta \qquad (4)$$

$$\dot{a} = -\lambda a + \mu\cos\zeta.$$

The critical point of this system is determined by the conditions

$$a = \frac{\mu}{\lambda}\cos\zeta, \qquad \zeta = \tan^{-1}\frac{\omega_d}{\lambda(\beta+1)} \qquad (5)$$

It is important to note that because of the possibility of going to low amplitudes there always exist critical points, regardless of the value of ω_d. This situation is in sharp contrast with the phase-locked oscillator. For a filterless phase-locked oscillator the equation corresponding to (4) would be

$$\dot{\zeta} = \omega_d - \mu\sin\zeta,$$

which has no critical point if $\omega_d > \mu$. Thus in phase-lock there is usually a critical frequency deviation above which locking is impossible for lack of critical points, and below which it may or may not occur. In the FMFB receiver, though, the critical points always exist but, as we shall see later, they are not always stable.

We determine the stability of the critical point (5) by the standard method of linearization. The matrix

$$\begin{bmatrix} \frac{\partial}{\partial \zeta} \dot{\zeta} & \frac{\partial}{\partial a} \dot{\zeta} \\ \frac{\partial}{\partial \zeta} \dot{a} & \frac{\partial}{\partial a} \dot{a} \end{bmatrix} = \begin{bmatrix} -\frac{\mu(\beta + 1)}{a} \cos \zeta & \frac{\mu(\beta + 1)}{a^2} \sin \zeta \\ -\mu \sin \zeta & -\lambda \end{bmatrix}$$

of partial derivatives, evaluated at the critical point, is the matrix A appropriate for the linearized system. The determinant of $(sI\text{-}A)$ turns out to be

$$s^2 + s\lambda(\beta + 2) + \lambda^2(\beta + 1) + \frac{\omega_d^2}{\beta + 1},$$

with roots in the left half-plane. Hence the critical point is stable; in a neighborhood of it the trajectories approach it.

Because of the symmetry of the equations, there is no loss of generality in assuming, as we do henceforth, that $\omega_d < 0$. This convention is used in Figs. 3 and 4.

Although we have not proved it, it is natural (and we conjecture) that a separatrix lies between solutions which pass around the origin in the upper half of the a, ζ plane, and solutions which just miss the origin as they go past it in the lower half of the plane. Roughly speaking, the former pick up an extra 2π of phase before settling. This separatrix may even be a fan[7] of solutions each of which goes into the origin, although in all likelihood it consists of a single trajectory. This conclusion is supported by a heuristic low-amplitude analysis of equations (4) suggested by J. A. Morrison. He writes (4) as the single equation

$$\frac{d\zeta}{da} = \frac{a\omega_d - \mu(\beta + 1) \sin \zeta}{\mu a \cos \zeta - \lambda a^2},$$

and for $|\cos \zeta| \cong 1$ drops terms of order a^2, obtaining

$$\mu a \cos \zeta \frac{d\zeta}{da} = a\omega_d - \mu(\beta + 1) \sin \zeta = \mu a \frac{d}{da} \sin \zeta,$$

whence by integration from a_0 to a

$$\sin \zeta - \frac{\omega_d a}{\mu(\beta + 2)} = \left(\frac{a_0}{a}\right)^{\beta+1} \left(\sin \zeta_0 - \frac{\omega_d a_0}{\mu(\beta + 2)}\right).$$

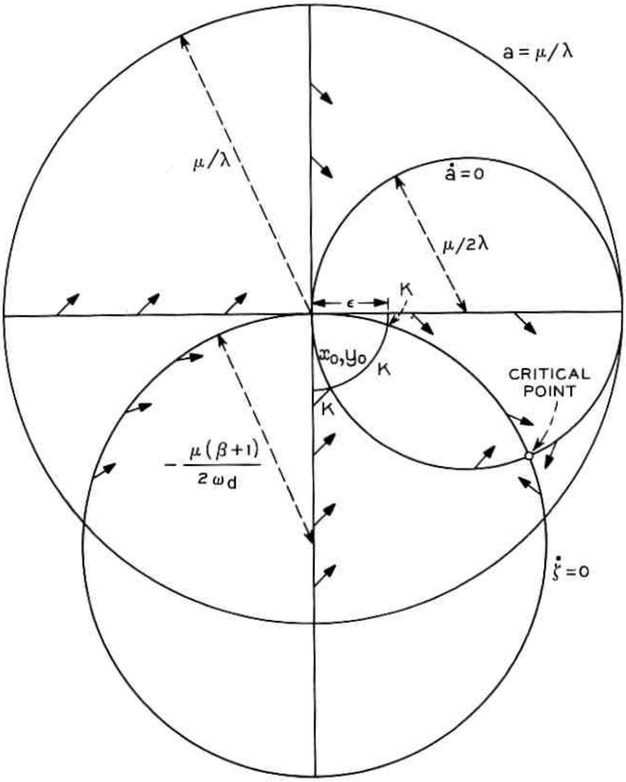

Fig. 3—Phase plane for mistuning.

This formula suggests that if a point (a_0, ζ_0), on the circle

$$a = \frac{\mu(\beta + 2)}{\omega_d} \sin \zeta$$

and near the origin, is on a trajectory then a nearby point on the circle is on the same trajectory. In other words, a trajectory going through the origin does so like the circle above, which is tangent to but outside the circle $\dot{\zeta} = 0$ with equation

$$a = \frac{\mu(\beta + 1)}{\omega_d} \sin \zeta.$$

To avoid difficulties we shall consider only trajectories which are bounded away from the origin.

We now address ourselves to the global stability of the mistuned equations (4). Since the system is two dimensional it is possible to try to use the topological Poincaré-Bendixson methods.[7] This is most conveniently done by passing to rectangular coordinates $x = a \cos \zeta$, $y = a \sin \zeta$ again, and calculating the divergence. We rewrite (4) as

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} -\lambda x + \dfrac{\mu x^2}{x^2 + y^2} + \mu(\beta + 1) \dfrac{y^2}{x^2 + y^2} - \omega_d y \\ -\lambda y - \dfrac{\mu \beta x y}{x^2 + y^2} + \omega_d x \end{pmatrix}$$

$$= v(x, y)$$

We find

$$\frac{\partial \dot{x}}{\partial x} = -\lambda - 2\mu\beta \frac{xy^2}{(x^2 + y^2)^2}$$

$$\frac{\partial \dot{y}}{\partial y} = -\lambda - \mu\beta x \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

$$\operatorname{div} v = -2\lambda - \frac{\mu\beta x}{x^2 + y^2}$$

$$= -2\lambda - \frac{\mu\beta}{a} \cos \zeta .$$

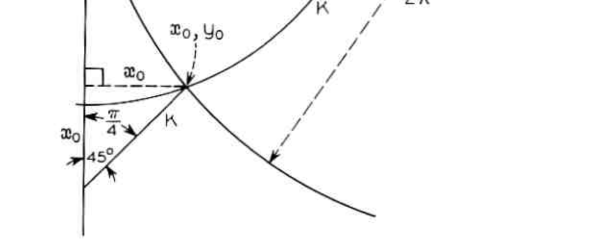

Fig. 4—Details of the curve K.

Lemma 1: In the portion of the fourth quadrant comprised by an arbitrary neighborhood of the origin there is always a curve K, joining the positive x-axis to the negative y-axis, such that on K trajectories of (4) cross K in the outward direction, i.e., out of the part cut off by K near the origin, and into the part separated by K from the origin. (*See Fig. 3.*)

Proof: Let K consist of the circle $a = \epsilon$ from the x-axis down to the point where this circle crosses $\dot{a} = 0$, i.e., until $\cos \zeta = \epsilon\lambda/\mu$. The Cartesian coordinates of this point are

$$x_0 = \frac{\lambda\epsilon^2}{\mu}, \qquad y_0 = -\frac{\epsilon}{\mu}\sqrt{\mu^2 - \epsilon^2\lambda^2}$$

From here let K continue to the y-axis at slope 1, i.e., let it consist of that portion of the line

$$y = x - \frac{\epsilon}{\mu}\sqrt{\mu^2 - \epsilon^2\lambda^2} - \frac{\lambda\epsilon^2}{\mu}$$

which is between its intercept $y_0 - x_0$ on the y-axis and the circle $\dot{a} = 0$.

Now on $a = \epsilon$ inside $\dot{a} = 0$ we have $\dot{a} > 0$, so on the circular part of K all trajectories are entering $a > \epsilon$, even at x_0, y_0. At x_0, y_0 the trajectory is actually tangent to $a = \epsilon$, but for small enough ϵ it is pointed in the direction of increasing x and so there too it must enter $a > \epsilon$.

On the linear part of K we want to verify that

$$\frac{dy}{dx} = \frac{-\lambda y + \omega_d x - \mu\beta \dfrac{xy}{x^2 + y^2}}{-\lambda x + \mu - \omega_d y + \mu\beta \dfrac{y^2}{x^2 + y^2}} < 1$$

if ϵ is small enough. This is true because on the linear part of K we have $x \rightarrow 0$, $y \rightarrow 0$, and

$$\cos \zeta = \frac{x}{\sqrt{x^2 + y^2}} \rightarrow 0$$

all monotonely and uniformly, as $\epsilon \rightarrow 0$; near $\epsilon = 0$ the denominator of dy/dx is close to $\mu(\beta + 1)$ for $(x, y) \in K$, so on the linear part of K the trajectories are moving in the direction of increasing x at a slope $dy/dx < 1$; hence they are crossing K in the direction of increasing amplitude. This proves the lemma.

Theorem 2: If $|\omega_d| < \lambda(\beta + 1)$, *then every trajectory of (4) that is bounded away from the origin approaches the critical point* $a = \mu/\lambda \cos \zeta$, $\zeta = \tan^{-1} \omega_d/\lambda(\beta + 1)$.

Proof: It suffices to consider only $\omega_d < 0$. All trajectories outside $a = \mu/\lambda$ have $\dot{a} < 0$, so it is enough to consider those starting inside, because the others get there eventually. Consider a path starting in $0 \leqq \zeta \leqq 3\pi/2$, $\dot{\zeta} \leqq 0$, $a \leqq \mu/\lambda$, and bounded away from the origin. Either it stays in this region forever, or it reaches the fourth quadrant, or else it moves into $\dot{\zeta} > 0$. If it stays there forever then there is a closed region, free of critical points and excluding the origin, in which it stays. By a result of Poincaré (Ref. 7, p. 232), this closed region also contains a closed path γ, of period say τ. Then because γ is closed, if $\zeta(0)$ is on γ,

$$\zeta(\tau) - \zeta(0) = \int_0^\tau \dot{\zeta}(t)\, dt = 0.$$

But this is impossible since $\dot{\zeta} < 0$ throughout the region in question.

In the third quadrant a trajectory can cross $\dot{\zeta} = 0$ only once. The argument just given also shows that no path bounded away from the origin can stay in the region $\pi \leqq \zeta \leqq 3\pi/2$, $\dot{\zeta} > 0$, $a \leqq \mu/\lambda$. Thus all paths starting in the region $a \leqq \mu/\lambda$ and bounded away from the origin reach the fourth quadrant.

The inequality in the hypothesis implies that the circle $\dot{\zeta} = 0$ intersects the circle $a = \mu/\lambda$. Away from the origin we have $\dot{\zeta} < 0$ for $x = 0$, $y > 0$ and $\dot{\zeta} > 0$ for $x = 0$, $y < 0$. On $0 < x \leqq \mu/\lambda$, $y = 0$ the trajectories enter the fourth quadrant intersected with $a \leqq \mu/\lambda$. Since a is nonincreasing on $a = \mu/\lambda$, it follows from Lemma 1 that there is a region R with these properties:

(i) R is closed.
(ii) $R \subseteq \{a \leqq \mu/\lambda\} \cap$ fourth quadrant.
(iii) $(\{a \leqq \mu/\lambda\} \cap$ fourth quadrant $- R)$ is in an arbitrarily small neighborhood of the origin.
(iv) R is entered by the path under consideration.
(v) R is invariant, i.e. maps into itself under the motion.

Indeed R can be chosen to be a 2-cell (homeomorph of a disc). We note that the divergence is negative throughout R. It follows from the criterium of Bendixson (Ref. 7, p. 238) that R contains no limit cycles nor even an oval going to and from a critical point. Since R contains only one critical point, it can contain no path-polygon. Thus two of the three alternatives in Bendixson's theorem (Ref. 7, p. 230) are ruled out, and all paths starting in R go to the critical point. Since we can associate a region like R with any trajectory bounded away from the origin, the theorem is proved.

We remark that when $\lambda = \mu$ the condition of the theorem can be rendered in physical terms as

| mistuning | ≦ (half-power IF bandwidth)(1 + feedback gain).

We next show that a result similar to Theorem 2 can be obtained by a Lyapunov function argument.

Theorem 3: If

$$|\omega_d| < \frac{2\lambda(\beta + 1)^{\frac{3}{2}}}{\beta},$$

then every trajectory of (4) that is bounded away from the origin approaches the critical point $a = \mu/\lambda$, $\xi = \tan^{-1} \omega_d/\lambda(\beta + 1)$.

Proof: Consider the scalar function V defined by

$$\begin{aligned} 2V &= (\lambda a - \mu \cos \zeta)^2 + (\beta + 1)^{-2}(a\omega_d - \mu(\beta + 1) \sin \zeta)^2 \\ &= \dot{a}^2 + (\beta + 1)^{-2}(a\dot{\xi})^2. \end{aligned}$$

We find

$$\begin{aligned} \dot{V} = &-\lambda(\lambda a - \mu \cos \zeta)^2 \\ &+ \frac{\beta\omega_d}{(\beta + 1)^2}(\lambda a - \mu \cos \zeta)(a\omega_d - \mu(\beta + 1) \sin \zeta) \\ &- \frac{\lambda}{\beta + 1}(a\omega_d - \mu(\beta + 1) \sin \zeta)^2. \end{aligned}$$

$-\dot{V}$ is a quadratic form in $\dot{a}$ and $a\dot{\xi}$ with determinant

$$\begin{vmatrix} \lambda & \dfrac{-\beta\omega_d}{2(\beta + 1)^2} \\ \dfrac{-\beta\omega_d}{2(\beta + 1)^2} & \dfrac{\lambda}{\beta + 1} \end{vmatrix}$$

which is positive whenever $|\omega_d| < 2\lambda(\beta + 1)^{3/2}/\beta$.

Consider now a trajectory bounded away from the origin. It is clearly bounded, so it has a positive limiting set Γ^+ which is invariant and to which it tends. There is a constant k such that the trajectory is entirely contained in a bounded subregion Ω of $\{V < k\}$. Hence $\Gamma^+ \subseteq \Omega$ and $\dot{V} = 0$ on Γ^+. Since a is bounded away from 0 on the trajectory, it follows that $\dot{a} = 0$ and $\dot{\zeta} = 0$ on Γ^+. Thus the trajectory tends to the critical point. (This is a variant of the argument for Theorem VI, p. 58

of Ref. 6.) Again, the condition of the theorem is that $|\omega_d|$ not be too big, viz.,

$$|\omega_d| < 2(\text{half-power IF bandwidth}) \times (\beta + 1)^{3/2}/\beta,$$

where β is the feedback gain.

VI. ONE-POLE FILTER IN THE FEEDBACK

After the filterless case considered so far, the next simplest model for the FMFB receiver would have one-pole, no-zero filters both as the baseband equivalent $f(\cdot)$ of the IF response, and as the response $k(\cdot)$ in the feedback. This is the simplest case that has appreciable practical import: the IF filter corresponds closely to the one-mesh design described by Giger and Chaffee (loc. cit., p. 1119 and Fig. 5, p. 1120); the feedback filter and the gain β are a rudimentary version of the dc-and-baseband amplifier sketched by these authors (loc. cit. pp. 1121–22.)

In this case the differential equations for the system are

$$\begin{aligned}
\dot u_c &= -\lambda u_c + \mu(x_c \cos \beta\varphi + x_s \sin \beta\varphi) \\
\dot u_s &= -\lambda\mu_s + \mu(x_s \cos \beta\varphi - x_c \sin \beta\varphi) \\
\dot\theta &= a^{-2}(u_c \dot u_s - u_s \dot u_c) = \frac{d}{dt} \tan^{-1} \frac{u_s}{u_c} \\
\dot x &= -\gamma x + \delta\dot\theta \\
\dot\varphi &= x
\end{aligned}$$

with $a = (u_c^2 + u_s^2)^{\frac{1}{2}}$ as before, and $\delta/(\gamma + s)$ the transfer function of the feedback filter. Upon setting $\xi = \theta + \beta\varphi$ these simplify to

$$\begin{aligned}
\dot\xi &= \beta x + \frac{\mu}{a}(x_s \cos \xi - x_c \sin \xi) \\
\dot a &= -\lambda a + \mu(x_c \cos \xi + x_s \sin \xi) \\
\dot x &= -\gamma x + \frac{\delta\mu}{a}(x_s \cos \xi - x_c \sin \xi).
\end{aligned}$$

When the modulating signal is a constant ω_d, then $x_s(t) = \sin \omega_d t$, $x_c(t) = \cos \omega_d t$, and with $\zeta(t) = \omega_d t - \xi(t)$ the equations become

$$\dot\zeta = \omega_d - \beta x - \frac{\mu}{a} \sin \zeta$$

$$\begin{aligned} \dot{a} &= -\lambda a + \mu \cos \zeta \\ \dot{x} &= -\gamma x + \frac{\delta\mu}{a} \sin \zeta . \end{aligned} \tag{6}$$

Let us note heuristically and physically that if $\delta = \gamma \to \infty$ then the feedback bandwidth goes to ∞ and we obtain the equations (4) of the simplest case.

We start with a study of the stability of the critical point ζ_0 , a_0 , x_0 defined by

$$\begin{aligned} \zeta_0 &= \tan^{-1} \frac{\omega_d}{\lambda\left(1 + \dfrac{\beta\delta}{\gamma}\right)} \\ a_0 &= \frac{\mu}{\lambda} \cos \zeta_0 \\ x_0 &= \frac{\delta\omega_d}{\gamma + \beta\delta} . \end{aligned} \tag{7}$$

The matrix $A = (\partial f_i/\partial x_j)$ of partial derivatives evaluated at the critical point is

$$\begin{array}{c|ccc|} & \zeta & a & x \\ \dot{\zeta} & -\lambda & \dfrac{\lambda^2}{\mu} \dfrac{\tan \zeta_0}{\cos \zeta_0} & -\beta \\ \dot{a} & -\mu \sin \zeta_0 & -\lambda & 0 \\ \dot{x} & \delta\lambda & \dfrac{\delta\lambda^2}{\mu} \dfrac{\tan \zeta_0}{\cos \zeta_0} & -\gamma \end{array} .$$

The determinant of $(sI\text{-}A)$ is

$$\begin{aligned} (s + \lambda)((s + \lambda)(s + \gamma) + \beta\delta\lambda) + \delta^2\lambda^2\beta \tan^2 \zeta_0 + \lambda^2 \tan^2 \zeta_0(s + \gamma) \\ = s^3 + (2\lambda + \gamma)s^2 + (\lambda\gamma + \beta\delta\lambda + \lambda^2 \tan^2 \zeta_0)s \\ + \lambda^2(\beta\delta + \tan^2 \zeta_0(\beta\delta + \gamma)) \\ = s^3 + a_2 s^2 + a_1 s + a_0 . \end{aligned}$$

A necessary and sufficient condition for stability is that

$$a_1 , a_2 , a_0 > 0 \quad \text{and} \quad a_2 a_1 > a_0 .$$

The first three are clearly true, and the last amounts to

$$(2\lambda + \gamma)(2\lambda\gamma + \lambda^2 + \lambda\beta\delta) + \left(\frac{\gamma\omega_d}{\gamma + \beta\delta}\right)^2 > \lambda^2\beta\delta + \frac{\gamma^2\omega_d^2}{\gamma + \beta\delta} .$$

This is symmetric in $\pm\omega_d$, and is true for $|\omega_d|$ small enough. It becomes false for large $|\omega_d|$ if $2\lambda < \beta\delta$. If $\lambda = \mu$ and $\gamma = \delta$, then these numbers are the half-power bandwidths of the IF and feedback filter respectively, and we may say in physical terms that if

$$\beta = \text{feedback gain} < 2 \times \frac{\text{IF bandwidth}}{\text{feedback bandwidth}} = \frac{2\lambda}{\delta}$$

then a very large mistuning cannot affect the local stability of the system, but if β exceeds twice the ratio of IF to baseband widths then sufficiently large mistuning will make the system unstable. This result was first observed in an unpublished work (although with some errors) of T. R. Williams.

The global stability of the equations (6) for the case with a one-pole in the feedback is a far more difficult topic than the local. Naturally, as more complicated filters are assumed for the IF and the feedback, the dimension of the problem goes up, and the kind of geometric analysis we are using here becomes virtually impossible. In particular, the Poincaré-Bendixson theory used earlier is already unavailable in three dimensions, and also there seems to be no ready way to prove the boundedness of solutions. Nevertheless some information can be obtained from the construction of a Lyapunov function for the case of no mistuning; all attempts to extend the method to the case of mistuning have failed.

Theorem 4: If $\omega_d = 0$ (no mistuning) then every trajectory of (6) that is bounded away from the line $a = 0$ approaches the critical point given by (7).

Proof: Consider the scalar function V defined by

$$2V = \frac{\dot{a}^2}{\lambda} + \frac{a^2\dot{\zeta}^2}{\lambda + \gamma + \delta\beta} + \frac{\mu^2}{\lambda}\sin^2\zeta + \frac{\mu^2\cos^2\zeta}{\lambda + \gamma + \delta\beta}.$$

V is certainly positive along any trajectory satisfying the hypothesis. We find after a lot of elementary calculus that

$$\dot{V} = -(\dot{a})^2 - \frac{(\lambda + \gamma)a^2(\dot{\zeta})^2}{\lambda + \gamma + \delta\beta}.$$

Since the trajectory assumed in the theorem is bounded away from the line $a = 0$, it is easy to see from the equations (6) that it is bounded. Thus the positive limiting set of this trajectory is a nonempty, compact invariant set Γ^+, to which it tends. There is a constant k such that the trajectory is eventually in a bounded subregion Ω of $\{V < k\}$. Thus $\Gamma^+ \subseteq$ and $\dot{V} = 0$ on Γ^+. Consider now the largest invariant subset M of $\{\dot{V} = 0\} \cap \Omega$. Clearly $\Gamma^+ \subseteq M$. Thus the trajectory tends to M. On M, $\dot{V} = 0$; hence since the trajectory is bounded away from $a = 0$, we see also that $\dot{a} = 0$ and $\dot{\zeta} = 0$ on M. Now the equations $\dot{a} = 0$, $\dot{\zeta} = 0$ define a spiral curve C on the cylinder $\lambda a = \mu \cos \zeta$ by the formula

$$\psi = -\frac{\lambda}{\beta} \tan \zeta,$$

and M is an invariant subregion of this curve, bounded away from $a = 0$ (which C is not). On C the vector field defined by the equations must either vanish, or else must point in the $+\psi$ direction, or else point in the $-\psi$ direction; this is because there can be no motion in the a, ζ plane on $\dot{a} = 0$, $\dot{\zeta} = 0$. If either of the second two alternatives holds at a point of C, that point cannot belong to M, because the trajectory through it would move off C and $M \subseteq$ C. Hence M consists of C-points at which the field vanishes, i.e., $M = \{\text{critical point}\}$. (Cf. Theorem VI, p. 58 of Ref. 6).

Try as we might (and we tried many Vs) we have not succeeded in proving a version of Theorem 4 in which there was mistuning. If the same V is used with $\omega_d \neq 0$ as was used in Theorem 4, then it is no longer true that $\dot{V} < 0$; thus the results we feel are there still elude proof.

VII. COMPENSATED ATTENUATOR IN FEEDBACK

In private communication, L. H. Enloe and B. R. Davis have suggested that probably the most important practical FMFB receiver is one having a single-pole (as the baseband equivalent of the) IF filter, and a single-pole—zero-feedback filter, i.e., one with transfer

$$\frac{s + a}{s + b}. \tag{8}$$

In the time domain this acts like a delta-function plus an exponential. From formula (6) we see that the right-hand side of the equation for $\dot{\zeta}$ can be thought of as the output of a filter whose response is a delta-function plus an exponential. This suggests that the analysis in Section 6 can be made to cover the filter transfer (8) as well as the one-pole, because the differential equations are such as naturally to supply the constant in (8) if it is not present.

Writing the transfer function as

$$\frac{s + a}{s + b} = 1 + \frac{c}{s + b}$$

with $c = a - b$, the differential equations for the system become

$$\begin{aligned}
\dot{u}_c &= -\lambda u_c + \mu(x_c \cos \beta\varphi + x_s \sin \beta\varphi) \\
\dot{u}_s &= -\lambda u_s + \mu(x_s \cos \beta\varphi - x_c \sin \beta\varphi) \\
\dot{\theta} &= a^{-2}(u_c \dot{u}_s - u_s \dot{u}_c) = \frac{d}{dt} \tan^{-1} \frac{u_s}{u_c} \\
\dot{\varphi} &= \dot{\theta} + y \\
\dot{y} &= -by + c\dot{\theta},
\end{aligned}$$

with $a = (u_c^2 + u_s^2)^{1/2}$ again. We note that

$$\dot{\theta} = \frac{\mu}{a} (x_s \cos (\beta\varphi + \theta) - x_c \sin (\beta\varphi + \theta));$$

now we set $\xi = \beta\varphi + \theta$ and simplify the equation to

$$\begin{aligned}
\dot{\xi} &= \beta y + \frac{(\beta + 1)\mu}{a} (x_s \cos \xi - x_c \sin \xi) \\
\dot{a} &= -\lambda a + \mu(x_c \cos \xi + x_s \sin \xi) \\
\dot{y} &= -by + \frac{c\mu}{a} (x_s \cos \xi - x_c \sin \xi).
\end{aligned}$$

With the modulating signal a constant ω_d, we have as for equation (6), $x_s(t) = \sin \omega_d t$, $x_c(t) = \cos \omega_d t$, and we can set $\zeta(t) = \omega_d t - \xi(t)$ to obtain the equations

$$\begin{aligned}
\dot{\zeta} &= \omega_d - \beta y - \frac{(\beta + 1)\mu \sin \zeta}{a} \\
\dot{a} &= -\lambda a + \mu \cos \zeta \qquad (9) \\
\dot{y} &= -by + \frac{c\mu}{a} \sin \zeta.
\end{aligned}$$

These equations have the same form as (6) except that c replaces δ, b replaces γ, and there is an extra $(\beta + 1)$ coefficient in the sin term of $\dot{\zeta}$. The critical point is at

$$\begin{aligned}
a_0 &= \frac{\mu}{\lambda} \cos \zeta_0 \\
\zeta_0 &= \tan^{-1} \frac{\omega_d}{\lambda(\beta + 1)\left(1 + \frac{c\beta}{b}\right)} \\
y_0 &= \frac{\lambda c}{b} (\beta + 1) \tan \zeta_0 .
\end{aligned}$$

The matrix of partial derivatives evaluated at the critical point is

$$\begin{array}{c|ccc|} & \zeta & a & y \\ \dot{\zeta} & -(\beta+1)\lambda & (\beta+1)\dfrac{\lambda^2}{\mu}\dfrac{\tan\zeta_0}{\cos\zeta_0} & -\beta \\ \dot{a} & -\mu\sin\zeta_0 & -\lambda & 0 \\ \dot{y} & c(\beta+1)\lambda & -c(\beta+1)\dfrac{\lambda^2}{\mu}\dfrac{\tan\zeta_0}{\cos\zeta_0} & -b \end{array}.$$

The appropriate polynomial is

$$\begin{aligned} s^3 &+ s^2(\lambda(\beta+1)+b+\lambda)+s(\lambda b(\beta+2) \\ &\quad + \lambda^2(\beta+1)(1+\tan^2\zeta_0)+c\lambda\beta(\beta+1)) \\ &\quad + b(\beta+1)\lambda^2\tan^2\zeta_0+c\lambda^2\beta(\beta+1)+c(\beta+1)\beta\lambda^2\tan^2\zeta_0 \\ &= s^3+a_2s^2+a_1s+a_0 \,. \end{aligned}$$

A necessary and sufficient condition for local stability is then that a_0, a_1, and a_2 all > 0, and that $a_2a_1 > a_0$. It is clear that the first three conditions are met whenever $c > 0$, i.e., $a > b$. The case $a = b$ is degenerate and reduces in dimension to the filterless case of Section 4. Also, a_2 is always positive. However, since

$$\tan\zeta_0 = \frac{\omega_d}{\lambda(\beta+1)\left(1+\dfrac{c\beta}{b}\right)}$$

we can write a_1 as

$$a_1 = \lambda b+\lambda^2(\beta+1)+\lambda(\beta+1)(b+c\beta)+\frac{|\,\omega_d\,|^2\,b^2}{(b+c\beta)^2(\beta+1)}.$$

If

$$b\left(\beta-1-\frac{1}{\beta+1}\right) > \lambda+\beta a$$

then the sum of the first three terms is negative, and $a_1 < 0$ for $|\omega_d|$ *sufficiently small.* Similarly

$$a_0 = c\lambda^2\beta(\beta+1)+\frac{|\,\omega_d\,|^2\,(b+c\beta)}{\left(1+\dfrac{c\beta}{b}\right)^2(\beta+1)^2}$$

which is negative for *any* ω_d if $b + c\beta < 0$, and is also negative for $|\omega_d|$ sufficiently small if $c < 0$. The case $c < 0$ is physically a bit strange, because it is equivalent to having positive feedback in one loop of the feedback path; thus it is not surprising that in this case there can be instability even for $\omega_d = 0$.

In the case $c > 0$ only the condition $a_2a_1 > a_0$ is of concern and this is

$$(\lambda\beta + 2\lambda + b)\lambda b(\beta + 2) + \lambda^2(\beta + 1)(1 + \tan^2 \zeta_0) + c\lambda\beta(\beta + 1) > (b + \beta c)(\beta + 1)\lambda \tan^2 \zeta_0 + c\lambda^2\beta(\beta + 1),$$

which simplifies somewhat to

$$\lambda^2 b(\beta^2 + 7\beta + 9) + \lambda^3(\beta + 1)(\beta + 2) + c\lambda\beta(\beta + 1)(b + \lambda\beta + \lambda) + \lambda b^2(\beta + 2) > \frac{|\omega_d|^2}{(\beta + 1)^2\left(1 + \frac{c\beta}{b}\right)^2} (\lambda^2(\beta + 1)\beta c - \lambda^3(\beta + 1)(\beta + 2)).$$

Again, this is symmetric in $\pm\omega_d$, and is true for $|\omega_d|$ small enough. It becomes false for $|\omega_d|$ large if

$$\beta c > \lambda(\beta + 2).$$

We can think of the feedback path in this example as consisting of two parallel branches whose sum is added, one being an amplifier with gain, 1, the other being a single-pole filter S with dc gain 1 and (half-power) bandwidth b, in series with an amplifier A of gain c/b. We can then say in physical terms, assuming $\lambda = \mu$, that if

$$\beta < (\beta + 2) \times \frac{IF \text{ bandwidth}}{S \text{ bandwidth}} \times \text{gain of } A$$

then even a very large mistuning cannot affect local system stability, but if not, then a sufficiently large mistuning can. The fact that $(\beta + 2)$ appears as a factor on the right shows how much the zero in the compensating attenuator helps prevent instability, in agreement with what has been observed in practice by L. H. Enloe and B. R. Davis (private communication).

The global stability of the equations (9) may be studied by the same methods as were used in Section 6 for that of equations (6), but this topic is not pursued further here.

VIII. ACKNOWLEDGMENTS

The author is indebted to B. R. Davis, L. H. Enloe, A. J. Giger, S. O. Rice, and C. L. Ruthroff for helpful discussions, and to J. A. Morrison for the low-amplitude analysis in Section V.

REFERENCES

. Chaffee, J. G., United States Patent 2,075,503, granted March 30, 1937; see also Chaffee, J. G., "The Application of Negative Feedback to Frequency Modulation Systems," Proc. I.R.E., *27* (1939), May 1939, pp. 317–331.

2. B.S.T.J., Project Echo issue, *40,* No. 4, part 2 (July 1961), pp. 975–1238.
3. B.S.T.J., The Telstar Experiment (part 1), *42,* No. 4, part 2 (July 1963), pp. 739–1135.
4. Enloe, L. H., "Decreasing the Threshold in FM by Frequency Feedback," Proc. I.R.E., *50* (1962), pp. 18–30.
5. Giger, A. J., and Chaffee, J. G., The FM Demodulator with Negative Feedback," B.S.T.J., *42,* No. 4, part 2 (July 1963), pp. 1109–1135.
6. La Salle J., and Lefschetz, S., Stability by Liapunov's Direct Method, New York: Academic Press, 1961.
7. Lefschetz, S., Differential Equations: Geometric Theory, Second Edition, New York: John Wiley (Interscience), 1962.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Some Computer Experiments in Picture Processing for Bandwidth Reduction

By H. J. LANDAU and D. SLEPIAN

(Manuscript received January 4, 1971)

Some computer experiments in processing still pictures for bandwidth reduction are described. In the scheme studied, a picture is partitioned into subpictures each of which is encoded separately. A subpicture is expressed as a linear combination of a finite set of specially chosen basis subpictures. Quantized versions of the coefficients of this expansion are transmitted as binary digits. Using this procedure, we were able to obtain pictures of good quality using approximately 2 bits per picture-element; we were unable to do so at lower bit-rates.

Some general comments on the encoding of pictures are included.

I. INTRODUCTION

This paper describes some computer experiments in picture processing carried out by us during the winter of 1969 and the spring of 1970. Our goal was to explore a particular method for the efficient encoding of typical *Picturephone*® scenes into binary digits. The experiments involved still pictures only. We first describe the experiments and their results, then follow with some general comments on the encoding of pictures. These comments are intended to explain our motivation for the particular investigation undertaken.

II. THE EXPERIMENTS

By means of the TAPEX unit at Murray Hill,[1,2,3] a photograph can be represented in digital form suitable for handling by the GE-635 computer. Specifically, the picture is scanned from top to bottom along n_1 horizontal lines, the light intensity being sampled n_2 times along each line; every sample is then quantized to the nearest one of 2^k equally spaced amplitude levels, and recorded on a digital tape as a single k-bit integer. Conversely, given the digital tape, each k-bit integer is replaced by its corresponding amplitude value, which is then regarded as a sample, taken at the Nyquist rate, of a bandlimited waveform. The reconstructed waveform controls the beam intensity of successive lines traced by a scanning cathode-ray oscilloscope. The oscilloscope face is photographed.

In all our experiments, the values used for the above quantities were $n_1 = n_2 = 256$, and $k = 10$. Figure 1a is an original photograph; Fig. 1b is the result of converting the picture into binary digits on tape and reconstructing via TAPEX. Comparison shows that, with the parameters as chosen, the digital representation is of customary television quality. It has, of course, the inevitable raster lines.*

In processing a picture, we first converted each k-tuple of binary digits into the corresponding integer, and subtracted 2^{k-1}. The resulting integers, lying in the range $(-2^{k-1}, 2^{k-1} -1)$, are called *picture elements,* and we denote by X_{ij} the picture element obtained from the jth sample of the ith line of the picture. For computer processing, we regard a picture as an $n_1 \times n_2$ matrix of picture-elements. In our experiments, we further partitioned the $n_1 \times n_2$-element picture by a square grid (as in Fig. 2) into $n_1 \times n_2/m^2$ square subpictures, each having m picture-elements on a side. These subpictures were encoded independently, one at a time, by the scheme described below.

We view the $M = m^2$ picture-elements of a subpicture, when read out row by row from left to right, as the components of an M-dimensional vector $\mathbf{Y}$, which represents the subpicture. For example,

$$\mathbf{Y} = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{m1} \\ \vdots \\ X_{mm} \end{bmatrix}$$

is the vector representing the top-left subpicture of Fig. 2. To describe such vectors, it is natural to introduce a basis. We therefore choose M orthonormal M-dimensional basis vectors $\mathbf{b}_1, \cdots, \mathbf{b}_M$. These remain fixed, and determine the particular encoding scheme under discussion.

* (*Added in proof.*) The half-tone method of picture reproduction used in printing this article of necessity obscures some of the detail visible in the original TAPEX photographs. Copies of these photographs will be sent on request.

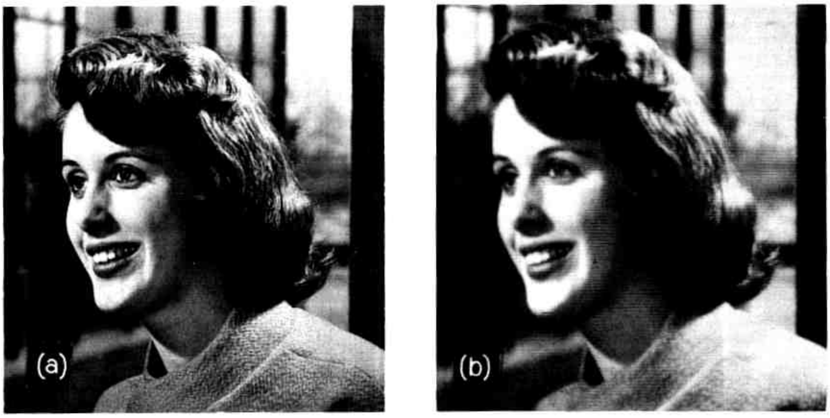

Fig. 1—(a) Original photograph; (b) Reconstruction of (a) via TAPEX, using 10 bits per picture-element.

We may now expand **Y** in terms of the basis vectors, to obtain

$$\mathbf{Y} = \sum_{1}^{M} c_j \mathbf{b}_j \tag{1}$$

where, by orthonormality of the $\mathbf{b}_j$,

$$c_j = \mathbf{b}_j \cdot \mathbf{Y}, \qquad j = 1, \cdots, M. \tag{2}$$

We then quantize c_j into one of r_j different values, denoting by $\hat{c}_j$ the quantized version of c_j. To transmit these quantized coefficients of a subpicture in the simplest possible way, i.e., by encoding them independently without exploiting the statistical distribution of their values, requires $r = \sum_1^M [\log_2 r_j]$ binary digits. We take the number

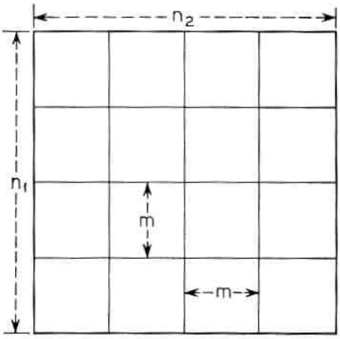

Fig. 2—Partition of pictures into subpictures.

$R = r/m^2$, which is the number of bits used per picture-element, as a measure of the efficiency (or bandwidth) of the encoding scheme.

To reconstruct a picture, we suppose the quantized coefficients are known, and obtain a reconstructed subpicture code vector $\hat{\mathbf{Y}}$ from the recipe

$$\hat{\mathbf{Y}} = \sum \hat{c}_j \mathbf{b}_j \ .$$

The components of $\hat{\mathbf{Y}}$ are quantized to the nearest integer value in the range $[-2^{k-1}, 2^{k-1} - 1]$. These are the picture-elements $\hat{X}_{ij}$ of the reconstituted picture. A photograph is obtained from these values using the TAPEX unit in the manner already described.

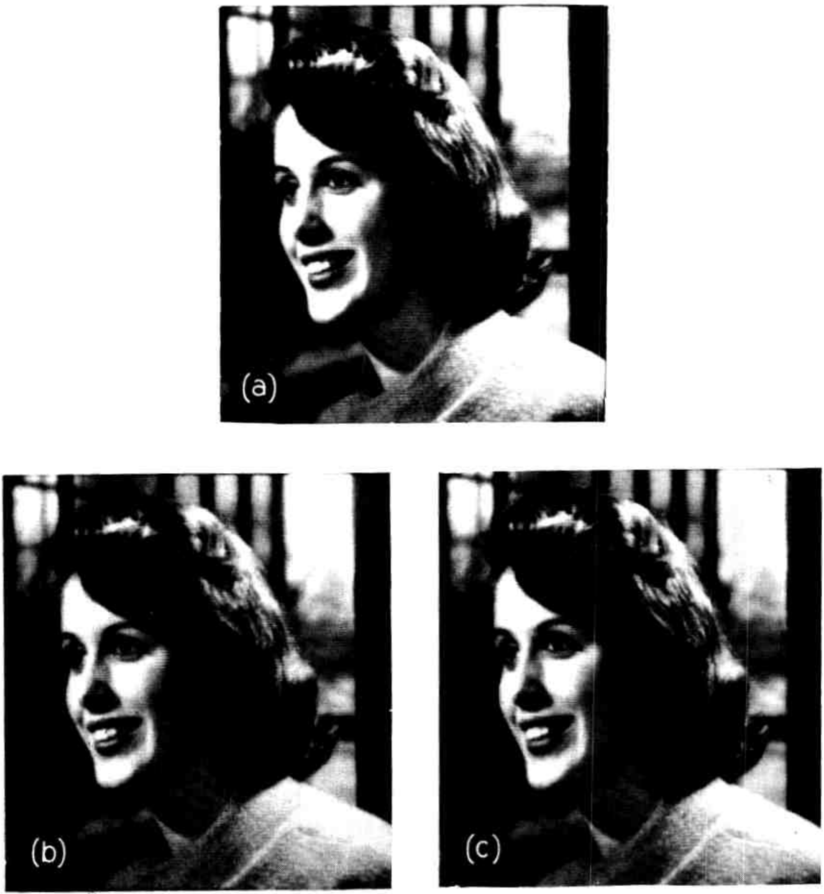

Fig. 3—(a) 10 bits per picture-element; (b) Reconstruction of (a) by means of the Hadamard basis, using 2 bits per picture-element; (c) Reconstruction of (a) by means of differential PCM, using 3 bits per picture-element.

Our experiment consisted of choosing various basis vectors and various quantization rules for the expansion coefficients c_j. We also experimented with making nonlinear transformations on the picture-elements before and after bandwidth compression processing. None of the various types of companding we tried yielded better results than were obtained without companding. In most of)ur work, subpictures of $M = 16$ picture-elements ($m = 4$) were used; a few experiments were run with $m = 8$.

We were able to obtain pictures of good quality with a rate $R = 2$ bits per picture-element, but were unable to do so at lower rates. Figure 3a repeats the 10 bit per picture-element photograph of Fig. 1b. Figure 3b shows a reconstructed picture with $R = 2$ bits per picture-element obtained with a scheme using $m = 4$ and the Hadamard basis described below. We also simulated on the computer the differential PCM scheme employed in *Picturephone* coding which uses 3 bits per picture-element.[4] Figure 3c shows the result of this simulation; it compares favorably with Fig. 3b. Two different subjects are treated analogously in Figs. 4 and 5.

Although the subpicture encoding achieves a one-third decrease in rate, the differential PCM scheme is far easier to instrument. From our experience, it seems unlikely that good pictures can be obtained with the subpicture scheme at rates much less than 2 bits per picture-element.

III. COMMENTS ON COMPRESSION

To avoid needless complications, in all that follows we shall think of a picture in discrete terms, i.e., as a finite collection of picture-elements, each of which can assume finitely many different values.

How many bits must one use to transmit the picture of Fig. 1b? The answer is, of course, zero. It is a single picture. The question is not an interesting one. More pertinently, we can ask how many bits per picture are required on the average to transmit long strings of pictures drawn from a given ensemble of pictures. Since a picture source can be regarded as producing sequences of picture elements, each of which can assume one of K different values, evidently we can transmit all possible pictures perfectly by using $[\log K]$ bits per picture-element. For any reduction of the bit-rate below this value, we must capitalize on one or both of these facts:

(*i*) not all pictures are produced with equal probability by the source, nor are they produced independently;

(*ii*) the observer does not require all pictures to be reproduced exactly.

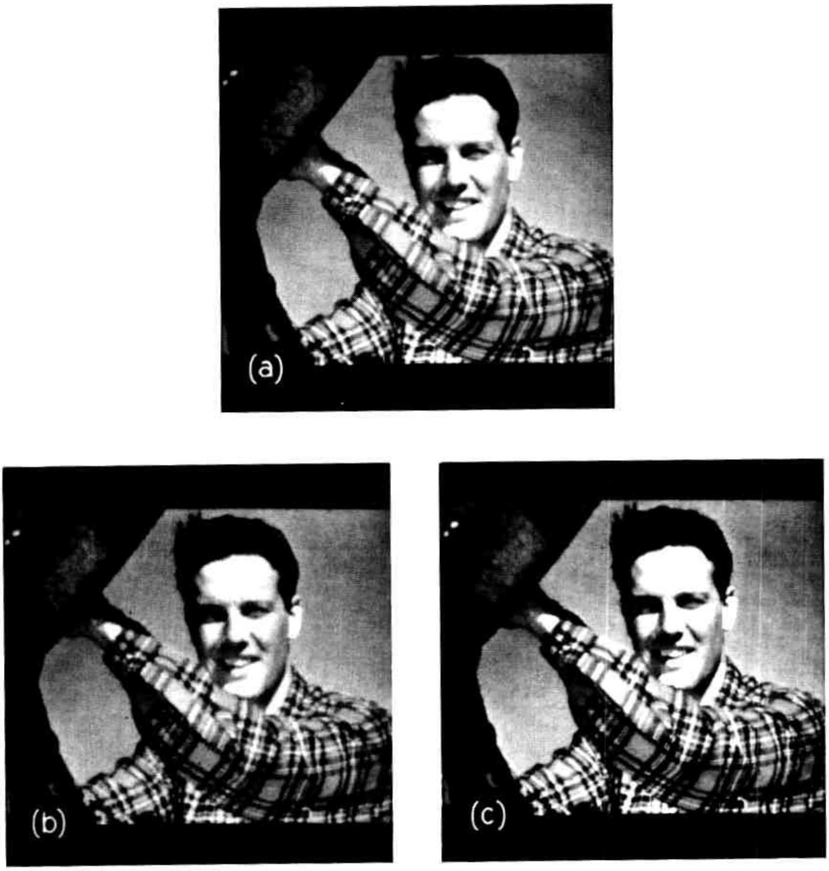

Fig. 4—(a) 10 bits per picture-element; (b) Reconstruction of (a) by means of the Hadamard basis, using 2 bits per picture-element; (c) Reconstruction of (a) by means of differential PCM, using 3 bits per picture-element.

The question of how to take advantage of such considerations has been much studied in information theory, and methods are known in principle for computing the answer. A calculation of the *entropy* of the picture ensemble describes how far it is possible to reduce the bit-rate, and still maintain perfect reconstruction, by exploiting source redundancies; this minimum rate is determined solely by the statistics of the ensemble, and has nothing to do with the nature of pictures, vision, or the observer.

Determination of the entropy of a picture source does not solve the problem of real interest to workers in picture transmission; for pictures, as they are usually presented by a source, have more detail and resolution than the observer can utilize. Thus pictures that are counted as different in the source ensemble may be indistinguishable to the viewer who wishes to reconstruct them only to some set limit of accuracy, i.e., to achieve some "level of fidelity". The minimum bit-rate, given as a function of both the particular fidelity criterion adopted and the source statistics, may also be computed, and is called the *rate distortion function*. As with entropy, this minimum rate is achievable only in the limit of more and more complicated encoding processes.

Conceptually, rate distortion theory formulates carefully and answers completely the foremost question in the TV coder's mind: "How many bits do I need?" In actuality, it doesn't do very much for him

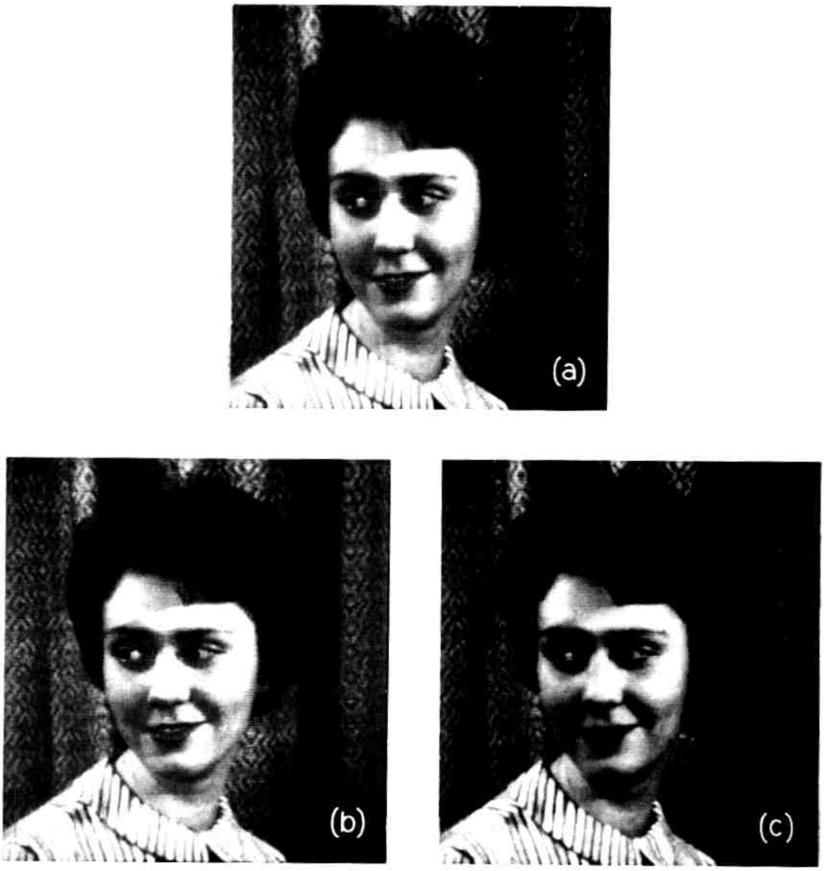

Fig. 5—(a) 10 bits per picture-element; (b) Reconstruction of (a) by means of the Hadamard basis, using 2 bits per picture-element; (c) Reconstruction of (a) by means of differential PCM, using 3 bits per picture-element.

at all. To see why this is so, we must look just a bit closer at the formalism of rate distortion theory.

The general theory presupposes a source that produces infinite strings of symbols, each symbol being drawn from a K-letter alphabet. The "values" of the letters play no role in the theory, so for convenience we suppose that they are the integers $1, 2, \cdots, K$. Denote a typical string produced by the source by $\cdots X_{-1}, X_0, X_1, X_2 \cdots$, where each X is one of the integers from 1 to K. A measure is placed upon the set of such infinite strings in such a way that we can regard the Xs as random variables and meaningfully ask and answer such questions as "What is the probability that $X_0 = 3$, $X_2 = 1$, and $X_3 = 5$?" There are many technicalities involved in specifying this measure, but they need not concern us here. We are also given a numerical-valued distortion function $\delta(j, k) \geqq 0$ which gives the distortion when a transmitted letter j is reconstructed as letter k.

Let us now consider transmitting the strings produced by the source by encoding them in the following manner. We break the source strings up into blocks of n successive symbols. Since each block is composed of n source symbols and each symbol can be one of K different integers, there are $B = K^n$ different blocks possible. We suppose that a dictionary is provided, which lists for each one of these B blocks a special block called its representative block. As successive strings of n source letters are produced by the source, each is looked up in the dictionary and encoded into its representative block. If the letters of a block are $x_1, x_2, \cdots, x_n$ and the letters of the corresponding representative block taken from the dictionary are $y_1, y_2, \cdots, y_n$, we take the quantity $D = \sum_1^n \delta(x_i, y_i)$ as the distortion per block. We take $d =$ average D/n as the level of distortion achieved with the given code book, where the average is over all source strings.

Let us now fix the number of representative blocks in the dictionary at 2^L. Some code books translating the B blocks into 2^L representative blocks will yield smaller values for the distortion d than will others. We denote by $d(L, n)$ the smallest distortion obtainable by any such code book. Note now that since there are only 2^L representative blocks in these code books, we could use L binary digits to transmit each representative block name to a destination. We would achieve distortion $d(L, n)$ and be transmitting at a rate

$$R = (L/n) \text{ bits/(source symbol).}$$

Now fix R and write $\bar{d}(R) = \lim_{n \to \infty} d(nR, n)$. This function gives the smallest distortion obtainable for a fixed binary rate R that can be had in the limit of arbitrarily large code books. The inverse function $R(\bar{d})$ which gives the smallest binary rate per source letter that will yield a given distortion $\bar{d}$ is called the *rate distortion function*. Information theory shows how $R(\bar{d})$ can be calculated in principle from the symbol distortion function $\delta(j, k)$ and the measure assigned to the source. We do not display these complex formulas here.

How can we apply this to picture transmission? There are two obvious different methods of identifying the source symbol X with a quantity of interest in picture coding. The method most satisfying conceptually is to identify the random variable X with an entire picture. This is possible since there are only finitely many different pictures, due to our assumption that a picture is composed of $n_1 \times n_2$ picture elements, each taking one of 2^k values. We number the possible pictures and take the picture numbers as the values of X. The distortion function $\delta(j, k)$, which we must now describe to apply the theory, measures our dissatisfaction at having picture j reproduced as picture k. Conceptually such a measure exists, but we know little about it. In our experiments, we would have to prescribe it for $(2^{10\times256\times256})^2$ pairs of values of j and k.

To compute a value for the rate distortion function, we require in addition a measure on the source symbols: at a minimum, this involves assigning a probability distribution, bearing some relation to what will be observed in practical transmission, to the $2^{10\times256\times256} = 10^{197,283}$ different possible pictures. This task seems quite beyond us now. For to obtain a histogram empirically is out of the question: at 30 frames per second, one sees only 10^9 frames per year, and if the different possible pictures were run off in sequence at this rate, it would take $10^{197,274}$ years to view them all. On the other hand, to specify the distribution theoretically requires more understanding of the situation than we now have.

Indeed, being able to describe a reasonable distribution for the possible pictures goes a long way towards solving the problem of efficient coding. We suspect that a reasonably good description would assign probability $1/N$ to each of N of the pictures, and zero to the rest, with N small indeed compared to $10^{197,283}$. If we could describe this set well, we could encode using log N bits/picture. But which are the "likely" pictures? For *Picturephone* service or commercial broadcast television, intuition suggests that chaotic pictures, in which adjacent picture-elements jump about between extreme values, would be classified as unlikely. Likely pictures are, roughly speaking, made up of regions of nearly constant brightness. The brightness might change considerably from one region to the next, but there cannot be too many small regions, or we are back to unlikely chaos, nor can the boundaries of the region be too wild or fast-turning. However, the enormous number of possibilities involved prevents an accurate description.

A more tractable application of the general theory comes from associating the source symbol X with a picture-element. Now, however, the distortion function $\delta(j, k)$ measures our displeasure at having the jth level of brightness for a picture-element reproduced as the kth level of brightness. This is an excessively local measure of picture fidelity and is probably quite remote from the criteria used by human observers.

In summary, rate distortion theory tells us that to encode efficiently we must pay attention to the more likely pictures (or sequences of pictures), and that we must replace these in groups by representative values which yield an acceptable distortion. The theory tells us how to calculate the minimum bit-rate needed to achieve a given level of fidelity, but to carry out the calculation we need to know the distortion function $\delta(j, k)$ and the measure that gives a statistical description of the source. In picture coding we have at present very meager knowledge concerning these quantities. Any new understanding of either will undoubtedly lead to improved practical coding schemes. It will take a great deal of understanding, however, to know these quantities well enough to allow a calculation of a rate distortion function in which one can have much confidence.

IV. MOTIVATION FOR THE EXPERIMENTS

As we have argued, we lack the information required to bring the full force of rate distortion theory to bear on picture coding. Nevertheless, it was the general approach of rate distortion theory that led to the encoding scheme of our experiments. We wanted an encoding procedure that also would derive from considerations of likelihood and fidelity, but that would be manageable in practice. Accordingly, we began by focusing on subsections of the picture.

Although we cannot characterize adequately those entire pictures that are likely, perhaps we can do so for subpictures. How large must a section of a TV picture be before we can describe it as a likely subpicture or an unlikely subpicture? If we look at a single picture-element, every value is "likely." If we look at two adjacent picture-elements, again we must say that any pair of values is "likely." If we consider square subpictures of $m \times m$ picture-elements, most observers feel that for $m = 4$ they can already classify some subpictures as likely parts of the *Picturephone* or TV ensemble and others as less likely. The 4×4 checkerboard pattern, where adjacent picture elements oscillate between extreme values, seems unlikely: the uniform 4×4 subpicture seems highly likely.

We decided then to break a picture into $m \times m$ subpictures and to encode each subpicture independently. If m is large enough, not much compression potential will be lost by neglecting the correlation between subpictures, for the chaos of structure that we intuitively feel to be unlikely in the TV ensemble is of a within-subpicture scale. The factor driving us to choose a small value of m is the need for a reasonable number of pictures to distinguish among on a probabilistic basis.

Even with $m = 4$ and $k = 10$, there are $2^{10 \times 16} = 10^{48}$ different subpictures, so that it is out of the question to use a Huffman-Fano code, or other dictionary-like code, to take advantage of the unequal probabilities of the various subpictures. We seek some other scheme. A natural idea here is to represent the subpicture in terms of some coordinates that can be treated independently and that are related to the probability measure on the subpictures.

We begin by interpreting the subpicture as a vector, in the manner described in Section II. Suppose that an M-dimensional subpicture vector $\mathbf{Y}$ is to be expanded on M linearly independent basis vectors $\mathbf{b}_i$ as in equation (1), but that only $J < M$ of the cs will be used (exactly) to reconstruct an approximation $\hat{\mathbf{Y}}$ to $\mathbf{Y}$. Thus

$$\mathbf{Y} = \sum_1^M c_i \mathbf{b}_i \,, \qquad \hat{\mathbf{Y}} = \sum_1^J c_{\alpha_i} \mathbf{b}_{\alpha_i} \,.$$

where $\alpha_1, \alpha_2, \ldots, \alpha_J$ are distinct integers from the list $1, 2, \ldots, M$. What basis vectors should be used, and which J coefficients retained, in order to minimize the mean squared error between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$? The answer to this problem is well known. Let $\mathbf{Y} = (y_1, y_2, \cdots, y_M)$ and denote the covariances of the components of $\mathbf{Y}$ by $\rho_{ij} = Ey_iy_j$. Let ρ be the $M \times M$ matrix with elements ρ_{ij}. The basis vectors that solve the above problem are the eigenvectors of ρ having the J largest eigenvalues. Basis vectors chosen in this way are known as a Karhunen-Loève basis. The fidelity criterion implicit here is that of mean-square error—one not very adequate when applied to pictures.

We began our experiments with finding the Karhunen-Loève basis for the picture of Fig. 1, by determining empirically the 16×16 covariance matrix for the 4×4 subpictures. We discovered, as expected, that the ρ_{ij} were all extremely close to 1, expressing the high likelihood of uniform brightness over so small a square. When each $\rho_{ij} = 1$, the eigenvalue problem is degenerate: the first eigenvector has all components equal and an eigenvalue of 1, while the remaining eigenvectors are indeterminate and correspond to an eigenvalue of 0. Thus we felt no great confidence in our determination of the Karhunen-Loève vectors, believing that it was probably unstable, and turned instead to the following intuitive justification for introducing another basis, which we call the Hadamard basis.

Intuitively, a subpicture with constant brightness for all picture-elements is a very likely subpicture. Part (1) of Figure 6 depicts a

4×4 array of picture elements all having value $+ 1/4$. Another likely subpicture has a vertical edge running down its middle. Part (2) of Figure 6 depicts such a case, where the picture elements on the left have value $+ 1/4$ and those on the right have value $- 1/4$. If now we form a basis vector, $\mathbf{b}_1$, from Fig. 6(1) and $\mathbf{b}_2$ from Fig. 6(2), we see that linear combinations $\mathbf{Y} = c_1\mathbf{b}_1 + c_2\mathbf{b}_2$ give all possible subpictures having a center vertical transition between two regions of uniform brightness.

Continuing this train of thought, we are led to seek 16 linearly independent subpictures of decreasing likelihood that will serve as a basis on which to expand an arbitrary subpicture. Such a basis, chosen so that the vectors are orthonormal, is shown in Fig. 6. If each subpicture $\mathbf{Y}$ of a large picture is expanded on this basis, so that

$$\mathbf{Y} = \sum_1^{16} c_i\mathbf{b}_i ,$$

we would expect frequently to find the higher coefficients, say c_{10}, c_{11}, etc., to have values near zero. The coefficient c_1, which gives the average brightness of the pictures, would be expected to have a large variance—higher coefficients, a much smaller variance. Table I lists the ratios $\xi_i = \sigma_i^2/\sigma_1^2$, where σ_i^2 is the variance of c_i, as determined empirically from all the subpictures of Fig. 1b. The results agree remarkably well with intuition. That c_1, which has more than ten times the variance of any other coefficient, does indeed contain a great deal of the essence of the original picture is seen from Fig. 7b, which shows the picture resulting from the reconstruction $\hat{\mathbf{Y}} = c_1\mathbf{b}_1$, next to the original photo-

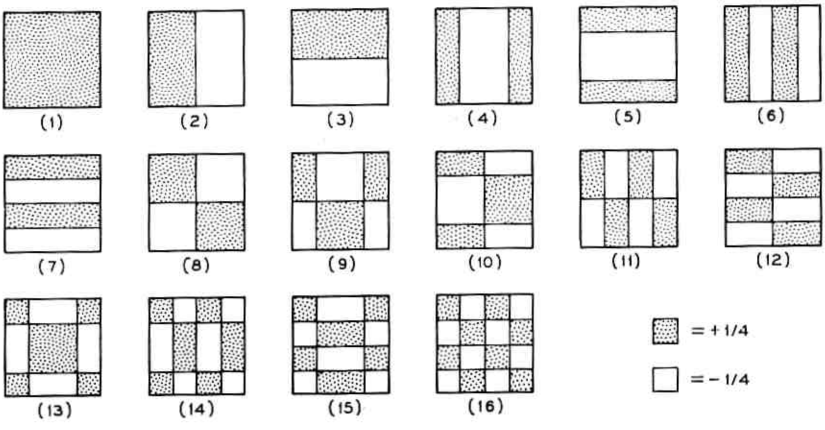

Fig. 6—The Hadamard basis.

Table I—Coefficient Variances

i	ξ_i	i	ξ_i
1	1.00	9	0.024
2	0.098	10	0.024
3	0.087	11	0.020
4	0.035	12	0.022
5	0.038	13	0.019
6	0.051	14	0.015
7	0.048	15	0.016
8	0.034	16	0.014

graph 7a. Figure 7b allows one to see clearly the size of the subpictures used in the experiments.

Our choice of the Hadamard basis is thus dictated by plausible guesses about the probabilities of subpictures. Furthermore, it is not inconsistent with the Kahunen-Loève procedure, since the correlation matrix is so nearly singular. Finally, it has an important practical advantage: since its components each have value ± 1/4, the computation of the coefficients can be carried out by simple switching. In our experiments, we found the results of processing with the Karhunen-Loève basis to be no better than those obtained with the Hadamard basis, and so, for reasons of simplicity, judged the latter to be superior.

Since our object is to reduce the bit-rate, we must adopt some scheme of quantization for the coefficients. This will lead to an approximate reconstruction of the subpicture, and considerations of fidelity must guide us in our choice of rules. One such quantization scheme—keeping some of the coefficients exactly and dropping the remainder altogether—gives rise to the Karhunen-Loève problem. We adopted a different procedure, based on quantizing successive coefficients more

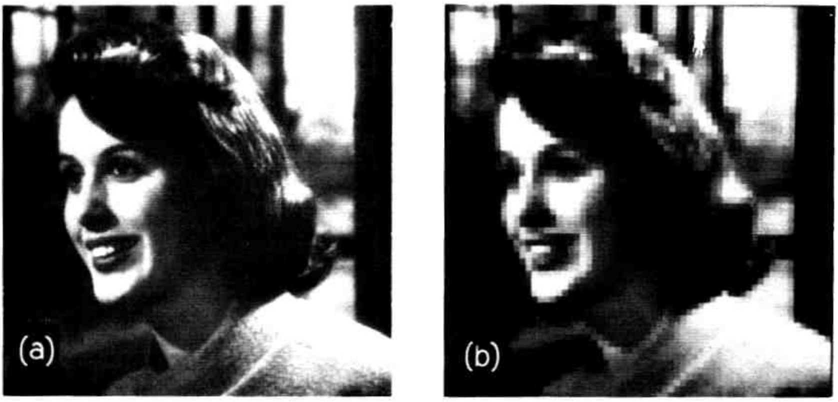

Fig. 7—(a) Original photograph; (b) Reconstruction of (a) by means of c_1 only.

TABLE II—QUANTIZATION OF COEFFICIENTS

i	r_i	i	r_i
1	64	9	4
2	16	10	4
3	16	11	0
4	8	12	0
5	8	13	0
6	8	14	0
7	8	15	0
8	4	16	0

and more coarsely. Two arguments led us to this. Firstly, since the lower coefficients have more variability, reproducing these more accurately helps reduce the mean-square error for the more probable pictures. Secondly, the higher coefficients tend to be large mainly when the subpicture has a very "busy" or chaotic nature; we guessed the detail of that chaos to be less important to the viewer than the existence of chaos. Thus the fidelity criterion behind our encoding contains an element of the characteristics of observers, in addition to considerations of mean-square error.

Much experimenting bore out the general truth of these suppositions. Table II gives the number of quanta, r_i, used for c_i in Figs. 3b, 4b, and 5b. The quantization of a given c was carried out by dividing its range into disjoint intervals whose endpoints are called *cut points* and by associating with each interval a *representative value*. The quantized value of c is the representative value associated with the interval in which c lies. Table III lists the cut points and representative values used to obtain Figs. 3b, 4b, and 5b.

We carried out over 100 experiments in which the r_i, the cut points, and representative values were varied over considerable ranges; details are available on request. Although we ultimately settled on the configuration described in Tables II and III, the number of possibilities to be explored is so large that we have no great confidence that we have found the best values for the parameters. On the other hand, based on our experience we would judge it unlikely that significant improvements can be made with this scheme by further changes of parameter values.

V. PICTURE REPRODUCTION AND QUALITY JUDGMENT*

The development of ordinary photographic film, as well as the characteristics of analog devices such as scanners and picture tubes,

* (*Added in proof.*) The comments of this section refer to the original TAPEX photographs, rather than to their reproductions in this article. See footnote on p. 1526.

Table III—Cut Points and Representative Values used for Quantizing the Coefficients $c_1, \cdots, c_{16}$.

Possible values for the c_k lie in the range ± 2048.0, and are integer multiples of 0.25.

c_1: The range between −1940 and +1900 was divided into 64 intervals of length 60, the midpoint of each serving as representative point for the interval. The first and last representative points also represented any values of c_1 outside this range.

The cut points and representative values for $c_2 - c_{10}$ were generated from curves of the form $y = k \sqrt{x}$, as follows:

c_2, c_3:	Cut points	±	0		8.6		34.4		77.4		137.5
	Repr. values	±		2.1		19.3		53.7		105	
	Cut points	±		215		309		421			
	Repr. values	±	174		260		363		483.		
c_4, c_5, c_6, c_7:	Cut points	±	0		15.6		62.4		140.4		
	Repr. values	±		3.9		35.2		97.6		191.	
c_8, c_9, c_{10}:	Cut points	±	0		62.5						
	Repr. values	±		15.6		141.					

$c_{11}, \cdots, c_{16}$: dropped

are sensitive to many parameters and can vary noticeably over time. The result is that in processing and reproducing photographs it is extremely difficult to maintain rigid control of contrast and average gray level. Yet these two quantities strongly influence the viewer's judgment of the quality of a picture.

In our subpicture encoding scheme, information about overall contrast and gray level is contained almost entirely in the values of c_1. We are convinced by experiments performed that the quantization of this coefficient, as described by Tables II and III, is sufficiently fine to render these characteristics faithfully. We therefore believe that what variation in contrast exists on Figs. 1, 3, 4, 5, and 7 is attributable to vagaries of the reproduction process and not to failures of the encoding, which are evidenced by inaccuracies in edges and texture. Accordingly, the reader should attempt to subtract out the differences in contrast among the photographs and should judge the quality of our scheme by examination of detail in Figs. 3b, 4b, and 5b.

VI. RELATED WORK

At about the same time that the present experiments were carried out, rather similar investigations were independently conducted elsewhere by other workers.[5,6,7] While related to our work, these studies differ from it somewhat in detail of execution and very much in theoretical approach.

VII. ACKNOWLEDGMENT

We are greatly indebted to C. A. Sjursen for his continual help in scanning and reproducing pictures on the TAPEX unit.

REFERENCES

1. Anderson, W. A., Swartzwelder, J. D., Weller, D. R., "System and Programming Aspects of TAPEX," unpublished work, March 1, 1966.
2. Camlet, J. V., "Remote High-Speed Data Station," unpublished work, January 16, 1967.
3. Giordano, P. P., "Variable Speed Scanner," unpublished work, March 7, 1962.
4. Limb, J. O., and Mounts, F. W., "Digital Differential Quantizer for Television," B.S.T.J., *48*, No. 7, part 3 (September 1969), pp. 2583–2599.
5. Tasto, M., and Wintz, P. A., "Picture Bandwidth Compression by Adaptive Block Quantization," Report TR-EE 70-14, School of Electrical Engineering, Purdue University, Lafayette, Indiana, July 1970.
6. Huang, T. S., and Woods, J. W., "Picture Bandwidth Compression by Block Quantization," presented at 1969 IEEE International Symposium on Information Theory, Ellenville, N. Y.
7. Pratt, W. W., and Andrews, H. C., "Transform Processing and Coding of Images," USCEE Report 341, University of Southern California, Los Angeles, Calif., March 1969.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Computer Synthesis of Speech by Concatenation of Formant-Coded Words

By L. R. RABINER, R. W. SCHAFER and J. L. FLANAGAN

(Manuscript received January 13, 1971)

Speech signals can be described in terms of the resonances of the vocal tract. These resonances, or formants, change at rates comparable to the motions of the vocal tract. They therefore can be sampled and quantized to low bit-rates, and hence constitute an economical form for digital storage of speech information. Formant coding also permits flexible arrangement of speech elements into various contexts. This report describes a computer technique for synthesizing continuous messages by concatenating formant data for word-length utterances. The stored data for the synthesis corresponds to a bit-rate of 533 b/s. A Honeywell DDP-516 computer is used to experimentally evaluate a voice response system. In an initial application, the system is used to synthesize 7-digit telephone numbers. To assess the synthesis an interactive dialing experiment, also conducted by the computer, is described. The results show the synthesized numbers to be comparable in communicative effectiveness to naturally spoken digits.

I. INTRODUCTION

If computers could speak with sophisticated vocabularies they could provide a variety of automatic information services. Machines could be interrogated from conventional *Touch-Tone*® telephones and stored data could be accessed by voice.

Naturally spoken speech messages can of course be prerecorded and stored. However, the digital storage required for sizeable amounts of natural speech is inordinate. Further, elements of natural speech in one context cannot be realistically assembled into a different message. With individual pieces of the signal waveform there is no practical way of making natural transitions from one element to the next. In certain messages of highly limited context—notably the Automatic Intercept System—individual words are adequately abutted by having more than one spoken version of each word. In general, however, sentence-length material cannot be satisfactorily produced in this manner.

For answer-back purposes, requiring sizeable vocabularies, an efficient means of storing and accessing speech information is required. This requirement implies low bit-rate representation of vocabulary elements *and* a flexible means for assembling the vocabulary elements into any message specified by the answer-back program. Toward this requirement, we have devised a synthesis method based upon formant-coded vocabulary elements.

Formants are the resonances, or eigenfrequencies, of the vocal tract. They change at relatively slow rates, comparable in speed to the articulatory motions. Their variations with time can consequently be sampled and quantized to low bit-rates. Furthermore, this description of the speech signal permits separation of information about vocal-tract excitation (i.e., voiced/unvoiced distinctions and voice pitch) from the resonance information. The formant description therefore provides a flexible means for smoothly assembling vocabulary elements into connected speech.

Toward this goal of low bit-rate storage and flexible assembly of computer speech, we have implemented and experimentally evaluated a formant-synthesis answer-back system. In the subsequent discussion we outline principles of the implementation and offer results of an initial application to the synthesis of telephone numbers.

II. SYNTHESIS MODEL

The model for formant synthesis of speech is shown in Fig. 1. The voiced sounds of speech (i.e., those generated by vocal-cord vibration) are produced by the upper branch of the system. An impulse generator produces a sequence of impulses whose spacing is controlled by the "pitch period" parameter, P, which corresponds to the period of vocal cord vibration. This impulse train is modulated in amplitude by a control parameter, A_V, which represents the intensity of voiced sounds. The resulting signal excites a time-varying digital filter composed of four cascaded resonators. Three of the resonators have time-varying resonant frequencies (F_1, F_2, F_3)—which correspond to the first three resonances, or formants, of the vocal tract. The output of this system is passed to a fixed, second-order digital filter which approximates the source spectrum and mouth radiation characteristics of human speech.[1]

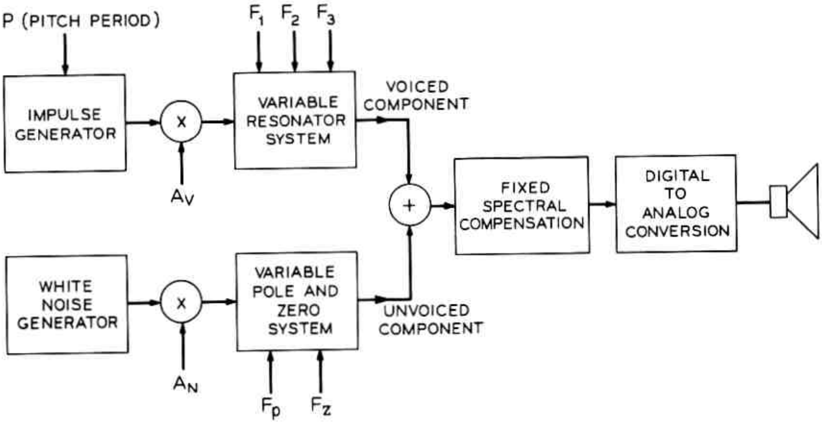

Fig. 1—Digital formant synthesizer, block diagram.

Unvoiced speech is produced by the lower branch of the system in Fig. 1. A random number generator, representative of the fricative noise source in unvoiced speech, produces samples of uniformly distributed white noise. The noise amplitude is modulated by the control parameter, A_N, which represents the intensity of unvoiced sounds. This signal excites another time-varying digital filter composed of one time-varying resonator (F_p) and one time-varying antiresonator (F_z). This pole-zero pair constitute an approximation to the formant structure of unvoiced speech sounds.[1] The output of this system also is passed to the fixed spectral-shaping filter. Digital-to-analog conversion provides an audible output.

All the parameters required by the synthesis system of Fig. 1 can be estimated automatically from natural speech by recently developed digital signal processing techniques.[2,3]

III. CONCATENATION MODEL

The Acoustics Research DDP-516 computer facility has been used to implement a complete answer-back system. A block diagram of the system used for synthesis of connected speech from a vocabulary of formant-coded words is shown in Fig. 2. Naturally spoken, isolated words (or phrases) are analyzed by a formant analyzer to give three formants (F_1, F_2, F_3) voiced and unvoiced amplitude (A_V, A_N), pitch period (P), and unvoiced pole and zero (F_p, F_z) once every 10 ms. These control parameters are smoothed by programmed digital

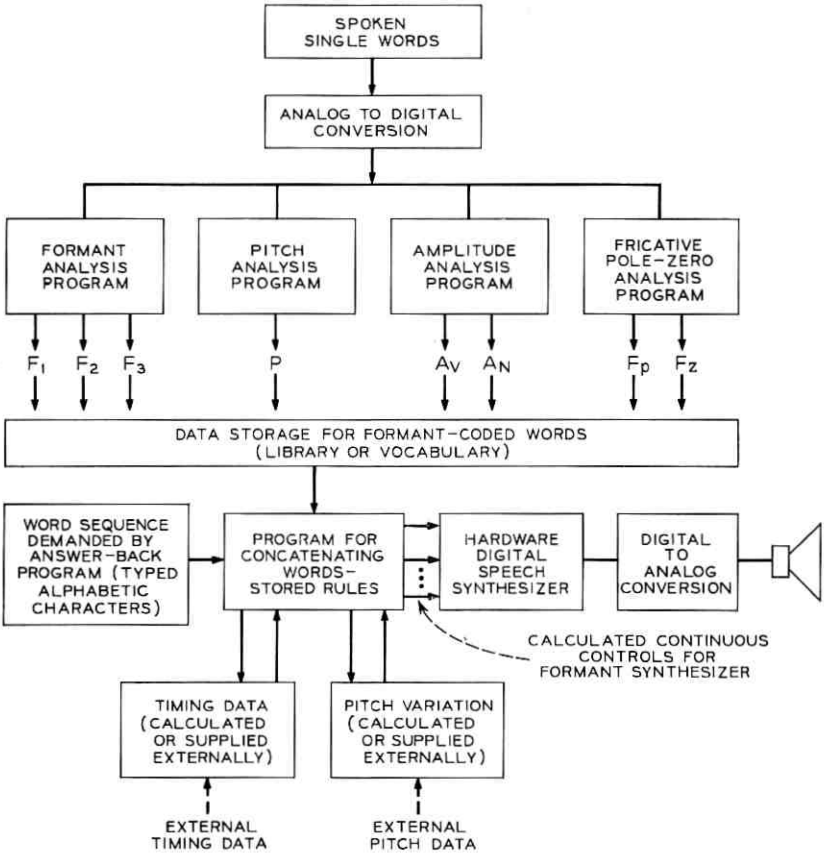
SPOKEN SINGLE WORDS
ANALOG TO DIGITAL CONVERSION
FORMANT ANALYSIS PROGRAM
PITCH ANALYSIS PROGRAM
AMPLITUDE ANALYSIS PROGRAM
FRICATIVE POLE-ZERO ANALYSIS PROGRAM
F1 F2 F3
P
AV AN
Fp Fz
DATA STORAGE FOR FORMANT-CODED WORDS (LIBRARY OR VOCABULARY)
WORD SEQUENCE DEMANDED BY ANSWER-BACK PROGRAM (TYPED ALPHABETIC CHARACTERS)
PROGRAM FOR CONCATENATING WORDS-STORED RULES
HARDWARE DIGITAL SPEECH SYNTHESIZER
DIGITAL TO ANALOG CONVERSION
CALCULATED CONTINUOUS CONTROLS FOR FORMANT SYNTHESIZER
TIMING DATA (CALCULATED OR SUPPLIED EXTERNALLY)
PITCH VARIATION (CALCULATED OR SUPPLIED EXTERNALLY)
EXTERNAL TIMING DATA
EXTERNAL PITCH DATA

Fig. 2—Overall synthesis system, block diagram.

filters, sampled at their Nyquist rates (typically 33-1/3 s^{-1}), quantized, and stored in the word catalog as the reference library. The typical bit-rate used for storage of these data is 700 b/s when the pitch signal is saved. When pitch is not saved (the usual situation since it is normally calculated by the concatenation program) the bit-rate for the stored data is 533 b/s. Table I shows a breakdown of how these bit-rates are achieved. The data in this table were derived from experimental investigation of the effects of smoothing and quantization on the perception of the synthetic output.[4]

As shown in Table I, at every 10-ms interval the speech is classified as voiced or unvoiced (V/U) by a 1-bit signal. Thus, for each frame, storage is required for either voiced parameters or for unvoiced parameters, but not for both. It should be noted that the control parameter frame rate (33-1/3 s^{-1}) is one-third the rate of the V/U signal. The manner in which the raw data (which are obtained at a 100 s^{-1} rate) are coded to the lower rate, consistent with the frame rate of the V/U signal, is described in Ref. 3.

Once input words and phrases are coded in terms of the formant representation, they can easily be modified for use with the synthesis program. Words can be lengthened or shortened, formants can be changed easily, and a pitch contour, different from the one originally spoken, can be superimposed on the data. Thus the vocal resonance data is available to the synthesis program in a form flexible enough to conform to the timing and pitch generated by the concatenation program.

The lower portion of Fig. 2 shows how the system assembles a synthetic message composed of words and phrases from the reference library. First, the answer-back program requests the word sequence for a specific message. The word concatenation program first determines timing data for the message (in one of several ways to be explained below) from an auxiliary program. The timing data is in the form of a word duration for each word in the output message. The concatenation program then accesses, in sequence, the control parameters for each of the words in the string. A duration modification adjustment on each word is first made, so the word duration in context matches the duration specified by the timing rules. Next the concatenation program smoothly interpolates the formant control parameters when the final part of any word and the initial part of the following word are both voiced. An interpolation algorithm designed to

TABLE I—CODING OF FORMANT PARAMETERS

Parameter	No. bits/frame	No. frames/second	No. bits/second
F_1 or F_p	3	33-1/3	100
F_2 or F_z	4	33-1/3	133-1/3
F_3	3	33-1/3	100
P	5	33-1/3	166-2/3
A_V or A_N	3	33-1/3	100
V/U	1	100	100
		Total	700
		Pitch	−166-2/3
Data rate for synthesis using calculated pitch data			533-1/3

produce physiologically realistic formant transitions is used. Finally a continuous function for pitch variation is produced for the whole message. All computed control parameters are outputted to a hardware digital speech synthesizer designed in accordance with Fig. 1. Digital-to-analog conversion produces a continuous synthetic speech output.

In the remainder of this section we will detail the way each of the above operations is carried out. In Section IV we will give an illustrative example of the use of the system for the synthesis of 7-digit telephone numbers. Further, we will describe a dialing experiment, using the synthesized numbers and the DDP-516 in an interactive manner, to estimate the communicative effectiveness of the synthetic speech.

The duration computations and the interpolation algorithm of the concatenation program depend upon a measure we call *"spectral derivative."*

3.1 *Spectral Derivative*

The control parameters are stored in the catalog at a sampling rate of 33-1/3 per second. When accessed and used in the synthesis program, however, they are interpolated to a rate of 100 per second; i.e., 10 ms between frames. For each 10-ms frame of a given word, a calculation is made of the absolute rate of change of the formant data from the previous frame. We call this calculation the spectral derivative, since it is a measure of how rapidly the spectrum is changing. The spectral derivative is used to determine where to lengthen or shorten a word, and is also used to determine at what rate a formant transition is made from one voiced interval to the next.

For each voiced 10-ms interval, the spectral derivative, SD_i, is computed as:

$$SD_i = \sum_{j=1}^{3} | F_j(i) - F_j(i-1) | \tag{1}$$

where i is the ith 10-ms interval in the word, and $F_j(i)$ is the value of the jth formant in the ith time interval. This measure of spectral change is an arbitrarily chosen one; several others could be considered. For instance a weighted sum of absolute values of formant change:

$$SD_i = \sum_{j=1}^{3} a_j \, | F_j(i) - F_j(i-1) | \tag{2}$$

might be a suitable replacement for equation (1) above. By adjusting the weights, a_j, the influence of changes in individual formants can be made large or small. For example, by making a_2 much larger than a_1, or a_3, the spectral derivative is essentially the absolute change in the second formant. A more reasonable choice for the weight, a_j, might be the average value of the jth formant. The spectral derivative would then be the sum of relative changes in the formants. Although there are several possibilities for spectral derivative, the measure of equation (1) is the one we use throughout.

3.2 *Timing Calculation*

The timing calculation essentially consists of determining the duration of each of the words and phrases in the context of the message to be produced. There are several possible methods we have considered for determining these durations—ranging from fully automatic rules, which use syntactic and grammatical information, to manual insertion into the program of the desired timing sequence.

One technique, and the most accurate way of obtaining timing data, is to make measurements from a naturally spoken version of the message and manually supply these data to the program. This possibility is indicated at the bottom of Fig. 2 as the external timing data input to the timing subroutine. The timing data obtained in this manner are optimum and can be used to evaluate the efficacy of other aspects of the synthesis rules. This form of input is therefore important for evaluational purposes.

A second technique for obtaining word duration data is to make the duration of each word be some fixed percentage of its duration in isolation, independent of the message context. The motivation here is that the duration of the word in isolation is an overbound of its duration in context because of the unusually long vowels when spoken in isolation. Hence some shortened version of the word would suffice in many contextual situations. Clearly, the more limited the context of the message, the more applicable is the above approximation.

Another technique for obtaining durational data is by simple table-lookup procedures. Here the duration of every possible input, in every possible contextual position must be tabulated. For limited context messages, such as telephone number generation, this table-lookup procedure is an attractive way of generating timing information because of the limited number of situations which arise in practice. For more general situations, the amount of storage necessary would often become prohibitive.

The most sophisticated way of generating timing data is to make calculations based on language rules. A syntactic and phonetic analysis of the printed text of the message is converted by rules into durational data about each of the phonemes in the message. For the most general cases of speech synthesis, i.e. unrestricted context, this kind of procedure is an absolute necessity to give good timing data. A computer program for such sophisticated analysis has recently been developed.[5,6] Continuing work is aimed at combining this program with the concatenation system.

3.3 *Word Duration Modification*

Once the duration of the jth word in the message has been determined by one of the methods discussed in the previous section, it is then necessary to modify the set of control signals of the reference version of the word to match the desired duration. Assume the duration of the reference version of the jth word is w_j frames and the desired duration is d_j frames where a frame is 10 ms long. If we define the symbols:

$$I_P(j) = \begin{cases} 1 & \text{if the end of the } (j-1)\text{st word is voiced, and the beginning} \\ & \text{of the } j\text{th word is voiced.} \\ 0 & \text{otherwise} \end{cases}$$

$$I_F(j) = \begin{cases} 1 & \text{if the end of the } j\text{th word is voiced, and the beginning of} \\ & \text{the } (j+1)\text{st word is voiced.} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$b_j = w_j - d_j + \frac{t_c}{2} \cdot (I_P(j) + I_F(j)), \tag{3}$$

where

t_c = duration (in frames) over which voiced intervals are concatenated

and

b_j = number of frames to be eliminated from (if $b_j > 0$) or added to (if $b_j < 0$) the jth reference word.

The reason for the last term in equation (3) is that whenever adjacent voiced intervals occur between words they are smoothly merged together. Hence their durations overlap and it is this last term which accounts for the overlap. Typical values of t_c are 4 to 10; i.e., 40 to 100 ms overlap between voiced words.

The manner in which the b_j frames are eliminated, or added in, is based solely on the spectral derivative. To eliminate frames, the b_j frames in the word having the smallest spectral derivatives are removed. To add frames, the region of the word having the smallest spectral derivative is located, and b_j consecutive frames are inserted in the middle of this region. The parameter values during the inserted frames are identical to those of the frame nearest the middle of the region. The rationale behind this technique is that to lengthen or shorten a word, by any significant amount, it is most desirable to do this in parts of the word where the spectrum is changing the least. Thus the dynamics of the word are always unaltered by this method. A linear compression, or expansion, of the whole word is a useful technique only when the compression or expansion ratio is close to 1.0. This is not always the case in synthesis, and so the above technique is used instead.

3.4 *Merging of Isolated Words*

Generally, the manner in which the control signals from isolated words are combined is by abutting them directly, once the timing modifications described above have been made. However when the words to be combined have a common voiced interval (i.e. the end of the one word and the beginning of the next word are both voiced), a more complicated procedure is used to merge the words. This is because merely abutting the words would often produce cases where formants on one side of the word boundary would be vastly different from formants on the other side of the boundary. If such data were merely abutted, then in synthesizing the message objectionable transients would be present at the boundary. To alleviate this problem, a merging interpolation algorithm is used. The algorithm is based on the spectral derivative, and provides smooth formant transitions from one word to the next.

The merging procedure combines data over the last t_c frames of the first word and the first t_c frames of the second word. The duration of t_c frames is called the overlap region of the words. The average spectral derivative during this region, for both words, is calculated as:

$$\overline{SD1} = \frac{1}{t_c} \sum_{i=1}^{t_c} SD1\,(i) \tag{4}$$

$$\overline{SD2} = \frac{1}{t_c} \sum_{i=1}^{t_c} SD2\,(i) \tag{5}$$

where $SD1(i)$, and $SD2(i)$ are the spectral derivatives for the two words during the t_c overlap frames. Using the notation:

$F_j(i)$ = value of the jth formant at frame i during the overlap region

$F_j^k(i)$ = value of the jth formant at frame i during the overlap region' for word k

then the interpolation function used is:

$$F_j(i) = \frac{F_j^1(i)\cdot(t_c - i - 1)\cdot\overline{SD1} + F_j^2(i)\cdot i\cdot\overline{SD2}}{(t_c - i - 1)\cdot\overline{SD1} + i\cdot\overline{SD2}}, \qquad i = 1, 2, \cdots, t_c \tag{6}$$

Figure 3 illustrates the type of interpolation performed for four simple cases. (Although all three formants are interpolated in the program, for simplicity just one formant is drawn in Fig. 3 for each word.) The interpolated curve always begins at the formant of the first word, and terminates at the formant of the second word. The rate at which the interpolated curve makes the transition from the formants of the first word, to those of the second word, is determined by the average spectral derivatives $\overline{SD1}$, and $\overline{SD2}$. For case 1, in Fig. 3, $\overline{SD1} \approx 0$ so $\overline{SD2} \gg \overline{SD1}$; hence the interpolated curve makes a rapid transition to the formant of word 2. Case 2 is the reverse of case 1: here $\overline{SD1} \gg \overline{SD2}$,

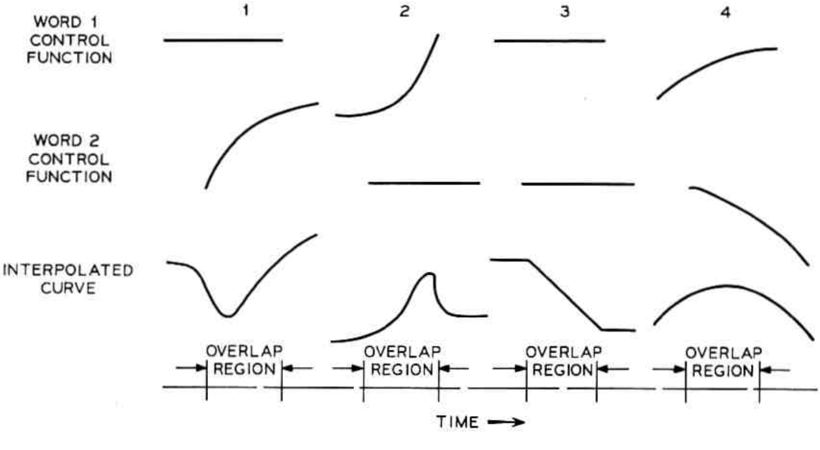

Fig. 3—Interpolation of control parameter contours for four typical cases.

so the transition does not occur until near the end of the overlap region. For case 3, both words have a small spectral derivative; hence the interpolation function degenerates into a linear transition. For case 4, both words have large spectral derivatives; hence the transition occurs about midway through the overlap region.

The data of Fig. 3 show that the interpolated formant function tends to be a smooth, continuous curve when the above technique is used. Values for t_c, the number of frames in the overlap region, have been from 4 to 10; i.e., 40 ms to 100 ms overlap.

3.5 *Pitch Calculation*

One of the most important aspects of speech synthesis is the determination of a suitable pitch variation for the message being produced. We have considered several ways of obtaining pitch information. These have included:

(i) Supplying a pitch contour extracted from a naturally spoken version of the message: These data, when used with similarly extracted timing data, give the most natural sounding messages that can be obtained with the technique. This form of input is most useful for evaluation purposes, but is not practical for an automatic system.

(ii) Using an archetypal pitch contour: For limited context applications this technique supplies a contour with realistic intonation, and hence is quite acceptable. The use of monotone pitch throughout the message is a special case of an archetypal contour, but such a contour gives an unacceptable drone to the speech, and hence would only be used in special situations.

(iii) Calculating a pitch contour by rule based on a stress analysis of the text of the message: This is a difficult task to do, but is most appropriate for an unlimited context, fully automatic system. Present research[5,6] on this topic makes it an attractive possibility for incorporating into a concatenation system.

(iv) Using the pitch variations associated with the isolated versions of the word, and concatenating them to give the overall pitch contour: This technique is unacceptable, unless several versions of each word are stored in the library, because the pitch contour of the isolated word tends to characterize the word only in isolation. The pitch usually rises sharply at the beginning of the word, and falls sharply at the end of the word. When concatenated, the words sound distinct, rather than merging into a continuous message.

3.6 *Gain Parameters*

The voiced gain parameter, A_V, is preserved on a frame-by-frame basis along with the formants. When formants are merged, the gain parameter is also merged. The unvoiced gain parameter, A_N, is also preserved on a frame-by-frame basis along with the fricative pole and zero. A_N is not required to be merged.

IV. AN ILLUSTRATIVE EXAMPLE

Figure 4 illustrates how these synthesis rules are applied in a typical case. At the top of this figure are shown the resonance data for four words spoken in isolation. The first and fourth words are entirely voiced, and the second and third words contain both voiced and unvoiced sections. The duration of each of the words spoken in isolation is shown by the w_i's in Fig. 4a. In order to form a message composed of these four words the following steps occur:

(*i*) The duration of each word in the specified context is determined.
(*ii*) Duration adjustments are made (frames removed or inserted) to match the timing of step *i*.
(*iii*) Since words 1 and 2 do not share a common voiced interval, the time adjusted control signals for word 1 are accessed.
(*iv*) Since words 2 and 3 do share a common voiced interval, all but the last t_c frames of the time adjusted control signals for word 2 are accessed and abutted to the controls from word 1.
(*v*) The last t_c frames from word 2 are interpolated with the first t_c frames from word 3, and added on to the previous control signals.
(*vi*) Since words 3 and 4 do not share a common voiced interval, the remaining control signals for word 3, and the time adjusted control signals for word 4 are added on to the previous control signals.
(*vii*) A pitch contour for the entire message is calculated.
(*viii*) The message is synthesized.

The resulting control signals and pitch contour are shown in Fig. 4b.

4.1 *Synthesis of Telephone Numbers*

For evaluation of this technique we chose the limited context situation of synthesis of the carrier phrase "The number is" followed by a 7-digit telephone number. Here, the timing was generated by a simple table-lookup procedure. The timing data we used are shown in Table

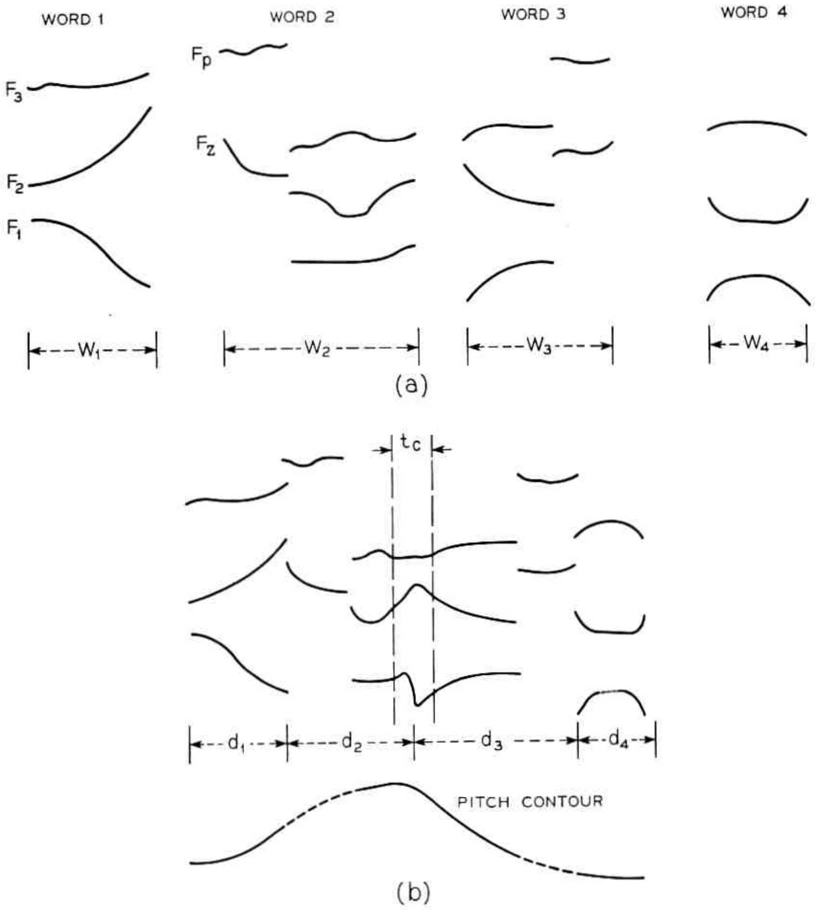

Fig. 4—Typical example of how control parameters are generated from the word library store. A message composed of four words is illustrated. All parameters are functions of time.

II. The table shows the digit duration (in milliseconds) as a function of the number of phonemes in that digit and its position in the string. The data in the table were obtained from measurements on real 7-digit numbers and, in effect, constitute first-order statistics on duration. The influence of context (position in the digit string) is easily seen in Table II. For example, any digit in the third position is from 50 to 90 percent longer than the same digit in the sixth position.

A single archetypal pitch contour was used in all cases. The arche-

TABLE II—SIMPLE TIMING RULES FOR 7-DIGIT NUMBERS

Position in Digit Sequence	Time, Milliseconds			
	Phonemes/Digit			
	1	2	3	4
1	250	330	410	490
2	280	330	390	450
3	450	500	560	610
4	260	300	340	380
5	340	370	410	440
6	230	280	340	390
7	290	380	460	550

typal form was taken to match as well as possible the general shape of the pitch contours measured in naturally spoken 7-digit numbers. This basic shape was used to calculate the pitch contour for each number string requested by the answer-back program. Informal listening suggested that this pitch contour was adequate as an initial estimate, and was a substantial improvement over the pitch information associated with individual isolated words.

The synthesis program ran on the Honeywell DDP-516 computer. The isolated digits were analyzed and stored in the computer memory at a data rate of 533-1/3 b/s. The concatenation program accepted an input sequence from the typewriter or card reader, computed the control signals for the message, smoothed them by programmed digital filters, and outputted the data to a hardware digital terminal analog synthesizer[7,8] in real time. Figure 5 shows a spectrographic comparison between a typical computer-generated 7-digit number, and a natural version of the same number. The timing and formant data of the synthesized example are seen to be reasonably good matches to those of the natural utterance.

4.2 *Dialing Experiment Using Synthetic Speech*

To evaluate the communicative effectiveness of the concatenated synthetic speech in a real dialing situation, we arranged for the DDP-516 computer to speak telephone numbers (both natural and synthetic) to a listener in a sound booth. The listener was provided a conven-

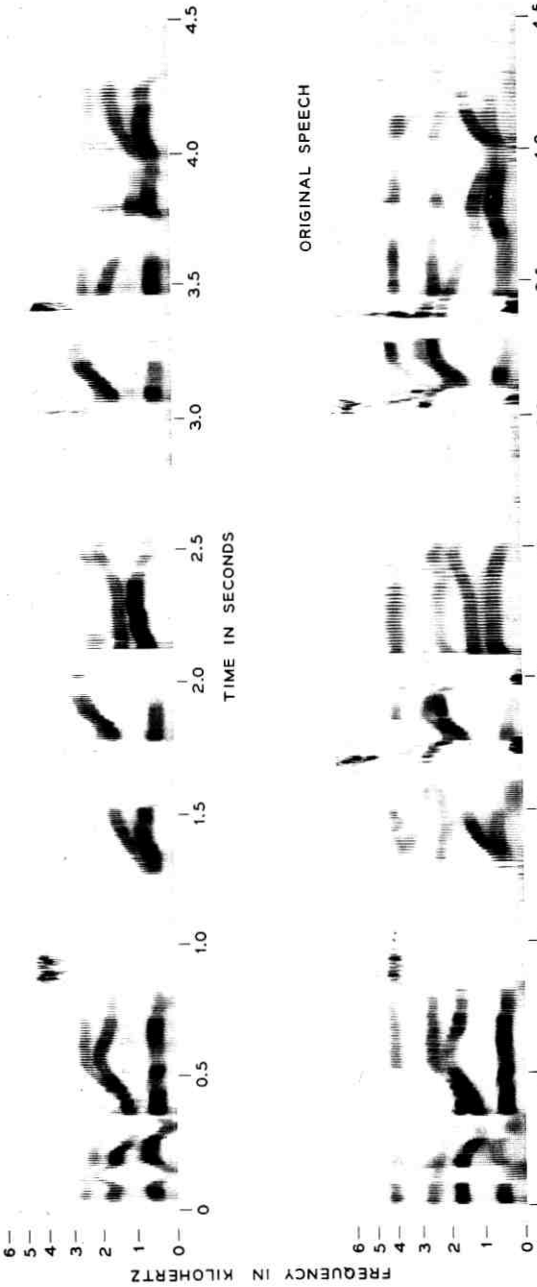

Fig. 5—Spectrogram comparison between synthetic and natural versions of a typical telephone number.

tional *Touch-Tone* telephone with which he could dial the numbers. A central office *Touch-Tone* decoder received the dial pulses, decoded them, and presented them via a data channel to the computer. The computer maintained a running analysis of the results. The experimental arrangement is shown in Fig. 6.

In the experiment we compared four types of speech. These included:

I. Naturally-spoken, 7-digit telephone numbers.
II. Naturally-spoken, isolated digits, abutted together.
III. Synthetic isolated digits, abutted together.
IV. Concatenated digits produced by the concatenation program method.

Listeners, seated in a sound booth, heard telephone numbers over the *Touch-Tone* telephone. After a prescribed delay, they were required to dial the number just heard. The DDP-516 computer generated the signal, read the number dialed, and tabulated the results.

Figure 7 shows the total number of dialing errors for 12 subjects. The dialing errors are broken down into digit errors (i.e., number of digits incorrectly dialed) and telephone number errors (i.e., number of phone numbers with one or more digit errors). The lower pair of curves shows the number of digit errors and the number of phone-number errors for 1-second delay in dialing of speech. The upper pair of curves shows the corresponding results for 5-seconds delay in dialing.

An analysis of variance of these data indicated that, at the 95 percent level of confidence, there existed no significant difference between dialing performances with the natural and the synthetic concatenated signals (i.e., between speech types I and IV). In other words, synthetic, concatenated speech is comparable to natural speech in dialing effectiveness.

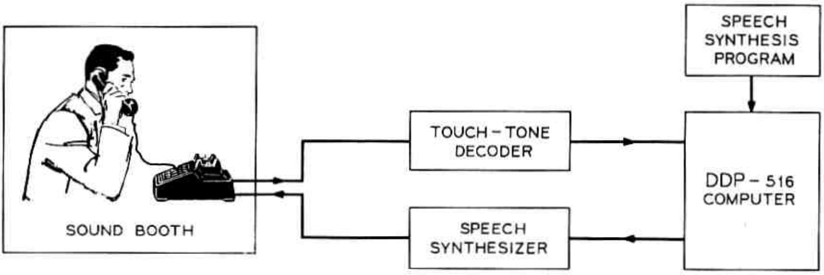

Fig. 6—Experimental arrangement used to measure the communicative effectiveness of several types of speech.

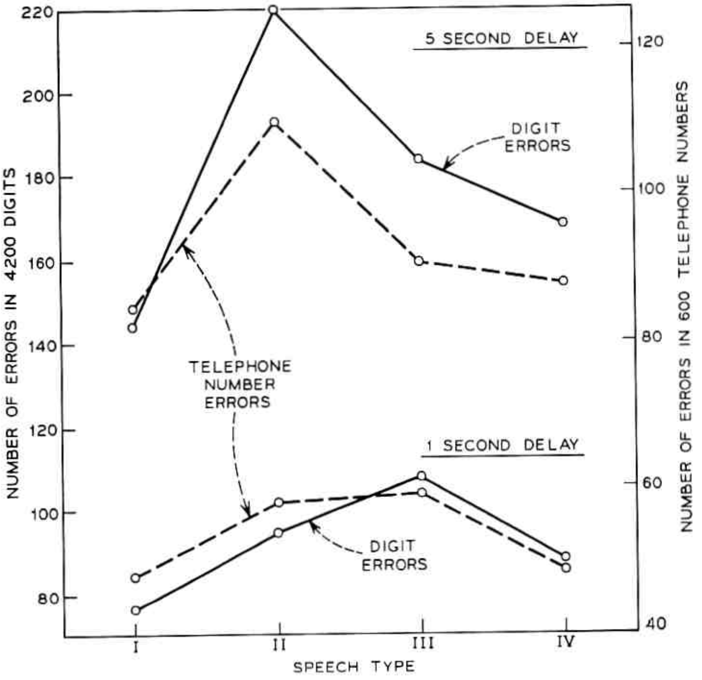

Fig. 7—Experimental results showing the total number of digit errors and telephone number errors. Four types of speech are tested: I, natural digits; II, natural, abutted digits; III, synthetic abutted digits; IV, concatenated digits. Response delays of 1 and 5 seconds are tested.

The differences between speech types I or IV and types II or III, however, was significant at the 95 percent level. That is, digital strings produced by simple abutting (II and III) led to a greater number of dialing errors. The suggestion is that the concatenation program is effective in reducing the dialing errors over that which would result from mere abutting of the digits.

Another factor of interest, of course, is the naturalness of the signal. Some preliminary informal experiments indicate that listeners rank the naturalness of these four signals in order of the "machine attributes," i.e., type I speech is ranked most natural, followed by types II, III, and IV. The synthetic concatenated signal has more machine-made features than any of the others—with pitch and duration both being calculated by machine. One might be willing to accept machine accent if the signal has attractive advantages in communicative accuracy and economy of storage. Formant synthesis using the concatenation technique appears to have both.

V. ACKNOWLEDGMENT

We wish to thank J. D. Robinson for conducting the dialing experiment described here, and D. Bock for designing and implementing the *Touch-Tone* decoder/DDP-516 interface.

REFERENCES

1. Flanagan, J. L., *Speech Analysis, Synthesis and Perception,* New York: Academic Press, 1965, Chapter III.
2. Schafer, R. W., and Rabiner, L. R., "System for Automatic Analysis of Voiced Speech," J. Acoust. Soc. Amer., *47,* (February 1970), pp. 634–648.
3. Schafer, R. W., Rabiner, L. R., and Graham, N., "Formant Analysis, Synthesis and Coding for Computer Voice Response," unpublished work.
4. Rosenberg, A. E., Schafer, R. W., and Rabiner, L. R., "Effects of Smoothing and Quantization of the Parameters of Formant-Coded Speech," unpublished work.
5. Coker, C. H. and Umeda, N., "Text to Speech Conversion," IEEE International Convention Record, New York, N. Y., March 1970.
6. Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N., "Synthetic Voices for Computers," IEEE Spectrum, *7,* (October 1970), pp. 22–45.
7. Rabiner, L. R. "Digital Formant Synthesizer for Speech-Synthesis Studies," J. Acoust. Soc. Am., *43,* (April 1968), pp. 822–828.
8. Rabiner, L. R., Jackson, L. B., Schafer, R. W., and Coker, C. H., "Digital Hardware for Speech Synthesis," Seventh International Congress on Acoustics, Budapest, August 1971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Listener Evaluation of Simulated Telephone Calling Signals

By P. D. BRICKER

(Manuscript received December 2, 1970)

This research concerns the judged pleasantness of a variety of electronic calling signals. Five experiments are reported in which type and frequency of carrier; type, frequency, and waveform of modulation; and spectral composition were varied. The results have aided in the selection of two signals for further trials.

I. INTRODUCTION

Technological considerations suggest that an electronic "ringer" may succeed electromechanical devices in future telephones. This possibility has generated interest within the telephone industry in the specification of desirable characteristics for electronic calling signals. Not the least important of these characteristics is that such signals be acceptable to the subscriber on purely aesthetic grounds. The literature on tone ringers includes some references[1,2] to the measurement of listeners' opinions, but only recently has any systematic work on what constitutes a pleasant signal begun to appear.[3] An earlier study by P. D. Bricker and J. L. Flanagan[4] marked the beginning of an attempt to chart the preference-relevant dimensions of a fairly large class of calling signals. The present paper reports five subsequent experiments that have clarified the effects on evaluative judgments of a half-dozen parameters. These experiments have interacted with studies of calling-signal detectability and with development work to produce two distinct realizable signals, which are scheduled for evaluation in a field trial.

II. EXPERIMENT 1

2.1 *Background*

The first experiment was identical in form to that reported by Bricker and Flanagan.[4] That is, listeners heard one signal at a time and assigned each a number that reflected their opinion of the signal. This technique, in conjunction with a means of analyzing the data in terms of perceptual attributes, had provided considerable information about a limited variety of signals in the earlier study. The purpose of the present experiment was to obtain a rough idea of the parameters important to evaluation in a much larger domain of signals, to serve as a guide to more detailed investigations.

2.2 *Signals*

There were 100 different signals in this experiment of three basic types:

(*i*) Thirty-four *a*mplitude-*m*odulated *p*ulse-train-*c*arrier (AMPC) signals, which were a subset of the signals studied by Bricker and Flanagan.[4] These represented selected combinations of four modulation frequencies, three duty factors, three carrier frequencies, and three harmonic compositions.

(*ii*) Six *a*mplitude-*m*odulated *s*inusoidal-*c*arrier (AMSC) signals, representing three modulation frequencies and two carrier frequencies.

(*iii*) Sixty *f*requency-*m*odulated *s*inusoidal-*c*arrier (FMSC) signals, representing selected combinations of six modulation frequencies, three carrier frequencies, three amounts of frequency deviation, and five modulation waveforms.

2.3 *Listeners*

Forty-three persons of various nonsupervisory employment classifications at Bell Laboratories served as listeners.

2.4 *Procedure*

Groups of four to six listeners, seated around a table in a carpeted room with draperies, listened to one of four permutations of the 100 signals reproduced over a high-quality magnetic tape playback system. They were instructed to record a positive number on the answer sheet for signals they liked and a negative number for signals they disliked; the greater the degree of liking or disliking, the larger the positive or negative number. A new signal occurred every 6 seconds, so that the entire procedure, including reading the instructions and rest periods, required about 15 minutes.

2.5 *Analysis*

The listeners' ratings were arranged in a matrix of 43 rows (listeners) by 100 columns (signals) and each row was normalized so that it had $\mu = 0$ and $\sigma = 1$. These data were analyzed by the MDPREF computer program of J. J. V. Chang and J. D. Carroll[5] so as to produce a spatial representation of both signals and listeners. MDPREF solutions represent the stimuli (calling signals, in this case) as points and the subjects (listeners) as vectors in multidimensional space in such a way that the projections of the points on each subject's vector correspond maximally, in a least squares sense, to his input data vector. Another property of these solutions is that successive dimensions are orthogonal to those preceding and account for as much of the residual variance as possible. It is left to the experimenter to determine how many solution dimensions will be considered significant and how he will rotate the axes to render the solution interpretable.

Another technique found useful in interpreting the results of this experiment was to regress the coordinates of the stimulus points on physical property vectors, so as to locate vectors maximally corresponding to the parameters that were varied to generate the stimuli. In the earlier experiment,[4] regression techniques had been used to find a three-dimensional structure in the data that was interpretable in terms of three parameters of signal design.

Finally, S. C. Johnson's hierarchical clustering analysis[6] was applied to the data quite independently of the multidimensional scaling. This technique groups stimuli (signals) according to their mutual closeness in terms of a distance measure provided by the user. The interstimulus distance measure used for these data was defined as follows:

$$d_{jk}^2 = \sum_i (R_{ij} - R_{ik})^2,$$

where d_{jk}^2 is the squared distance between stimuli j and k and R_{ij} is the rating given to the jth stimulus by the ith subject. This measure hopefully reflects the similarity of treatment of two signals by each listener. The computer-implemented Johnson technique produces a hierarchy of clustering of n objects, running all the way from n "clusters" of one object each to one cluster of n objects. It also computes a measure of compactness of the clusters at each level between these extremes. Using this measure, as defined by Johnson,[7] we traced the compactness of various clusters as they grew in size to maximum compactness, in an effort to define types of signals.

2.6 *Results*

The first six dimensions of the MDPREF solution accounted for 28, 22, 6, 4, 4, and 3 percent of the variance, respectively. The large drop after the second dimension suggests that only two dimensions of the solution are interpretable.

Of the parameters used for regression analysis, only modulation frequency (MF) could be located with sufficient confidence to identify it with a dimension of the solution. However, interpretation of the two-dimensional solution as a whole was greatly aided by the cluster analysis. Six clusters, including almost all the stimuli, were found to be at maximum compactness at about the same level in the hierarchy. Inspection of the membership of each cluster suggested a name for the type of signal comprising the cluster. Furthermore, when the stimulus points were projected on the plane of the first two MDPREF dimensions, a fairly simple closed curve could be drawn around each of the clusters without overlapping the others. This projection, along with the cluster contours and the subjects' vectors, is shown in Fig. 1. The pulse-carrier signals are shown as open squares, the sine-carrier AM signals as shaded circles, and the sine-carrier FM signals as filled circles. Subjects' vectors, normalized to unit length in two dimensions*, are shown as arrowheads pointing in the direction of higher evaluation. The vertical axis (Dimension I) corresponds to the MF vector, with low MF (5, 7, 10 Hz) at the bottom of the figure and high MF (20, 40, 80 Hz) toward the top.

The import of the results for calling-signal design is conveyed by consideration of the gross characteristics of the signal clusters. Although a more detailed examination of the results with respect to parameter levels reveals interesting information about tone perception, these findings are deemed inappropriate for present purposes. Complete details are available from the author.[8]

The signals in both clusters below the horizontal axis of Fig. 1 ("gliding pitch" and "trills") are, with three exceptions, FM signals modulated at a slow enough rate that the pitch can be heard to change. In the case of "gliding pitch" signals, the pitch changes in a continuous manner, while the tone of the "trills" jumps from one pitch to another. The three exceptions are AMSC signals of which the single pitch alternates with silence. Note that no listener's vector is located so as to indicate a clear preference for gliding pitch signals, whereas quite a few listeners evaluate trills higher than any other type of signal.

* This normalization was adopted in the interest of presenting an uncluttered picture. A similar plot in which overall percentage of variance accounted for was represented by the distance of each arrowhead from the origin revealed no systematic relation between orientation and length of vectors.

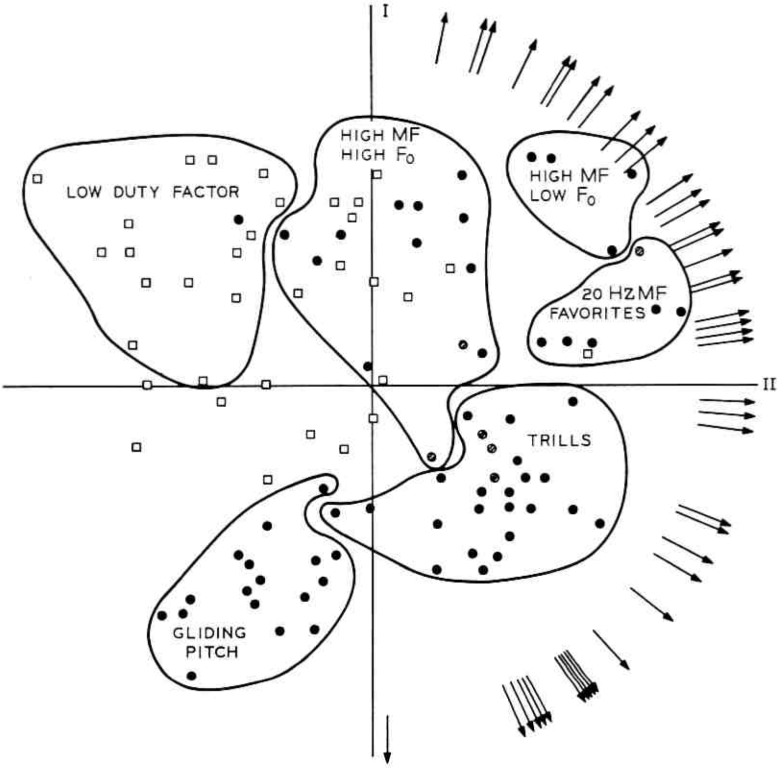

Fig. 1—Projections of 100 signals and 43 listeners' vectors on Dimensions I and II of the MDPREF solution for Experiment 1. Arrowheads represent listeners' vectors, filled circles FMSC signals, shaded circles AMSC signals, and open squares AMPC signals.

The one characteristic shared by all of the signals in the "favorites" cluster is an MF of 20 Hz. They also have low or medium carrier frequencies. This cluster is positioned so as to receive high ratings from many listeners who prefer 10- or 40-Hz MF somewhat more. Its name derives from the fact that it includes the signals with the three highest mean ratings and three others in the top ten. The single AMPC signal to reach this distinction was also first in the earlier study[4]; it appears again in Experiment 4.

The four signals in the "high MF-low F_o" cluster all have MF = 40 or 80 Hz and a carrier frequency of 400 Hz. Some listeners clearly prefer these signals but not many other high-MF signals, to those with lower MF.

The name of the "high MF-high F_o" group is self-explanatory ("high F_o" means carrier frequencies of 1,600 Hz or 800 Hz for AMSC and FMSC signals, and 900 or 700 Hz for AMPC signals). This group receives few high ratings. The last group ("low duty factor") is almost entirely AMPC signals with a low or medium duty factor, which renders harmonics of the modulating frequency prominent. No listener's vector is located so as to indicate a preference for these signals.

2.7 *Conclusions*

Some tentative principles of good calling-signal design suggested by these results are:

(*i*) There seems to be an optimum modulation frequency around 20 Hz, even though individuals vary widely with respect to this parameter. Both the superiority of 20 Hz and the diversity of listeners were observed by Bricker and Flanagan.[4]

(*ii*) Pitch should change abruptly rather than gradually at low modulation frequencies.

(*iii*) Smooth amplitude modulation (high duty factor) is superior to abrupt amplitude modulation (low duty factor).

(*iv*) Low carrier frequency and a narrow, low-centered spectrum are advantages, while high-frequency energy is a disadvantage, whether it arises from a high carrier frequency or the presence of harmonics of the carrier.

The first two principles receive further support from subsequent experiments, the third is not investigated further, and the fourth is explored and refined in the next three experiments.

III. EXPERIMENT 2

3.1 *Background*

The problem most in need of attention after Experiment 1 was the status of pulse-carrier signals. It is clear that most of the AMPC signals and the "gliding pitch" FMSC signals received low ratings, and that the two types are separated in the solution space (Fig. 1). However, they are separated chiefly on Dimension I, which corresponds to modulation frequency, and for very good reason: the AMPC signals in the study all had MF $\geqq$ 10 Hz, while the "gliding pitch" signals all had MF $\leqq$ 10 Hz. The two types project to similar points on Dimension II, which is enigmatic: this dimension could reflect either a perceptual characteristic or merely the confounding in the experimental design. The purpose of Experiment 2, then, was to determine whether AMPC signals required a dimension of their own to describe their perceptual relations with FMSC signals. The physical correlate of the dimension thought to be useful for this purpose is the bandwidth of the signals: AM *sine*-carrier signals have a narrower bandwidth than FMSC signals, which in turn are narrower than AMPC signals. Experiment 2 includes signals of all three types, each represented at the same values of MF so as to remove that source of confounding. Note that a narrow bandwidth was listed as a desirable characteristic in the fourth conclusion of Experiment 1, and that Experiment 2 is designed to provide additional information on this point.

Since this experiment and those that follow use a novel method of collecting data, the procedure and some results it has produced will be briefly reviewed. The technique, called auditory sorting, has been described in the literature.[9] Briefly, it provides the listener with an array of movable pushbuttons, each of which evokes a distinctive sound. The listener arranges the buttons (sounds) in groups or in order according to instructions. In an early experiment with this technique, listeners were asked to group 24 three-parameter frequency-modulated tones according to similarity. Using an appropriate multidimensional scaling technique, it was possible to recover a perceptual space that closely resembled that recovered from the much more laborious pair-comparison procedure in a companion experiment.[10] In another experiment, listeners ordered a subset of the tones used in Experiment 1 according to preference. From these data, MDPREF constructed a space very much like that based on the rating data for the same subset. The strategy in this experiment, then, was to include enough FMSC signals to recover a three-dimensional perceptual space and then observe whether the AMPC signals, the AMSC signals, or both required an additional dimension to account for their evaluations.

3.2 *Signals*

Twenty-four signals, all of which were derived from an 800-Hz carrier frequency, were used in this experiment. They were of three types: FMSC ($n = 18$), AMSC ($n = 3$), and AMPC ($n = 3$). Each type was represented at the same three modulation frequencies: 10 Hz, 20 Hz, and 40 Hz. There were six FMSC signals at each MF, realizing all combinations of two FM waveforms (sinusoidal and rectangular) and three amounts of frequency deviation (±3, ±10, and ±25 percent). The AM signals were modulated with a symmetrical sinusoidal envelope, and the PC signals employed a carrier with approximately equal-amplitude components at 800, 1,600, and 2,400 Hz.

3.3 *Listeners*

Thirty Bell Laboratories employees, 17 male and 13 female, served as listeners in this experiment.

3.4 *Procedure*

The listener was seated in a small sound-attenuating booth with the sorting apparatus in front of him and a loudspeaker on the wall above it. He was shown how each button evoked a different sound (signal) from the speaker, and instructed to arrange the signals (buttons) from right to left according to "how much [he] would like each of these tones if it replaced the telephone bell." Listeners were permitted to produce partial orderings, i.e., to arrange the signals in groups such that the evaluative ordering obtained between groups, but signals within a group were tied.

3.5 *Analysis*

The vectors were each normalized and arranged in a 30 (listeners) by 24 (signals) matrix to serve as input for MDPREF. Regression techniques were used to find, in the resulting space, three orthogonal directions that best corresponded to the three parameters of the FMSC signals. A fourth dimension was located so as to satisfy three conditions: (*i*) mutual orthogonality to the first three, (*ii*) maximization of residual variance accounted for, and (*iii*) close nonlinear correspondence to a property defined by three levels of spectral width—narrow (AMSC), medium (FM), and wide (AMPC).

3.6 *Results*

In the figures that present the results, plotting symbols have been used that suggest the parameter values of the signals they represent. Thus, large symbols are used for 25-percent frequency modulation, medium for 10-percent, and small for 3-percent; round symbols for sinusoidal FM, square for rectangular FM. Modulation frequency in Hz is given by arabic numerals. For the AM signals, sinusoidal and pulse carriers are distinguished by symbols representing one cycle of their respective waveforms.

The projections of all 24 signals on the plane of the first two solution dimensions are shown in Fig. 2a, and their projections on the plane of the third and fourth dimensions are shown in Fig. 2b. For the 18 FMSC signals, the first three dimensions represent modulation frequency, modulation percentage (MP), and modulation waveform (WF), re-

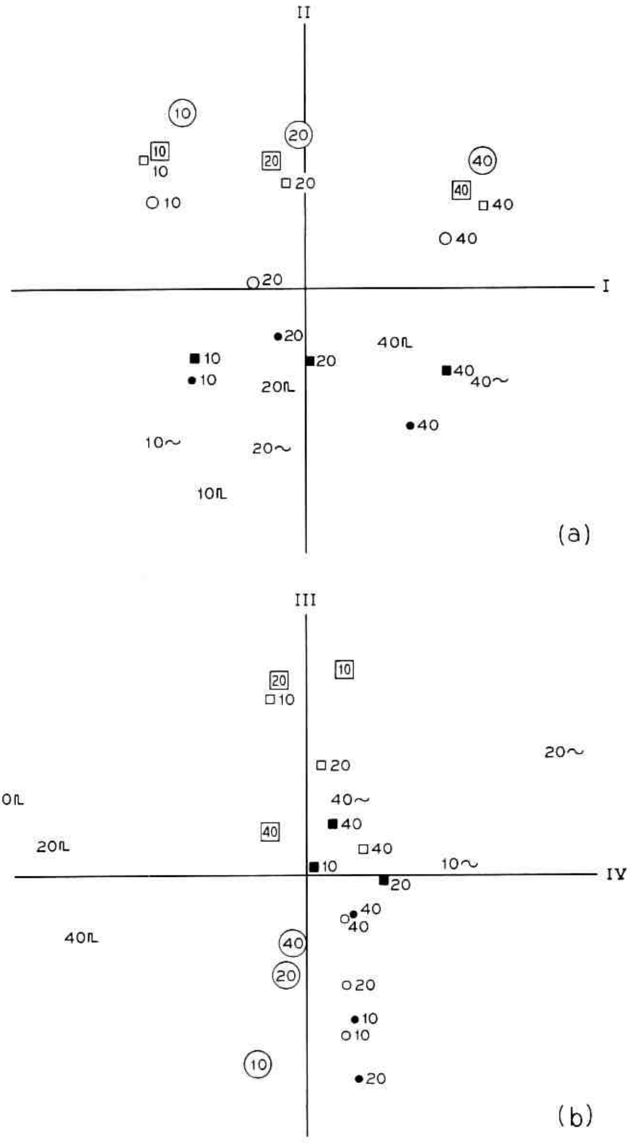

Fig. 2—Projections of 24 signals on the MDPREF solution dimensions for Experiment 2: (a) Dimensions I and II; (b) Dimensions III and IV.

spectively. This configuration resembles that obtained from the aforementioned similarity-judgment experiments[10] in sufficient detail to support the conclusion that the present experiment reveals dimensions of perceptual significance.

Note that the AM signals fall in appropriate places on Dimension I (MF) and that they are just beyond the 3-percent FM signals on Dimension II (MP); this latter location is consistent with their having 0 percent frequency deviation. The chief function of Dimension IV is to separate the AM signals, with the AMPC signals to the left and the AMSC signals to the right. This dimension accounts for 10 percent of the variance, which compares favorably with 28, 13, and 16 percent for the first three, respectively.

The listeners' vectors are shown in Fig. 3a and b. Whereas Fig. 3a shows the usual diversity of opinion with respect to MF (and MP as well), Fig. 3b shows a considerable concentration of vectors so as to reflect low ratings for both AMPC and low-rate sinusoidal FM signals. It is not clear from these figures whether evaluation continues to improve as bandwidth is reduced. Subjects are in fact evenly divided as to whether their highest-ranked AMSC signal is ranked above their highest-ranked FMSC signal, and mean normalized rank is slightly higher for FM. Thus, although bandwidth is established as a perceptually significant parameter, listener evaluation is not a monotonic function of it.

3.7 *Conclusion*

A reasonable accounting of the data demands that AM signals be regarded as differing from the (three-dimensional) FM signals on a fourth dimension. Although signal bandwidth provides a satisfactory interpretation of this dimension, its relation to listener evaluation is other than monotonic.

IV. EXPERIMENT 3

4.1 *Background*

While the experiments to this point had considerably clarified the nature of the perceptual space, they had not explored all combinations of parameter values. In particular, Experiment 2 suffers from confounding of carrier type with modulation type (FM or AM), even though it served to un-confound these parameters with modulation frequency. Thus we are left with the formal possibility that Dimension IV of Experiment 2 could be interpreted as "modulation type" rather

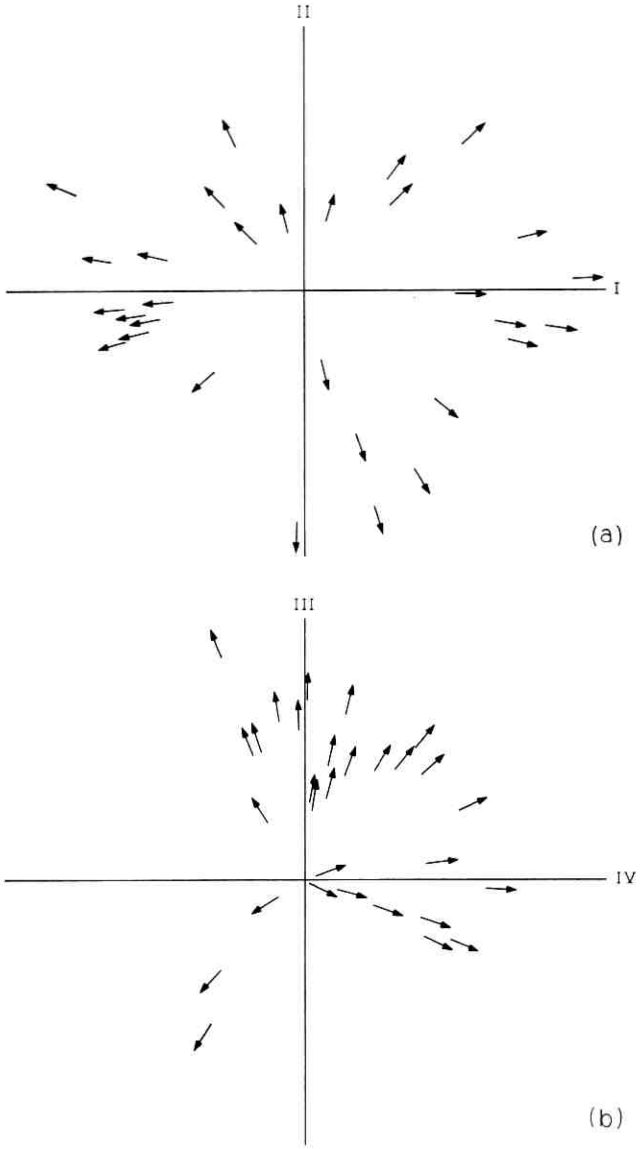

Fig. 3—Locations of listeners' vectors with respect to the dimensions shown in Fig. 2: (a) Dimensions I and II; (b) Dimensions III and IV.

than "bandwidth," and with the practical possibility that pulse-carrier signals might receive higher ratings if frequency modulated. Experiment 3 allows for complete factorial arrangement of both types of carrier with both types of modulation.

4.2 *Signals*

All possible combinations of three modulation frequencies (10, 20, and 40 Hz), two modulation types (AM and FM), and two carrier types (sinusoidal and pulse) were generated, for a total of twelve signals. The carrier frequency was 800 Hz, and the FM modulating waveform was rectangular with ±10-percent deviation. The pulse carrier contained the first three harmonics of the carrier and the AM parameters were as in Experiment 2.

4.3 *Listeners*

Forty-four Bell Laboratories employees served as listeners. They were each selected with approximately equal frequency from two categories each of age (over or under 30) and sex. In addition, each listener submitted to an audiometric screening test; none was found to exhibit a clinically significant loss in the range of the signals under test.

4.4 *Procedure*

Each listener produced an evaluative ordering or partial ordering of the 12 signals, using the sorting apparatus under the same instructions as in Experiment 2. At the termination of the ranking procedure, each listener was asked to state whether he liked "the bell on his home telephone" more than none, some, or all of the tones, and to locate it in the hierarchy if "some".

4.5 *Analysis*

The objective in this and subsequent experiments was more to identify optimum parameter combinations than to discover perceptual dimensions. Consequently, the results are presented in terms of conventional statistical summaries and tests of significance rather than multidimensional analysis. The basic datum is the rank assigned each of the 12 signals by each of the 44 listeners.

4.6 *Results*

The median rank assigned each of the 12 signals is shown in Table I, where lower numbers indicate higher evaluations. It is immediately apparent that sine carrier ranks higher than pulse carrier in each of the six comparisons with the other factors held constant. Neither of the

TABLE I—MEDIAN RANK ASSIGNED EACH OF TWELVE SIGNALS, EXPERIMENT 3

Modulation Type		FM		AM	
Carrier Type		Sine	Pulse	Sine	Pulse
Modulation	10	4.5	5.6	5.5	7.5
Frequency	20	2.8	6.9	3.8	6.9
(Hz)	40	5.8	9.4	4.7	9.3

other parameters shows quite so consistent an effect. An analysis of variance showed all three parameters to have statistically significant effects, with carrier type the largest and modulation type the smallest. Again 20 Hz is the highest-ranking modulation frequency; the rank distributions for most of the 10-Hz and 40-Hz signals reflect the diversity of opinion about MF observed in the earlier experiments, in that they are generally bimodal or broadly dispersed. Overall, FM is somewhat higher ranked than AM. Another finding of the analysis of variance was that there was no systematic difference among the four age-sex groups in their patterns of evaluation.

Table II shows, for each signal, the number of listeners who ranked it above or equal to their remembered concept of the bell. The pattern of preferences here is much the same as that shown by median rank (in fact, it would be statistically the same if there were no correlation between electronic signal evaluation and "bell" evaluation). The main value of this measure is to give the signal ratings some external reference, however tenuous.

4.7 *Conclusion*

The data show clearly that at least for a carrier frequency of 800 Hz, a signal containing three harmonics of the carrier is less well liked than its single-frequency counterpart, regardless of whether amplitude or frequency modulation is employed. Furthermore, there is no practi-

TABLE II—NUMBER OF SUBJECTS (OUT OF 44) RANKING EACH SIGNAL ABOVE OR EQUAL TO THE "BELL", EXPERIMENT 3

Modulation Type		FM		AM	
Carrier Type		Sine	Pulse	Sine	Pulse
Modulation	10	11	12	17	7
Frequency	20	23	11	19	5
(Hz)	40	14	5	12	4

cal advantage to frequency-modulating a pulse-carrier signal, although FMSC signals are again slightly superior to AMSC signals.

Discovery of a way to improve evaluations of pulse-carrier signals would be valuable, because such broadband signals have a well-established advantage in detectability. A search for such a means gave rise to the next experiment which although brief and unavailing was informative in other respects.

V. EXPERIMENT 4

5.1 *Background*

The best-liked FM signal (20-Hz, 10-percent rectangular FM) has a musical aspect worth noting: the two alternating frequencies (720 Hz and 880 Hz) stand in a relation close to a major third, generally thought to be a pleasing musical interval. The pulse-carrier signals used so far also have a musical aspect: the three components (e.g., 800, 1,600, and 2,400 Hz) establish two intervals—an octave and a major fifth. While not dissonant, the fifth is generally considered harsher and less pleasing than the third. But certain members of the harmonic series other than the first three can be chosen so as to generate thirds (fourth and fifth harmonic) and pleasing inverted (or open) triads. This experiment was an attempt to improve the rating of pulse-carrier signals by selecting such combinations.

5.2 *Signals*

Each of the eight signals in this experiment had three components whose relative amplitudes decreased at a rate of 3 dB per octave from 500 to 4,000 Hz. All signals were amplitude modulated as before at 20 Hz. The frequencies of all the components in kHz are shown in Table III. The more important musical aspects of this set are: (*i*) Signals 1, 2, 3, and 6 are *compact,* in that adjacent harmonics of the respective fundamentals are selected. Of the compact signals, only number 3 represents a major triad (second inversion); (*ii*) Signals 4, 5, 7, and 8 are *open,* in that certain harmonics are suppressed in between those that are passed. Each of the open tones constitutes a major triad, inverted and opened to span more than an octave.

5.3 *Listeners*

Bell Laboratories employees were selected to represent wide variation in musical skill and training, and to have normal hearing over the range of component frequencies. The experiment was terminated after six listeners had been run.

TABLE III—FREQUENCIES IN KHz OF COMPONENTS OF EIGHT SIGNALS, EXPERIMENT 4

Signal No.	1	2	3	4	5	6	7	8
First Component	0.5	1.0	1.5	1.5	0.5	0.75	1.0	1.25
Second Component	1.0	1.5	2.0	2.5	1.5	1.5	1.25	2.0
Third Component	1.5	2.0	2.5	4.0	2.5	2.25	3.0	3.0

5.4 *Procedure*

Listeners were asked to use the sorting apparatus to arrange the eight signals in evaluative order, as before.

5.5 *Results*

Regardless of the musical background, no subject ranked any of the five triad-producing signals first. The signal most often ranked first was number 1, which had the lowest frequencies for all three components. Number 6 was the only other signal ranked above the median by all listeners. Listeners described the low-ranked "musical" signals as "high-pitched," "tinny," and "jarring." One listener who was sophisticated in both music and acoustics recognized the differences in musicalness, but averred that the high-frequency components were irritating, even though they served to complete a chord. Since even those listeners expected to be most favorably disposed toward chord-signals rejected them in favor of signals with low component frequencies, the experiment was terminated after only six listeners.

In addition to rejecting the notion that musicalness of component intervals might "rescue" broad-spectrum signals, this experiment affords certain informative comparisons with earlier studies. For example, the first three signals have the same bandwidth, but differ in location of their spectra. Rankings plummet from first through average to last as frequency of components increases across this set. Signal number 6 has a greater bandwidth than any of these and a highest component between those of signals 2 and 3, yet its overall rank was equivalent to that of signal number 1. These facts suggest that listener evaluation depends in a complex way on the frequencies of the components, so that both the average frequency and the highest frequency can be determining. Bandwidth per se, it seems, is much less important. Searching through the details of Experiment 1 produces support for this notion, and further suggests that the region between 2,000 and 2,500 Hz is critical for upper component, 1,500 to 2,000 Hz for average. One would interpret the poor showing of pulse-carrier signals in Experiments 2 and 3 in retrospect as an invasion of a critical frequency region by the highest component, rather than as an effect of bandwidth or the mere presence of harmonics.

It happens that signal number 2 in Experiment 4 has exactly the same specifications as the lone AMPC signal to join the "favorites" cluster in Experiment 1 (Fig. 1). Since signal number 2 was outranked in Experiment 4 by signal number 6, even though the latter has a component above 2,000 Hz, we might expect signal number 6 to compete well with sine-carrier signals. However, the most similar signal in Experiment 3 (AMPC, 20 Hz, components at 800, 1,600 and 2,400 Hz) was considered equal to or better than "the bell" by only 5 out of the 44 listeners, compared to 23 out of 44 for the best FMSC signal (see Table II). Although it is possible that three harmonics of 750 Hz (especially with 3 dB per octave attenuation) could be much more pleasant than three harmonics of 800 Hz, it is safer, in the absence of a complete map of component-frequency effects, to take these Experiment 3 findings as a guide to how signal number 6 would fare against FMSC signals. The reason for the attention given here to signal number 6 is a practical one: its acoustic specifications are exactly the same as those of a ringer now under development.[11] This ringer has been shown[12] to be satisfactorily detectable in typical room noise, and to be superior in this respect to a narrow-spectrum FMSC signal. The present experiments suggest that while a three-harmonic 750-Hz signal is a good one of its type, it is likely to be less well liked by listeners than a good FMSC signal.

Since the laboratory affords no way to equate pleasantness and detectability, evaluation of both leading signals under operating conditions seemed an appropriate means of resolving the conflict between the criteria. To make this evaluation possible, the Telephone Laboratory at Indianapolis modified its basic design so that it could generate an FM signal. Questions that arose in the course of this redesigning effort prompted the next and last experiment.

VI. EXPERIMENT 5

6.1 *Background*

The ringer consists mainly of an electromagnetic transducer and a resonant cavity. The resonator must be small enough to fit inside a telephone station set and large enough to resonate the lowest frequency component of the desired signal. The higher the carrier frequency, the smaller the ringer could be, so the listener-evaluation function of carrier frequency (F_o, or average) is important design information Experiment 1 had indicated that signals with $F_o = 1{,}600$ Hz were not as well liked as those with $F_o = 800$ Hz, but there was also a suggestion that evaluation was not monotonic with F_o. In any event, the parameter F_o had not been explored in sufficiently small steps to guide the design of a ringer for narrow-band signals.

Another purpose of Experiment 5 was to assess the effects of superimposed amplitude modulation on listener evaluation. This question arose because such amplitude modulation was found to be technically difficult to eliminate from the design under consideration.

6.2 *Signals*

There were eight FMSC signals involved in this experiment. Each was rectangularly frequency-modulated at 20 Hz, with the upper and lower components standing in the ratio of 5 to 4 in frequency—a major third. There were four values of F_o, as shown in Table IV, along with the upper and lower component frequencies. There were two signals at each F_o, one of which was pure FM, the other of which was amplitude modulated at 20 Hz in such a way as to imitate the effect found in the practical design.

6.3 *Listeners*

Thirty listeners, 10 male and 20 female, were recruited from among Bell Laboratories' clerical, shop, and technical employees.

6.4 *Procedure*

Each listener used the auditory sorting apparatus to rank or partially rank the eight signals, as in Experiments 3 and 4.

6.5 *Results*

The number of listeners who assigned each rank to each of the eight signals is shown in Fig. 4. This method of presenting the data is resorted

TABLE IV—FREQUENCIES IN Hz OF F_o AND BOTH COMPONENTS OF SIGNALS TESTED IN EXPERIMENT 5

F_o	Upper Component	Lower Component
1,350	1,500	1,200
1,125	1,250	1,000
900	1,000	800
675	750	600

to here because the extreme biomodality of the ranking for $F_0 = 675$ Hz renders any measure of central tendency misleading. For practical purposes, one would wish to avoid a signal about which listener opinion is so divided. There is more of a consensus on low ranks for the signals with $F_o = 1{,}350$ Hz, and there is little to choose among the signals with $F_o = 1{,}125$ or 900 Hz.

Detailed analyses confirm the impression given by the figure that there is little difference between signals with or without superimposed AM. A tally was made of the outcome of each of the four same-carrier-frequency comparisons for each listener, the possible outcomes being FM > AM, AM > FM, and FM = AM. The results were that AM made essentially no difference to 11 listeners, while 6 listeners preferred FM only three or four out of four times, and 5 listeners preferred AM to FM only three or four times.

The results of this experiment are viewed as supporting two engineering decisions taken subsequently, rather than as evidence for more general conclusions. The first was to develop a narrow-band FM ringer with $F_o = 1{,}012.5$ Hz, which is midway between the two generally acceptable F_o in the experiment. The second was not to attempt to eliminate the superimposed AM, in the light of the indifference to it apparent in the experiment.

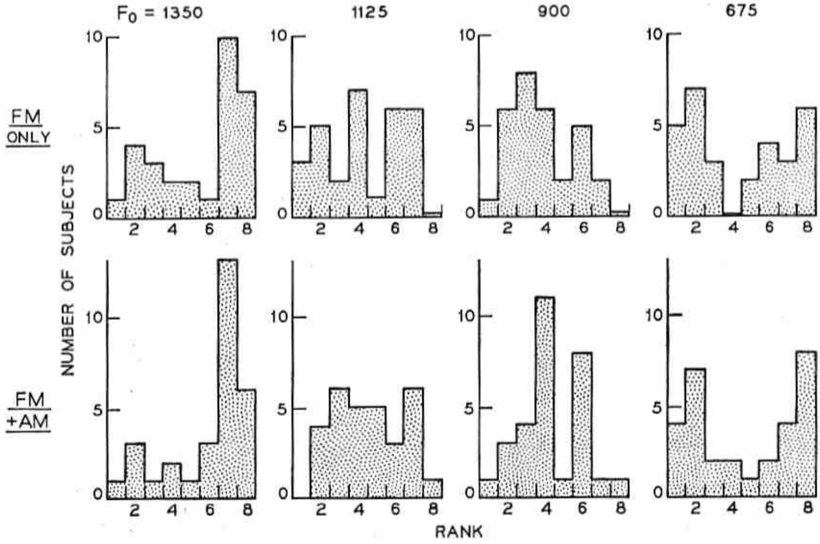

Fig. 4—Histograms of number of listeners assigning each of eight ranks to each of eight signals, Experiment 5; $N = 30$ for each histogram.

VII. SUMMARY OF CONCLUSIONS

Results of these experiments permit the specification of two ringers —one narrow-band and one broad-band—that are promising with regard to listener evaluation. The narrow-band signal is frequency modulated at 20 Hz, with its upper and lower components tuned to a frequency ratio of 5 to 4, averaging around 1 kHz. The broad-band signal is amplitude modulated at 20 Hz, and has three roughly equal-amplitude components at 750, 1,500, and 2,250 Hz. The evidence indicates that the narrow-band signal will be preferred to the broad-band signal by a majority of listeners when the two are in direct contest, and that the broad-band signal is more detectable in typical room noise when the two are equated for power.

Laboratory studies do not reveal, however, how listeners will evaluate either signal after some experience with it in actual use. Neither do they tell us how effective these signals will be in practice. A field trial involving residential subscribers is planned to gather information on these points. This trial also makes it possible for subscribers to tell us something about the tradeoff between the pleasantness advantage of the narrow-band signal and the detectability advantage of the broad-band signal by adjusting their volume controls. Informal experiments have indicated that listeners attenuate an unpleasant signal, when given a volume control in the laboratory. If subscribers behave similarly in the field, we shall have as one of the data the amount by which they offset the detectability advantage of the less pleasant signal. In any event, we shall collect data on answering times, no-answer rates, and opinions-after-experience. As useful as the laboratory studies have been, they could not have provided information on these crucial points.

The field trial will also provide an opportunity for a direct comparison of these two tone ringers with a widely used gong (C4). There are so many differences between tone ringers and gong ringers, ranging from the obvious difference in excitation to the factor of familiarity, that interpretation of this aspect of the study will be difficult. Nevertheless, these results may require modification of some of the principles (e.g., those pertaining to bandwidth) derived from this series of experiments on tone ringers alone.

VIII. ACKNOWLEDGMENTS

For their contributions to one or more phases of this work, the author is grateful to J. D. Carroll, Jih-Jie Chang, J. L. Flanagan, Barbara McDermott, Sandra Pruzansky, and M. Wish.

REFERENCES

1. Archbold, R. B., Ithell, A. H., and Johnson, E. G. T., "The Ideal Characteristics for the Calling Signal of a Subscriber's Telephone Set," Research Report No. 21143, Post Office Research Station, Dollis Hill, London, November 22, 1967.
2. Mevissen, H. M. J., and Stremmelaar, H., "A Study on the Appreciation of Tone Ringing by the Subscribers of a Fully Electronic Exchange," Fourth Int. Symp. on Human Factors in Telephony, Bad Wiessee, Germany, September 22–27, 1968.
3. Gale, J., "Human Factors and the Telephone," Northern Electric Telesis, *1,* No. 9 (October 1970), pp. 287–293.
4. Bricker, P. D., and Flanagan, J. L., "Subjective Assessment of Computer-Simulated Telephone Calling Signals," IEEE Trans. Audio and Electroacoustics, *AU-18,* No. 1, (March 1970), pp. 19–25.
5. Chang, J. J. V., and Carroll, J. D., "How to Use MDPREF, a Computer Program for Multidimensional Analysis of Preference Data," unpublished work.
6. Johnson, S. C., "Hierarchical Clustering Schemes," Psychometrika, *32* (1967), pp. 241–254.
7. Johnson, S. C., "A Simple Cluster Statistic," unpublished work.
8. Bricker, P. D., Carroll, J. D., McDermott, B., Pruzansky, S., and Wish, M., unpublished work.
9. Bricker, P. D., Johnson, S. C., and Mattke, C. F., "Apparatus for Auditory Stimulus Sorting," Behaviorial Research Methods and Instrumentation, *1* (1969), pp. 148–149.
10. Bricker, P. D., and Pruzansky, Sandra, "Comparison of Sorting and Pairwise Similarity Judgment Techniques for Scaling Auditory Stimuli," J. Acoust. Soc. Amer. *47* (1970), p. 96 (A).
11. Hunt, R. M., "Determination of an Effective Tone Ringer Signal," Preprint No. 722, 38th Conv. Audio Eng. Soc., Los Angeles, May 1970.
12. Cooper, P. T., unpublished work.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

On a Class of Rearrangeable Switching Networks

Part I: Control Algorithm

By D. C. OPFERMAN and N. T. TSAO-WU

(Manuscript received December 1, 1970)

An algorithm is developed to control a class of rearrangeable switching networks, particularly with the base-2 structure. Various methods of implementing this algorithm are also described. System organization and processing time for rearranging the network are studied and are shown to be practical.

I. INTRODUCTION

One type of a switching network which has drawn considerable interest lately is the class of rearrangeable switching networks (RSN). With these networks, any idle input terminal of the network can always be connected to any idle output terminal by rerouting the existing connections if necessary. These networks can be used where one-to-one full access and nonblocking features are required, and rerouting is feasible, e.g., main distribution frames[1] and facility switches[2] in telephone systems and data transfer networks in a multiprocessor computer system.[3]

Most of the earlier efforts, notably by C. Clos,[4] V. E. Beneš,[5] and A. E. Joel, Jr.,[6] have been made in the context of telephone switching networks. Their emphasis has been on the network structure, on its combinatorial properties and on bounds on the number of connections that require rerouting. Recently, this type of network has been of interest in such computer areas as data-sorting systems[7] and self-repairing multiprocessors.[3] The network structure is also applicable for cellular arrays.[8] However, very few reports[9,10] have been made on the control aspect of these networks.

This paper will begin with a brief discussion of the general structure of RSN's, followed by the development of a method for the control of these networks and its practical implementation. The relationship between the network structure and the ease (or difficulty) with which it can be controlled will also be discussed.

II. THE NETWORK STRUCTURE

Discussion will be limited to a class of rearrangeable switching networks connecting N input terminals and N output terminals [abbreviated as $(N \times N)$ networks]. Extension to the more general case for $(N \times M)$ networks, $N \neq M$, can be readily made.

Let $N = dq$, where d and q are integer factors of N. A $(N \times N)$ network can be decomposed into an input stage and an output stage having altogether $2N/d$ $(d \times d)$ networks (one of which can be eliminated) and a middle stage having d $(N/d \times N/d)$ networks, as shown in Fig. 1. This network is said to have a base-d structure, and the smaller networks are called subnetworks. This type of network structure falls into the general class considered by Clos and Beneš.

The network with base-2 structure is of great importance, for two primary reasons. First, it yields the most efficient network (or the least number of two-state switching elements) and, secondly, its control is relatively simple. It consists of (2×2) networks in the input and output stages, altogether $(N - 1)$ in number where N is even or odd. Two $(N/2 \times N/2)$ networks are in the middle stage if N is even, or one $[(N-1)/2 \times (N-1)/2]$ network and one $[(N+1)/2 \times (N+1)/2]$ network are in the middle stage if N is odd, as shown in Fig. 2a and b. Clearly, further decomposition of the networks in the middle stage is possible, and if the base-2 structure is carried throughout, one may show by an iterative process that the total number of basic switch-

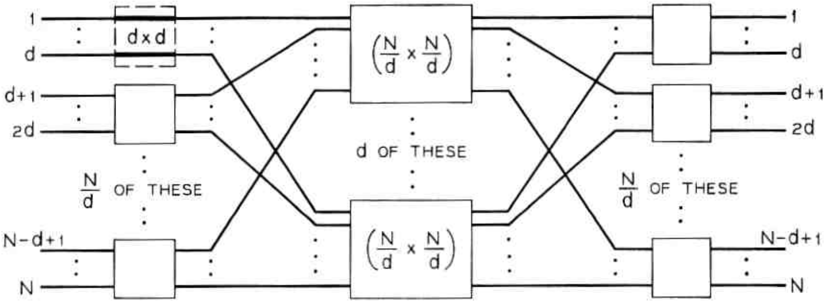

Fig. 1—Rearrangeable $(N \times N)$ network of a general base-d structure.

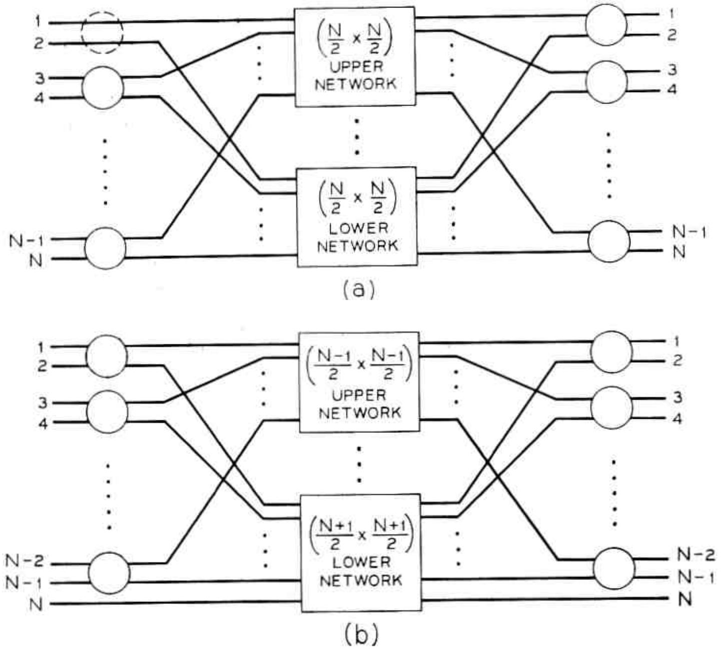
1
2
3
4
N-1
N
$\left(\frac{N}{2} \times \frac{N}{2}\right)$
UPPER NETWORK
$\left(\frac{N}{2} \times \frac{N}{2}\right)$
LOWER NETWORK
1
2
3
4
N-1
N
(a)
1
2
3
4
N-2
N-1
N
$\left(\frac{N-1}{2} \times \frac{N-1}{2}\right)$
UPPER NETWORK
$\left(\frac{N+1}{2} \times \frac{N+1}{2}\right)$
LOWER NETWORK
1
2
3
4
N-2
N-1
N
(b)

Fig. 2—An $(N \times N)$ network with base-2 structure for: (a) N even; (b) N odd.

ing elements or (2×2) networks, named β-elements by Joel,[6] is given by*

$$N\langle\log_2 N\rangle - 2^{\langle\log_2 N\rangle} + 1.$$

It can easily be seen that the number of these β-elements is bounded by $\langle\log_2 N!\rangle$. A (11×11) network consisting entirely of β-elements is shown in Fig. 3. It is a very efficient network, since there are 29 such elements and one must have 26 $(>\log_2(11!)>25)$ two-state devices to accommodate all possible permutations. Some of the enumeration studies given in Part II will account for the additional states.

III. THE NETWORK CONTROL

The control algorithm is first developed for the general $(N \times N)$ network, having a base-d structure, as it is shown in Fig. 1. The special case where $d = 2$ will then be considered, and practical implementations of the control algorithm will be given in Section IV. First, some definitions are needed.

* $\langle z\rangle$ is the smallest integer greater than z.

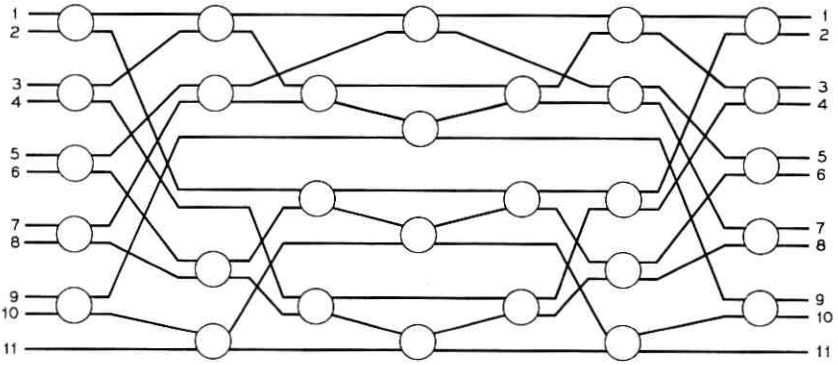

Fig. 3—An (11 × 11) network with base-2 structure.

3.1 *Definitions*

(*i*) An input-output pair $\begin{bmatrix} x \\ \pi(x) \end{bmatrix}$ defines a connection between input terminal x and output terminal $\pi(x)$ of a $(N \times N)$ network, where $1 \leqq x \leqq N, 1 \leqq \pi(x) \leqq N$

(*ii*) A connection set C is a collection of m such input-output pairs, $m \leqq N$, expressed as follows:

$$C = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ \pi(x_1) & \pi(x_2) & \cdots & \pi(x_m) \end{bmatrix}.$$

If $m = N$, C is then in the form of the familiar permutation. This may be denoted by P, with $C \subseteq P$, describing all the connections (or the traffic pattern) through the $(N \times N)$ network.

(*iii*) Let $I = \{1, 2, \cdots, N\}$ be a set of positive distinct integers; a subset $J(l, d)$ is defined by

$$J(l, d) = \left\{a \mid \left[\frac{a + d - 1}{d}\right]^* = l\right\}, \quad \text{and} \quad a \in I$$

where d divides N, and l is some constant integer, $1 \leqq l \leqq N/d$. $J(l, d)$ is called an integer set of order d and of characteristic l, and any element belonging to it is said to have characteristic l relative to the base-d. This is merely a formal way of grouping all those terminals associated with the same subnetwork in the input (or output) stage.

(*iv*) A connection set C having m input-output pairs is said to be

* $[w]$ denotes the integral value of w.

reducible if and only if on replacing every integer by its characteristic, C becomes a permutation on m distinct integers.

3.2 *The Generalized Control Algorithm*

Any given set of connections through the network can be described as:

$$P = \begin{pmatrix} x_1 & x_2 & \cdots & x_N \\ \pi(x_1) & \pi(x_2) & \cdots & \pi(x_N) \end{pmatrix}.$$

The objective of the control algorithm is to derive from P permutations to be realized by each of the subnetworks. This is accomplished by first decomposing P into reducible connection sets $C_1, C_2, \cdots, C_d$. Permutations for the middle subnetworks are defined by using the characteristics of the elements in these sets, and permutations for the input and output stages are determined directly from these elements.

3.2.1 *To Decompose a Permutation into Reducible Connection Sets*

Let the output terminals be partitioned into sets denoted by S_l where

$$S_l = \{\pi(x_i) \mid x_i \in J(l, d)\}, \qquad 1 \leqq l \leqq \frac{N}{d}.$$

The reducible connection sets C_i, $1 \leqq i \leqq d$, are constructed by grouping N/d input-output pairs of which the output integers are selected, one from each S_l, such that no two output integers have the same characteristic.

Let C_i be expressed as

$$C_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,N/d} \\ \pi(x_{i,1}) & \pi(x_{i,2}) & \cdots & \pi(x_{i,N/d}) \end{bmatrix}, \qquad 1 \leqq i \leqq d;$$

when it is reduced, one has the permutation

$$P_i = \begin{pmatrix} p_{i,1} & p_{i,2} & \cdots & p_{i,N/d} \\ \pi(p_{i,1}) & \pi(p_{i,2}) & & \pi(p_{i,N/d}) \end{pmatrix}$$

where $p_{i,j}$ and $\pi(p_{i,j})$ are the characteristics of the integers $x_{i,j}$ and $\pi(x_{i,j})$, respectively, $1 \leqq i \leqq d$, $1 \leqq j \leqq N/d$. Each of these d permutations P_i can be realized by any of the d $(N/d \times N/d)$ subnetworks. However, if each C_i is assigned to a particular $(N/d \times N/d)$ subnetwork, one of the $(d \times d)$ subnetworks, say the one at the upper-left corner of the input stage (Fig. 1), can be eliminated. This elimination implies that the connection set C_i, and hence P_i, is assigned to the first $(N/d \times N/d)$ subnetwork in the middle stage if and only if $x_{i,j} = 1$ for some j, $1 \leqq j \leqq N/d$. In general, C_i, and hence P_i, is assigned to the k^{th} $(N/d \times N/d)$ subnetwork in the middle stage, $1 \leqq k \leqq d$, (counting from the top) if and only if $x_{i,j} = k$ for some j, $1 \leqq j \leqq N/d$. One may then reorder the indices i such that C_i contains the input-output pair $\begin{bmatrix} x_{i,j=i} \\ \pi(x_{i,j}) \end{bmatrix}$, and it follows that the permutations $P_1, P_2, \cdots, P_d$ are similarly ordered for the d $(N/d \times N/d)$ subnetworks in the middle stage.

3.2.2 *To Obtain Permutations for Subnetworks in the Input and Output Stages*

Let $P_{I,1}, P_{I,2}, \cdots, P_{I,N/d}$ and $P_{0,1}, P_{0,2}, \cdots, P_{0,N/d}$ be the sets of permutations to be realized by the 1st, 2nd, $\cdots$, $(N/d)^{\text{th}}$ $(d \times d)$ subnetworks in the input and output stage, respectively. Moreover, let the reducible connection set C_i be written as

$$C_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,N/d} \\ \pi(x_{i,1}) & \pi(x_{i,2}) & \cdots & \pi(x_{i,N/d}) \end{bmatrix}, \qquad 1 \leqq i \leqq d$$

such that $x_{i,1} < x_{i,2} \cdots < \cdots < x_{i,N/d}$.

Then, from the network structure, one can see simply that the permutations

$$P_{I,j} = \begin{pmatrix} \pi_j^{-1}(1) & \pi_j^{-1}(2) & \cdots & \pi_j^{-1}(d) \\ 1 & 2 & \cdots & d \end{pmatrix}$$

and

$$P_{0,k} = \begin{pmatrix} 1 & 2 & \cdots & d \\ \pi_k(1) & \pi_k(2) & \cdots & \pi_k(d) \end{pmatrix}$$

can be obtained from C_i with

$$\pi_j^{-1}(a) = x_{a,j} - (j-1)d, \quad 1 \leqq j \leqq N/d, \quad 1 \leqq a \leqq d$$

where $\pi^{-1}(a)$ denotes the input terminal to be connected to the output terminal a, and $\pi_k(a) = \pi(x_{a,t}) - (k-1)\,d$, $1 \leqq k \leqq N/d$, for some t such that $\pi(x_{a,t}) \in J(k, d)$.

3.2.3 *An Example on Decomposing a Permutation*

Consider a (15 × 15) network, having a base $d = 5$ structure. Such a network is shown in Fig. 4, with two (5 × 5) subnetworks in the input stage, three (5 × 5) subnetworks in the output stage, and five

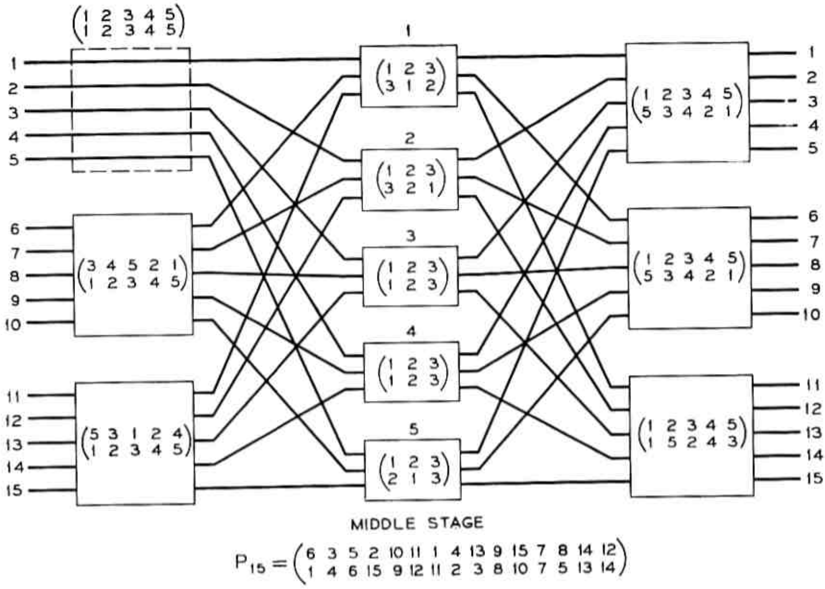

Fig. 4—A (15 × 15) network with permutations assignment.

(3 × 3) subnetworks in the middle stage. Let the connections through the network be described by the following permutation:

$$P = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 11 & 15 & 4 & 2 & 6 & 1 & 7 & 5 & 8 & 9 & 12 & 14 & 3 & 13 & 10 \end{pmatrix}.$$

The output integers are partitioned as follows:

$$S_1 = \{11, 15, 4, 2, 6\},$$

$$S_2 = \{1, 7, 5, 8, 9\},$$

$$S_3 = \{12, 14, 3, 13, 10\}.$$

From these, the reducible connection sets C_1, C_2, C_3, C_4, and C_5 are constructed as shown in Table I and, in general, they are not unique. The corresponding permutations (P_i), which are also shown in Table I, are obtained from C_i by replacing each element with its characteristic. Note that connection sets are ordered according to the input integers (1, 2, 3, 4, 5) and that $x_{i,1} < x_{i,2} < x_{i,3}$ for each i. Table II shows the permutations derived from C_i for subnetworks in the input and output stages using the relations for $\pi_i^{-1}(a)$ and $\pi_k(a)$ as given in Section 3.2.2.

TABLE I—REDUCED CONNECTION SETS AND THEIR CORRESPONDING PERMUTATIONS

i	C_i	P_i
1	$\begin{bmatrix} 1 & 8 & 15 \\ 11 & 5 & 10 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$
2	$\begin{bmatrix} 2 & 9 & 13 \\ 15 & 8 & 3 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$
3	$\begin{bmatrix} 3 & 10 & 11 \\ 4 & 9 & 12 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$
4	$\begin{bmatrix} 4 & 7 & 12 \\ 2 & 7 & 14 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$
5	$\begin{bmatrix} 5 & 6 & 14 \\ 6 & 1 & 13 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$

TABLE II—PERMUTATIONS FOR SUBNETWORKS IN THE INPUT AND OUTPUT STAGES

j	$P_{I,j}$	$P_{O,j}$
1	$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 4 & 2 & 1 \end{pmatrix}$
2	$\begin{bmatrix} 3 & 4 & 5 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 4 & 2 & 1 \end{pmatrix}$
3	$\begin{bmatrix} 5 & 3 & 1 & 2 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 5 & 2 & 4 & 3 \end{pmatrix}$

IV. THE CONTROL ALGORITHM FOR THE BASE-2 NETWORK

The selection of output integers from the set S_l, $1 \leqq l \leqq N/d$, to form the connection sets C_i, $1 \leqq i \leqq d$, is by no means simple. One procedure has been reported by V. I. Neiman.[10] By modifying the previous example, it can be shown that the selection can *not* be made on a strictly sequential basis. Let the given permutation be modified such that the sets S_1, S_2, and S_3 are as follows:

$$S_1 = \{11, 10, 4, 2, 6\},$$
$$S_2 = \{1, 7, 5, 8, 9\},$$
$$S_3 = \{12, 14, 3, 13, 15\}.$$

If one had chosen $\binom{1}{11}$ and $\binom{6}{1}$ input-output pairs to form C_1, where the output integers 11 and 1 are from S_1 and S_2 respectively, there is no output integer in S_3 which would have a characteristic different from that of the integer 11 or 1. Thus, in general, simultaneous selections must be made in the construction of the connection sets. For networks with base-2 structure, however, the selection is reduced to a mere binary choice, resulting in a much simpler algorithm. For the other extreme case,[11] where $d = N/2$, the difficulty described above will not arise because there are only two sets S_1 and S_2, each containing exactly d integers.

For a $(N \times N)$ network with a base-2 structure, the control algorithm for setting the β-elements to realize a given permutation P consists of three parts: (*i*) decomposing of P into reducible connection sets C_1 and C_2; (*ii*) reducing C_1 and C_2 to P_1 and P_2 respectively; and (*iii*) setting the β-elements in the input and output stages. Since the network has an iterative structure, the same procedure is applied to each of the $(N/2 \times N/2)$, $(N/4 \times N/4)$, $\cdots$, (2×2) subnetworks. There are $\log_2 N$ levels of an $(N \times N)$ network with the last level of (2×2) subnetworks being a trivial case, assuming N is a power of 2.

A coding scheme for the input and output integers which facilitates the required operations and two methods for decomposing P are described in the following sections.

4.1 *Coding Scheme*

It is clear from what has been discussed so far that the control algorithm essentially accepts the connection requirements as input data and, after processing, generates a set of output data which are used to rearrange the network. It is necessary, therefore, to have an input/output (I/O) memory, which stores the output terminal $\pi(x_i)$ at the address determined by the numerical value of the input terminal x_i. A simple coding scheme for these integers proves to simplify the implementation of the algorithm.

Referring again to the network, one can, of course, use the set of integers $(0, 1, 2, \cdots, N - 1)$ to number the input and output terminals, without loss of generality in all the previous discussions.* Then, the familiar binary code can be used directly, both for the input x_i as address and for the output $\pi(x_i)$ as the contents at x_i. We shall now show how this code can be used at all $\log_2 N$ levels. Let the binary representation be

$$b_{n-1}b_{n-2} \cdots b_1 b_0 ,$$

where $n = \log_2 N$, and assuming N a power of 2. Beginning at the first level of the network, the least significant bit of the address is used to set the input β-element defined by the remaining $n - 1$ bits, and that of the contents to set the output β-element defined by the remaining $n - 1$ bits of the contents. Moreover, the conversion from C_1 (or C_2) to P_1 (or P_2) is accomplished by merely eliminating b_0 for each coded output integer. Finally, for an output integer $\pi(x_i)$, the bit b_0 is set to identify whether the particular $\pi(x_i)$ belongs to P_1 or P_2. However, for an input integer, the most significant bit of the address, b_{n-1}, indicates whether x_i belongs to P_1 or P_2.

The same coding procedure is applied at each subnetwork, and the I/O memory is partitioned (part of algorithm) in the appropriate manner. In general, at the i^{th} level of the network, $i < \log_2 N$, the $(i - 1)$ least significant bits of a word in memory define the particular subnetwork of size $N(2^{1-i})$, and the $\log_2 N - (i - 1)$ most significant bits define the output integer. The $(i-1)$ most significant bits, however, of the address designate the subnetwork, and the remaining bits define the input integer. An example will be given in detail to illustrate this in Section 4.2.1.

4.2 *Decomposition by Looping*

With $d = 2$, an integer set is reduced to an integer pair, consisting of only two elements, and one is said to be the dual of the other. If one continues to use the integers $(0, 1, \cdots, N - 1)$ to number the input and output terminals of an $(N \times N)$ network, then the integers a and b constitute an integer pair if

$$\left[\frac{a}{2}\right] = \left[\frac{b}{2}\right] = l$$

for some integer l. The dual of a (or b) is denoted by $\hat{a}$ (or $\hat{b}$), and,

* Except that the definition of the integer set $J(l, d)$ needs to be slightly modified, as follows:

$$J(l, d) = \{a | [a/d] = l\}, \qquad 0 \leqq l \leqq (N/d) - 1$$

therefore,

$$\hat{a} = b \quad \text{and} \quad \hat{b} = a.$$

Moreover, any permutation is decomposed into only two reducible connection sets C_1 and C_2, and if $\left[{x \atop \pi(x)}\right] \in C_1$, then $\left[{\hat{x} \atop \pi(\hat{x})}\right]$ and $\left[{x_i \atop \pi(x_i)}\right] \in C_2$ for some x_i, where $\pi(x_i)$ is the dual of $\pi(x)$. In coded form, the dual is obtained merely by complementing the least significant bit.

One method, called the looping procedure, of constructing C_1 and C_2 from P is to search the I/O memory for the required outputs. The sequence starts by selecting the output $\pi(0)$ in the first location of the memory (or $\left[{0 \atop \pi(0)}\right]$). The first $(n - 1)$ bits $(a_{n-1}a_{n-2} \cdots a_1)$ of the address which are all zeros define the first β-element (β_{i1}) in the input stage, and the last bit (a_0) which is also zero defines β_{i1} to be set to the "straight through" state (see Fig. 5). The bits $m_{n-1}m_{n-2} \cdots m_1$ and m_0 at this address define one of the β-elements (β_{oi}) in the output stage and its setting respectively. Bit m_0 is now reset to zero to designate that this particular output $(m_{n-1}m_{n-2} \cdots m_1)$ is for P_1. There is also another bit m_n which is set to one when that particular output has been placed in P_1 or P_2 so that an unused output can be selected when it is necessary. The example in Section 4.2.1 will clarify this.

The memory is next scanned for the dual of $\pi(0)$; this output and its address x_k define $\left[{x_k \atop \pi(x_k)}\right]$ for C_2. For the input-output pairs in C_2

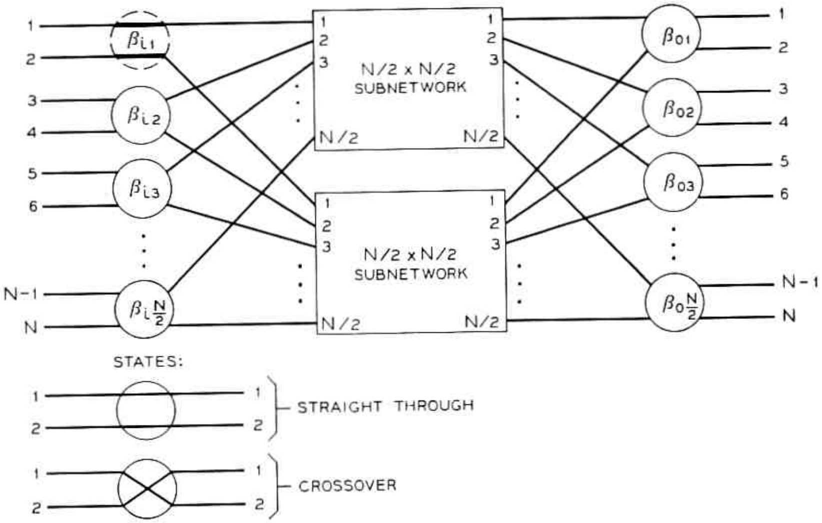

Fig. 5—Base-2 $(N \times N)$ rearrangeable switching network.

it is only necessary to set m_n and m_0 of the memory word to '1' to signify that this word (output) has been used and that it is for P_2. The output $\pi(\hat{x}_k)$ at the address $\hat{x}_k$ defines $\left[\begin{smallmatrix}\hat{x}_k\\ \pi(\hat{x}_k)\end{smallmatrix}\right]$ and is designated for C_1. The same looping procedure is continued until all $\pi(x_i)$, $1 \leqq i \leqq N$ are assigned to either C_1 or C_2. In most cases, however, the looping will end before all output integers are used. The procedure is started again by arbitrarily selecting an unassigned output for P_1, by examining m_n. Because of this arbitrariness, C_1 is, in general, not unique. This fact is used to advantage in the other procedure to be described in Section 4.3.

After the looping procedure is completed, the memory is reorganized to have P_1 and P_2 in locations 0 to $N/2 - 1$ and $N/2$ to $N - 1$, respectively. (A small scratch pad memory may be necessary.) Then the same procedure is applied to each of these $(N/2 \times N/2)$ subnetworks, and it is continued until all of the β-elements in the $(N \times N)$ network are set.

The searching can be eliminated by employing two memories, one of which is the I/O memory. The additional one is an output/input (O/I) memory that stores the input x_i at the address corresponding to the numerical value of the output $\pi(x_i)$. The decomposition of P into P_1 and P_2 is achieved by crisscrossing between the two memories. For example, output $\pi(x_i)$ and its corresponding address x_i in the I/O memory define $\left[\begin{smallmatrix}x_i\\ \pi(x_i)\end{smallmatrix}\right]$ for C_1. Now the dual of $\pi(x_i)$, say $\pi(x_j)$, is the address for the O/I memory and the corresponding word x_j defines $\left[\begin{smallmatrix}x_j\\ \pi(x_j)\end{smallmatrix}\right]$ for C_2. Then $\hat{x}_j$ is used for the address in the I/O memory. If the I/O memory is a content addressable memory (CAM),[12] the required $\pi(x_i)$'s for C_1 and C_2 are determined directly in the content addressable mode, without the use of a second memory.

4.2.1 *An Example of the Looping Procedure*

In order to illustrate in a meaningful way the looping procedure using the coding scheme at various levels, a permutation for a (32 × 32) network, as given in Table III, will be utilized.

The looping procedure begins with $\pi(0) = 14$ and continues to select input-output pairs for the connection set C_1. (The sequence of this selection is indicated by the number in the "looping sequence" column.) As each output integer is selected for C_1, the last bit m_0 is used to set the output β-element designated by $(m_4m_3m_2m_1)$ (see Fig. 6). Then the bit m_0 is set to '0' to indicate that it is for C_1, and the bit m_5 is set to '1' to indicate that it has been used. The bits m_0 and m_5 of the output integers for C_2 are merely set to '1'. In this particular

TABLE III—MEMORY CONTENTS ON THE INTERCONNECTIONS

Input Terminals	Outputs Terminals	Coded Output Integers m_5	m_4	m_3	m_2	m_1	m_0	Looping Sequence
0	14	0	0	1	1	1	0	1
1	23	0	1	0	1	1	1	
2	3	0	0	0	0	1	1	7
3	20	0	1	0	1	0	0	
4	28	0	1	1	1	0	0	8
5	2	0	0	0	0	1	0	
6	16	0	1	0	0	0	0	
7	11	0	0	1	0	1	1	
8	31	0	1	1	1	1	1	9
9	29	0	1	1	1	0	1	
10	1	0	0	0	0	0	1	
11	27	0	1	1	0	1	1	5
12	12	0	0	1	1	0	0	3
13	18	0	1	0	0	1	0	
14	5	0	0	0	1	0	1	
15	24	0	1	1	0	0	0	11
16	26	0	1	1	0	1	0	
17	21	0	1	0	1	0	1	6
18	0	0	0	0	0	0	0	4
19	13	0	0	1	1	0	1	
20	8	0	0	1	0	0	0	
21	6	0	0	0	1	1	0	13
22	19	0	1	0	0	1	1	2
23	15	0	0	1	1	1	1	
24	4	0	0	0	1	0	0	10
25	30	0	1	1	1	1	0	
26	7	0	0	0	1	1	1	
27	22	0	1	0	1	1	0	14
28	17	0	1	0	0	0	1	
29	10	0	0	1	0	1	0	
30	25	0	1	1	0	0	1	
31	9	0	0	1	0	0	1	12

example, the looping procedure ends prematurely and leaves the connection set

$$\begin{Bmatrix} 6 & 7 & 28 & 29 \\ 16 & 11 & 17 & 10 \end{Bmatrix}$$

as indicated on Table IV.

One then repeats the looping procedure by starting arbitrarily at some remaining output integers (as indicated by $m_5 = 0$). After every output integer has been assigned to C_1 (or C_2), one rearranges the memory such that all the integers with $m_0 = 0$ occupy the upper half of the memory, corresponding to the connections through the upper network, with the remaining output integers for the lower network. This is shown in Table V.

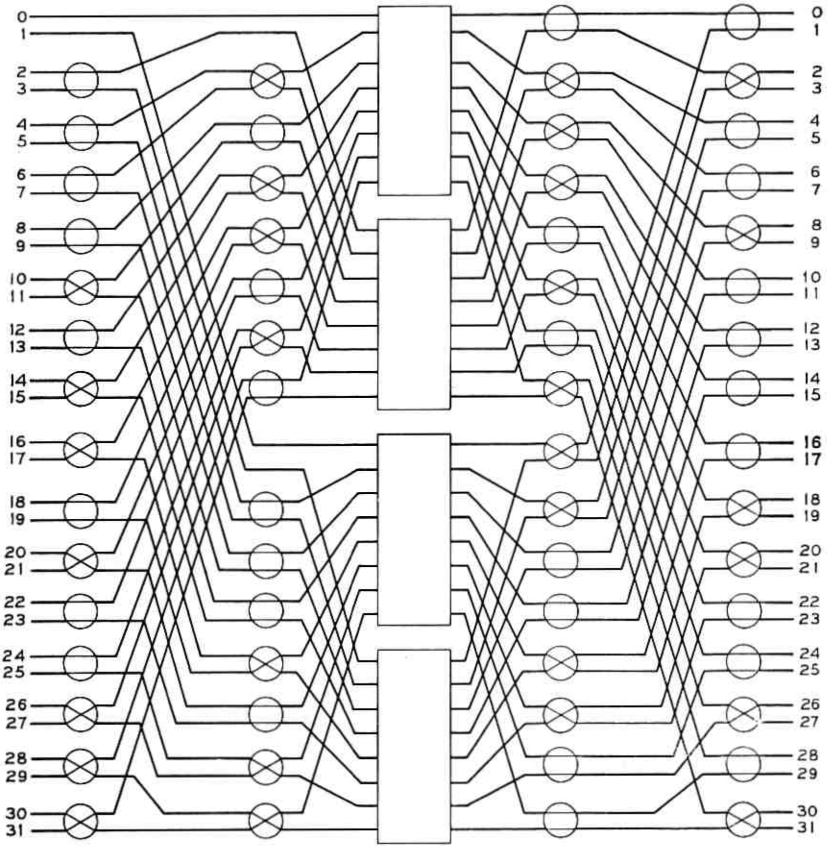

Fig. 6—A (32 × 32) rearrangeable network with partial setting of β-elements.

Table VI shows the memory contents after another looping procedure is applied to the output integers (for both the upper and lower networks) and subsequent rearranging. At this level, m_1m_0 denotes the (8 × 8) subnetwork, and if one reverses the order to m_0m_1, it is just the binary representation of natural ascending numbers, 0 being the upmost (8 × 8) subnetwork and 3 being the lowest (8 × 8) subnetwork. One could also obtain the same information from the addresses, since the contents are rearranged into this order.

4.3 *Decomposition with a Trial-Partition Procedure*

A second method for decomposing P which incorporates a trial-and-error procedure is presented. Although this method is practical only for small networks, it is very important because of the simple implementation, and intelligence can be included to reduce the average processing time by taking advantage of the fact that the decomposition may not be unique.

From the derivation of the reducible connection sets C_1 and C_2, it is seen that they must contain one and only one $\pi(x_i)$ from each S_l, $0 \leqq l \leqq (N/2) - 1$, and the corresponding input x_i. Since $\binom{0}{\pi(0)}$ and $\binom{1}{\pi(1)}$ are defined to be in C_1 and C_2, respectively, only $(N/2 - 1)$ additional input-output pairs must be selected for C_1; the remaining pairs are for C_2. After C_1 and C_2 are determined, P_1 and P_2 and the rearranging of the I/O memory are derived in the same manner as in the looping procedure. Let Ψ be the set of connection sets; each includes $(N/2 - 1)$ input-output pairs formed by having one $\pi(x_i)$ from

TABLE IV—MEMORY CONTENTS AFTER SEQUENCING THROUGH ONE LOOP

m_5	m_4	m_3	m_2	m_1	m_0	
1	0	1	1	1	0	
1	1	0	1	1	1	
1	0	0	0	1	0	
1	1	0	1	0	1	these belong
1	1	1	1	0	0	to another loop
1	0	0	0	1	1	
0	1	0	0	0	0	→ 1 1 0 0 0 0
0	0	1	0	1	1	→ 1 0 1 0 1 1
1	1	1	1	1	0	
1	1	1	1	0	1	
1	0	0	0	0	1	
1	1	1	0	1	0	
1	0	1	1	0	0	
1	1	0	0	1	1	
1	0	0	1	0	1	
1	1	1	0	0	0	
1	1	1	0	1	1	
1	1	0	1	0	0	
1	0	0	0	0	0	
1	0	1	1	0	1	
1	0	1	0	0	1	
1	0	0	1	1	0	
1	1	0	0	1	0	
1	0	1	1	1	1	
1	0	0	1	0	0	
1	1	1	1	1	1	
1	0	0	1	1	1	
1	1	0	1	1	0	
0	1	0	0	0	1	→ 1 1 0 0 0 1
0	0	1	0	1	0	→ 1 0 1 0 1 0
1	1	1	0	0	1	
1	0	1	0	0	0	

TABLE V—MEMORY CONTENTS AFTER REARRANGING

m_5	m_4	m_3	m_2	m_1	m_0	
0	0	1	1	1	0	
0	0	0	0	1	0	
0	1	1	1	0	0	
0	1	0	0	0	0	upper subnetwork
0	1	1	1	1	0	
0	1	1	0	1	0	
0	0	1	1	0	0	
0	1	1	0	0	0	
0	1	0	1	0	0	
0	0	0	0	0	0	
0	0	0	1	1	0	
0	1	0	0	1	0	
0	0	0	1	0	0	
0	1	0	1	1	0	
0	0	1	0	1	0	
0	0	1	0	0	0	
0	1	0	1	1	1	
0	1	0	1	0	1	
0	0	0	0	1	1	lower subnetwork
0	0	1	0	1	1	
0	1	1	1	0	1	
0	0	0	0	0	1	
0	1	0	0	1	1	
0	0	0	1	0	1	
0	1	1	0	1	1	
0	0	1	1	0	1	
0	0	1	0	0	1	
0	0	1	1	1	1	
0	1	1	1	1	1	
0	0	0	1	1	1	
0	1	0	0	0	1	
0	1	1	0	0	1	

each S_l , $l > 0$. Since there are two elements in each S_l , Ψ consists of $2^{(N/2)-1}$ connection sets. For any arbitrarily selected $C \in \Psi$, the test will be only on the output integers. Therefore, C defines C_1 if and only if for any $\pi(x_i) \in C$, its dual is not in C.

These ideas lead to the application of a finite state machine with $2^{(N/2)-1}$ states which are used to generate Ψ. This machine, called the Trial-Partition Machine (TPM), is composed of $(N/2 - 1)$ two-state storage devices (flip-flops), each of which represent an input integer pair. The "0" or "1" state of the flip-flop designates that the odd or even input, respectively, of the input integer pair and the corresponding output is in C. If σ_i is a state of the TPM, then it defines C, and the $\pi(x_i) \in C$ are tested for an integer pair (two outputs with the same characteristic). If C contains no pairs, then it is reducible and can be used as C_1. However, if there is at least one output integer pair in C, the TPM is advanced to σ_{j+1}. Since the outputs are serially stored in the I/O memory, the memory must be sequenced in order to perform the test for each σ_j.

This type of a TPM can be easily implemented with a $(N/2 - 1)$ bit binary counter. However, it may be desired to have a more intelligent machine so that C_1 can be determined in less time (with fewer trials). Of course, the complexity and, consequently, the cost of the TPM will increase as the intelligence of the machine grows. One way to

Table VI—Memory Contents at the Third Level (after two looping procedures)

m_5	m_4	m_3	m_2	m_1	m_0
0	0	1	1	0	0
0	1	0	0	0	0
0	1	1	1	0	0
0	1	1	0	0	0
0	0	0	0	0	0
0	0	0	1	0	0
0	1	0	1	0	0
0	0	1	0	0	0
0	0	0	0	1	0
0	1	1	1	1	0
0	1	1	0	1	0
0	0	1	1	1	0
0	1	0	1	1	0
0	1	0	0	1	0
0	0	0	1	1	0
0	0	1	0	1	0
0	1	0	1	0	1
0	0	0	0	0	1
0	1	1	1	0	1
0	1	0	0	0	1
0	0	1	1	0	1
0	0	1	0	0	1
0	0	0	1	0	1
0	1	1	0	0	1
0	1	0	1	1	1
0	0	1	0	1	1
0	0	0	0	1	1
0	0	0	1	1	1
0	1	1	0	1	1
0	0	1	1	1	1
0	1	1	1	1	1
0	1	0	0	1	1

enhance the intelligence is to count the number of output integer pairs in C. This number (Z) specifies how many $\begin{bmatrix} x_i \\ \pi(x_i) \end{bmatrix}$ must be changed in order to have all $\pi(x_i)$ from different output integer pairs. Then the next set (C') is selected from those sets which contain at least Z different pairs and have not been previously tested. Another method is to observe the input-output pairs $\begin{bmatrix} x_i \\ \pi(x_i) \end{bmatrix}$ and $\begin{bmatrix} x_j \\ \pi(x_j) \end{bmatrix}$ that form an output integer pair, and select a C' that contains either $\hat{x}_i$ or $\hat{x}_j$. For an example of a TPM, see Ref. 11.

The TPM is applicable only for small networks because the number of tests becomes prohibitive as N increases. There are $2^{(N/2)-1}$ connection sets in Ψ and, on the average, half of these must be tried. A TPM with intelligence, however, will reduce the number of sets to be tested, but for networks larger than (64 × 64) this will still be too time consuming.

4.4 *Using Combinational Logic*

The control algorithm for the (4 × 4) network can be implemented with combinational logic. This is achieved by assigning the states of the five β-elements for each of the 24 possible input-output permutations. The combinational logic, which consists of 13 NAND gates, determines the correct setting from the outputs which are coded in the binary code. This method is practical only for very small networks because the number of permutations grows very rapidly. The (4 × 4) network is important, however, because it could be used as a building block to construct larger networks, and the combinational logic and the β-elements could be on the same semiconductor chip. This same idea also applies to an (8 × 8) or (16 × 16) network using the TPM.

4.5 *System Description*

The block diagram of a rearrangeable switching system with only one memory (I/O Memory) is shown in Fig. 7. The Network Control realizes the control algorithm by employing one or more of the above methods. For example, it may be advantageous to use the combinational logic and TPM for the small networks (or subnetworks) because they are relatively inexpensive to implement. However, as the size of the network grows, the processing time becomes critical. Then the looping procedure with one and two memories (or a content addressable memory) may be used for less than 1000 terminals and more than 1000 terminals, respectively. The timing considerations for these methods are discussed in the next section.

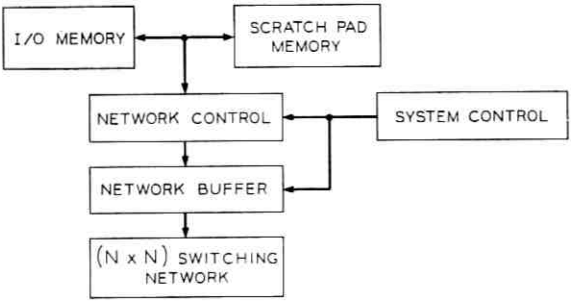

Fig. 7—System block diagram.

The Scratch Pad Memory temporarily stores the outputs during decomposition of P and also while the I/O Memory is being partitioned. The System Control generates the timing and control sequences for all of the operations. This unit, as well as the Network Control, could be implemented with stored program techniques if it is economical and if there is sufficient real time.

The other important unit, called the Network Buffer, isolates the network from the Network Control so that the existing traffic will not be affected during the processing time of the control algorithm. As the settings of the β-elements are determined, they are stored in the Network Buffer. After the control algorithm is completed, the states of the β-elements are set within the time required to insure the quality of the transmission. This time is dictated by the switching transients of the β-elements, network terminations, and the application. Serial shift-registers can provide economical buffers, if the β-elements can be set in a stage-by-stage sequence.

V. TIMING CONSIDERATIONS

In this section, the processing time and the necessary equipment for the various methods of implementing the control algorithm are discussed. The processing time for the algorithm must be sufficiently short to accommodate the traffic changes. The combinational logic method is the fastest; however, it is applicable only for very small networks. The TPM has the next order of complexity, and it is applicable for (64 × 64) networks or smaller. The processing time, which increases exponentially, is derived as follows: For an $(N \times N)$ network, there are $2^{(N/2)-1}$ states of the TPM. If the TPM is a binary counter (no intelligence), on the average about half of the states or $2^{(N/2)-2}$ must be tested before the given P is decomposed to C_1 and C_2. Also, on the average, one-half of the I/O memory is scanned before an output integer pair is detected. Therefore, the number (A_T) of times that the I/O memory is accessed for $\log_2 N$ levels [the $\log_2 N^{\text{th}}$ level is the trivial (2×2) network, and the number of times the memory is accessed is simply $N/2$] is:

$$A_T = \frac{N}{8} \sum_{i=1}^{\log_2 N-1} 2^{N/2^i} + \frac{N}{2}. \tag{1}$$

The processing time can be greatly reduced by using the looping procedure. If a random-access I/O memory is employed, it is necessary to search for one-half of the outputs (the remaining outputs are obtained directly from memory) on each level of the network. For an $(N \times N)$ network, $N/2$ outputs are determined by accessing the I/O memory, on the average, $N/2$ times. Then

$$A_T = N^2 \sum_{i=1}^{\log_2 N} \left(\frac{1}{2}\right)^{i+1} = \tfrac{1}{2}N(N - 1). \tag{2}$$

In addition, the access time required to partition the memory is $2N \log_2 N$ and should be included in equations (1) and (2) to give the total processing time.

If two memories (in the crisscross manner) or a CAM is utilized, no searching is necessary, and access times for the decomposition of P and partitioning of memories are:

$$A_T = 4N \log_2 N$$

and (3)

$$A_T = 3N \log_2 N \quad \text{respectively.}$$

For $N = 16{,}384$ and a CAM with 1-μsec access time, the control algorithm, implemented with wired logic and a 10-MHz basic clock, can be accomplished in approximately *750 msec.* If two memories (random access) with 1-μsec cycle time are used, a processor with an instruction execution time of 3 μsec can implement the control algorithm in approximately 50 seconds.

VI. CONCLUSION

In this paper, the network structure and control algorithm for certain $(N \times N)$ rearrangeable switching networks are described. The algorithm consists of decomposing a given permutation into d (where

d is the base of the network) permutations for the d $(N/d \times N/d)$ subnetworks, and determining the connections for the N/d $(d \times d)$ networks in the input and output stages. The same procedure is applied to the subnetworks in an iterative manner until all of the connections in the $(N \times N)$ network are defined.

Although the network can be constructed with various building blocks (bases), the base-2 structure is the most important because it requires the least number of two-state devices (β-elements) and the control algorithm is relatively simple. The algorithm is implemented by performing the decomposition either by the looping procedure or by a Trial-Partition Machine. Also, an efficient coding scheme is defined to facilitate the decomposition of the permutation and the partitioning of the memory. With the base-2 structure, the control algorithm and the coding scheme can be used in a consistent manner at each level of the network. There are other classes of network structures with the same number of β-elements, such as the nested-tree networks,[6] that do not have this property.

The processing time and equipment complexity vary with the methods of implementation. The combinational logic is the fastest and least expensive; however, it is only applicable for very small networks. The Trial-Partition Machine is economical, but it is too slow for large networks; however, intelligence can be designed into the machine taking advantage of the fact that for a large number of permutations there are more ways than one of setting the network. Consequently, the processing time is dependent on the permutations given, as well as the amount of intelligence built in. The most suitable method for networks larger than (64 × 64) is the looping procedure with a content addressable memory to store the outputs. The processing time is independent of the permutations given. With this method, it is possible to determine the setting of all the β-elements for a (16,384 × 16,384) network in less than one second for *any* number of new connections or terminations. During the processing time, the new settings for β-elements are stored in a buffer; then their states are changed. The memory required for this system is about 300k bits or approximately 20 bits per input terminal.

This paper has described a control algorithm for a rearrangeable switching network that is practical from both the system and processing time viewpoints. The application of this network should be considered where full access and nonblocking is required, and rerouting is possible.

REFERENCES

1. Joel, A. E., Jr., unpublished work.
2. Case, C. C., unpublished work.
3. Levitt, K. N., et al., "A Study of the Data Communication Problems in a Self-Repairable Multiprocessor," Spring 1968 Joint Computer Conf., AFIPS Proc., *32,* pp. 515–527.
4. Clos, C., "A Study of Non-Blocking Switching Networks," B.S.T.J., *32,* No. 2 (March 1953), pp. 406–424.
5. Beneš, V. E., "Optimal Rearrangeable Multistage Connecting Networks," B.S.T.J., *43,* No. 4, Part 2 (July 1964), pp. 1641–1656.
6. Joel, A. E., Jr., "On Permutation Switching Networks," B.S.T.J., *47,* No. 5 (June 1968), pp. 813–822.
7. Batcher, K. E., "Sorting Networks and Their Applications," Spring 1968 Joint Computer Conf., AFIPS Proc., *32,* pp. 307–314.
8. Kautz, W. H., et al., "Cellular Interconnection Arrays," IEEE Trans. Computers, *EC-17* (May 1968), pp. 443–451.
9. Waksman, A., "A Permutation Network," JACM, *15,* No. 1 (January 1968), pp. 159–163.
10. Neiman, V. I., "Structure et Commande Optimales de Réseaux de Connexion Sans Blocage," Annales des Tèlècommunications, July/August 1969, pp. 232–238.
11. Tsao-Wu, N. T., and Opferman, D. C., "On Permutation Algorithms for Rearrangeable Switching Networks," IEEE Int. Conf. Commun., 1969, Conf. Record, pp. 10–29—10–34.
12. Koo, J., "Integrated Circuit Content Addressable Memories," IEEE Int. Solid-State Circuit Conf., 1970, Digest of Technical Papers, pp. 72–73.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

On a Class of Rearrangeable Switching Networks

Part II: Enumeration Studies and Fault Diagnosis

By D. C. OPFERMAN and N. T. TSAO–WU

(Manuscript received December 1, 1970)

The decomposition of permutations as used in the control algorithm for a class of rearrangeable switching networks is proved. Enumeration studies on permutations related to the network are presented. Theorems for constructing a set of traffic patterns for diagnostic purposes are also given. Finally, a procedure for detecting and locating faulty switching elements in the network is described.

I. INTRODUCTION

This part of the paper will cover some of the theoretical considerations related to the rearrangeable switching networks discussed in Part I. For the general $(N \times N)$ network with base-d structure, it is shown that it can indeed accommodate any of the $N!$ connection patterns. A thorough study is then made of the $(N \times N)$ network having a base-2 structure. It was pointed out in Part I that the setting of the β-element is, in general, not unique for an arbitrary input-output permutation. Furthermore, the number of β-elements for an $(N \times N)$ network exceeds $\langle \log_2 (N!) \rangle$, for $N > 4$. Some enumeration studies are given to account for this. Finally, fault diagnostic studies are given in relation to the base-2 network. A method to construct a set of permutations useful for testing the network is developed. This is then followed by discussing a procedure to detect and/or locate faulty β-elements in the network.

II. PERMUTATION PROPERTY OF THE NETWORK

In this section it will be shown that the decomposition of the given permutation into reducible connection sets (as used in the control algorithm) is always possible. From Section 3.2.1 of Part I, it is evident that the decomposition is equivalent to the selection of d sets of output integers, $\pi(x_i)$, one from each S_l, such that all the output integers in any of the d sets have distinct characteristics, where S_l, as previously defined, is

$$S_l = \{\pi(x_i) \mid x_i \in J(l, d)\} \qquad 1 \leqq l \leqq N/d;$$

and there are d elements in each S_l.

To show that this selection and, therefore, the decomposition can always be done, P. Hall's Theorem[1] on Distinct Representatives is used and is stated as follows:

P. Hall's Theorem: Let L be a finite set of indices $L = \{1, 2, \cdots n\}$. *For each* $l \in L$, *let* T_l *be a subset of a set T. A necessary and sufficient condition for the existence of distinct representatives* t_l, $l = 1, 2, \cdots, n$, $t_l \in T_l$, $t_i \neq t_j$ *when* $i \neq j$, *is that for every* $k = 1, 2, \cdots, n$ *and every choice of k distinct indices* $l_1, l_2, \cdots, l_k$, *the subsets* $T_{l_1}, T_{l_2}, \cdots, T_{l_k}$ *contain between them at least k distinct elements.*

This theorem can be used directly if a mapping ϕ is defined on the sets S_l as follows:

$$S_l = \{\pi(x_i)\} \xrightarrow{\phi} T_l = \{t\} \qquad 1 \leqq l \leqq N/d,$$

where

$$t = \left[\frac{\pi(x_i) + d - 1}{d}\right]^*.$$

This simply means that each integer in S_l is replaced by its characteristic t, with T_l having exactly the same number of elements as S_l. Thus, the selection of d sets of $\pi(x_i)$, one from each S_l, such that the integers in each of the d sets have distinct characteristics, is equivalent to the selection of d sets of N/d *distinct* representatives, one from each T_l, such that in each of the d sets $t_i \neq t_j$ for $i \neq j$.

By Hall's theorem, it is sufficient to show that for every $k = 1, 2, \cdots, N/d$ and choice of k distinct indices $l_1, l_2, \cdots, l_k$, the sets $T_{l_1}, T_{l_2}, \cdots, T_{l_k}$ contain between them at least k distinct elements. But this is clearly the case here, since each set S_l, and, therefore, each set T_l, contains exactly $(d - j)$ elements after j sets have been so selected. $0 \leqq j \leqq d - 1$. Thus, there are $k(d - j)$ elements in the sets T_{l_1}, $T_{l_2}, \cdots, T_{l_k}$, of which *at most* $(d - j)$ elements are identical (derived from the fact that there are *at most* $(d - j)$ output integers belonging

* $[z]$ is the integral value of z.

to the same integer set after j sets have been so selected). Therefore, there are *at least* k distinct elements. The index j is introduced to show that the selection of d sets of $\pi(x_i)$ can be made on a sequential basis.

III. SOME RESULTS ON ENUMERATIONS

For the remaining sections, the discussion will be restricted to the $(N \times N)$ network with base-2 structure. Some definitions (in addition to those in Part I) relevant to the enumeration study as well as the network diagnosis are given first.

3.1 *Definitions*

(*i*) For any given connection set C, $C \subseteq P$, having input-output pairs (the outputs are denoted as y_i instead of $\pi(x_i)$ to simplify the notations),

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ y_1 & y_2 & \cdots & y_m \end{bmatrix} \qquad 1 \leqq m \leqq N,$$

there exists an inverse of C, denoted by C^{-1}, which is a connection set

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_m \\ x_1 & x_2 & \cdots & x_m \end{bmatrix}.$$

(*ii*) For any two connection sets C_i and C_j that have the same set of input (x_i) and output (y_i) integers, the product $C_iC_j^{-1}$ and its cycle can be defined in the standard manner, similar to that usually associated with permutations.[2]

(*iii*) A loop is a connection set where, for any

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} \in L, \quad \begin{bmatrix} x_j = \hat{x}_i \\ y_j \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_k \\ y_k = \hat{y}_i \end{bmatrix} \in L.$$

The number of these input-output pairs in L is called the order of L, which is necessarily even. Moreover, all loops are distinct, i.e., any two loops do not have any common input-output pair.

(*iv*) A proper loop is a loop in which the input-output pairs are arranged so that both x and $\hat{x}$ and y and $\hat{y}$ are adjacent in a circular sense, e.g.,

$$L = \begin{bmatrix} x_1 & x_2 = \hat{x}_1 & x_3 & x_4 = \hat{x}_3 & x_5 & x_6 = \hat{x}_5 \\ y_1 = \hat{y}_6 & y_2 & y_3 = \hat{y}_2 & y_4 & y_5 = \hat{y}_4 & y_6 \end{bmatrix}.$$

This can always be done, and any loop is considered to be a proper loop, unless otherwise specified. Any permutation P on N integers can be written as

$$P = (L_1 , L_2 , \cdots , L_m)$$

and is said to have m loops, $1 \leqq m \leqq N/2$.

(*v*) A loop L, of order $2k$, is said to be decomposed into two independent connection sets C_1 and C_2, $C_1 , C_2 \subseteq L$, if for any pair $\binom{x_i}{y_i} \in C_1$,

$$\begin{bmatrix} x_j = \hat{x}_i \\ \\ y_j \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_k \\ \\ y_k = \hat{y}_i \end{bmatrix} \in C_2 .$$

(*vi*) The derived sets Q_1 and Q_2 (obtained from independent connection sets C_1 and C_2 respectively by replacing every integer by its characteristic) are denoted by (Q_1 , Q_2). If C_1 and C_2 are reducible, Q_1 and Q_2 are permutations P_1 and P_2 respectively, and they are referred to as derived permutations.

3.2 *Enumeration of Permutations by Loops*

In terms of the definitions just given, the looping procedure for the control of the $(N \times N)$ network, described in Section 4.2 of Part I, is equivalent to arranging the given permutation having m loops into the form

$$P = (L_1 , L_2 , \cdots , L_m) \qquad 1 \leqq m \leqq N/2,$$

and decomposing it to two reducible connection sets C_1 and C_2 by grouping the alternate input-output pairs from each loop into C_1 and the remaining into C_2. Since the decomposition is not unique if $m > 1$, it is readily seen that for any permutation with m loops, there are 2^{m-1} possible ways of decomposition. This leads naturally to the question of how many of the $N!$ permutations have m loops, $1 \leq m \leq N/2$.

The following lemmas and theorems will establish a natural relation between cycles and loops. The enumeration of permutations with m loops $(1 \leqq m \leqq N/2)$ can be expressed in terms of that of cycles, which have been well studied.[3]

Lemma 1: The derived sets (Q_1 , Q_2) of a loop L of order $2k$ have the same set of integers x_i and y_i , and the product $Q_1Q_2^{-1}$ has one cycle of length k.

Proof: Indeed, by definition (*v*), for every input (or output) integer in C_1 , its dual is in C_2 ; and since they have the same characteristic,

Q_1 and Q_2 have the same set of input (or output) integers. Furthermore, since L is a loop, C_1 and C_2 are of the following form

$$C_1 = \begin{bmatrix} x'_1 & x'_2 & \cdots & x_k \\ \hat{y}'_k & \hat{y}'_1 & \cdots & \hat{y}'_{k-1} \end{bmatrix}, \qquad C_2 = \begin{bmatrix} \hat{x}'_1 & \hat{x}'_2 & \cdots & \hat{x}'_k \\ y'_1 & y'_2 & \cdots & y'_k \end{bmatrix},$$

where $[(x'_i + 1)/2] = [(\hat{x}'_i + 1)/2] = x_i$. Thus, Q_1 and Q_2 are of the form

$$Q_1 = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ y_k & y_1 & \cdots & y_{k-1} \end{bmatrix}, \qquad Q_2 = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \end{bmatrix}$$

and, clearly,

$$Q_1 Q_2^{-1} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ x_k & x_1 & \cdots & x_{k-1} \end{bmatrix}$$

and has one cycle of length k.

Q.E.D.

Corollary 1.1: There are 2^{2k-1} loops that give identical derived sets (Q_1, Q_2).

This is clear from the fact that there are $2k$ integers (input and output) in Q_1, and for each (Q_1, Q_2), there are 2^{2k} possible pairs C_1 and C_2. For any given pair C_1 and C_2, the pair C_2 and C_1 reduces to the same (Q_1, Q_2); therefore, 2^{2k} is divided by two.

From definition *(iv)*, any P with m loops can be written as

$$P = (L_1, L_2, \cdots, L_m),$$

where $L_1, L_2, \cdots, L_m$ are disjoint loops. Applying Lemma 1 repeatedly on L_i, the following important theorem that establishes the relation between loops and cycles is obtained.

Theorem 2: The product $P_1 P_2^{-1}$, where P_1 and P_2 are obtained by grouping one Q from each L, has m cycles if and only if P has m loops $(1 \leqq m \leqq N/2)$. As defined in Section 3.1, P_1 and P_2 thus obtained are the derived permutations.

Corollary 2.1: There are 2^{N-m} permutations, having m loops, that will give the same derived permutations (P_1, P_2).

This is proved by repeatedly using Corollary 1.1, and it leads to another enumeration on the number of permutations P that have m loops.

Lemma 3: Define R to be the set of all the derived permutations (P_1, P_2) which have the same product $P_1P_2^{-1}$. Then there are $(N/2)!$ (P_1, P_2) in R.

This is true because both P_1 and P_2 are permutations on $N/2$ integers.

Let $C(n, m)$ denote the number of permutations on n integers that have m cycles. Then there are $C(N/2, m)$ distinct products $P_1P_2^{-1}$ that have m cycles. As a direct consequence of Corollary 2.1 and Lemma 3, the following theorem is established.

Theorem 4. There are exactly $2^{N-m}(N/2)!\ C(N/2, m)$ permutations P which have m loops.

Thus, the enumeration of permutations by loops is related, in a simple manner, to the enumeration of permutations by cycles. The latter problem has been well studied,[3] and the enumeration is generally expressed in terms of the Stirling numbers[4] of the first kind, $s(n, m)$, as follows:

$$C(n, m) = (-1)^{n+m}s(n, m),$$

where $s(n, m)$ can be evaluated from the following generating function

$$\sum_{m=0}^{n} s(n, m)t^m = t(t-1)\cdots(t-n+1)$$

and $(-1)^{n+m}s(n, m)$ is always positive. For the interesting case $m = 1$, the number of permutations with one loop is

$$(N/2)!(N/2-1)!2^{N-1}.$$

3.3 *An Example*

This enumeration is illustrated with the case $N = 8$. If the number of permutations P that have m loops is denoted by $D(N, m)$, Table I

TABLE I—THE NUMBER OF PERMUTATIONS $D(N, m)$ THAT HAVE m LOOPS

m	$C(4, m)$	$D(8, m)$
1	6	$2^7.4!\cdot 6 = 18{,}432$
2	11	$2^6.4!\cdot 11 = 16{,}896$
3	6	$2^5.4!\cdot 6 = 4{,}608$
4	1	$2^4.4!\cdot 1 = 384$
		Total $8! = 40{,}320$

accounts for all the permutations. Wherever $m > 1$, there are more than one setting of the β-elements in the input and output stages that will satisfy the same permutation. This applies to all the stages as each subnetwork is taken into consideration, and, thus, the total number of states provided by all the β-elements exceeds $\langle \log_2(N!) \rangle$.

IV. CONSTRUCTION OF TEST PERMUTATIONS

It has been pointed out above that for any P having m loops, certain m β-elements at the input and output stages can be arbitrarily set. To detect faulty β-elements, one must find a class of input-output permutations, or test permutations which are realized by a unique setting of the β-elements. The property of such a permutation is that it and all its derived permutations, at every level of the network, have exactly one loop. To show that they do exist and can be generated, one proceeds as follows:

Lemma 5: If a loop L is given, then any loop L' formed from L by taking the dual of one or more of the integer pairs (input or output) in L will give the same derived sets (Q_1, Q_2), (Q_2, Q_1) being considered the same as (Q_1, Q_2) for the remaining discussion.

This is obvious from the fact that the characteristic of an integer is not changed by taking its dual.

Theorem 6: Let L and L' be two loops having the same (Q_1, Q_2). Then the product $L(L')^{-1}$ has one cycle if and only if L' is obtained from L by replacing every integer except one integer pair (input or output) in L by its dual.

Proof: That L and L' do have the same (Q_1, Q_2) is a direct consequence of Lemma 5. Moreover, the loops L and L' have the same set of x_i and y_i, since x_i, $\hat{x}_i$, y_i, $\hat{y}_i$ are all in L. Thus the product $L(L')^{-1}$ is defined. Now, let L of order $2k$ be written as follows:

$$L = \begin{pmatrix} \hat{x}_1 & x_2 & \hat{x}_2 & \cdots & x_k & \hat{x}_k & x_1 \\ y_1 & \hat{y}_1 & y_2 & \cdots & \hat{y}_{k-1} & y_k & \hat{y}_k \end{pmatrix}$$

and

$$L' = \begin{pmatrix} x_1 & \hat{x}_2 & x_2 & \cdots & \hat{x}_k & x_k & \hat{x}_1 \\ \hat{y}_1 & y_1 & \hat{y}_2 & \cdots & y_{k-1} & y_k & \hat{y}_k \end{pmatrix},$$

where, without loss of generality, the only unchanged pair is y_k and $\hat{y}_k$, since the ordering of subscripts is immaterial. The transformation from L to L' can be expressed in terms of the input-output pairs, namely, $\left[\begin{smallmatrix}\hat{x}_i\\ y_i\end{smallmatrix}\right]$ is replaced by $\left[\begin{smallmatrix}x_i\\ \hat{y}_i\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}x_i\\ \hat{y}_{i-1}\end{smallmatrix}\right]$ by $\left[\begin{smallmatrix}\hat{x}_i\\ y_{i-1}\end{smallmatrix}\right]$ except $\left[\begin{smallmatrix}\hat{x}_k\\ y_k\end{smallmatrix}\right]$ is replaced by $\left[\begin{smallmatrix}x_k\\ y_k\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}x_1\\ \hat{y}_k\end{smallmatrix}\right]$ by $\left[\begin{smallmatrix}\hat{x}_1\\ \hat{y}_k\end{smallmatrix}\right]$. The input-output pair for $L(L')^{-1}$ will be, in general, $\left[\begin{smallmatrix}\hat{x}_i\\ \hat{x}_{i+1}\end{smallmatrix}\right]$ for $1 \leqq i < k$ and $\left[\begin{smallmatrix}x_i\\ x_{i-1}\end{smallmatrix}\right]$ for $1 < i \leqq k$ with the exception of $\left[\begin{smallmatrix}\hat{x}_k\\ \hat{x}_k\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}x_1\\ \hat{x}_1\end{smallmatrix}\right]$. Therefore,

$$L(L')^{-1} = (\hat{x}_1\hat{x}_2 \cdots \hat{x}_k x_k \cdots x_1),$$

where the product written in the familiar cycle form has one cycle of length $2k$.

To show the converse, it is sufficient to show that the loop L', obtained by either taking the dual of every integer in L or taking the dual of every integer except two or more integer pairs, will not satisfy the second property. Referring to the above, it is seen that if every integer is replaced by its dual, then the product

$$L(L')^{-1} = (\hat{x}_1\hat{x}_2 \cdots \hat{x}_k)(x_k x_{k-1} \cdots x_1)$$

has two cycles, each of length k.

If there are more than two integer-pairs unchanged, one can always write

$$L' = \begin{bmatrix} x_1 & \hat{x}_2 & x_2 & \cdots & x_j & \hat{x}_{j+1} & \cdots & \hat{x}_k & x_k & \hat{x}_1 \\ \hat{y}_1 & y_1 & \hat{y}_2 & \cdots & y_j & \hat{y}_j & \cdots & y_{k-1} & y_k & \hat{y}_k \end{bmatrix},$$

where the first other unchanged integer pair is y_j and $\hat{y}_j$, $j < k$. Then, by the same argument given above, the product $L(L')^{-1}$ has at least one cycle of length $2j$, namely,

$$(\hat{x}_1\hat{x}_2 \cdots \hat{x}_j x_j x_{j-1} \cdots x_1) \qquad 2j < 2k. \qquad \text{Q.E.D.}$$

Corollary 6:1: If L is a loop of order k, $k \geqq 4$, there are k such loops L′ where L and L′ give identical derived sets (Q_1 , Q_2) and $L(L')^{-1}$ has one cycle.

This is obvious since there are $k/2$ input integer pairs and $k/2$ output integer pairs. The case $k = 2$ is a degenerate one, since taking the dual of the input pair only yields the same loop as the one obtained by taking the dual of the output pair only. Hence, only one L' is possible.

By repeatedly using the above theorem, one can show the following.

Theorem 7: If $P = (L_1 , L_2 , \cdots , L_m)$ and its derived permutations are (P_1 , P_2), then another permutation P′ will have the same derived permutations and $P(P')^{-1}$ will have m cycles if and only if P′ is ob-

tained by taking the dual of every integer except one integer pair (input or output) in each L_i, $1 \leqq i \leqq m$, in P.

Corollary 7.1: There are $k_1 k_2 \cdots k_m$ ways of deriving P' such that P and P' give identical derived permutations (P_1, P_2) and $P(P')^{-1}$ has m cycles, where k_i is the order of L_i and $k_i \geqq 4$. If $k_i = 2$ for some L_i, it will be taken as unity.

The following example illustrates what has been discussed. P has two loops, and the derived permutations (P_1, P_2) have the property that $P_1 P_2^{-1}$ has two cycles.

$$P = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 8 & 5 & 1 & 9 & 7 & 11 & 3 & 12 & 2 & 10 & 4 & 6 \end{pmatrix}$$

$$= \left(\underbrace{\begin{matrix} 1 & 5 & 6 & 8 & 7 & 11 & 12 & 2 \\ 8 & 7 & 11 & 12 & 3 & 4 & 6 & 5 \end{matrix}}_{\text{a loop}} \; \underbrace{\begin{matrix} 3 & 9 & 10 & 4 \\ 1 & 2 & 10 & 9 \end{matrix}}_{\text{a loop}} \right).$$

The decomposition of P yields:

$$P_1 = \begin{pmatrix} 1 & 3 & 4 & 6 & 2 & 5 \\ 4 & 6 & 2 & 3 & 1 & 5 \end{pmatrix}; \qquad P_2 = \begin{pmatrix} 3 & 4 & 6 & 1 & 5 & 2 \\ 4 & 6 & 2 & 3 & 1 & 5 \end{pmatrix};$$

and $P_1 P_2^{-1} = (1\ 3\ 4\ 6)\ (2\ 5)$ has two cycles. P', which gives the same pair (P_1, P_2), is obtained from P, and one of the 32 possibilities is

$$P' = \begin{pmatrix} 2 & 6 & 5 & 7 & 8 & 12 & 11 & 1 & 4 & 10 & 9 & 3 \\ 7 & 8 & 12 & 11 & 4 & 3 & \underline{6} & \underline{5} & \underline{1} & \underline{2} & 9 & 10 \end{pmatrix}$$

(The underlined integers are the unchanged ones in each loop.) Furthermore, $P(P')^{-1} = (1\ 6\ 7\ 12\ 11\ 8\ 5\ 2)\ (3\ 4\ 9\ 10)$.

The permutations for which $m = 1$ can be used in the generation of the test permutations. This is achieved, for any N, by starting with any permutation on 4 integers that has one loop and applying Theorem 7 repeatedly in an iterative manner. One can show that there are

$$2^{(2N-3+(\log_2 N(\log_2 N-3))/2)}$$

such test permutations by repeatedly using Corollary 2.1 and Corollary 7.1, with $m = 1$. The construction of one of these, based on Theorem 7, is illustrated as follows.

In order to clarify the following discussion, test permutations on N

integers and their derived permutations are denoted as $T(N)$ and $(T_1(N/2), T_2(N/2))$ respectively. If it is desired to construct a $T(16)$, then the first step is to select a $T_1(4)$ and generate $T_2(4)$ such that $T_1(4)(T_2(4))^{-1}$ has one cycle. There are 16 $T(4)$ that have one loop; one of these is

$$T_1(4) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}.$$

$T_2(4)$ is obtained by taking the dual of every element except one integer pair (by Theorem 7). One of the four choices is

$$T_2(4) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}.$$

Any permutation that decomposes into $T_1(4)$ and $T_2(4)$ can be used for $T_1(8)$; one of the 128 possible permutations (by Corollary 2.1) is

$$T_1(8) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 3 & 8 & 5 & 4 & 7 & 6 & 2 \end{pmatrix},$$

where the connection set corresponding to $T_1(4)$ is taken as

$$\begin{bmatrix} 1 & 3 & 5 & 7 \\ 1 & 8 & 4 & 6 \end{bmatrix}.$$

There are eight choices for $T_2(8)$, and one of these is

$$T_2(8) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 4 & 6 & 7 & 8 & 3 & 1 & 5 \end{pmatrix}.$$

Similarly, one of the possible $T(16)$'s is

$$T(16) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 6 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 1 & 4 & 5 & 8 & 12 & 16 & 13 & 10 & 15 & 7 & 6 & 14 & 2 & 11 & 9 & 3 \end{pmatrix}.$$

It will now be shown that a permutation that is realized by complementing every setting of the β-elements that realizes $T(N)$ except the one corresponding to inputs 1 and 2 (see network structure) is also a test permutation, $T^c(N)$, and that it can be generated parallel to $T(N)$. These two permutations are used for fault detection in a manner to be described later.

Theorem 8: Let $T_1(N)$ be a test permutation that has $(T_1(N/2), T_2(N/2))$ as the derived permutations. And if there exist two other test permutations

$T_1^c(N/2)$ and $T_2^c(N/2)$ such that the product $T_1^c(N/2)(T_2^c(N/2))^{-1}$ has one cycle, then one can construct a test permutation which will have $(T_1^c(N/2), T_2^c(N/2))$ as the derived permutations.

Proof: Let $T_1(N)$ be decomposed into two reducible connection sets, C_1 and C_2, where C_1 can be written as

$$C_1 = \begin{bmatrix} x_1 = 1 & x_2 & \cdots & x_{N/2} \\ y_1 & y_2 & \cdots & y_{N/2} \end{bmatrix}$$

($x_1 = 1$ is arbitrarily defined by the network structure).

The connection set C_1^c is formed by replacing each input and output integer $z \in T_1^c(N/2)$, except $x_1 = 1$, by x (or y) that has the characteristic z and has its dual $\hat{x} = x_i$, for some i, belonging to C_1, $1 < i \leqq N/2$ (or $\hat{y} = y_i \in C_1, 1 \leqq i \leqq N/2$). Clearly, this can always be done because every $z \in T_1^c(N/2)$ has two integers with z as their characteristic, and only one of them is in C_1. Similarly, a connection set C_2^c can be formed from $T_2^c(N/2)$, based on C_2. The permutation, obtained simply by combining C_1^c and C_2^c, has the derived permutations $(T_1^c(N/2), T_2^c(N/2))$, and it has only one loop. Furthermore, it results in the complementary setting of β-elements by the looping algorithm, since any integer in C_1^c is the dual of some integer in C_1. Q.E.D.

Theorem 9: One can construct a test permutation $T_2^c(N)$ from $T_1^c(N)$ in the same way as $T_2(N)$ is obtained from $T_1(N)$ as given in Theorem 7, and the product has also one cycle.

Proof: Let $T_2(N)$ be obtained from $T_1(N)$ by taking the dual of every integer except one pair, say, $(x_i, \hat{x}_i)$. The permutation, obtained from $T_1^c(N)$ by taking the dual of every integer pair except the same pair $(x_i, \hat{x}_i)$, is indeed a test permutation by Theorem 7. Using the same argument as in Theorem 8, it is easily seen that the setting of β-element to realize this permutation is complementary to that for $T_2(N)$. Hence it is $T_2^c(N)$. Q.E.D.

Since the permutations $T_1(2)$ and $T_2(2)$, and the corresponding $T_1^c(2)$ and $T_2^c(2)$, can always be constructed, one can, by induction on N, construct the two test permutations $T(N)$ and $T^c(N)$ for arbitrary value of N. One can, in fact, generalize Theorems 8 and 9 to establish arbitrary relations between two permutations in the β-element settings in addition to $T(N)$ and $T^c(N)$ for which the settings are complementary.

The construction procedure for $T^c(N)$ is illustrated by determining the $T^c(16)$ as related to $T(16)$ given in the previous example. In that example, $C_1 = \begin{bmatrix} 1 & 4 \\ 1 & 3 \end{bmatrix}$ and $C_2 = \begin{bmatrix} 2 & 3 \\ 4 & 2 \end{bmatrix}$. Also $T_1(2) = \binom{1\ 2}{1\ 2}$ and $T_2(2) = \binom{1\ 2}{2\ 1}$;

therefore, $T_1^c(2) = \left(\begin{smallmatrix}1 & 2\\ 2 & 1\end{smallmatrix}\right)$ and $T_2^c(2) = \left(\begin{smallmatrix}1 & 2\\ 1 & 2\end{smallmatrix}\right)$. All elements except input 1 in C_1^c must have their dual in C_1; also $T_1^c(2)$ and $T_2^c(2)$ must be satisfied; therefore,

$$C_1^c = \begin{bmatrix}1 & 3\\ 4 & 2\end{bmatrix}, \quad \text{and} \quad C_2^c = \begin{bmatrix}2 & 4\\ 1 & 3\end{bmatrix}$$

and

$$T_1^c(4) = \begin{pmatrix}1 & 2 & 3 & 4\\ 4 & 1 & 2 & 3\end{pmatrix}.$$

Keeping the input integer pair (1, 2) unchanged, one has

$$T_2^c(4) = \begin{pmatrix}1 & 2 & 3 & 4\\ 3 & 2 & 4 & 1\end{pmatrix}.$$

Repeating the procedure, one obtains

$$T_1^c(8) = \begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8\\ 7 & 6 & 4 & 2 & 8 & 3 & 1 & 5\end{pmatrix},$$

$$T_2^c(8) = \begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8\\ 8 & 5 & 1 & 3 & 4 & 7 & 6 & 2\end{pmatrix},$$

and

$$T^c(16) = \begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16\\ 13 & 16 & 10 & 12 & 8 & 1 & 4 & 5 & 15 & 7 & 6 & 14 & 2 & 11 & 9 & 3\end{pmatrix}.$$

V. INVERSE CONNECTING EQUATIONS FOR BASE-2 STRUCTURE

The looping procedure for setting the β-elements to realize a given P is described in Part I. The inverse problem of defining P from the states of the β-elements is also of some interest. If P can be derived from the β-element setting, then it is not necessary to store the connections in another memory. Also the inverse connecting equations are used in the location of faulty β-elements.

In the control algorithm for the base-2 structure, the states of the β-elements are derived in an iterative manner from the outside (first) level to the center $(\log_2 N)^{\text{th}}$ level. Therefore, to obtain the inverse connecting equations, the states of β-elements in the center stage are considered first.

The β-elements in the network are numbered (see Fig. 1) such that

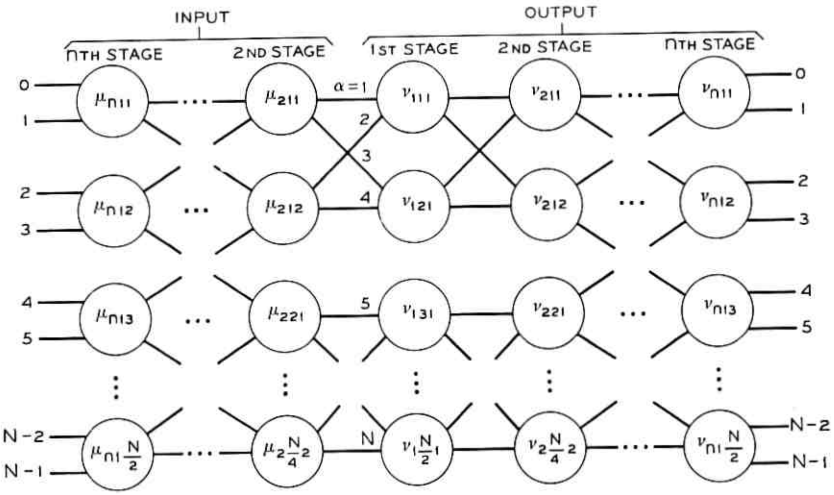

Fig. 1—The numbering of β-elements in an $(N \times N)$ network.

the defining equations for each input-output pair appear in a simple form. μ_{jkl} and ν_{jkl} are the input and output β-elements respectively. They are located in the jth stage (counting from the center stage), the kth $(2^j \times 2^j)$ network, and the lth position at the input (or output) stage. The center stage is considered as the first output stage, and the β-elements are denoted by ν_{1k1}. They are defined as '0' or '1' when set to straight-through state or crossover state, respectively.

The code for the input and output integers is the same as given in Section 4.1 of Part I. For any input integer x_i (or output integer y_i), the normal binary representation is its coded form, having a code length of $n = \log_2 N$. And it is expressed as follows:

$$x_i = x_{i1}x_{i2} \cdots x_{in} .$$

The inputs to the center stage are designated by α, $\alpha = 1, 2, \cdots, N$, as shown in Fig. 1, and the input-output pairs are ordered according to α. For each input-output pair $\binom{x_\alpha}{y_\alpha}$, the code words for $x_\alpha = x_{\alpha 1}x_{\alpha 2} \cdots x_{\alpha n}$ and $y_\alpha = y_{\alpha 1}y_{\alpha 2} \cdots y_{\alpha n}$ can be calculated from the following inverse connecting equations:

$$x_{\alpha 1} = (\alpha + 1) \bmod 2,$$

$$x_{\alpha j} = \mu_{jkl_x}^{\rho \bmod 2*} \qquad 1 < j \leqq n;$$

* $z^1 \equiv z$ and $z^0 \equiv \bar{z}$, the complement of z, and $z = 0$ or 1.

and

$$
\begin{aligned}
y_{\alpha 1} &= \nu_{1[(\alpha+1)/2]1}^{\alpha \bmod 2} , \\
y_{\alpha j} &= \nu_{jkl_y}^{\rho \bmod 2} \qquad 1 < j \leqq n;
\end{aligned}
\tag{1}
$$

where

$$
\rho = \left[\frac{\alpha - 1 + 2^{i-1}}{2^{i-1}} \right], \qquad k = \left[\frac{\rho + 1}{2} \right]
$$

and l_x and l_y are the integers represented in the coded form by $x_{\alpha 1} x_{\alpha 2} \cdots x_{\alpha (j-1)}$ and $y_{\alpha 1} y_{\alpha 2} \cdots y_{\alpha (j-1)}$ respectively. The equations for $x_{\alpha j}$ (or $y_{\alpha j}$), $1 < j \leqq n$, can be obtained in a recursive manner from the following Boolean equations:

If α and j are such that ρ is even,

$$
\begin{aligned}
x_{\alpha j} = (\bar{x}_{\alpha 1} \bar{x}_{\alpha 2} \cdots \bar{x}_{\alpha (j-1)}) \mu_{jk1} + (\bar{x}_{\alpha 1} \bar{x}_{\alpha 2} \cdots x_{\alpha (j-1)}) \mu_{jk2} + \cdots \\
+ (x_{\alpha 1} x_{\alpha 2} \cdots x_{\alpha (j-1)}) \mu_{jk(2^{i-1})} .
\end{aligned}
$$

And, if ρ is odd,

$$
\begin{aligned}
x_{\alpha j} = (\bar{x}_{\alpha 1} \bar{x}_{\alpha 2} \cdots \bar{x}_{\alpha (j-1)}) \bar{\mu}_{jk1} + (\bar{x}_{\alpha 1} \bar{x}_{\alpha 2} \cdots x_{\alpha (j-1)}) \bar{\mu}_{jk2} + \cdots \\
+ (x_{\alpha 1} x_{\alpha 2} \cdots x_{\alpha (j-1)}) \bar{\mu}_{jk(2^{i-1})} .
\end{aligned}
$$

The following example of the (8 × 8) network shown in Fig. 2 illustrates the inverse procedure. For each $\alpha = 1, 2, \cdots, 8$, the coded

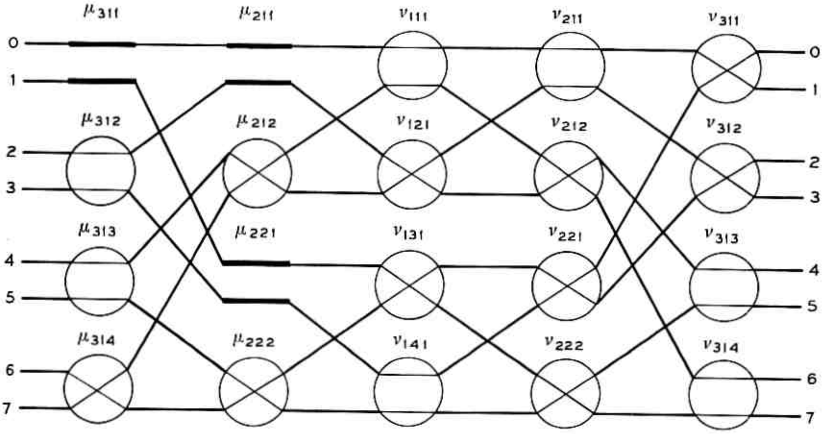

Fig. 2—β-element setting for $P = \binom{01234567}{17403526}$.

inputs and outputs are calculated. If $\alpha = 5$, then

$$x_{51} = (5 + 1) \bmod 2 = 0;$$

$$x_{52} = \mu_{2[(3+1)/2]1}^{3 \bmod 2} = \mu_{221} = 0;$$

and

$$x_{53} = \mu_{3[(2+1)/2]1}^{2 \bmod 2} = \bar{\mu}_{311} = 1.$$

Also

$$y_{51} = \nu_{1[(5+1)/2]1}^{5 \bmod 2} = \nu_{131} = 1;$$

$$y_{52} = \nu_{2[(3+1)/2]2}^{3 \bmod 2} = \nu_{222} = 1;$$

and

$$y_{53} = \nu_{3[(2+1)/2]4}^{2 \bmod 2} = \bar{\nu}_{314} = 1.$$

Therefore, the input-output pair

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \begin{bmatrix} 001 \\ 111 \end{bmatrix} = \begin{bmatrix} 1 \\ 7 \end{bmatrix}.$$

The remaining

$$\begin{bmatrix} x_\alpha \\ y_\alpha \end{bmatrix} (\alpha = 1, 2, 3, 4, 6, 7, 8)$$

are determined in the same manner, and they are:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 000 \\ 001 \end{bmatrix}; \quad \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 111 \\ 110 \end{bmatrix}; \quad \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 010 \\ 100 \end{bmatrix}; \quad \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} 100 \\ 011 \end{bmatrix};$$

$$\begin{bmatrix} x_6 \\ y_6 \end{bmatrix} = \begin{bmatrix} 110 \\ 010 \end{bmatrix}; \quad \begin{bmatrix} x_7 \\ y_7 \end{bmatrix} = \begin{bmatrix} 011 \\ 000 \end{bmatrix}; \text{ and } \begin{bmatrix} x_8 \\ y_8 \end{bmatrix} = \begin{bmatrix} 101 \\ 101 \end{bmatrix}.$$

Then the input-output permutation is

$$P = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 7 & 4 & 0 & 3 & 5 & 2 & 6 \end{pmatrix}.$$

VI. DIAGNOSIS OF FAULTY β-ELEMENTS

The physical design of the β-element is the major factor in determining the method of detecting and locating the faulty elements in the network. For example, the detection of a faulty β-element which either opens or shorts to ground is trivial. If one has access to the actual state of each β-element, then the location of faulty elements is also trivial. In this paper it is assumed that the individual β-element is not accessible, and it is considered to fail when it remains in one of the two states.

6.1 *Detection*

By using the test permutations $T(N)$ and $T^c(N)$, each β-element is checked for the two possible states. Failure in any number of β-elements will be detected by the fact that either $T(N)$ or $T^c(N)$ or both will not be realized. It is to be noted that any permutation having more than one loop cannot be used because the setting of some two or more β-elements is arbitrary, and, therefore, failure of these elements in certain states may not be detected.

6.2 *Location*

With any test permutation $T(N)$, failure of one β-element will result in a permutation different from $T(N)$ by only two input-output pairs, that is, the input-output pairs $\binom{x_i}{y_i}$ and $\binom{x_j}{y_j}$ become $\binom{x_i}{y_j}$ and $\binom{x_j}{y_i}$ for some i and j. The inverse connecting algorithm discussed in Section V can be used to locate the particular (or the faulty) β-element common to $\binom{x_i}{y_i}$ and $\binom{x_j}{y_j}$ with their associated α's which are stored in the memory.

The following example illustrates this procedure. If the test permutation for an (8×8) network is

$$T(8) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 2 & 7 & 4 & 3 & 6 & 5 & 1 \end{pmatrix},$$

then, using the same coding scheme as in Section V, the setting of the β-elements for it is shown in Fig. 3. Also, the α's corresponding to each input-output pair are calculated by using the inversing connecting equations, and they are given as follows: α_1, α_2, α_3, α_4, α_5 α_6, α_7, and and α_8 correspond to input-output pairs

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \text{ and } \begin{bmatrix} 7 \\ 1 \end{bmatrix} \text{ respectively.}$$

Assume that β-element ν_{212} is faulty, and it is fixed in the crossover position. Then the actual permutation realized is

$$P = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 2 & 5 & 4 & 3 & 6 & 7 & 1 \end{pmatrix},$$

and the incorrect pairs of $T(8)$ are $\binom{2}{7}$ and $\binom{6}{5}$ or in coded form $\binom{010}{111}$ and $\binom{110}{101}$. The α's for these pairs are 3 and 2 respectively. By using

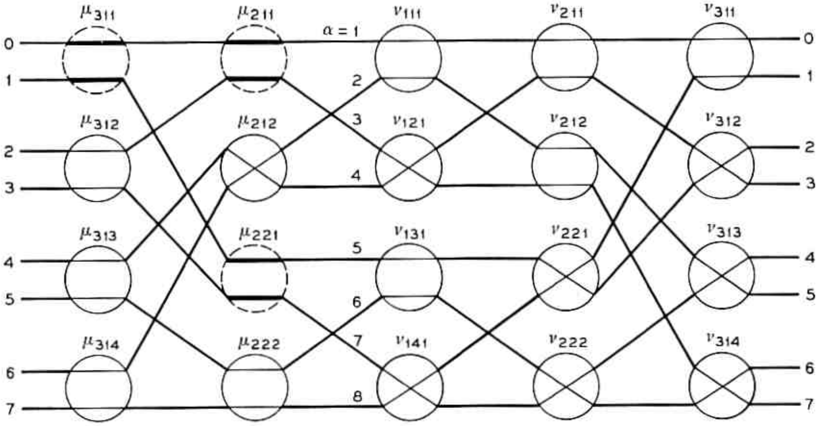

Fig. 3—β-element setting for $T(8) = \binom{01234567}{02743651}$.

the inverse connecting equations (1), the β-elements through which $\binom{2}{7}$ and $\binom{6}{5}$ are connected are found to be μ_{312} , μ_{211} , ν_{121} , ν_{212} , and ν_{314} , and μ_{212} , ν_{111} , ν_{212} , and ν_{313} respectively. It is seen that β-element ν_{212} is common to both input-output pairs; therefore, it is the faulty β-element.

If two β-elements are faulty, there are either three or four pairs in $T(N)$ not realized. If four input-output pairs are wrong, the locations of faulty elements can be determined in the same manner as described above. Three pairs are incorrect when one particular $\binom{x_i}{y_i}$ is connected through both of the faulty elements. For this case, it is necessary to change $T(N)$ so that one of the faulty elements is in the proper state, and then the other one can be located. This is achieved by having $2 \log_2 N - 1$ test permutations with each one constructed (using a generalized form of Theorems 8 and 9) to complement different stages of input or output β-elements, one stage at a time. The same procedure as above is used to locate the faulty β-element.

If the faulty β-elements are restricted to one stage of the network, then this stage can be located in a manner similar to the above. For this case, a set of $T(N)$, $\log_2 N$ in number, is used to complement the β-elements on each stage (input and output) of the network. The α's corresponding to the incorrect $\binom{x_i}{y_i}$'s will remain the same until the faulty elements are complemented. Therefore, the stage containing the faulty β-elements is determined.

6.3 *Adaptive "Looping" Algorithm*

In the looping algorithm as given in Part I, the derived permutations P_1 and P_2 are always routed through the upper and lower

$(N/2 \times N/2)$ networks respectively. This is because in the most efficient network structure (see Fig. 1 of Part I) one β-element in the first stage of each subnetwork (i.e., μ_{jk1}, $1 < j \leqq \log_2 N$, $1 \leqq k \leqq N/2^j$) is fixed in the straight-through state. However, by introducing redundant β-elements μ_{jk1}, (as in Fig. 1) one can change (or adapt) the control algorithm at certain stages to realize a particular permutation if there is one faulty β-element per subnetwork at each stage.

VII. CONCLUSION

Important relationships between the loops of input-output permutations and cycles of permutations are established. These properties are used to enumerate the input-output permutations in terms of loops and to construct special test permutations which require unique β-element settings. Also, inverse connecting equations which define the input-output permutation from the states of the β-elements are derived. These ideas are utilized in the diagnosis of faulty β-elements.

It is clear that network failure due to any number of faulty elements, which may be distributed over many stages, can be easily detected by using only a pair of test permutations. If these faulty elements are limited to only one stage of the network, this stage can be located by employing the inverse equations and a set of test permutations. Furthermore, if only one or two elements fail, their exact positions in the network can be located by employing a similar procedure.

If the faulty elements are limited to only one in the first stage of each subnetwork, then any input-output permutation can be realized correctly by adding a redundant β-element in the first stage of each subnetwork and adapting the looping algorithm at the appropriate subnetworks.

The fact that this type of rearrangeable switching network has some attractive diagnostic properties should enhance the possibility of it being used in some practical switching systems.

REFERENCES

1. Hall, P., Jr., *Combinatorial Theory,* Waltham, Mass.: Blaisdell Publishing Company, 1967, Chapter 5.
2. Ledermann, W., *Introduction to the Theory of Finite Groups,* New York: Interscience Publishers, 1964, Chapter 3.
3. Riordan, J., *Introduction to Combinatorial Analysis,* New York: John Wiley and Sons, 1958, Chapter 4.
4. National Bureau of Standards, *Handbook of Mathematical Functions,* AMS 55, Washington: National Bureau of Standards, March 1965, p. 824.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

A Full-Duplex Echo Suppressor Using Center-Clipping

By O. M. MRACEK MITCHELL and DAVID A. BERKLEY

(Manuscript received October 9, 1970)

For telephone circuits which include synchronous satellites, conventional echo suppressors of the voice-switching type are less than satisfactory because of speech mutilation and the presence of echo during double talking.[1] *We have found that a multiband center-clipping process may be used as an echo suppressor. This echo suppressor is unique in that no double-talking decision has to be made. The near-end signal, plus echo of the far-end signal, is divided into several contiguous bands with each filter output going to a center clipper. A control circuit sets each clipping level equal to or greater than the echo level in that band. A preliminary analogue implementation of this echo suppressor, in which control circuit gains were manually adjusted to match the experimental return loss, was informally demonstrated using a simulated satellite circuit. Although no attempt at quantitative evaluation has yet been carried out and further evaluation is necessary, no echo was reported during this demonstration, even during double talking, for return losses approaching 0 dB. Operation appeared to be full-duplex at all times with little distortion of the speech. For return losses greater than about 15 dB, the center-clipping system was almost indistinguishable from a 4-wire connection with no echo path. In practice, adaptive setting of control circuit gains as a function of return loss would be desirable if this technique is used as a replacement for conventional echo suppressors.*

I. INTRODUCTION

During investigations of a multiband center-clipping process for use in reverberation reduction[2] it occurred to us that this process, which can remove the effects of long-time reverberation or echoes in a room, could also be used to remove echoes in telephone lines resulting from imperfect hybrid junctions.[3] Independently, J. R. Pierce also suggested that this process could be applied to echo suppression and proposed a scheme for controlling the levels of the center clippers in a conventional split echo suppressor configuration.[4]

One end of a conventional split echo suppressor is shown in Fig. 1. It is located in the 4-wire section of line near the hybrid junction to the 2-wire loop of the near-end customer. A similar configuration is inserted at the other end of the 4-wire trunk. Because of imperfect balancing of the hybrid, part of the received signal from the far-end talker feeds through the hybrid to the transmit side of the 4-wire line. The return loss of the hybrid is typically 15 dB, that is, the echo level at the echo suppressor is 15 dB below the normal transmit signal level of the near-end talker measured at the same point. The conventional echo suppressor is a voice-operated switch. The logic and control circuit detects the presence of received signal and causes a loss of at least 50 dB to be inserted in the path of the echo signal on the transmit side. Since the loss would also attenuate the signal from the near-end talker, and temporarily make the connection one way, the logic and control circuit also detects the presence of double talking and puts the suppressor into a "break-in" mode which allows an interruption to take place.

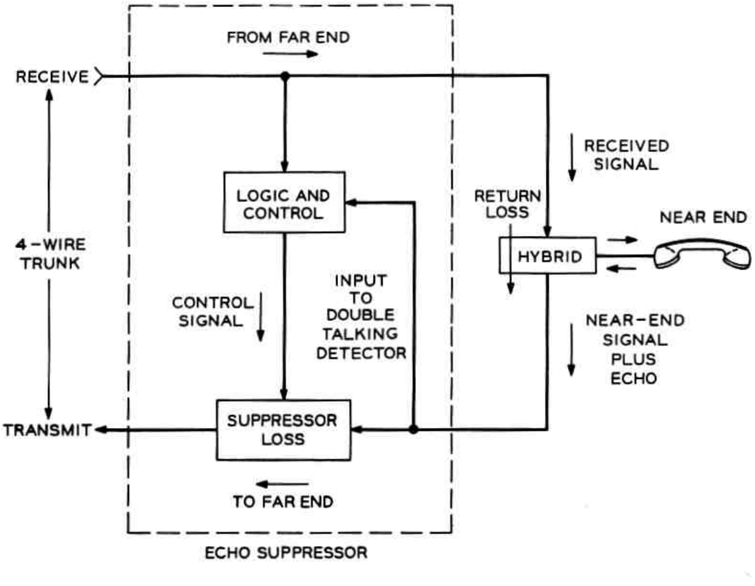

Fig. 1—One end of a conventional split echo suppressor.

Alternatively, we have found that the echo signal can be removed by replacing the voice switch with the multiband center-clipping process which is mentioned above, and which we have described previously.[2] This configuration is shown in Fig. 2. The outgoing signal from the hybrid is divided into a number of contiguous frequency bands by an input filter bank, each band is center clipped independently, and then the odd harmonic distortion products introduced by the center clippers are removed by an output filter bank generally identical to the input filter bank. For echo suppression, the center-clipping levels are controlled by the received signal. This signal is divided into contiguous bands by a control filter bank which is identical to the input filter bank. The attenuation in each band is adjusted to be equal to or less than the trans-hybrid loss in that band so that control signals identical to or larger than the filtered echo are obtained. The output of each band is peak detected and the detected output sets the clipping level in the corresponding center clipper so as to remove the echo signal in that particular band. In the absence of received signal, the clipping levels are zero. The clipping-level rise-times are comparable to the speech bandwidth and should have a hold time greater than the echo end-delay which may be up to 25 ms.

This center-clipping system has several advantages over existing echo suppressors of the voice-switching type. Since the frequency spectrum is divided into a number of bands, the near-end signal is unaffected in bands where there is no energy in the echo signal and the echo is completely removed in bands where there is no near-end signal component. However, the main advantage appears to come from the use of center clipping as opposed to voice switching. Break-in of the near-end talker can occur without a double-talking decision, even for a return loss approaching 0 dB, and no echo is heard during double talking. A comparison of the effect of center clipping and voice switching on signals will be discussed in the next section to show how these advantages come about.

II. CENTER CLIPPING AS AN ALTERNATIVE TO VOICE SWITCHING

The transfer function of the center clipper we will discuss is shown in Fig. 3. This center clipper completely eliminates signals below the clipping level, but leaves instantaneous signal values greater than the clipping level unaffected. In a sense, a center clipper is a voice switch operating on the instantaneous amplitude of the signal. However, it differs greatly from the process commonly referred to as voice switch-

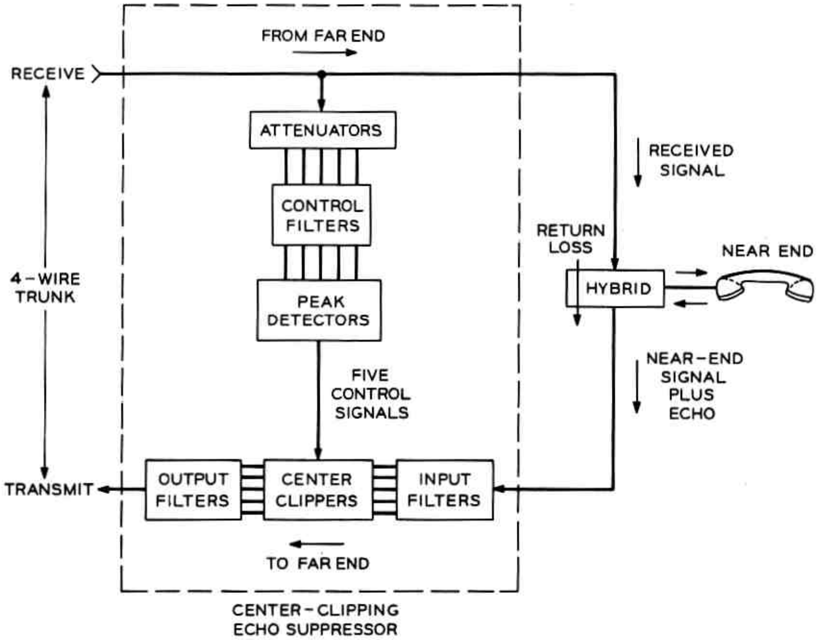

Fig. 2—One end of a split center-clipping echo suppressor.

ing. As we have mentioned in the preceding section, a large constant amount of attenuation (>50 dB) is generally switched into the transmit path in response to the control signal. In principle a more ideal kind of voice switching would be switching of only the amount of attenuation required, in addition to existing hybrid return loss, to

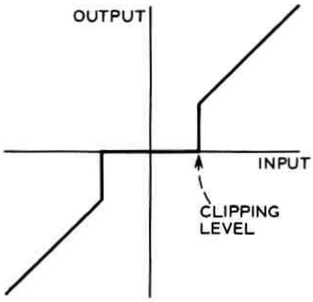

Fig. 3—Minimum distortion center-clipping transfer function.

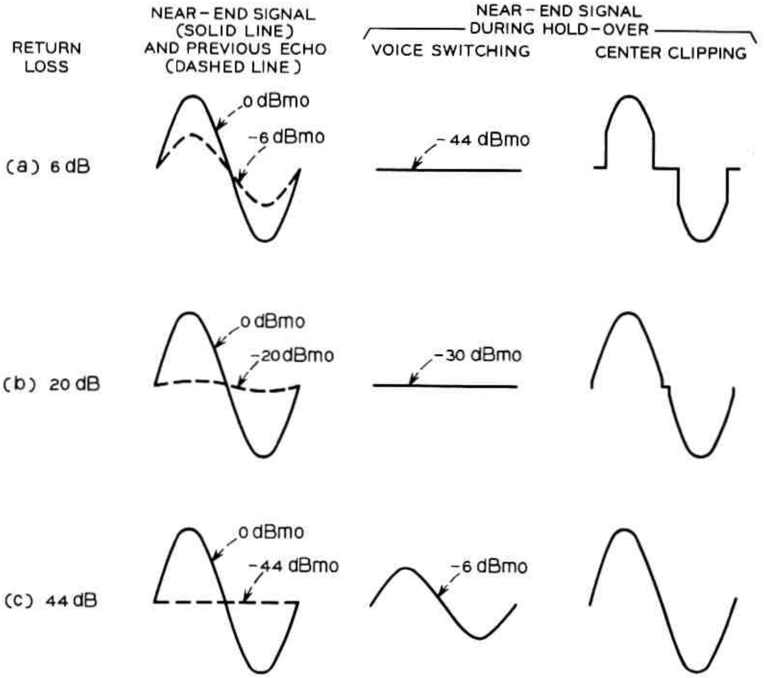

Fig. 4—Comparison of voice-switching and center-clipping necessary to produce 50 dB of echo suppression for return losses of: (a) 6 dB, (b) 20 dB, (c) 44 dB.

reduce the unwanted signal to a tolerable level. It is this kind of voice switch which we will compare with the center clipper of Fig. 3.

For satellite communications connections, a conservative estimate is that the echo signal level should be about 50 dB below the level of the near-end talker. Consider the situation depicted in Fig. 1, however, where the suppressor loss is replaced by either the minimum amount of attenuation or center clipping required, and where no double-talking detector is provided. The basic difference between these two hypothetical processing systems is shown in Fig. 4 for three values of return loss. The output of the echo suppressor for each case is shown in response to a sinusoidal signal, at 0 dBm0, from the near end into the echo suppressor. These graphs apply during the hold-over time after the voice switching or clipping level has been set by a previously received signal of the same transmission level as the near-end signal, and where the echo level has decreased to a negligible value.

In Fig. 4a, for a return loss of 6 dB, the echo signal is at −6 dB relative to the near-end signal, i.e., at −6 dBm0. Consequently, an attenuation of 44 dB has to be switched into the transmit path to achieve the desired 50 dB suppression. During the hold-over, this would drop the near-end signal by 44 dB. On the other hand, center clipping at one-half peak amplitude eliminates the echo and results in only 6 percent loss of fundamental signal energy.

In Fig. 4b, the signals for a return loss of 20 dB are shown. The echo signal is at −20 dBm0. Voice switching of 30 dB of attenuation reduces the near-end signal to −30 dBm0 while center clipping at 10 percent of peak, sufficient to remove the echo, produces very little distortion of the near-end signal.

Even when the unwanted signal is −44 dBm0 as in Fig. 4c, voice switching of 6 dB is necessary. This reduces the near-end signal to half amplitude while the corresponding center clipping at 1 percent of peak results in negligible effect on the near-end signal.

It is evident in Fig. 4 that, for reasonable return loss, center clipping is a much less severe form of processing than is voice switching, especially when narrow-band center clipping is used to avoid harmonic distortion products in the output. Because of the relatively slight mutilation of the near-end signal by the center clipping, the center clippers do not have to be removed during double talking. Thus no separate double talking detector has to be used. Echo suppression is also quite effective during double talking and will be discussed in more detail in Section V.

III. SIMULATION AND IMPLEMENTATION

Initially, we simulated the center-clipping echo suppressor on a CDC 3300—EAI 8800 hybrid computer. Double talking was simulated with return losses of 15 and 30 dB and the output of the center-clipping process was recorded for each condition. No echo was heard in either case. For 15 dB return loss, a small amount of degradation of the near-end speech was noticeable after processing. For 30 dB return loss, negligible degradation of the near-end speech resulted from the center clipping.

In order to study the center-clipping process under actual conditions of double talking, we needed a real-time processing system. The required center clippers and control circuits for the clipping levels were designed and built using analogue components. However, the clippers used were not the minimum distortion form shown in Fig. 3, but the somewhat less efficient form of Fig. 5.[5] The peak detectors had switchable decay times of 0 ms for alignment and 10 ms for use during echo suppression. Three General Radio (GR) Model 1925 filter banks composed of 1/3-octave 6th-order Butterworth filters were used to complete the center-clipping system.

We investigated the center-clipping system as an echo suppressor in a simulated toll circuit designed for evaluation of echo suppressors. Figure 6 is a simplified diagram of one end of the circuit. This circuit connects two 4-wire telephones, with active sidetone, via a 4-wire delay path. Hybrids are simulated by echo paths in which return loss can be set from 0 to 50 dB. Selection of various echo suppressors or a 4-wire line is provided between the two echo paths and the 4-wire network. For comparison, we had available the center-clipping echo suppressor (one end of a split system), a split 3A echo suppressor with speech compression, and a 4-wire connection. All systems were lowpass-filtered at 3200 Hz. The 3A units are echo suppressors employing voice switching, currently in use in the telephone plant. We also had available about 0.6 second of tape delay, which was introduced as shown in Fig. 6, for simulation of a satellite connection.

The control circuit attenuators in the center-clipping system (Fig. 2) were adjusted manually so that echoes of far-end sinusoidal signals were completely eliminated in each band for the selected return loss. This initial adjustment resulted in no echo being heard during single talking.

IV. RESULTS

Evaluation of the performance of an echo suppressor is a difficult task because most meaningful testing has to be done during normal

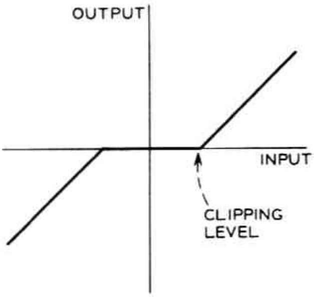

Fig. 5—Center-clipping transfer function implemented in analogue circuits.

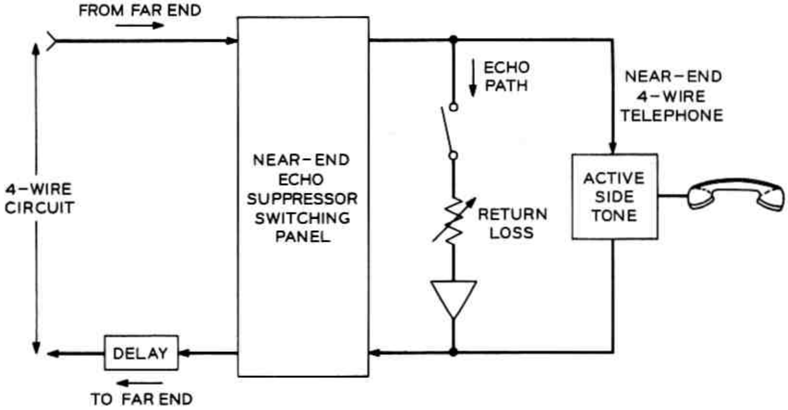

Fig. 6—One end of simulated toll circuit for testing echo suppressors.

conversations. No attempt at a quantitative evaluation of the center-clipping echo suppressor has as yet been carried out. However, in this section we present results of informal demonstrations using the simulated toll circuit.

The system initially used had six 2/3-octave bands. It performed very well in suppressing echoes in that no echo was heard by the far-end talker, even during double talking, for return losses down to 0 dB. However, even during single talking from the near end, some degradation was unexpectedly still present. This was due to a combination of phase distortion and coloration caused by passing the speech through two of the GR filter banks before recombining the bands. Each of the filter banks has a spectral ripple which is about ± 1 dB. However, the spectral ripple is several dB for two filter banks in series. In addition, phase distortion, which is not serious in one filter bank, is doubled for two filter banks and becomes objectionable. Because the phase delays correspond to those of the 1/3-octave filters combined to make 2/3-octave bands, the distortion is greater than was present in the original computer simulation.

In order to improve the speech quality in single talking, we substituted GR 1-octave filters, center frequencies 250, 500, 1000, and 2000 Hz, for the four lowest filters and used a 1/3-octave filter, center frequency 3150 Hz, at the top of the frequency band to make a 5-channel system. This system covered the same total bandwidth as the 6-channel system but had less phase distortion because of the wider filters used. Its performance is expected to be nearly identical to that of a 4-channel system since the same bandwidth could be covered by 4 filters, each only slightly wider than one octave.

The 5-channel system was as effective in suppressing echoes as the 6-channel system. As expected, the speech quality for single talking from the near end was improved but was still slightly degraded by coloration. As a result, we found that we could get better quality during single talking by removing the output filter bank. Because there is no clipping during single talking, output filters are unnecessary for this condition since no distortion products are generated. Surprisingly, however, distortion of the near-end speech during double talking was not very noticeable to the far-end talker who was simultaneously talking and listening. This was apparently due to masking.

We have demonstrated the systems to numerous people in different areas of Bell Laboratories. In these demonstrations, the speech quality of the center-clipping system was judged to be comparable to the simulated 4-wire satellite connection (or the 3A echo suppressors) for single talking conditions. When a comparison was made between the center-clipping system and the 3A echo suppressors during double talking, they differed in two respects. First, noticeable echo could be heard during double talking with the 3A echo suppressors since the 3A's offer little echo suppression in the break-in mode, while no echo was heard during double talking with the center-clipping echo suppressor. Second, the 3A's gave a chopped quality to the speech apparently independent of the return loss, as they switched between suppression and break-in, while this kind of switching sound was absent from the center-clipping system. (In the break-in mode of the 3A's during double talking, a variable amount of loss is introduced into the receive paths depending on the relative and absolute levels of the two end signals.) With the control circuits adjusted for return losses less than about 15 dB, the center-clipping system contributed some distortion to the speech during double talking which became more noticeable as the return loss was decreased to 0 dB. However, for return losses greater than 15 dB, the center-clipping system was almost indistinguishable from a 4-wire connection with no echo path.

V. DISCUSSION

The center-clipping process is a unique echo suppressor in that no decision between single talking and double talking has to be made. It is obvious how it operates under single-talking conditions. In single talking from the far end, the clipping levels are set with a rise time faster than any speech component so as just to remove the echo in each band. When the received signal ceases, the clipping levels fall to zero with a holding time greater than the end delay. For single talking from the near end, the clipping levels are zero and the speech is, in principle, unaffected.

It is not so apparent how echo is eliminated during double talking. In this case, the echo signal is added to the near-end signal and this composite signal is fed to the input filter bank. The clipping levels still follow the echo signal, and eliminate echo in bands where the two signals do not overlap and during gaps between words and sentences in the near-end speech. When energy from both signals appears in any band, clipping cannot remove the echo signal. However, it appears that, in this case, the echo is partially masked in that band. For these reasons, it is probably advantageous to have the bandwidths of the channels as small as possible compatible with other system requirements of speech quality and cost. These considerations indicate that the minimum number of channels possible may be determined by the effectiveness in echo suppression rather than by the avoidance of harmonic distortion. That is, a 3-channel system may not perform as well in echo suppression even though there is no harmonic distortion at the output. (A 3-channel system with bandwidths of individual filters just under two octaves includes no harmonic distortion products in the output since only odd-harmonic distortion products are produced by the center clippers). So far a 3-channel system has not been investigated.

In the demonstrations described, control signal levels were adjusted manually to match the trans-hybrid loss. In practice, this setting should either be permanently adjusted for worst case or adaptively controlled. If a center-clipping system is used as a back-up for an echo canceller,[6] worst-case setting will still yield almost perfect results. However, in the normal network, where 6 dB return loss is the worst case, adaptive setting, even if quite crude, would be desirable.

As mentioned in the preceding sections, several kinds of speech degradation occur in the center-clipping echo suppressor. Inherent in the process is the degradation observed in the computer simulation where coloration and phase distortion of the filters and nonlinear distortion of the center clippers were minimized. In this case, degradation resulted mainly from loss of part of the signal caused by the center-clipping process. However, considerable loss of information can be tolerated without significant decrease in subjective quality because of the redundant nature of speech. In the analogue experiments, other distortions were present in addition to this inherent one. Because of this, optimum operation was realized with the output filter bank removed even though nonlinear distortion was present during double talking.

So far, all the discussion of evaluation of center clipping echo suppression has been for the case of a 0.6-second transmission delay. For shorter delays, the subjective effect of the degradations present during double talking is greatly reduced and speech of comparable quality is obtained for smaller return losses.

VI. CONCLUDING REMARKS

We have described demonstrations of an experimental center-clipping system for electrical echo suppression. This echo-suppressor principle is unique in that no double-talking decision has to be made, Echoes appear to be completely removed, even during double talking, for return losses as small as 0 dB. Speech communication is full-duplex at all times and, for return losses greater than about 15 dB, is almost indistinguishable from a 4-wire connection. The center-clipping echo suppressor would appear to be an excellent back-up for an echo canceller if the echo cancellation plus return loss reduces the echo level to -20 dBm0 or less.

We have also made tests of this echo suppressor using a "real" end section including an N3-carrier, 4-1/2 miles of simulated loaded cable, and a real telephone and hybrid. The results were similar to those already discussed when the attenuation in each band was manually adjusted to match the return loss characteristics of the carrier system. (Return loss varied from about 6 to 18 dB.)

In another experimental application, we have used the center-clipping system as a replacement for voice switching in the suppression of acoustical echo generated in an idealized 4-wire speakerphone.

VII. ACKNOWLEDGMENTS

We are indebted to C. A. Ross for the circuit design and initial construction of the analogue center-clipping system, to J. E. Unrue, Jr., and G. K. Helder for several stimulating discussions, to B. E. Crawford for assistance with the simulated toll circuitry, to G. H. Yates for help with some of the wiring and alignment, to B. J. Karafin for his cooperation in our use of the CDC 3300 computer, and to J. S. Courtney-Pratt for valuable discussions and continued encouragement.

REFERENCES

1. Helder, G. K., "Customer Evaluation of Telephone Circuits with Delay," B.S.T.J., *45,* No. 7 (September 1966), pp. 1157–1191.
2. Mitchell, O. M. M., and Berkley, D. A., "Reduction of Long-Time Reverberation by a Center-Clipping Process," J. Acoust. Soc. Amer., *47,* No. 1, Part 1 (January 1970), p. 84 (abstract).
3. Berkley, D. A., and Mitchell, O. M. M., "A Multiband Center-Clipping Process for Reverberation Reduction and Echo Suppression," patent applied for.
4. Pierce, J. R., private communication.
5. Trimble, D. C., "An Analytical Comparison of Two Types of Center Clipping," unpublished work.
6. Sondhi, M. M., "An Adaptive Echo Canceller," B.S.T.J., *46,* No. 3 (March 1967), pp. 497–511.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Low-Loss Modes in Dielectric Lined Waveguide

By J. W. CARLIN and P. D'AGOSTINO

(Manuscript received December 15, 1970)

Recent studies of the heat loss characteristics of the normal modes in dielectric lined circular waveguide have shown that modes other than those of the circular electric type may have low loss over wide frequency bands. This unexpected behavior of the mode loss characteristics is explained by utilizing the well-known duality relationships between the electric and magnetic fields. Specifically, it is shown that the lowest loss modes are alternately circular electric and circular magnetic as frequency (or lining thickness) increases, with low loss occurring at frequencies and lining thickness where the wall impedance of the dielectric coated guide approximates a short circuit (or electric wall) for circular electric modes, and an open circuit (or magnetic wall) for circular magnetic modes.

These findings will influence and aid in the selection and design of an appropriate waveguide(s) (employing the circular electric TE_{01} mode) for the WTS millimeter wave transmission system which is presently under development; they may also influence the design of future guided wave systems.

I. INTRODUCTION

The possibility of using a circular electric (TE_{01}) mode in circular waveguide has been of considerable interest since the initial discovery of its desirable low-loss characteristics (in the following, we are concerned only with the heat loss in the guide). One type of guide currently under consideration (Fig. 1) consists of a highly conducting outer wall to which a thin dielectric liner is bonded to break the degeneracy in phase velocities for the TE_{01} and TM_{11} modes in metallic circular waveguide.

The circular electric mode loss characteristics of thinly lined circular waveguide have been determined by H. G. Unger.[1-3] This work has

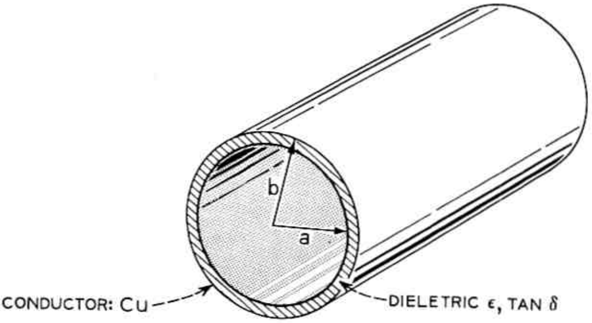

Fig. 1—Dielectric lined circular waveguide.

since been extended[4] and it was found that other modes also have very low heat loss over an appreciable frequency range. A full scale discussion of the analysis is beyond the scope of this paper but will be available in a forthcoming paper.[4] In the following, the analytical approach is indicated and some typical results for the heat loss (copper and dielectric loss) of the TE_{01}, TE_{02}, and TM_{02} modes in lined circular guide are given.

II. DISCUSSION

The problem of obtaining the normal mode loss for a perfectly straight circular waveguide with a uniform lining, as shown in Fig. 1, was approached in two ways. In one approach, the well-known induced current method was used. The field and wall currents in the guide were found for a lossless structure; they were then used to determine the heat losses of the waveguide. In the second approach, the impedance at the dielectric-free-space interface was prescribed as a boundary condition for the solution of the wave equation. This impedance was established by using a transmission line model to transform the surface impedance of the copper conducting wall through the dielectric lining, which was assumed to have a small but finite loss tangent. The complex eigenvalue equation was then solved and the overall losses thus determined.

The results of these two methods are in good agreement. The wall impedance approach aids in understanding the physical phenomena occurring in dielectric lined guide and is used in the following paragraphs to explain the loss characteristics of circular electric and magnetic modes in dielectric lined guide.

In Fig. 2, we have plotted the total heat loss for some circular electric and magnetic modes in lined waveguide with a 1-percent lining ($\rho = 1.01$). Here ρ is defined as the ratio of the waveguide conductor radius to dielectric radius. The losses of the circular electric modes initially decrease with increasing frequency, as in unlined waveguide, and reach minimum values at approximately 140 GHz (at this point, the dielectric loss accounts for 8 percent of the total heat loss). The TE_{0n} losses then increase rapidly with a further increase in frequency. We observe an interesting phenomenon at approximately 240 GHz. At this frequency the lining is approximately a quarter wavelength thick ($\lambda_{rad}/4$) in the radial direction. The equivalent wavelength (λ_{rad})* in the radial direction for the dielectric region is related to the free-space wavelength (λ) by

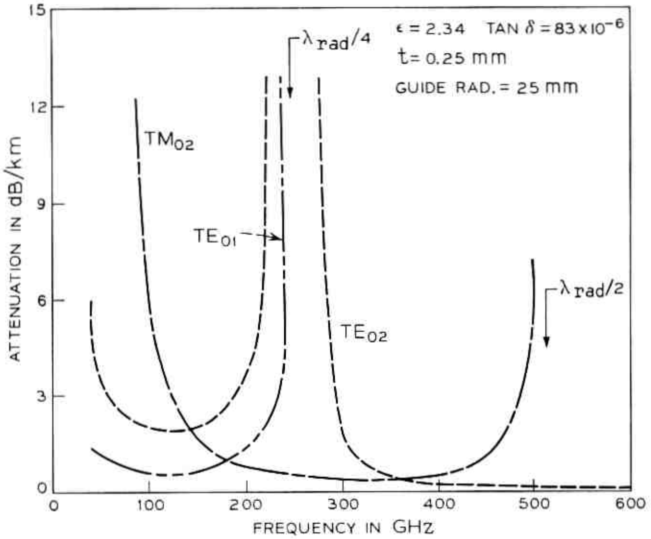

Fig. 2—Loss characteristics of circular symmetric modes in lined waveguide (wall impedance model).

$$\lambda_{rad} = \lambda / \sqrt{\epsilon - 1}. \qquad (1)$$

As the frequency increases beyond 240 GHz, the TE_{01} loss increases indefinitely. This is attributable to a surface-wave phenomenon (the

* In highly overmoded guide, it can be shown that the radial propagation constant in the dielectric region is $(2\pi/\lambda)\sqrt{\epsilon - 1}$ by considering plane-wave reflection by a grounded slab at grazing incidence.

field is bound to the dielectric region), and the fact that it occurs only when the lining thickness is greater than a quarter wavelength is in agreement with the minimum thickness ($n = 1$) required to propagate a TE surface wave on a grounded slab:[5]

$$t_{\min} = \frac{n\lambda}{4\sqrt{\epsilon - 1}} \qquad n = 1, 3, 5, \cdots . \tag{2}$$

The TE_{02} loss (20 percent of which is due to dielectric losses in the lining at this frequency), conversely, decreases as the frequency becomes greater than 240 GHz and reaches a minimum when the lining is a half wavelength thick ($\lambda_{\mathrm{rad}}/2$), as shown in Fig. 2. We would find the TE_{02} loss increasing rapidly with a further increase in frequency as it eventually propagates as a surface wave when the lining is three-quarters of a wavelength [corresponding to $n = 3$ in equation (2)] thick.

The most interesting feature of Fig. 2 is the steadily decreasing loss of the TM_{02} mode as the frequency increases. At the upper end of the proposed WTS frequency band (110 GHz), the TM_{02} loss (10 percent is due to dielectric loss) is a surprisingly low 4.7 dB/km for this example. The TM_{02} mode also has lower loss than *any* circular electric mode over the frequency range of 180–350 GHz. At 240 GHz, the TM_{02} loss is 0.54 dB/km and dielectric losses account for 13 percent of this.

From Fig. 2 we observe that the TM_{0n} loss characteristics for a lining of thickness t are similar to those of a TE_{0n} mode in a guide with a lining thickness a quarter wavelength greater. The TM_{0n} modes have loss minima for linings $\lambda_{\mathrm{rad}}/4$, $3\,\lambda_{\mathrm{rad}}/4$, $\cdots$ thick while the TE_{0n} modes have loss minima for 0, $\lambda_{\mathrm{rad}}/2$, $\cdots$ thick linings. The minimum dielectric thickness required for a TM_{0n} mode to propagate as a surface wave is[5]

$$t_{\min} = \frac{n\lambda}{4\sqrt{\epsilon - 1}} \qquad n = 0, 2, 4, \cdots , \tag{3}$$

which differs from the minimum thickness for a TE_{0n} surface-wave mode in (2) by a quarter wavelength. From (3) we see that the TM_{01} mode propagates as a surface wave for very thin linings ($n = 0$); the TM_{01} loss curve was not shown in Fig. 2, since the loss steadily increased with increasing frequency from an initial value of 47 dB/km at 40 GHz.

The preceding loss characteristics can be explained by a simple physical argument and the use of duality. Let us consider Fig. 3. Here

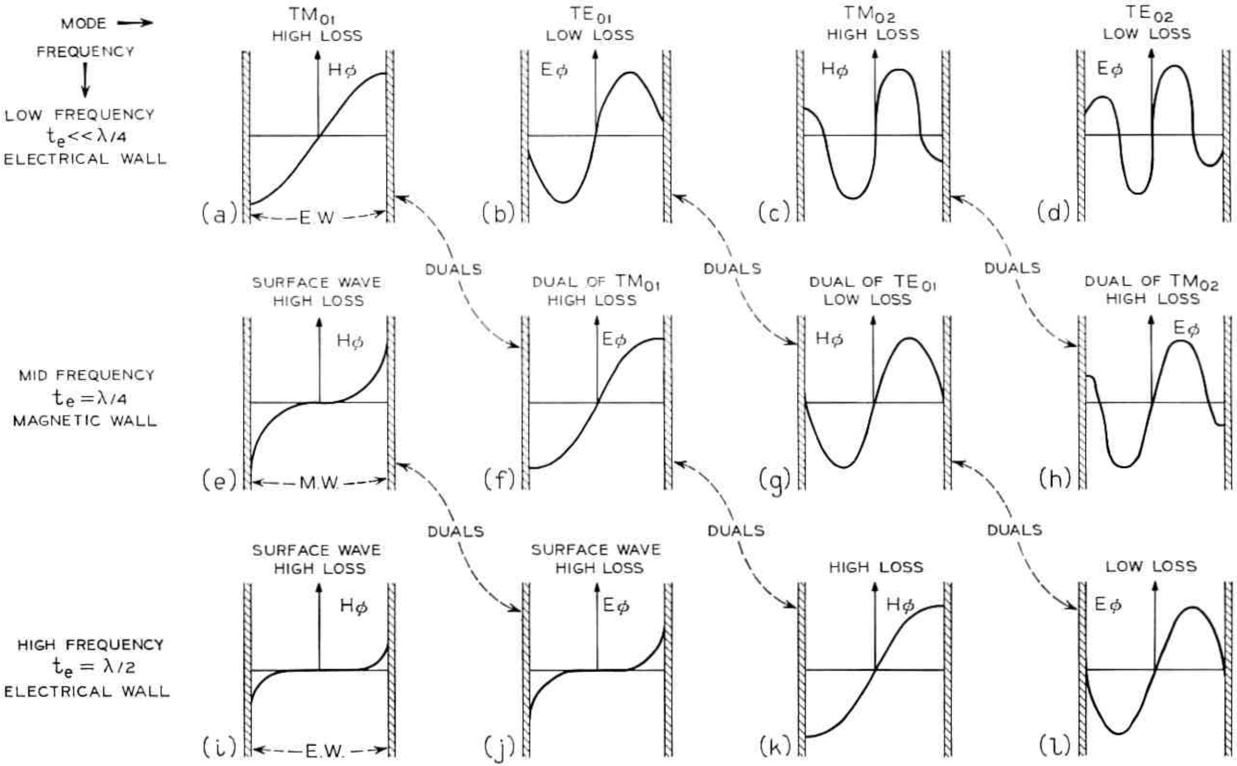

Fig. 3—Field distributions of the circular symmetric modes in unlined and lined waveguide.

we have sketched the field distributions of the circular electric and circular magnetic modes for no lining, a quarter-wave lining, and a half-wave lining. For the TM_{01} and TM_{02} modes with no lining (Fig. 3a and c), we have a strong normal electric field at the wall of the guide with a high induced electric charge density and hence high loss. For the TE_{01} and TE_{02} modes (Fig. 3b and d) the losses are quite low.

For a quarter-wave lining, the impedance at the dielectric inner face is approximately an open circuit or magnetic wall. The TM_{01} mode is trapped in the lining and decays in an exponential fashion towards the center of the guide (Fig. 3e). The field configuration for the TE_{01} in the air region (Fig. 3f) is the dual of that for the TM_{01} mode in Fig. 3a. Hence it will have high losses. The field configuration for the TM_{02} mode (Fig. 3g) is now the dual of the TE_{01} mode in Fig. 3b and hence it is a low-loss mode. On a further increase in the lining thickness to $\lambda_{rad}/2$, we find TE_{01} and TM_{01} both propagate as a surface wave bound to the lining (Fig. 3i and j), while TM_{02} has high loss (Fig. 3k). TE_{02} (Fig. 3l) and TE_{03} (not shown in Fig. 3), conversely, have low losses for this lining thickness.

In Fig. 4, we have plotted the loss for several TE_{0n} and TM_{0n} modes at 100 GHz for lining thickness up to 5 percent. We find the results

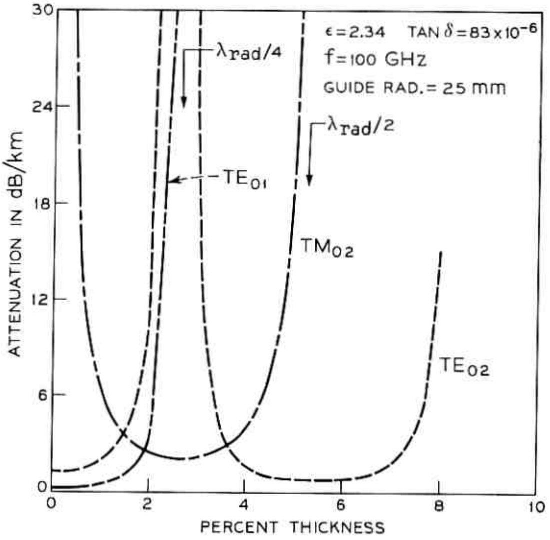

Fig. 4—Loss characteristics of circular symmetric modes at 100 GHz for waveguide lining thicknesses up to 5 percent (wall impedance model of lined waveguide).

for a change in lining thickness are similar to those for a change in frequency. The TE_{0n} and TM_{0n} modes go through relative loss maxima and minima for a $\lambda_{rad}/4$ change in lining thickness. An additional mode also becomes trapped and propagates as a surface wave for every $\lambda_{rad}/4$ change in thickness.

In Fig. 5, some representative solutions of the appropriate eigenvalue equations for TE_{0n} and TM_{0n} modes in dielectric lined guide are given. The eigenvalue (k_n) is defined as

$$k_n = \chi_n a$$

where a is the dielectric radius and χ_n the radial propagation constant in the air region. The eigenvalues for liners which are integral multiples of a quarter wavelength thick in the radial direction are approximately the same as the eigenvalues for circular symmetric modes in empty guide with an electric or magnetic wall. The eigenvalue also tends to zero as the thickness increases and eventually becomes pure imaginary which is indicative of a surface-wave phenomenon with very large heat losses.

III. CONCLUSION

In the preceding sections we have seen that modes other than those of the circular electric type have low loss in dielectric lined guide. In order to transmit a circular electric (not necessarily TE_{01}) mode with low loss, the lining must be significantly less than a quarter wavelength or approximately an integral multiple of a half wavelength thick. On the other hand, it is possible to use a dielectric lined guide with a quarter-wavelength-thick liner (also $3\lambda/4$, $5\lambda/4$, etc.) as a low-loss circular magnetic mode transmission medium. The tolerances on such a system for mode conversion loss would be similar to those on the appropriate dual circular electric guide.

The results also indicate there is a range of thicknesses or frequencies for which both TM_{0n} and TE_{0n} modes have low loss. Since the local character of the fields for any mode near the wall of a metallic waveguide must be similar to that of either a TE_{0n} or TM_{0n} mode, there will be a range of thicknesses and frequencies for which many quasi TM and TE modes have low loss (on the order of 6 dB/km or less) in dielectric lined guide. (This has been confirmed by recent results.[4]) This implies that the ohmic losses in route bends will be reduced. Further, since the dielectric liner not only reduces the heat loss for the spurious quasi TE and TM modes generated by a route bend or other guide deformation but also alters their field distributions relative

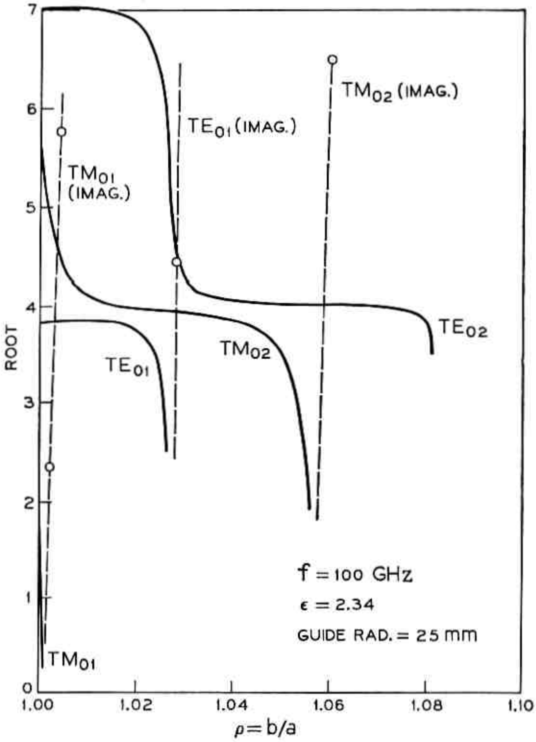

Fig. 5—Eigenvalues for circular symmetric modes in lined waveguide.

to those of unlined waveguide, it will be necessary to design mode filters with this in mind.

IV. ACKNOWLEDGMENT

The authors wish to thank W. DeLang for assistance with the computer programming.

REFERENCES

1. Unger, H.-G., "Circular Electric Wave Transmission in a Dielectric-Coated Waveguide," B.S.T.J., *36,* No. 5 (September 1957), pp. 1253–1278.
2. Unger, H.-G., "Lined Waveguide," B.S.T.J., *41,* No. 2 (March 1962), pp. 745–768.
3. Kreipe, H. L., and Unger, H.-G., "Imperfections in Lined Waveguide," B.S.T.J., *41,* No. 5. (September 1962), pp. 1589–1619.
4. Carlin, J. W., and D'Agostino, P., "Normal Modes in Dielectric Lined Waveguide," unpublished work.
5. Collin, R. E., *Field Theory of Guided Waves,* New York: McGraw-Hill, 1960, p. 473.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

A Relation for the Loss Characteristics of Circular Electric and Magnetic Modes in Dielectric Lined Waveguide

By J. W. CARLIN

(Manuscript received December 28, 1970)

Recent studies have shown that modes not of the circular electric type have low-loss characteristics in dielectric lined circular waveguide. It was determined that circular electric modes are low-loss for linings approximately 0, $\lambda/2$, λ, $\cdots$ wavelengths thick while circular magnetic modes are low-loss for $\lambda/4$, $3\lambda/4$, $\cdots$ thick linings. In this paper we derive a simple relationship between the loss characteristics of circular electric waves with 0, $\lambda/2$, $\cdots$ thick linings and circular magnetic waves with $\lambda/4$, $3\lambda/4$, $\cdots$ thick linings.

Specifically, we show that the minimum obtainable circular magnetic mode loss is at least four times greater than the minimum obtainable circular electric mode loss. We also show that the minimum loss for successively higher order circular electric (magnetic) modes corresponding to approximately 0, $\lambda/2$, $\cdots$ ($\lambda/4$, $3\lambda/4$, $\cdots$) thick linings is approximately the same if we neglect the dielectric losses.

I. INTRODUCTION

Recent studies[1] indicate many modes not of the circular electric type may have very low loss in dielectric lined circular waveguide. In these studies the duality principle was used to explain the low-loss characteristics of circular magnetic modes for linings having thicknesses which are an odd multiple of a quarter wavelength.

In this paper we extend this use of the duality principle and derive a simple relation between the loss characteristics of circular electric and magnetic modes in dielectric lined circular waveguide. In all cases, we find the minimum heat loss obtainable as the dielectric thickness varies is greater for circular magnetic modes in comparison with the minimum circular electric mode heat loss.

II. DISCUSSION

The waveguide under consideration is shown in Fig. 1. It consists of a highly conducting outer wall to which a thin layer of dielectric (of relative permittivity ϵ) is bonded. The liner is usually "electrically" thin and its sole function is to break the phase velocity degeneracy between the TE_{01} and TM_{11} modes in hollow metal-walled waveguide. In this paper we are concerned with the effect of thicker linings (linings which are an integral multiple of a quarter wavelength thick) on the conducting wall losses of the waveguide. We will assume that the dielectric is lossless in this study.

The metal walls of the waveguide in Fig. 1a may be modeled[2] as a low-impedance termination ($Z_t \ll \eta$) for the fields interior to the walls. Here $\eta = \sqrt{\mu_0/\epsilon_0}$ is the characteristic impedance of the interior filler (free space) in the guide. The dielectric liner is equivalent to a short section of transmission line. This transmission line has a characteristic impedance*

$$Z_{0\phi} = \eta/\sqrt{\epsilon - 1} \tag{1}$$

for the electric field polarized parallel to the wall and a characteristic impedance

$$Z_{0z} = \frac{\eta\sqrt{\epsilon - 1}}{\epsilon} \tag{2}$$

for the magnetic field polarized parallel to the wall. The appropriate transmission line propagation constant is

$$\beta_e = k_0\sqrt{\epsilon - 1}, \tag{3}$$

where k_0 is the free-space propagation constant. The equivalent impedance conditions (Z_ϕ and Z_z) at the inner face of the dielectric in Fig. 1b may be obtained from the transmission line parameters in (1), (2), and (3) in the usual manner. The wall impedance guide in Fig. 1b is equivalent to the lined guide in Fig. 1a and may be used to predict its electrical properties with a small error.[3]

R. E. Collin[4] has shown that if the fields $\bar{E}_1$, $\bar{H}_1$ are solutions of Maxwell's equations in a source-free region of free space, the dual fields $\eta\bar{H}_2$, $-(1/\eta)\bar{E}_2$ are also a solution. The same transformation is applicable to the wall impedance guide in Fig. 1b but we must also transform the wall impedances. The appropriate dual is shown in

* The expressions in (1), (2), and (3) were derived from the plane-wave scattering at grazing incidence by a grounded slab.

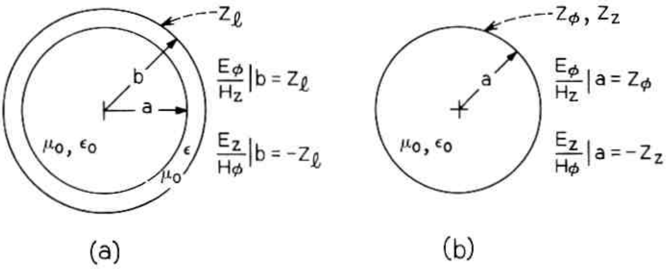

Fig. 1—(a) Dielectric lined waveguide. (b) Equivalent wall impedance waveguide.

Fig. 2. The fields are transformed as in Collin and the impedance wall is replaced by an equivalent admittance wall. We may summarize the duality principle for a guide of radius a, as in Fig. 2, as follows:

$$\bar{E}_1 \leftrightarrow \eta \bar{H}_2 .$$

$$\bar{H}_1 \leftrightarrow -\frac{1}{\eta} \bar{E}_2 .$$

$$\left.\frac{E_{1\phi}}{H_{1z}}\right| a = Z_{1\phi} \leftrightarrow \eta^2 Y_{2\phi} , \qquad -\left.\frac{H_{2\phi}}{E_{2z}}\right| a = Y_{2\phi} . \tag{4}$$

$$-\left.\frac{E_{1z}}{H_{1\phi}}\right| a = Z_{1z} \leftrightarrow \eta^2 Y_{2z} , \qquad \left.\frac{H_{2z}}{E_{2\phi}}\right| a = Y_{2z} .$$

From (4) we see that the dual of a circular electric mode in a guide with a low-impedance wall ($Z_{1\phi} \ll \eta$) is a circular magnetic mode in a guide with a high-impedance or low-admittance ($Y_{2\phi} = (1/\eta^2)Z_{1\phi} \ll 1/\eta$) wall. Since the $\bar{E} \times \bar{H}$ product is invariant under the above transformation, the loss characteristics of the two duals are identical.

Using the duality relations in (4), we can now obtain a simple relation

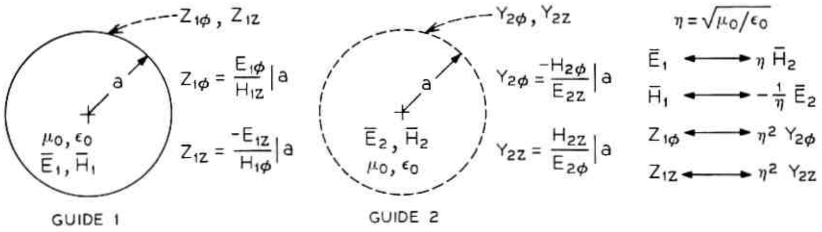

Fig. 2—Dual waveguides.

between the losses of circular electric and magnetic modes in dielectric lined guide as shown in Fig. 3. We first consider the loss for a TE_{01} mode in Fig. 3a. Since there is no lining at $r = a$, we have $Z_{1\phi} = Z_l$ and the loss $\alpha_0^{TE_{01}}$ for a TE_{01} mode in hollow guide is thus[5]

$$\alpha_0^{TE_{01}} = C \operatorname{Re}(Z_l). \tag{5}$$

We now consider the loss for a TM_{02} mode in Fig. 3b. The input admittance $Y_{2\phi}$ at $r = \mathrm{a}$ is easily determined for a quarter-wave lining from the transmission line impedance in (2) as

$$Y_{2\phi} = \frac{H_{2\phi}}{E_{2z}} = \frac{\epsilon^2 Z_l}{\eta^2(\epsilon - 1)}. \tag{6}$$

But this circular magnetic mode will have the same loss [see equation (4)] as a circular electric mode in a guide with wall impedance

$$Z_{1\phi} = \frac{\epsilon^2}{(\epsilon - 1)} Z_l .$$

We see that the equivalent wall impedance for the circular electric mode dual is higher than the wall impedance of the unlined waveguide. Hence, the loss for the TM_{02} mode in the guide shown in Fig. 3b is greater than that for the TE_{01} mode in Fig. 3a and is given by

$$\alpha_{\lambda/4}^{TM_{02}} = C \frac{\epsilon^2}{\epsilon - 1} \operatorname{Re} Z_l . \tag{7}$$

If we now consider the TE_{02} mode in Fig. 3c we see it has the same fields for $r < a$ as the TE_{01} mode in Fig. 3a and, hence, its loss will be the same. We, thus, have

$$\alpha_{\lambda/2}^{TE_{02}} = C \operatorname{Re}(Z_l). \tag{8}$$

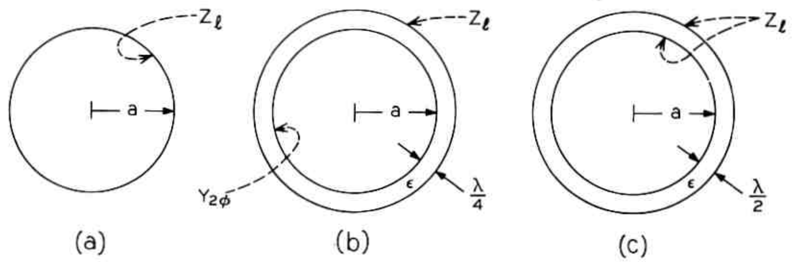

Fig. 3—Low-loss dielectric waveguides: (a) TE_{01} mode; (b) TM_{02} mode; (c) TE_{02} mode.

In the above analysis we have neglected the power carried in the dielectric regions of the guide. This is a reasonable approximation if the lining thickness is much less than the guide radius. We have also neglected the dielectric losses. Recent computer-generated results[1] indicate that this is a reasonable assumption since a quarter-wave polyethyline liner ($\epsilon = 2.34$, $\tan \delta = 83 \times 10^{-6}$) in 50mm-diameter circular guide at 100 GHz has a metal-wall loss for Cu walls of 1.79 dB/km, while the dielectric loss is 0.43 dB/km for the TM_{02} mode.

The present study shows that the minimum circular magnetic mode heat loss $\alpha_{\min}^{TM_0}$ and the minimum circular electric mode heat loss $\alpha_{\min}^{TE_0}$ are related by

$$\frac{\alpha_{\min}^{TM_0}}{\alpha_{\min}^{TE0}} = \epsilon^2/\epsilon - 1 \tag{9}$$

in dielectric lined guide. This ratio is a minimum at $\epsilon = 2$ and has a value of 4. Relation (9) was found to agree well with computer-generated results for the modal[1] heat loss in dielectric lined guide.

III. CONCLUSION

We have seen that circular magnetic modes have a higher minimum heat loss than circular electric modes in dielectric lined waveguide. The ratio of the two losses was shown to be a simple function of the linings relative permittivity. The ratio was shown to have a minimum value of 4 for a lossless dielectric of permittivity 2.

The results indicate a TM_0 mode has possibilities as a long-haul carrier, in dielectric lined guide. The heat loss, however, is always at least four times greater than that for a comparable TE_0 mode. The mode conversion loss for the two systems also can be shown to be comparable by use of the duality principle.[6]

REFERENCES

1. Carlin, J. W., and D'Agostino, P., "Low-Loss Modes in Dielectric Lined Waveguide," B.S.T.J., this issue, pp. 1631–1638.
2. Ramo, S., and Whinnery, J. R., *Fields and Waves in Modern Radio,* New York: John Wiley and Sons, 1960, p. 368.
3. Unger, H.-G., "Lined Waveguide," B.S.T.J., *41,* No. 2 (March 1962), pp. 745–768.
4. Collin, R. E., *Field Theory of Guided Waves,* New York: McGraw-Hill, 1960, p. 29.
5. Karbowiak, A. E., *Trunk Waveguide Communication,* London: Chapman and Hall, 1965, p. 54.
6. D'Agostino, P., "TM_{om} Carrier System in Dielectric Lined Circular Waveguide," unpublished work.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Timing Recovery in PAM Systems

By R. D. GITLIN and J. SALZ

(Manuscript received January 7, 1971)

It is shown how various timing recovery schemes are reasonable approximations of the maximum likelihood strategy for estimating an unknown timing parameter in additive white gaussian noise. These schemes derive an appropriate error signal from the received data which is then used in a closed-loop system to change the timing phase of a voltage-controlled oscillator. The technique of stochastic approximation is utilized to cast the synchronization problem as a regression problem and to develop an estimation algorithm which rapidly converges to the desired sampling time. This estimate does not depend upon knowledge of the system impulse response, is independent of the noise distribution, is computed in real time, and can be synthesized as a feedback structure. As is characteristic of stochastic approximation algorithms, the current estimate is the sum of the previous estimate and a time-varying weighted approximation of the estimation error. The error is approximated by sampling the derivative of the received signal, and the mean-square error of the resulting estimate is minimized by optimizing the choice of the gain sequence.

If the receiver is provided with an ideal reference (or if the data error rate is small) it is shown that both the bias and the jitter (mean-square error) of the estimator approach zero as the number of iterations becomes large. The rate of convergence of the algorithm is derived and examples are provided which indicate that reliable synchronization information can be quickly acquired.

I. INTRODUCTION

The problem of symbol synchronization in digital data transmission in the presence of intersymbol interference is extremely complicated. The best sampling instants are channel dependent and are in general difficult to determine. Consequently, the problem of timing recovery in high-speed data transmission is intimately tied in with adaptive equalization. Since general methods for simultaneous optimum determination of the receiver parameters are not known, these parameters are independently determined.

Timing information is usually obtained directly from the data wave in a variety of ways.[1–3] Our objectives in this paper are:

(*i*) To indicate the optimum method (maximum likelihood) for estimating an unknown timing parameter from random data for a certain class of PAM data transmission systems;

(*ii*) To show that a variety of timing recovery methods currently in use are reasonable approximations of the optimum method, and to note that the generation of an error signal from the received signal is a feature common to these methods;

(*iii*) To demonstrate that timing recovery dynamics can often be studied and controlled through the application of stochastic approximation theory.[4–6]

Identifying the desired timing parameter as the solution of a regression equation will allow us to apply stochastic approximation theory to the symbol synchronization problem. For purposes of illustration we analyze a stochastic approximation timing recovery procedure for square-wave modulation. For this example we derive asymptotic formulas for the probability of error as a function of signal-to-noise ratio and the number of iterations used in the timing recovery loop. Since the number of iterations is directly proportional to the number of signaling intervals, insight is provided into the setup time required to achieve reliable symbol synchronization.

We finally focus on the more difficult problem of timing recovery in bandlimited PAM systems. Here timing information must be obtained in the presence of intersymbol interference as well as additive noise. A stochastic approximation algorithm is presented which derives symbol synchronization (i.e., estimates the desired sampling time) from the received data in a quick and accurate manner. The estimation algorithm developed does not require explicit knowledge of the system impulse response or the noise distribution. If the impulse response of the channel satisfies certain conditions, then the algorithm will converge in mean-square provided the gain sequence is properly chosen. Symbol synchronization is obtained by adjusting the sampling time in the following manner: at the end of each symbol interval the current estimate is taken to be the sum of the previous estimate and a weighted approximation to the actual estimation error. The desired sampling time is assumed to be that instant when the system impulse response is a maximum. For this sampling time it is shown that a reasonable approximation to the estimation error is the sampled derivative of the received signal.[†] When the error is small, its evolution can be described by a first-order random difference equation. At every iteration the mean-square error (mse) can be minimized by optimizing the choice of the (time-varying) weighting sequence. The optimum weighting sequence is of the form $1/(\alpha + \beta n)$, where α and β are quantities which depend on the system impulse response and noise power, and n is the discrete time index. Since α and β are generally unknown at the receiver they may either be estimated (giving rise to an adaptive synchronization algorithm) or picked arbitrarily. In an effort to overcome the lack of knowledge of α and β (in addition to simplifying the algorithm) it is tempting to use the asymptotic form of the gain c/n, where c is a constant. However, if $\beta \ll \alpha$ then the optimum gain is essentially a constant $(1/\alpha)$ for many iterations, and for a wide range of c the estimate obtained using c/n is shown to be unreliable. Hence it appears that in order to obtain satisfactory performance some adaptivity to determine α and β should be used in any realization of the algorithm.

Under the assumptions that the receiver error rate is small (so that an ideal reference can be assumed) and that the "eye" of the differentiated impulse response is open, the optimum mse is asymptotically of the form $1/\rho n$, where ρ is a "signal-to-noise" ratio. The "signal" term is the value of the slope of the differentiated impulse response near the origin, the "noise" term is the sum of the actual noise variance and two intersymbol interference type terms. Thus the mse can be driven to zero and an example is given to illustrate how an accurate estimate can be obtained in a few signaling intervals. We show that for a sin x/x impulse response, ten iterations will drive the mean-square error to less than 0.01 of a signaling interval.

In Section II we determine the maximum likelihood estimate of an unknown timing parameter for a baseband PAM data signal which has been contaminated by white gaussian noise. Several approximations to the optimum estimator are described in Section III. The theory of stochastic approximation is introduced in Section IV, and is used both to cast the synchronization problem as a regression problem,

[†] B. R. Saltzberg[7] has suggested a technique for timing recovery which uses this approximation. His investigation is restricted to algorithms which can be realized using time-invariant devices. The algorithm we develop exploits the advantages of using time-varying elements.

and to analyze and control the dynamics of timing recovery. In Section V we discuss a timing recovery algorithm for bandlimited PAM.

II. THE MAXIMUM LIKELIHOOD ESTIMATOR OF AN UNKNOWN TIMING PARAMETER

Consider the L level data wave in additive white gaussian noise $\nu(t)$ of double-sided spectral density N_0,

$$V(t) = \sum_n a_n h(t - nT - \tau^*) + \nu(t), \tag{1}$$

where $\{a_n\}$ are the data symbols taking on values $\pm d$, $\pm 3d$, $\cdots$ $\pm(L - 1)d$ with equal probability, $h(t)$ is a bandlimited pulse whose peak value occurs at τ^*, and $-T/2 \leqq \tau^* \leqq T/2$ is an unknown timing parameter.[†]

Detection of the data symbols $\{a_n\}$ is usually accomplished by first suitably filtering $V(t)$ and then sampling the output at time instants $\tau + kT$, $k = \pm 1, \pm 2, \cdots$. The resulting error rate is a function of τ in addition to other parameters. An ideal timing recovery system would supply the detector with τ which minimizes the probability of error. While this problem is conceptually straightforward, it is not analytically tractable and the structure of such an optimum timing recovery system is not generally yet known. We therefore must resort to a less utopian criterion.

Much simpler evaluation functions often used in data transmission[8] are

$$D_j(\tau - \tau^*) = \frac{1}{|\, h(\tau - \tau^*)\,|^j} \sum_{\substack{k \\ k \neq 0}} |\, h(\tau - \tau^* - kT)\,|^j \qquad j = 1 \quad \text{or} \quad 2. \tag{2}$$

Even for these relatively simple evaluation functions it is generally difficult to find the optimum τ. R. W. Chang[9] derives timing recovery procedures based on minimizing a particular version of equation (2). However, for a certain class of linear distortions, namely the type that gives rise to symmetrical pulse shapes, the best τ, which minimizes (2), is equal to the unknown parameter τ^*. For this class of channels the problem of optimal timing recovery procedures can be cast in the language of statistical estimation theory. This is the situation treated in this section.

[†] We assume throughout that τ^* is independent of time.

The statistical problem we pose is this: determine an estimation procedure for the parameter τ based on observations made on the received signal $V(t$ [equation (1)]. The more detailed question we wish to answer is the following. How should the observed signal, say for T_s seconds, be processed such that a "good" estimate of τ^* is obtained? The answer of course depends on what one means by good. A reasonable measure of goodness is to require that the estimate maximize the likelihood function of the unknown parameter. For binary transmission this is a classical problem for which a solution is known. (See for example Ref. 3, 10, and 11.)† The extension to multilevel signaling is straightforward and we now briefly sketch the derivation. The likelihood function of the received signal is proportional to (superfluous constants are omitted)

$$L[V] \sim E\left\{\exp - \frac{1}{2N_0} \int_0^{T_s} [V(t) - s(t;\tau)]^2 \, dt\right\}_a, \tag{3}$$

where $s(t;\tau) = \sum a_n h(t - nT - \tau)$ and $E\{\cdot\}_a$ denotes expectation with respect to the data symbols. The expectation indicated in (2) can be carried out provided the reasonable assumption is made that the power in the data signal $s(t;\tau)$ when measured over an interval $[0, T_s]$ (large compared with a symbol duration) is independent of the data sequence and the unknown parameter τ. This assumption leads to a simplified version of (3)

$$L[V] \sim E\left\{\exp\left\{\frac{1}{N_0}\int_0^{T_s} V(t)s(t;\tau)\,dt\right\}\right\}_a \tag{4}$$

$$L(V) \sim \prod_n \left\{\frac{2}{L} \sum_{\substack{k=1 \\ k \text{ odd}}}^{L-1} \cosh\left(\frac{kd}{N_0} z_n(\tau)\right)\right\}, \tag{5}$$

where

$$z_n(\tau) = \int_0^{T_s} V(t)h(t - nT - \tau)\,dt \tag{6}$$

is recognized as the sampled (at times $nT + \tau$) output of a filter matched to $h(t)$, whose input is $V(t)$.

The maximum likelihood estimate (MLE) is obtained by differentiating $L[V]$ with respect to τ and setting the resulting expression to zero. An equivalent strategy may be obtained by differentiating any monotonic function of L and a convenient such function in this appli-

† None of the references cited claims originality. It is difficult to determine where the result was written down first.

cation is the logarithmic function. From equation (5)

$$\Lambda[V] = \ln L[V] \sim \sum_n \left\{ \ln \left[\sum_{\substack{k=1\\ k \text{ odd}}}^{L-1} \cosh\left(\frac{kd}{N_0} z_n(\tau)\right)\right]\right\}, \tag{7a}$$

and upon differentiation we obtain

$$\frac{\partial \Lambda}{\partial \tau} = \sum_n \left\{ \frac{\displaystyle\sum_{k=1}^{L/2} (2k-1) \sinh\left(\frac{(2k-1)\,d}{N_0} z_n(\tau)\right)}{\displaystyle\sum_{k=1}^{L/2} \cosh\left(\frac{(2k-1)\,d}{N_0} z_n(\tau)\right)} \right\} \frac{d}{N_0} \frac{dz_n(\tau)}{d\tau}, \tag{7b}$$

where the bracketed term can be shown to be[†]

$$\frac{(L-1)\sinh\left(\frac{(L+1)\,d}{N_0} z_n(\tau)\right) - (L+1)\sinh\left(\frac{(L-1)\,d}{N_0} z_n(\tau)\right)}{\cosh\left(\frac{(L+1)\,d}{N_0} z_n(\tau)\right) - \cosh\left(\frac{(L-1)\,d}{N_0} z_n(\tau)\right)}; \tag{7c}$$

and for the typical data communication environment of a large signal-to-noise ratio the above expression becomes proportional to

$$(L-1)\tanh\left(\frac{(L+1)\,d}{N_0} z_n(\tau)\right).$$

Thus we finally have that

$$\frac{\partial \Lambda}{\partial \tau} \sim \sum_n \frac{dz_n(\tau)}{d\tau} \tanh\left(\frac{(L+1)\,d}{N_0} z_n(\tau)\right). \tag{8}$$

The optimum estimation strategy is exhibited in equation (8). The best value of τ (i.e., the MLE) makes the right-hand side of equation (8) as small as possible. The mathematical operations exhibited in equation (8) can readily be instrumented. The implementation objective would be to use the right-hand side of (8) as an error signal in a closed-loop system that iteratively adjusts τ to determine the MLE. A block diagram of this implementation is shown in Fig. 1. The received signal and its derivative are first passed through filters with identical impulse responses $h(-t)$ whose outputs are periodically sampled at times $nT + \tau$. In the undifferentiated branch, the samples are first multiplied by $(L+1)d/N_0$ and are then passed through the memoryless nonlinearity tanh $(\cdot)$ which resembles an infinite clipper for large input values. The output from the two branches are mul-

[†] Note that for $L = 2$, equation (7c) becomes $(\sinh 3y - 3\sinh y)/(\cosh 3y - \cosh y) = \tanh y$, which agrees with the bracketed term in (7b).

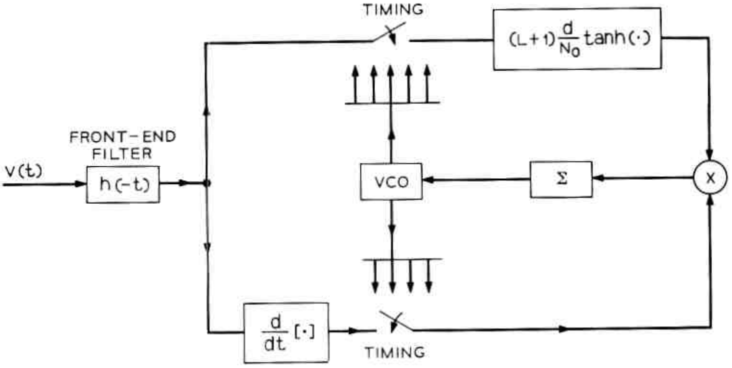

Fig. 1—Implementation of maximum likelihood strategy.

tiplied and averaged as indicated by the sum in equation (8). This then is the error signal driving a voltage-controlled oscillator which in turn determines the new timing phase.

III. IMPLEMENTATIONS APPROXIMATING THE OPTIMUM

We now examine approximations of equations (7) and (8) leading to several simplified implementations of timing recovery systems. The first approach is to approximate $\tanh(x)$ in equation (8) by the limiter function $\operatorname{sgn}(x)$. This approximation yields

$$\tanh\left(\frac{(L+1)\,d}{N_0}\,z_n(\tau)\right) \sim \operatorname{sgn} z_n(\tau) \equiv \operatorname{sgn} \hat{a}_n \,, \tag{9}$$

where $\hat{a}_n$ is the nth decision, or the estimate of the nth data symbol. The approximation (9) is a good one at large signal-to-noise ratio and in this case $\hat{a}_n$ will equal a_n most of the time. When this approximation is made, the detection circuit which computes $\hat{a}_n$ from $z_n(\tau)$ is separated from the timing circuit. In the timing branch the received signal is first passed through the filter with impulse response $h(-t)$ and the output is differentiated or equivalently passed through a high-pass filter and then sampled. These samples are multiplied by the sign of the respective decisions and summed to form an error signal. The multiplication of the respective derivative samples by the sign of the decisions is clearly necessary so as to convert all the error samples to the same polarity.[†] Figure 2 shows this simplified version of detec-

[†] This is a decision directed estimation procedure. As the timing phase is acquired, the decisions become more reliable.

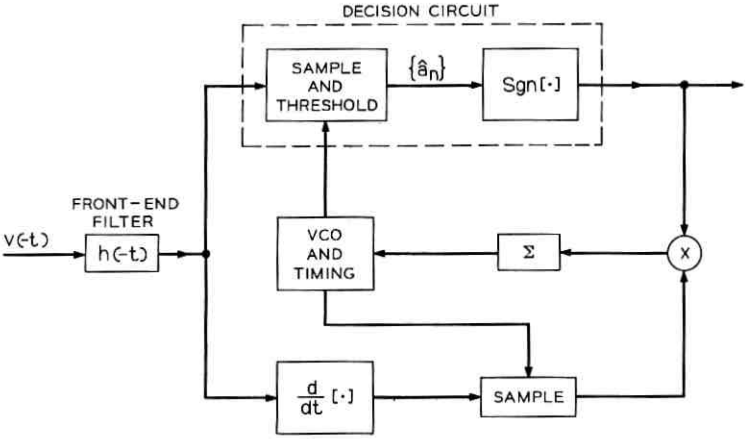

Fig. 2—An implementation approximating the ideal.

tion and timing recovery circuit. Deriving an error signal from the derivative of the received signal is very reasonable and a timing circuit based on this idea has been built and analyzed by Saltzberg.[7]

Another technique suggested from (7) is dubbed "early-late" timing recovery.[2,11] The approximations involved here are the following. First the derivative of $\Lambda[V]$ is approximated by the difference

$$\sum_k \{\ln \cosh ((kd/N_0)z_n(\tau + \Delta)) - \ln \cosh ((kd/N_0)z_n(\tau - \Delta))\}$$

$$\Delta \ll T. \qquad (10)$$

Next the nonlinear function ln [cosh (x)] is approximated by $|x|$. This again is a good approximation at large signal-to-noise ratio since for large $|x|$, cosh $x \rightarrow e^{|x|}$. This implementation is shown in Fig. 3. Here two clock pulses separated by 2Δ sample the received wave after appropriate filtering. The respective samples are then full-wave rectified and substracted from one another. The error signal is formed by adding a number of successive differences. It appears that any even Nth-law device may be used in place of the ln(cosh) nonlinearity in equation (7). Successful results for instance were obtained with a square-law device.[12]

A feature common to the above timing recovery systems is the generation of an error signal from the received signal. The sampling instant is then adjusted so as to decrease the magnitude of the error, a new error is computed, and the estimation continues in this manner.

The fewer the number of iterations needed to obtain a reliable estimate, the better the system. Stochastic approximation is a technique which will enable us to study and control the dynamic behavior of such iterative estimation algorithms by viewing the synchronization problem as a regression problem.

IV. THE APPLICATION OF STOCHASTIC APPROXIMATION TO SYMBOL SYNCHRONIZATION

4.1 *Stochastic Approximation*

We will briefly describe the salient features of stochastic approximation, in particular the Robbins-Monro algorithm. Stochastic approximation[4–6] is a technique employed to iteratively solve regression problems. The method is an extension of the Newton-Raphson technique to a random environment, and is especially useful when the regression function is unknown. More precisely, suppose z_n is a sequence of independent observations of a stationary random process and it is desired to find the value of the (non-random) parameter τ such that the regression equation,

$$E[f(z_n\ ;\ \tau)] \triangleq m(\tau) = m_o\ , \tag{11}$$

is satisfied; where E denotes expectation, $f(\cdot)$ is a given function, and $m(\cdot)$ is called the regression function. As mentioned above, $m(\cdot)$ is typically unknown, and we desire an algorithm which uses the data to sequentially estimate the value of τ, say τ^*, which satisfies (11). Robbins and Monro have shown that if (11) has a unique solution

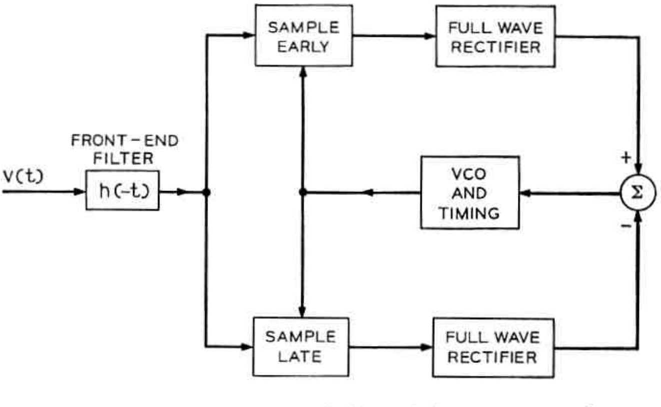

Fig. 3—Implementation of early-late timing recovery scheme.

then the estimate τ_n, given by

$$\tau_{n+1} = \tau_n + c_n[f(z_n\ ;\ \tau_n) - m_o] \qquad n = 1, 2, \cdots ,$$

will converge in mean-square and with probability one to τ^*, under some general conditions[4] on both the observations z_n and on the positive scalar time-varying weighting sequence c_n. A useful interpretation of the Robbins-Monro algorithm is that the current estimate is the sum of the previous estimate and a weighted correction term, where the average (with respect to the observations) correction is the error term $m(\tau_n) - m_o$. Thus the correction term will, on the average, give an increment in the correct direction, and the estimate will converge. Alternatively, if we regard the correction term as an approximation (in a stochastic sense) to an error term, we are reminded of the deterministic error or gradient search type of algorithms. The weighting sequence c_n is chosen to converge to zero fast enough so as to suppress the correction term as the estimate converges,† but slow enough so that large corrections are possible for many iterations (frequently c_n is of the form $1/n$).

We now cast the synchronization problem as a regression problem, and then use the theory of stochastic approximation to develop a synchronization algorithm which has desirable dynamic properties. From (8) the optimum (maximum likelihood) timing parameter is the solution of

$$\frac{\partial}{\partial \tau}[\Lambda(z_n\ ;\ \tau)] = 0.$$

If we make the identification

$$\frac{\partial}{\partial \tau}[\Lambda(z_n\ ;\ \tau)] \leftrightarrow f(z_n\ ;\ \tau), \tag{12a}$$

and now ask for the value of τ which satisfies

$$m(\tau) \equiv E\left[\frac{\partial}{\partial \tau}\Lambda(z_n\ ;\ \tau)\right] = 0, \tag{12b}$$

then the desired [i.e., the solution of (12b)] timing parameter will be the solution of a regression equation. It is important to note that the solutions of (8) and (12b) will not, in general, be the same. However the solution of (8) is a random variable, which as the observation

† Note that even when τ_n is close to τ^*, the variance of the correction term can be quite large due to the randomness of the data.

time T_s becomes large converges to τ^*; while the solution[†] of (12b) is in fact τ^*. Thus if we use a Robbins-Monro algorithm to iteratively solve (12b) we are indeed generating the maximum likelihood estimate.

4.2 *Binary Square-Wave Modulation*

Consider applying this method to analyze a timing recovery procedure when $h(t)$ in equation (1) is a rectangular pulse of T seconds duration and height A, where binary transmission is assumed for convenience. In this case, the observable function, equation (6), becomes

$$z_n(\tau) = \int_0^{T_s} V(t)h(t - nT - \tau)\,dt = \int_{nT+\tau}^{(n+1)T+\tau} V(t)\,dt. \qquad (13)$$

As mentioned earlier, we can use a square-law device to approximate the ln cosh $(\cdot)$ nonlinearity for mathematical convenience. Thus the MLE is obtained by finding a τ such that the derivative of $\sum_n z_n^2(\tau)$ is zero. From (7a) and (13) we obtain

$$\frac{d}{d\tau}\sum_n z_n^2(\tau) = 2\sum_n [V((n+1)T+\tau) - V(nT+\tau)]z_n(\tau). \qquad (14)$$

At large signal-to-noise ratio, symbol transition information is obtained from

$$d_n = V((n+1)T+\tau) - V(nT+\tau) \sim \begin{cases} 0, & a_{n+1}\cdot a_n = 1 \\ \pm 1, & a_{n+1}\cdot a_n = -1 \end{cases}. \qquad (15)$$

The Robbins-Monro procedure for recursively estimating τ can now be applied by using the regression function

$$m(\tau) = E\{d_n z_n(\tau)\}. \qquad (16)$$

For convenience we center the pulse $h(t)$ at $t = 0$ such that

$$h(t) = \begin{cases} A, & |t| \leqq T/2 \\ 0, & \text{elsewhere} \end{cases}.$$

and calculate

$$d_n z_n(\tau) = d_n A \sum_m a_m \int_{nT+\tau}^{(n+1)T+\tau} h(t - mT)\,dt + \int_{nT+\tau}^{(n+1)T+\tau} \nu(t)\,dt$$

[†] For a high signal-to-noise ratio.

$$= d_nA\left\{a_n \int_{nT+\tau}^{nT+T/2} dt + a_{n+1} \int_{nT+T/2}^{(n+1)T+\tau} dt\right\} + \nu_n$$
$$= d_nA\{a_n(T/2 - \tau) + a_{n+1}(T/2 + \tau)\} + \nu_n , \tag{17}$$

where

$$\nu_n = \int_{nT+\tau}^{(n+1)T+\tau} \nu(t)\, dt.$$

In the absence of data transitions, (17) is independent of τ while when transition occurs, i.e., $a_n \neq a_{n+1}$,

$$d_nz_n(\tau) = 2A\tau + \nu_n , \qquad -T/2 \leqq \tau \leqq T/2. \tag{18}$$

Using (13) the recursive procedure for estimating the unknown timing parameter, τ, is now as follows: Pick an arbitrary sampling phase τ_0 , $|\tau_0| \leqq T/2$, and compute the next sampling phase τ_1 from the relation (assuming that a data transition occurs)

$$\begin{aligned}\tau_1 &= \tau_0 - \tfrac{1}{2}(d_0z_0(\tau_0)) \\ &= \tau_0 - \tfrac{1}{2}(2A\tau_0 + \nu_0).\end{aligned} \tag{19}$$

The $(n + 1)$th sampling phase is then related to the nth by the recursion relation

$$\tau_{n+1} = \tau_n - \frac{1}{n+2}(d_nz_n(\tau_n)), \tag{20}$$

where we have taken c_n to be $1/n+1$. For numerical evaluation purposes it is convenient to normalize (18) and work with the regression function†

$$m(\tau_n) = E[f(x_n , \tau_n)] = E[\tau_n + x_n], \tag{21}$$

where $\{x_n\}$ is a sequence of gaussian random variables with

$$E\{x_n\} = 0 \tag{22}$$

and

$$E\{x_n^2\} = \frac{N_0T}{4A^2} = \frac{T^2}{8\rho}$$

where $\rho = A^2/2N_01/T$ is the signal-to-noise ratio in a bit-rate bandwidth.

† We are assuming that a linear theory applies, i.e., the sequence of $\{\tau_n\}$ rarely exceeds $|T/2|$. In practice no values of τ_n which exceeds $|T/2|$ will be accepted. Including these restrictions in the mathematical model will render equations (19), (20), and (21) nonlinear and thus mathematically intractable.

Upon substituting (21) into (20) a linear recursion relation is obtained with the well-known solution

$$\tau_n = \frac{\tau_0}{n+1} - \frac{1}{n+1} \sum_{k=0}^{k=n-1} x_k \qquad | \tau_n | \leqq T/2. \tag{23}$$

By inspection the following pertinent parameters are computed

$$\mu_n = E[\tau_n] = \frac{\tau_0}{n+1} \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty$$

and

$$\sigma_n^2 = E\{\tau_n - E^2[\tau_n]\} = \text{var } \tau_n = \frac{T^2}{8\rho} \frac{n}{(n+1)^2} \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty. \tag{24}$$

In evaluating (24) we assumed that the sequence of random variables $\{x_n\}$ is independent. This is not strictly true. We see from (17) that the sequence of random variables $\{x_n\}$ for fixed τ is indeed independent since each x_n represents nonoverlapping integrals of the white-noise process $\nu(t)$. However, as τ_n is changed according to equation (20) the noise integrals may overlap. To include this dependence in the analysis would render this seemingly simple problem untractable mathematically. Physically we feel, however, that this dependence is weak and therefore can be neglected.

From (23) we see that τ_n possesses a truncated gaussian probability density

$$\begin{aligned} P(\tau_n) &= p_1 \, \delta(\tau_n - T/2) + p_2 \, \delta(\tau_n + T/2) + G(\tau_n) \qquad && | \tau_n | \leqq T/2 \\ &= 0 && | \tau_n | > T/2 \end{aligned} \tag{25}$$

where

$$G(\tau_n) = \frac{1}{\sqrt{2\pi} \, \sigma_n} \exp \left\{ -\frac{1}{2\sigma_n^2} (\tau_n - \mu_n)^2 \right\}$$

and

$$p_1 = \int_{-\infty}^{-T/2} G(\tau_n) \, d\tau_n$$

$$p_2 = \int_{T/2}^{\infty} G(\tau_n) \, d\tau_n \, .$$

Using this probability density we can compute the system error rate.

Dispensing with tedious computational details, and focusing attention on essentials, we find that the conditional error rate (conditioned on the unknown parameter τ_n) for this simple system is asymptotically (large signal-to-noise ratio)

$$P_e(\tau_n) \sim \exp - \left\{\frac{A^2[T - 2 \mid \tau_n \mid]^2}{2\sigma^2}\right\} \qquad \mid \tau \mid \leqq T/2, \tag{26}$$

where

$$\sigma^2 = N_0 T.$$

When $\tau_n = 0$, we have ideal performance, as we should. When $\tau_n = \pm T/2$, we have disaster. To obtain the actual error rate we must average (26) over the permissable values of τ_n. This calculation yields

$$P_e = E\{P_e(\tau_n)\} \sim p_1 + p_2 + \int_{-T/2}^{T/2} P_e(\tau_n) G(\tau_n)\, d\tau_n\,. \tag{27}$$

The evaluation of (27) is straightforward. In terms of the normalized random variable $\alpha = \tau_n/T$, we express (26) in the form

$$P_e(\alpha) \sim e^{-(1-2|\alpha|)^2} \qquad |\alpha| < 1/2. \tag{28}$$

In terms of the same normalized variables and the explicit values of μ_n and α_n [equation (24)] we write

$$G(\alpha) \sim \exp\left[-4n\left(\alpha - \frac{1}{2n}\right)^2\right] \tag{29}$$

which is valid when n is large. In writing down (29) we set $\tau_0 = T/2$ (a worst initial guess).

Asymptotically, p_1 and p_2 behave as $e^{-n\rho}$ and, as we shall see shortly, can be neglected compared with the last term in (27). To conclude the error rate calculation we evaluate

$$\int_{-T/2}^{T/2} P_e(\tau_n) G(\tau_n)\, d\tau_n = \xi_n(\rho) \sim \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-\rho(1-2|\alpha|)^2 - \rho 4n(\alpha - 1/2n)^2}\, d\alpha$$

$$= \int_{-\frac{1}{2}}^{0} e^{-\rho E_1(\alpha)}\, d\alpha + \int_{0}^{\frac{1}{2}} e^{-\rho E_2(\alpha)}\, d\alpha, \tag{30}$$

where

$$E_1(\alpha) = (1 + 2\alpha)^2 + 4n\left(\alpha - \frac{1}{2n}\right)^2$$

and

$$E_2(\alpha) = (1 - 2\alpha)^2 + 4n\left(\alpha - \frac{1}{2n}\right)^2.$$

Using a saddle-point technique to obtain an asymptotic approximation for the integrals, we find that

$$P_e(\rho) \sim e^{-\rho M_1(n)} + e^{-\rho M_2(n)},$$

where

$$M_1(n) \sim 1 + \frac{1}{n}$$

and

$$M_2(n) \sim 1 - \frac{3}{n}. \tag{31}$$

Combining the above asymptotic results with p_1 and p_2 we obtain finally

$$P_e \sim \exp\left\{-\rho\left(1 - \frac{3}{n}\right)\right\} \tag{32}$$

for n and ρ large. All the other terms have exponents larger than (32) and therefore can be neglected. For example when $n = 30$, the degradation from ideal $(n \rightarrow \infty)$ is only 0.5 dB approximately.

What this example shows is that for square-wave modulations in the presence of additive white gaussian noise, bit timing can reliably be derived in approximately 30-bit intervals.

4.3 *Synchronization of Bandlimited PAM*

We now consider a timing recovery algorithm for a bandlimited PAM signal. As in the previous section the synchronization problem will be cast as a regression problem. Our received signal is given by equation (1)

$$V(t) = \sum_m a_m h(t - mT - \tau^*) + \nu(t), \tag{33}$$

and as before the objective of the synchronizer is to accurately and rapidly estimate τ^*. In order to extract information about τ^* we low-pass filter, differentiate, and sample the received signal. Hence the error signal is similar to that shown in Fig. 2, with the matched filter replaced by a low-pass filter. Thus the receiver does not need knowledge of the pulse $h(t)$. If we denote the derivative of $h(\cdot)$ by $g(\cdot)$, then the differentiated and sampled received signal is given by

$$\begin{aligned} V'(kT + \tau) &= \sum_m a_m g[(k - m)T + \tau - \tau^*] + \nu(kT + \tau) \\ &= \sum_m a_m g_{k-m}(\tau - \tau^*) + \nu_k , \end{aligned} \tag{34}$$

where τ is an arbitrary sampling time such that $|\tau| < T/2$, g_{k-m} denotes $g((k - m)T)$, and ν_k are samples[†] of the differentiated noise process $\nu(t)$. As before we let $\hat{a}_k$ denote the decision made at time $kT + \tau$. Assuming that the error rate is low enough so that with high probability $\hat{a}_k = a_k$, we then have that

$$\begin{aligned} \hat{a}_k V'(kT + \tau) &= \hat{a}_k a_k g_o(\tau - \tau^*) + \hat{a}_k \sum_{m \neq k} a_m g_{k-m}(\tau - \tau^*) + \nu_k \\ &= a_k^2 g(\tau - \tau^*) + \hat{a}_k \sum_{m \neq k} a_m g_{k-m}(\tau - \tau^*) + \nu_k , \end{aligned} \tag{35}$$

where we have noted that $g_o(\tau - \tau^*) = g(\tau - \tau^*)$. If we further assume that $\hat{a}_k$ is uncorrelated with a_j,[‡] for $j \neq k$, then averaging (35) gives

$$m(\tau) \triangleq E[\hat{a}_k V'(kT + \tau)] = \overline{a^2} g(\tau - \tau^*), \tag{36}$$

where

$$\overline{a^2} = \frac{d^2}{3}(L^2 - 1).$$

Now for the typical impulse response $h(t)$ and its derivative $g(t)$, shown in Figs. 4 and 5, respectively, it is true that the (regression) equation

$$g(\tau - \tau^*) = 0, \qquad |\tau - \tau^*| \leqq T/2 \tag{37}$$

has the unique solution

$$\tau = \tau^*. \tag{38}$$

Since the synchronization problem has been modeled as a regression problem, we again use a Robbins-Monro algorithm to sequentially estimate τ^*. Denoting the kth estimate by τ_k, we have the modified Robbins-Monro algorithm

$$\tau_{k+1} = \begin{cases} \tau_k + c_k[\hat{a}_k V'(kT + \tau_k)], & | \tau_k + c_k[\hat{a}_k V'(kT + \tau_k)] \,| < T/2 \\ \tau_k , & \text{otherwise.} \end{cases} \tag{39}$$

A feedback implementation of the above algorithm is shown in Fig. 6, with D denoting a delay. It is again noted that the algorithm con-

[†] The dependence of the noise sample on the sampling offset τ is not shown, since it is assumed that the noise is stationary.

[‡] As it will be if the a_k's are independent and the receiver is supplied with an ideal reference.

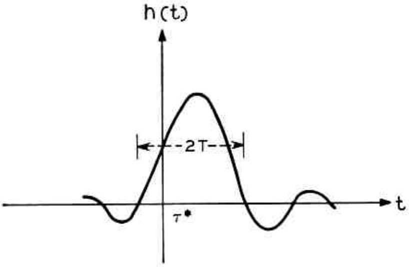

Fig. 4—A typical impulse response $h(t)$.

strains the estimate to a region of width T. This is consistent with the observation that any actual sampling instant will always be within $T/2$ seconds of the desired instant τ^*, i.e., we may "slip" T seconds but this is immaterial as far as estimating τ^* is concerned. It is by no means clear, *a priori*, that the above algorithm will converge rapidly or will converge at all. In fact the rest of this paper will consider the conditions which must be satisfied for the above algorithm to converge and the resulting rate of convergence.

V. ANALYSIS OF THE SYNCHRONIZATION ALGORITHM

5.1 *The Error Equation*

In order to evaluate the proposed synchronization algorithm we will derive a difference equation for the mean-square estimation error $\overline{e_k^2}$, where

$$e_k \equiv \tau_k - \tau^*, \tag{40}$$

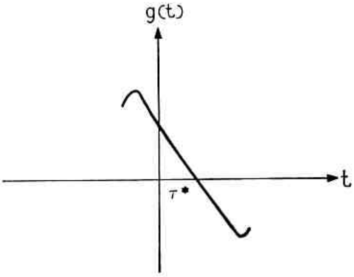

Fig. 5—The derivative of $h(t)$.

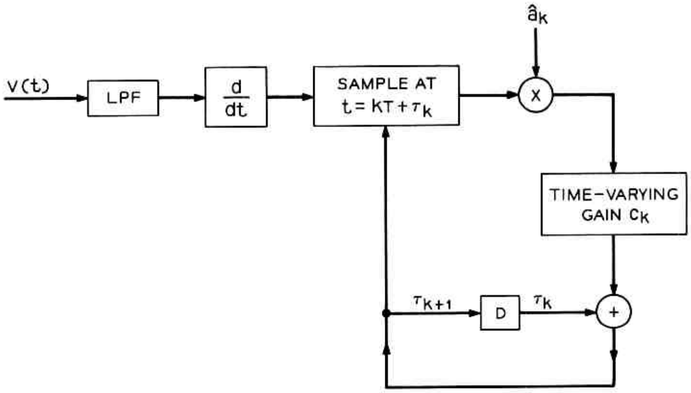

Fig. 6—A realization of the synchronization algorithm.

and the overbar denotes expectation.[†] In order to do this we see that from (39), and neglecting for the moment the constraining portion of the algorithm, we have

$$\begin{aligned} e_{k+1} &= \tau_{k+1} - \tau^* = \tau_k - \tau^* + c_k[\hat{a}_k V'(hT + \tau_k)] \\ &= \tau_k - \tau^* + c_k[g(\tau_k - \tau^*) + \hat{a}_k \sum_{m \neq k} a_m g_{k-m}(\tau_k - \tau^*) + \nu_k] \\ &= e_k + c_k[g(e_k) + \hat{a}_k \sum_{m \neq k} a_m g_{k-m}(e_k) + \nu_k]. \qquad (41) \end{aligned}$$

We note that $g(\cdot)$ is such that, on the average, the error is decreased at each iteration, and once the estimation error is small[‡] we need only keep first-order terms in a Taylor Series expansion of $g_{k-m}(e_k)$ about $(k - m)T$, i.e.,

$$g_{k-m}(e_k) \approx g_{k-m} + g'_{k-m} e_k \,, \qquad (42)$$

where g'_{k-m} denotes the derivative of $g(\cdot)$ evaluated at $(k - m)T$. Combining (41) and (42) yields the (approximate) first-order stochastic difference equation for the evolution of the error,

$$e_{k+1} = [1 + g'_o c_k + c_k \hat{a}_k \sum_{m \neq k} a_m g'_{k-m}] e_k + c_k \hat{a}_k [\sum_{m \neq k} a_m g_{k-m} + \nu_k]. \qquad (43)$$

Before studying the behavior of (43) we introduce the following

[†] We use the mean-square estimation error as a measure of performance. This is because the estimate is a nonlinear one, and thus the probability of error cannot be computed.

[‡] Under this assumption we can certainly neglect the possibility that $\tau_{k+1} = \tau_k$.

notation:

$$g'_o = -\alpha \tag{44a}$$

$$\beta_k = c_k(g'_o + \hat{a}_k \sum_{m \neq k} a_m g'_{k-m}) \tag{44b}$$

$$\gamma_k = 1 + \beta_k \tag{44c}$$

$$Q_k = \hat{a}_k \sum_{m \neq k} a_m g_{k-m} + \nu_k \ , \tag{44d}$$

and using the above we rewrite (43) as

$$e_{k+1} = \gamma_k e_k + c_k Q_k \ . \tag{45}$$

Thus the error obeys a stochastic difference equation where the gain (γ_k) and the driving term (Q_k) are correlated. It is important to note that for the system described by (45) the probability density of the present error e_k does not depend solely on a finite number of past data symbols, a_k, but depends on all past and future values. This renders impossible an exact analysis of the mean-square error. However, if we assume that both γ_k and Q_k are independent sequences, then e_k depends solely on past γ_k and Q_k, and we can obtain a bound on the mean-square error.† Squaring and averaging both sides of (45) gives

$$E[e_{k+1}^2] = E[\gamma_k^2 e_k^2] + 2c_k E[\gamma_k e_k Q_k] + c_k^2 E[Q_k^2]. \tag{46}$$

We now proceed to bound each of the terms on the right-hand side of (46). If we assume that the "eye" of the twice-differentiated impulse response is open, i.e.,

$$\alpha > \sum_{m \neq 0} | \, g'_m \, |, \tag{47}$$

then

$$\gamma_k = 1 - c_k(\alpha - \hat{a}_k \sum_{m \neq k} a_m g'_{k-m}) \leqq 1 - c_k(\alpha - \sum_{m \neq 0} | \, g'_m \, |) \tag{48a}$$

$$= 1 - c_k \beta, \tag{48b}$$

where β denotes $\alpha - \sum_{m \neq 0} |g'_m|$. Using the above assumption, and the boundedness of the error, we have that

$$|E[\gamma_k Q_k e_k]| = |E[e_k] E[\gamma_k Q_k]| \leqq T/2 \ |E[\gamma_k Q_k]|, \tag{49}$$

† Despite much effort we have been unable to proceed without this assumption, but since the results which follow are intuitively satisfying and provide insight into this difficult problem they have been included in the paper.

and due to the independence of the data bits

$$\begin{aligned} E[\gamma_k Q_k] &= E[((1 - c_k)\alpha + c_k \hat{a}_k \sum_{m \neq k} a_m g'_{k-m})\hat{a}_k(\sum_{i \neq k} a_i g_{k-i} + \nu_k)] \\ &= c_k \sum_{m \neq 0} g'_m g_m = c_k 2/TG, \end{aligned} \tag{50}$$

where[†] G denotes $T/2 \sum_{m \neq 0} g'_m g_m$. Finally we have

$$E[Q_k^2] = \sigma^2 + \sum_{m \neq 0} g_m^2 = \sigma^2 + P, \tag{51}$$

where P denotes $\sum_{m \neq 0} g_m^2$. Letting

$$\Delta_k = E[e_k^2], \tag{52}$$

and combining (46)–(52) we have the iterative bound

$$\Delta_{k+1} \leqq (1 - \beta c_k)^2 \Delta_k + c_k^2 M \tag{53}$$

on the mean-square error, where M is the sum of G and $\sigma^2 + P$. Although several assumptions have been made in obtaining (53) it is believed that the effect of the salient quantities upon the synchronization algorithm have been preserved. We now proceed to find the gain sequence which minimizes the bound of (53).

5.2 *The Optimum Gain Sequence*

We now find the sequence of gains, c_k^*, which minimize the right-hand side (RHS) of (53) for fixed Δ_k . Since we minimize a bound on the mean-square error at every iteration, this is a min-max procedure. We first find the optimum gain sequence in terms of Δ_k , and then by simultaneously iterating this equation and the bound of (53) we show that c_k^* is proportional to $1/k$ for large k. We begin by setting to zero the derivative of the RHS of (53) with respect to c_k , i.e.,

$$-\beta(1 - \beta c_k)\Delta_k + M c_k = 0$$

or

$$c_k^* = \frac{\beta \Delta_k}{M + \beta^2 \Delta_k}. \tag{54}$$

Using (54) in (53) we have

$$\begin{aligned} \Delta_{k+1} &\leqq \left(1 - \frac{\beta^2 \Delta_k}{M + \beta^2 \Delta_k}\right)^2 \Delta_k + M\left(\frac{\beta \Delta_k}{M + \beta^2 \Delta_k}\right)^2 \\ &\leqq \frac{M}{\beta} c_k^*, \end{aligned} \tag{55}$$

[†] It should be noted that if $h(t)$ is an even function of time (with respect to the origin), then $g(t)$ and $g'(t)$ will be respectively odd and even time functions and G will be zero.

or

$$\frac{c_{k+1}^*}{(1 - \beta c_{k+1}^*)} \leqq c_k^* . \tag{56}$$

Now if

$$(1 - \beta c_{k+1}^*) \geqq 0 \tag{57}$$

then we have the relation

$$c_{k+1}^* \leqq \frac{c_k^*}{1 + \beta c_k^*} , \tag{58}$$

which can be iterated to give[†]

$$c_k^* \leqq \frac{c_o^*}{1 + \beta c_o^* k} \tag{59a}$$

$$= \frac{\beta \Delta_o}{M + \beta^2 \Delta_o (k + 1)} , \tag{59b}$$

where

$$c_o^* = \frac{\beta \Delta_o}{M + \beta^2 \Delta_o} , \tag{60}$$

and Δ_o is the initial error variance. Henceforth we will interpret the sequence c_k^*, specified by (59) and (60) with the inequality replaced by an equality, as the optimum gain sequence. Combining (55), (59) and (60) we see that the mean-square error is bounded by

$$\Delta_k^* \leqq \frac{M \Delta_o}{M + \beta^2 \Delta_o k} \tag{61a}$$

which for large k becomes

$$\Delta_k^* \leqq \frac{M}{\beta^2} \cdot \frac{1}{k} \cdot \tag{61b}$$

Thus we see that asymptotically the minimized mean-square error is bounded by a term which decays as $1/k$, and is inversely proportional to signal-to-noise type ratio (β^2/M).

The optimum gain, as given by (59b), depends upon the parameters Δ_o, M, and β. Since these quantities are generally unknown it is tempting to replace c_k^* by its asymptotic (large k) value $1/\beta(k + 1)$. Caution must be exercised in making this approximation; since $M \gg \beta^2 \Delta_o$

[†] Note that $\beta c_k^* \leqq \beta c_o^*/(1 + \beta c_o^* k) \leqq 1$, thus satisfying (57).

implies that the optimum gain sequence is essentially constant for many iterations, substitution of a decaying sequence could lead to an unreliable estimate (we will consider this point in Section 5.4). However if[†] $\beta^2\Delta_o \gg M$, then $c_k^* \approx 1/\beta(k+1)$ and we have only one unknown parameter. A possibility is to replace β by an estimate—techniques of this sort are called adaptive estimation procedures. We now sketch a particular adaptive scheme.

5.3 *An Adaptive Synchronization Algorithm*

We now give a method for recursively estimating β, which can then be incorporated in an adaptive synchronization scheme. Since

$$\beta = \alpha - \sum_{m \neq 0} | g_m' |,$$

we desire a function of the received data which has β as its average value. We note that from (34) we have

$$E[\hat{a}_k V''(kT + \tau)] = g'(\tau - \tau^*) \approx -\alpha \tag{62a}$$

(where the approximation is for small $\tau - \tau^*$), and

$$E[\hat{a}_i V''(kT + \tau)] = g_{i-k}'(\tau - \tau^*) \approx g_{i-k}' . \tag{62b}$$

We can then estimate β by using a recursive stochastic approximation algorithm of the type discussed in Section 4.1. Such a scheme would twice differentiate the incoming data and then multiply the data sample by as many of the previous decisions as there are significant nonzero samples in the impulse response. Since even an approximate analysis of the above algorithm is hopelessly complex, we will consider the effect of using a gain of the form c/k, where c is a constant to be chosen.

5.4 *A Suboptimum Gain*

We consider the mean-square error, as given by (53), with $c_k = c/k$. This gain is chosen since the optimum gain is asymptotically of this form. Care must be taken in choosing c, since the mean-square error will be shown to be a sensitive function of this parameter. Iterating (53) gives

$$\Delta_{k+1} \leqq \prod_{i=0}^{k} (1 - \beta c_i)^2 \Delta_o + \sum_{i=0}^{k} \prod_{j=i+1}^{k} (1 - \beta c_j)^2 c_i^2 M . \tag{63}$$

[†] A condition one would expect to be satisfied in practice.

The inequality

$$1 - x \leqq e^{-x} \tag{64}$$

gives

$$\prod_{j=i+1}^{k} (1 - \beta c_j)^2 \leqq \exp\left(-2 \sum_{j=i+1}^{k} \beta c_j\right); \tag{65}$$

and noting that

$$\sum_{j=i+1}^{k} \beta \frac{c}{j} \approx \beta c \int_{i+1}^{k} \frac{1}{x}\, dx = \beta c \ln\left(\frac{k}{i+1}\right)$$

results in

$$\prod_{j=i+1}^{k} (1 - \beta c_j)^2 \leqq \left(\frac{i+1}{k}\right)^{2\beta c}. \tag{66}$$

We can see that the transient behavior of the mean-square error, which is specified by the first term on the RHS of (63), will be of the form $(1/k)^{2\beta c}$. The other component of the mean-square error will be (approximately) bounded by

$$\begin{aligned} \sum_{i=0}^{k} \left(\frac{i+1}{k}\right)^{2\beta c} M \frac{c^2}{i^2} &\leqq \frac{Mc^2}{k^{2\beta c}} \sum_{i=0}^{k} (i+1)^{2(\beta c - 1)} \\ &\leqq \frac{Mc^2}{(1+k)^{2\beta c}} \frac{1}{2\beta c - 1} (1+k)^{2\beta c - 1} \\ &= \frac{Mc^2}{(2\beta c - 1)(1+k)}, \end{aligned} \tag{67a}$$

which results in

$$\Delta_{k+1} \leqq \left(\frac{1}{k}\right)^{2\beta c} \Delta_o + \frac{Mc^2}{(2\beta c - 1)(1+k)} \tag{67b}$$

as a bound on the mean-square error. If $2\beta c > 1$, then for large k the above bound becomes

$$\Delta_{k+1} \leqq \frac{Mc^2}{(2\beta c - 1)(1+k)} \tag{67c}$$

and the mean-square error will converge at the optimum rate $(1/k)$. It is seen that care must be taken in selecting c, since for $c \geqq 1/2\beta$ (i.e., for $2\beta c > 1$) the quantity $Mc^2/2\beta c - 1$ has a minimum† at $c =$

† With $c = 1/\beta$, $\Delta_k \leqq M/\beta^2\ 1/k$ which is the optimum asymptotic rate of convergence.

$1/\beta$, and is infinite at both $c = 1/2\beta$ and $c = \infty$. Thus a very small step size ($c \ll 1/2\beta$) will result in an mse which converges at a less than optimum rate, while large step sizes ($c \gg 1/2\beta$) will result in a mean-square error which, while converging at the optimum rate, may be quite large for many iterations. The sensitivity of the above bound with respect to "c" may make the use of an adaptive procedure (which estimates β) advisable.

5.5 *An Example*

Consider the (minimum bandwidth) pulse

$$h(t) = A \frac{\sin \pi W t}{\pi W t} \tag{68}$$

where $W = 1/T$. It is easy to show that

$$\beta = \tfrac{1}{3} A(\pi W)^2$$

$$M = \sigma^2 + \frac{\pi^2}{3} A^2 W^2;$$

thus from (61b) the percentage minimized mean-square error is bounded by

$$\frac{\Delta_k}{T^2} < \frac{M}{T^2 \beta^2 k} = \frac{1 + \left(\frac{\sigma}{A}\right)^2 \left(\frac{\sqrt{3}}{\pi W}\right)^2}{\frac{1}{3}\pi^2 k}. \tag{69}$$

For a 30 dB signal-to-noise (A/σ) ratio, and with $W = 3000$ Hz, we see that Δ_k/T^2 is less than 0.01 for $k \geqq 10$. In other words, after 10 symbols have been received, the above synchronization algorithm reduces the mean-square error to less than 1/100 of a symbol interval.

REFERENCES

1. Bennett, W. R., and Davey, J. R., *Data Transmission,* New York: McGraw-Hill, 1965.
2. Sanger, David K., "Digital Demodulation with Data Subcarrier Tracking," Jet Propulsion Laboratory Technical Report 32–1314, August 1968.
3. Wintz, P. A., and Luecke, E. J., "Performance of Optimum and Supoptimum Synchronizers," IEEE Trans. Commun. Technology, *COM-17* (June 1969), pp. 380–389.
4. Robbins, H., and Monro, S., "A Stochastic Approximation Method," Ann. Math. Stat. *22* (1951), pp. 400–407.
5. Kiefer, J., and Wolfowitz, J., "Stochastic Estimation of the Maximum of a Regression Function," Ann. Math. Stat. *23* (1952), pp. 462–466.
6. Balakrishnan, A. V., ed., *Advances in Communication Systems,* Vol. 2, New York: Academic Press, 1966, pp. 51–106.

7. Saltzberg, B. R., "Timing Recovery for Synchronous Binary Data Transmission," B.S.T.J., *46,* No. 3 (March 1967), pp. 593–622.
8. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication,* New York: McGraw-Hill, 1968.
9. Chang, R. W., "Joint Equalization, Carrier Acquisition, and Timing Recovery for Data Communication," International Conf. Commun., San Francisco, June 1970.
10. Murphy, J. et al., "Ultra Long-Range Telemetry Study," AL-TDR 64–67, Avionics Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, April 1964.
11. Mallory, P., "A Maximum Likelihood Bit Synchronizer," International Telemetering Conf. Proc., *IV* (1968), pp. 1–16.
12. Hirsch, D., private communication.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 5, May-June, 1971
Printed in U.S.A.

Optimum Equalization and the Effect of Timing and Carrier Phase on Synchronous Data Systems

By E. Y. HO

(Manuscript received December 16, 1970)

The minimum mean-square error (M.M.S.E.) at the receiver output generally depends upon the sampling instant and demodulating carrier phase for synchronous data systems. In this study, it is shown that for certain single-sideband data systems with no excess bandwidth (e.g., class IV and class V partial-response systems), the M.M.S.E. is completely independent of the sampling instant and demodulating carrier phase if the receiver contains an infinitely long transversal filter equalizer. Practically speaking, computer calculations indicate that for a class IV system operating in the presence of typical received signal-to-noise ratios, a 19-tap equalizer is sufficient to make the M.M.S.E. relatively insensitive to the sampling instant and demodulating carrier phase. Thus, for such data systems, a significant reduction in the receiver complexity and possibly in the start-up time may be obtained, because no time is spent acquiring timing and carrier phase.

The optimum infinite-length equalizer for synchronous data systems with a fixed channel is also calculated for two different conditions. The conditions are: (i) the minimization of the output noise plus mean-square intersymbol interference and (ii) the minimization of the output noise subject to the constraint that the equalizer forces the intersymbol interference to zero. Explicit expressions for the optimum equalizer and the M.M.S.E. are obtained. Satisfying condition (i) results in the lower value of M.M.S.E.; however, the M.M.S.E.s for these two criteria are almost equivalent for either large signal-to-noise ratios or small slope of the amplitude-frequency characteristics of the channel.

I. INTRODUCTION

In synchronous data systems, the transmission rates are frequently limited by the intersymbol interference which is caused by the amplitude and phase distortion in the transmission channel. In order to reduce the effect of the intersymbol interference, it is necessary to equalize the channel before the data can be transmitted.

Several automatic equalization schemes using transversal filters have been devised for such data systems.[1–5] Chang[6] has investigated the effect of the sampling instant and carrier phase on the minimum mean-square intersymbol interference for a noiseless system with a finite-length transversal equalizer. In principle, we can make the mean-square intersymbol interference arbitrarily small by using an infinitely long transversal filter, provided that the tap-gain settings can be made arbitrarily accurate. However, the equalizer which forces the intersymbol interference to zero may not be the most desirable one when noise is present.

We have found, in this study, the optimum infinite-length mean-square equalizer for such synchronous data systems with a fixed channel. Two different cases are considered corresponding to the following optimality criteria: (*i*) the minimization of the output noise plus mean-square intersymbol interference, and (*ii*) the minimization of the output noise power subject to the constraint that the equalizer forces the intersymbol interference to zero.

Explicit expressions for the optimum equalizer and the M.M.S.E. are obtained. We also have found that for certain types of data systems (S.S.B. class IV or class V partial-response systems) the M.M.S.E. does not depend upon sampling instant and demodulating carrier phase. Thus, for such data systems, it may be possible to reduce significantly the receiver complexity and the start-up time.

A computer program has been written for a class IV system which is equipped with a finite-length mean-square equalizer. The number of taps needed to achieve near optimum performance in practical situations can then be determined. Throughout, additive Gaussian noise and independence of information digits are assumed.

II. GENERAL CONSIDERATIONS

The optimality criteria will be formulated in this section. A simplified block diagram of a general digital data system is shown in Fig. 1. We assume that every T seconds, an impulse of amplitude a_n ($a_n = \{2M + 1, \cdots 1, -1, \cdots -(2M + 1)\}$), is transmitted to the input of the system. The a_n are assumed to be identically distributed independent random variables.

In the absence of channel noise, for a sequence of input impulses

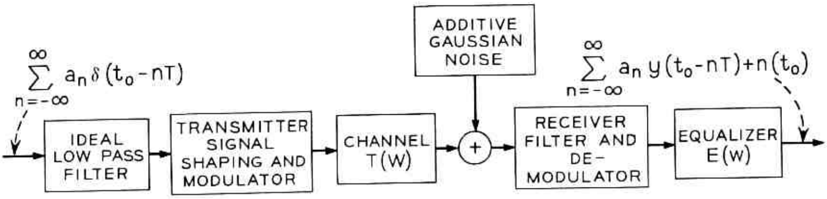
$\sum_{n=-\infty}^{\infty} a_n \delta(t_0 - nT)$
IDEAL LOW PASS FILTER
TRANSMITTER SIGNAL SHAPING AND MODULATOR
CHANNEL T(W)
ADDITIVE GAUSSIAN NOISE
+
RECEIVER FILTER AND DE-MODULATOR
EQUALIZER E(W)
$\sum_{n=-\infty}^{\infty} a_n y(t_0 - nT) + n(t_0)$

Fig. 1—Digital data system, simplified block diagram.

$$\sum_{n=-\infty}^{\infty} a_n \delta(t - nT), \tag{1}$$

the corresponding sequence at the receiver equalizer output is

$$\sum_{n=-\infty}^{\infty} a_n y(t - nT, \theta), \tag{2}$$

where $y(t, \theta)$ is the system impulse response with demodulating carrier phase θ.*

With noise, the output at the sampling instant t_0 with a demodulating carrier phase θ_0 is

$$V = a_0 y(t_0\,, \theta_0) + \sum_{n \neq 0} a_n y(t_0 - nT, \theta_0) + n(t_0) \tag{3a}$$

or

$$V = a_0 y_0(\theta_0) + \sum_{n \neq 0} a_n y_n(\theta_0) + n_0\,, \tag{3b}$$

where the terms $y_n(\theta_0)$, $n \neq 0$, represent intersymbol interference and $y_0(\theta_0)$ is the output value at the main sample point.

One useful measure of the performance of such a data system is mean-square error (M.S.E.). In this study, we define the normalized M.S.E. at the sampling instant t_0 with a demodulating carrier phase θ_0 to be

$$\begin{aligned} [\text{M.S.E.}]_{t_0,\theta_0} &= E\{[V - a_0 y_0(\theta_0)]^2\}/E\{a_0^2 y_0^2(\theta_0)\} \\ &= [E\{n_0^2\} + E[a_j^2] \sum_{n \neq 0} y_n^2(\theta_0)]/E\{a_0^2 y_0^2(\theta_0)\}, \end{aligned} \tag{4}$$

where $E[X]$ means expectation of random variable X.

For binary systems, $E[a_j^2]$ is equal to 1. The variance of the noise at

* $\theta = 0°$ corresponds to the phase of the frequency component of the received spectrum at the carrier frequency.

the equalizer output (see the Appendix) is

$$\sigma_0^2 = E(n_0^2) = \frac{\sigma_i^2 T}{\pi} \int_0^{\pi/T} \psi \text{ eq } (\omega) \cdot | E(\omega) |^2 \, d\omega, \tag{5}$$

where $\sigma_i^2 T$ is the power spectral density of the input white noise, ψ eq (ω) is the square of the equivalent baseband receiver filter characteristic and $E(\omega)$ is the transfer function of the equalizer.

Now we wish to design two optimum equalizers for two different conditions. In the first the $[\text{M.S.E.}]_{t_0, \theta_0}$ is minimized subject to the constraint that $y_0(\theta_0)$ is a constant,

$$C_1 = y_0(\theta_0) = \text{Re} \frac{T}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega, \theta_0) E(\omega) e^{j\omega t_0} \, d\omega, \tag{6}$$

where Y eq (ω, θ_0) (see the Appendix) is the equivalent baseband system transfer function for sampling instant t_0 and demodulating carrier phase θ_0 . In the second case the optimum equalizer is found by minimizing the variance of the output noise subject to the constraint equations, (6) and

$$0 = y_n(\theta_0) = \text{Re} \frac{T}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega, \theta_0) E(\omega) e^{j\omega(t_0 + nT)} \, d\omega, \quad n \neq 0. \tag{7}$$

III. MINIMIZATION OF NOISE PLUS INTERSYMBOL INTERFERENCE

3.1 *A General Binary Data System*

The details of the minimization of the $[\text{M.S.E.}]_{t_0, \theta_0}$ given by equation (4) subject to the constraint equation (6) are given in the Appendix.

For a binary data system, the optimum equalizer, $E_0(\omega)$, for sampling instant t_0 and demodulating carrier phase θ_0 is

$$[E_0(\omega)]_{t_0, \theta_0} = \frac{\{ Y \text{ eq } (\omega, \theta_0) e^{j\omega t_0} \}^*}{\{ \sigma_i^2 \psi \text{ eq } (\omega) + | Y \text{ eq } (\omega, \theta_0) |^2 \}} \cdot \frac{C_1}{\dfrac{T}{\pi} \displaystyle\int_0^{\pi/T} \frac{| Y \text{ eq } (\omega, \theta_0) |^2}{\sigma_i^2 \psi \text{ eq } (\omega) + | Y \text{ eq } (\omega, \theta_0) |^2} \, d\omega} \tag{8}$$

where $\{X\}^*$ means complex conjugate of X. It follows that the M.M.S.E. for the corresponding sampling instant and demodulating carrier phase is

$$[\text{M.M.S.E.}]_{t_0, \theta_0} = \frac{1}{\dfrac{T}{\pi} \displaystyle\int_0^{\pi/T} \frac{| Y \text{ eq } (\omega, \theta_0) |^2}{\sigma_i^2 \psi \text{ eq } (\omega) + | Y \text{ eq } (\omega, \theta_0) |^2} \, d\omega} - 1. \tag{9}$$

$|\, Y \text{ eq } (\omega,\, \theta_0) \,|^2$ is, in general, a function of t_0 and θ_0 . It follows that $[\text{M.M.S.E.}]_{t_0,\theta_0}$ depends generally upon t_0 and θ_0 . Hence, there exists an optimum sampling instant $\bar{t}_0$ and demodulating carrier phase $\bar{\theta}_0$, such that

$$[\text{M.M.S.E.}]_{\bar{t}_0,\bar{\theta}_0} = \min_{\text{all } t_0,\theta_0} \; [\text{M.M.S.E.}]_{t_0,\theta_0} \; . \tag{10}$$

However, there exist cases where $|Y \text{ eq } (\omega,\, \theta_0)|^2$ is not a function of t_0 and θ_0 ; for example, the $|Y \text{ eq } (\omega,\, \theta_0)|^2$ of a single-sideband system with no excess bandwidth is independent of the sampling instant, t_0 , and the demodulating carrier phase, θ_0 , (see the Appendix.) The SSB class IV partial-response system represents another example. Therefore, if an infinite equalizer is available, no loss in performance occurs when arbitrary t_0 and θ_0 are used.

3.2 *SSB Class IV Partial-Response System*

In this section we will show that the M.M.S.E. for a class IV partial-response system is independent of sampling instant and demodulating carrier phase. The M.M.S.E. will be computed for certain typical telephone channels. Such a data system has been fully described in Reference 7.

The transfer functions of the transmitter and receiver filters are

$$\begin{aligned} S(\omega_c - \omega) &= R(\omega_c - \omega) \\ &= \begin{cases} \sqrt{2T \sin |\, \omega \,|\, T}\; e^{-j(\omega/|\omega|)\pi/4} & 0 < |\, \omega \,| \leqq \dfrac{\pi}{T}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{11}$$

The M.S.E. for the partial-response system is defined to be

$$[\text{M.S.E.}]_{C,t_0,\theta_0} = \frac{[\sigma_0^2 + (y_1(\theta_0) + y_{-1}(\theta_0))^2 + \sum\limits_{n \neq 1,-1} y_n^2(\theta_0)]}{[(y_1(\theta_0) - y_{-1}(\theta_0))/2]^2} . \tag{12}$$

The constraint equations are

$$C = y_1(\theta_0) = \operatorname{Re} \frac{2T}{\pi} \int_0^{\pi/T} -j(\sin \omega T) T(\omega_c - \omega) e^{-j\theta_0} E(\omega) e^{j\omega(t_0 + T)} \, d\omega \tag{13}$$

and

$$\begin{aligned} C - 2 &= y_{-1}(\theta_0) \\ &= \operatorname{Re} \frac{2T}{\pi} \int_0^{\pi/T} -j(\sin \omega T) T(\omega_c - \omega) e^{-j\theta_0} E(\omega) e^{-j\omega(T - t_0)} \, d\omega \end{aligned} \tag{14}$$

where Re $[X]$ means real part of X, $T(\omega)$ is the transfer function of the channel, and C is a constant.

For a given constant C, sampling instant t_0, and demodulating carrier phase θ_0, the optimum equalizer is

$$[E_0(\omega)]_{t_0,\theta_0,C} = \begin{cases} \dfrac{\begin{bmatrix} \dfrac{C(A_1 - A_2) + 2A_2}{A_1^2 - A_2^2} [-j2T(\sin T \mid \omega \mid) \cdot T(\omega_c - \omega) \\ \cdot e^{-j\theta_0} \cdot e^{j\omega(t_0+T)}]^* \\ + \\ \dfrac{C(A_1 - A_2) - 2A_1}{A_1^2 - A_2^2} \cdot [-j2T(\sin T \mid \omega \mid) \cdot T(\omega_c - \omega) \\ \cdot e^{-j\theta_0} e^{j\omega(t_0-T)}]^* \end{bmatrix}}{\sigma_i^2 2T \sin T \mid \omega \mid + 4T^2 \sin^2 T \mid \omega \mid \cdot \mid T(\omega_c - \omega) \mid^2} & 0 \leqq \mid \omega \mid \leqq \dfrac{\pi}{T}, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where $[X]^*$ means complex conjugate of X, and, A_1 and A_2 are given by the equation (41a) and (41b) respectively.

It follows that, the corresponding M.M.S.E. is

$$\begin{aligned} &[\text{M.M.S.E.}]_{C,t_0,\theta_0} \\ &\quad = \frac{1}{\pi} \int_0^{\pi/T} [\sigma_i^2 2T \sin \omega T + 4T^2 \sin^2 \omega T \cdot \mid T(\omega_c - \omega) \mid^2] \\ &\qquad \cdot \mid [E_0(\omega)]_{C,t_0,\theta_0} \mid^2 d\omega + 2C^2 - 4C. \end{aligned} \tag{16}$$

Since $\mid [E_0(\omega)]_{C,t_0,\theta_0} \mid^2$ is independent of t_0 and θ_0, the M.M.S.E. does not depend upon t_0 and θ_0. Equation (16) can be further minimized over all possible values of C. The optimum solution, $C = 1$, results in the minimum of the M.S.E. for the class IV partial-response system equipped with a mean-square equalizer.

$$\begin{aligned} [\text{M.M.S.E.}]_{C=1} &= \frac{2}{A_1 - A_2} - 2 \\ &= \min_{\text{all } C} [\text{M.M.S.E.}]_C \end{aligned} \tag{17}$$

We may thus conclude that for the SSB class IV partial-response system, the M.M.S.E. does not depend upon t_0 and θ_0 since the constants

A_i do not depend on t_0 and θ_0. We may therefore arbitrarily choose the sampling instant and the demodulating carrier phase with no loss of optimality as long as the equalizer transversal filter length is infinite.

As an example, we assume that the equivalent baseband channel characteristics are linear in amplitude-frequency response and quadratic in delay-frequency response as shown in Fig. 2. The delay at the Nyquist frequency of π/T rad/s is taken to be $\beta_m T$ seconds. The transfer function of the channel is

$$T(\omega_c - \omega) = \left(1 - \alpha \frac{|\omega|}{\pi/T}\right) e^{-j(\beta_m T^3 \omega^3/3\pi^2)} \qquad 0 \leqq |\omega| \leqq \frac{\pi}{T}. \tag{18}$$

The M.M.S.E.s computed by equation (17) for various α, $\beta_m T$, and σ_i^2 are given in Table I. For these calculations the transmitted signal power is fixed at $4/\pi$ watts.

3.3 *Computer Results for Class IV Partial-Response System with a Finite Length Equalizer*

A particular case has been calculated by the computer for a class IV partial-response system with a finite length equalizer. The following assumptions are made:

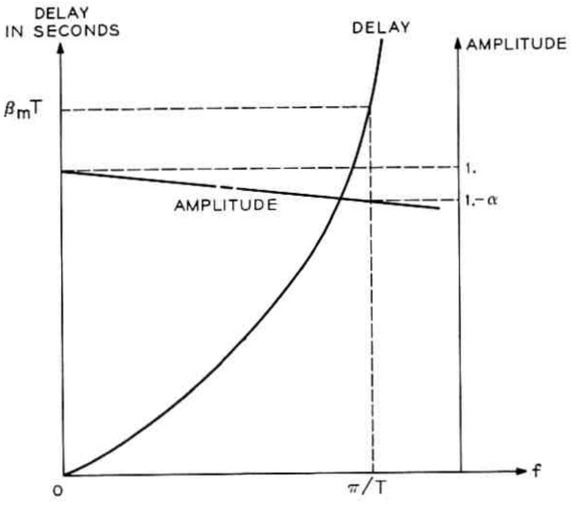

Fig. 2—Equivalent baseband channel characteristics.

TABLE I—M.M.S.E. COMPUTED BY EQUATION (17)

σ_i^2 (10^{-2} watts/Hz)	M.M.S.E. (10^{-2}) $\alpha = 0.1, \beta_m T = 1, T = 1$	M.M.S.E. (10^{-2}) $\alpha = 0.9, \beta_m T = 1, T = 1$
4	5.621	22.57
2	2.818	12.14
0.4	0.564	2.74
0.2	0.282	1.41

(*i*) The transfer functions of the channel is

$$T(\omega_c - \omega) = \left(1 - 0.1\,\frac{|\omega|}{\pi/T}\right)e^{-j(\omega^3\beta_m T^3/3\pi^2)} \qquad 0 \leqq |\omega| \leqq \pi/T. \quad (19)$$

(*ii*) The signal-to-noise ratio at the receiver input is assumed to be 21 dB.

(*iii*) The delay at the Nyquist frequency is taken to be 1 second and the baud is assumed to be 1 symbol/second.

Forty distinct combinations of sampling instants (0, 0.2, 0.4, 0.6, 0.8) and demodulating carrier phase (90°, 60°, 30°, 15°, 0°, −15°, −30°, −60°) have been tried with a 19-tap mean-square equalizer. The results and the minimum of the mean-square error are shown in Figures 3 through 10. It can be seen that the M.S.E. for most combinations is near the minimum achieved by the infinite equalizer. Practically speaking, in this example the system performance is acceptable (with error-rate upper bounded[8] by 10^{-8}) with a 19-tap mean-square equalizer for all 40 distinct combinations.

IV. MINIMIZATION OF NOISE SUBJECT TO THE CONSTRAINT THAT THE EQUALIZER FORCES THE INTERSYMBOL INTERFERENCE TO ZERO

4.1 *A General Binary Data System*

The minimization of the output noise power [see equation (5)],

$$\sigma_0^2 = \frac{\sigma_i^2 T}{\pi} \int_0^{\pi/T} \psi \text{ eq } (\omega) \cdot | E(\omega) |^2 \, d\omega,$$

subject to the constraint equations (6) and (7) can be solved through a straightforward application of the method of Lagrangian multipliers.

The expression of the optimum equalizer for sampling instant t_0 and demodulating carrier phase θ_0 is found to be

$$[E_0(\omega)]_{t_0,\theta_0} = \frac{C_1 R_d(\omega)}{Y \text{ eq } (\omega, \theta_0)} \qquad (20)$$

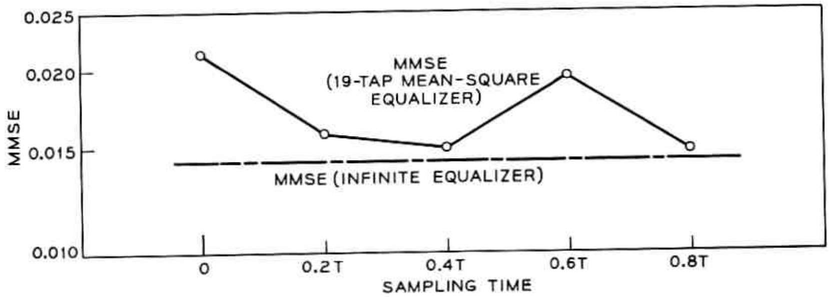

Fig. 3—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\}$ $\exp - j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = -60°$, $(\mathrm{S/N})_{\mathrm{Input}} =$ 21 dB.

where $R_d(\omega)$ is the desired received baseband equivalent signal spectrum.

It follows that the minimum output noise power is

$$[\min \sigma_0^2]_{t_0,\theta_0} = \frac{\sigma_i^2 T}{\pi} \int_0^{\pi/T} \psi \text{ eq } (\omega) C_1^2 \left| \frac{R_d(\omega)}{Y \text{ eq } (\omega,\ \theta_0)} \right|^2 d\omega. \qquad (21)$$

In general, Y eq $(\omega,\ \theta_0)$ is a function of the sampling instant, t_0, and the demodulating carrier phase, θ_0; therefore the minimum output noise power depends upon t_0 and θ_0.

4.2 *Class IV Partial-Response System*

It can be seen from equation (21) that the M.M.S.E. generally depends upon t_0 and θ_0. However, for the SSB class IV partial-

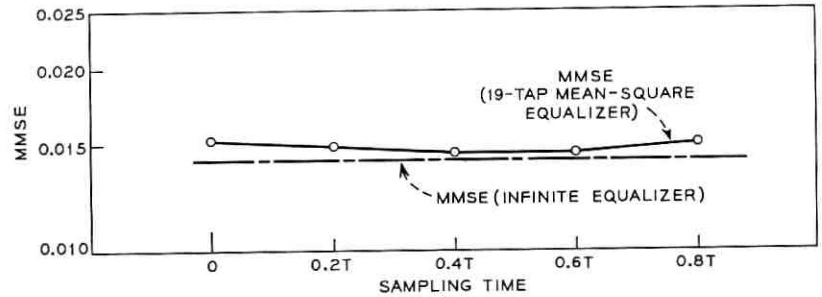

Fig. 4—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\}$ $\exp - j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = -30°$, $(\mathrm{S/N})_{\mathrm{Input}} =$ 21 dB.

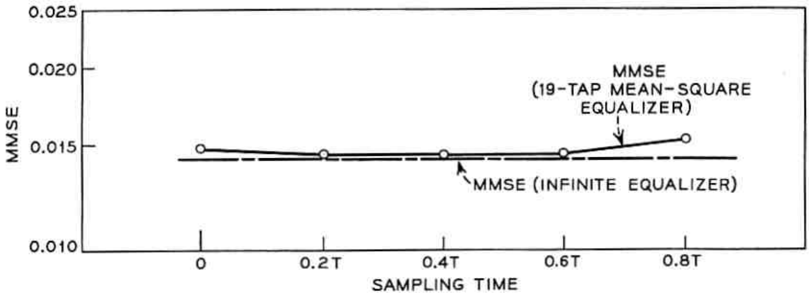

Fig. 5—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1 [|\omega|/(\pi/T)]\} \exp - j(\omega^3\beta_m T^3/3\pi^2)$; demodulating carrier phase, $\theta = -15°$, $(S/N)_{\text{input}} =$ 21 dB.

response system,

$$\left| \frac{R_d(\omega)}{Y \text{ eq } (\omega,\ \theta_0)} \right| = \left| \frac{1}{T(\omega_c - \omega)} \right|^2. \tag{22}$$

Therefore, the minimum output noise power is independent of t_0 and θ_0.

Table II shows the values of minimum output noise power computed by equations (21) and (22) for various α, $\beta_m T$, and σ_i^2 under the same assumptions made in Section III. For these calculations the transmitted signal power is fixed at $4/\pi$ watts.

Table III gives the difference in M.M.S.E. computed by equations (17) and (21).

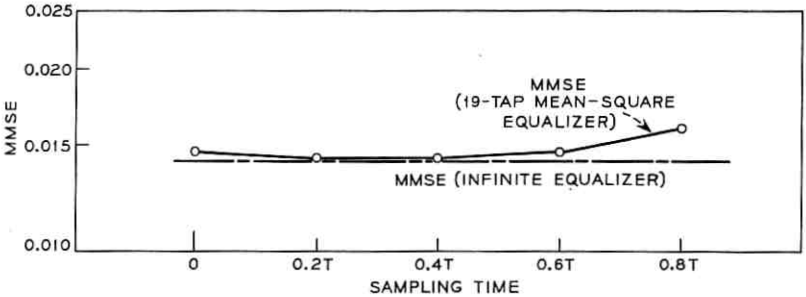

Fig. 6—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1 [|\omega|/(\pi/T)]\} \exp - j(\omega^3\beta_m T^3/3\pi^2)$; demodulating carrier phase, $\theta = 0°$, $(S/N)_{\text{input}} =$ 21 dB.

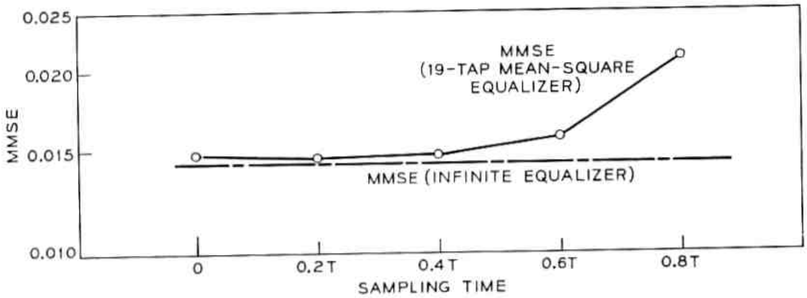

Fig. 7—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\}$ exp $- j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = 15°$, $(S/N)_{\text{Input}} =$ 21 dB.

The results show that the M.M.S.E.s computed by equations (17) and (21) are almost the same if either the signal-to-noise ratio is large or the slope of the amplitude-frequency characteristic of the channel is small (e.g., in this case the slope is 0.1). Notice that either decreasing the signal-to-noise ratio or increasing the slope of the amplitude-frequency characteristic of the channel increases the disparity of the M.M.S.E.s obtained by (17) and (21). As an example, if

$$\sigma_i^2 = 0.02 \text{ watts/Hz}$$

and

$$T(\omega_c - \omega) = \left(1 - 0.9\,\frac{|\omega|}{\pi}\right)e^{-j(\omega^3/3\pi^2)}, \tag{23}$$

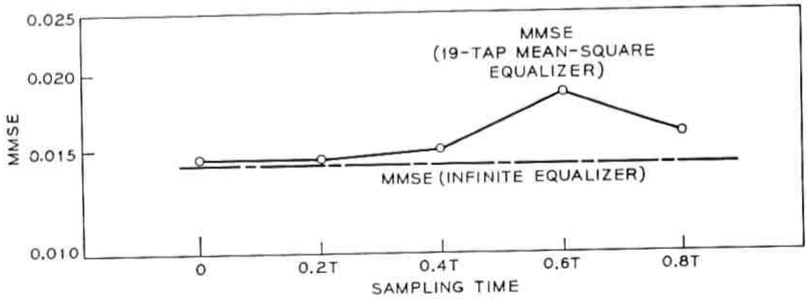

Fig. 8—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\}$ exp $- j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = 30°$, $(S/N)_{\text{Input}} =$ 21 dB.

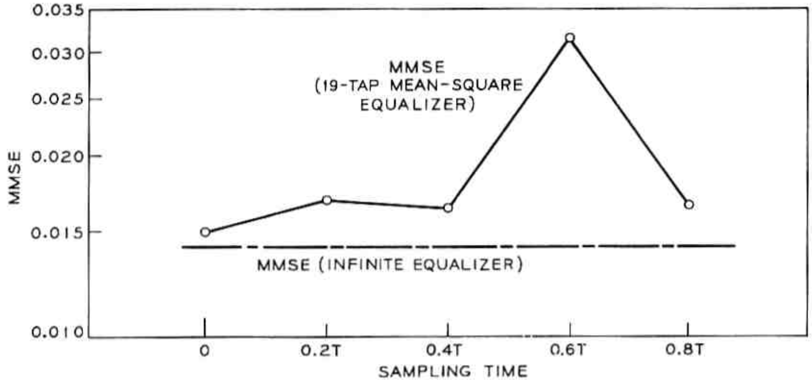

Fig. 9—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\} \exp - j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = 60°$, $(S/N)_{\text{input}} =$ 21 dB.

then the M.M.S.E.s obtained by equations (17) and (21) are 0.1214 and 0.1483 respectively. It can be seen that the M.M.S.E. is 16 percent less if the equation minimizing the mean-square intersymbol interference plus noise is used.

V. SUMMARY AND CONCLUSION

The optimum equalizer for a synchronous data system with a fixed channel is derived in this study. Two different optimality criteria are assumed: (*i*) the minimization of the output noise plus mean-square

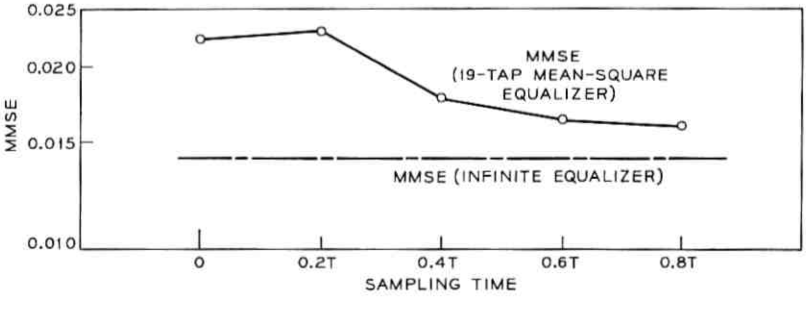

Fig. 10—M.M.S.E. versus sampling time; SSB class IV partial-response system; baseband equivalent channel transfer function, $\{1 - 0.1\ [|\omega|/(\pi/T)]\} \exp - j(\omega^3\beta_m\ T^3/3\pi^2)$; demodulating carrier phase, $\theta = 90°$, $(S/N)_{\text{input}} =$ 21 dB.

TABLE II—MINIMUM OUTPUT NOISE POWER COMPUTED BY EQUATION (21)

σ_i^2 (10^{-2} watts/Hz)	σ_0^2 (10^{-2}) $\alpha = 0.1, \beta_m T = 1, T = 1$	σ_0^2 (10^{-2}) $\alpha = 0.9, \beta_m T = 1, T = 1$
4	5.652	29.66
2	2.826	14.83
1	1.413	7.42
0.4	0.565	2.97
0.2	0.282	1.49

intersymbol interference and (*ii*) the minimization of the output noise subject to the constraint that the equalizer forces the intersymbol interference to zero.

Explicit expressions for the optimum equalizer and the corresponding M.M.S.E. are obtained. It is known that the M.M.S.E. at the equalizer output generally depends upon the sampling instant and the demodulating carrier phase. However, we have shown in this study that there exist cases where the M.M.S.E. is independent of the sampling instant and the demodulating carrier phase. The SSB class IV partial-response system represents a good example. Thus for such data systems, we may use arbitrary timing and carrier phase, thereby significantly reducing the receiver complexity and possibly the start-up time as well. The results calculated by the computer for an SSB class IV partial-response system equipped with a 19-tap mean-square equalizer show that the system error-rate for all 40 distinct combinations of sampling instants and carrier phases is less than 10^{-8}. The system is operated over a channel with linearly distorted amplitude-frequency characteristic and parabolically distorted delay-frequency characteristic (see Section III) which is worse than a worst-case C-2 line. The signal-to-noise ratio at the receiver input is assumed to be 21 dB. The results also show that with either small slope of the

TABLE III—DIFFERENCE IN M.M.S.E.

σ_i^2 (10^{-2} watts/Hz)	σ_0^2 − M.M.S.E. (10^{-2}) $\alpha = 0.1, \beta_m T = 1, T = 1$	σ_0^2 − M.M.S.E. (10^{-2}) $\alpha = 0.9, \beta_m T = 1, T = 1$
4	0.031	7.08
2	0.08	2.69
1	0.002	0.97
0.4	≈0	0.23
0.2	≈0	0.08

amplitude-frequency characteristic of the channel or large signal-to-noise ratio, the M.M.S.E.s obtained by the two different criteria considered in this study are almost the same. For example, with white-noise spectral density 0.02 watts/Hz (S/N at the receiver input is 18 dB) and the Fourier transform of the channel

$$\left(1 - 0.1\frac{|\omega|}{\pi}\right)e^{-j(\omega^3/3\pi^2)},$$

the M.M.S.E.s obtained by criteria (*i*) and (*ii*) are 0.02818 and 0.02826 respectively. However, either increasing the slope or decreasing the signal-to-noise ratio increases the disparity of the M.M.S.E.s obtained by criteria (*i*) and (*ii*). Under these situations, criterion (*i*) is much preferred. For example, with the same white-noise spectral density as before and the Fourier transform of the channel

$$\left(1 - 0.9\frac{|\omega|}{\pi}\right)e^{-j(\omega^3/3\pi^2)},$$

the M.M.S.E.s obtained by criteria (*i*) and (*ii*) are 0.121 and 0.148 respectively.

VI. ACKNOWLEDGMENTS

The author wishes to thank J. E. Mazo, J. Salz, and D. A. Spaulding for their helpful discussions.

APPENDIX

Minimization of Noise Plus Intersymbol Interference (Binary and Partial-Response Systems)

The dètails of the minimization procedure of noise-plus-intersymbol interference for the binary and partial-response systems will be given in this Appendix.

The block diagram of a general digital data system is shown in Fig. 1. The characteristics of the ideal low-pass filter and the equalizer are assumed to be

$$F_1(\omega) = \begin{cases} 1 & 0 \leqq |\omega| \leqq \omega_m, \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

and

$$E(\omega) = \sum_{n=-\infty}^{\infty} C_n e^{j\omega nT}.$$

With input white Gaussian noise having one-side power spectral density $T\sigma_i^2$ watts/Hz, the variance of the noise at the receiver equalizer output is,

$$
\begin{aligned}
\sigma_0^2 &= E(n_0^2) \\
&= \frac{T\sigma_i^2}{2\pi} \int_{-\omega m}^{\omega m} \{| R(\omega - \omega_c) |^2 + | R(\omega_c - \omega) |^2\}^{*} \cdot | E(\omega) |^2 \, d\omega \\
&= \frac{T\sigma_i^2}{2\pi} \int_{-\omega m}^{\omega m} \psi(\omega) \mid E(\omega) \mid^2 d\omega \\
&= \frac{\sigma_i^2 T}{2\pi} \left\{ \sum_{K=-N+1}^{N-1} \int_{(\pi/T)(2K-1)}^{(\pi/T)(2K+1)} \psi(\omega) \mid E(\omega) \mid^2 d\omega \right. \\
&\quad \left. + \int_{-\omega m}^{-(\pi/T)(+2N+1)} \psi(\omega) \mid E(\omega) \mid^2 d\omega + \int_{(\pi/T)(2N-1)}^{\omega m} \psi(\omega) \mid E(\omega) \mid^2 d\omega \right\} \\
&= \frac{T\sigma_i^2}{\pi} \int_0^{\pi/T} \psi \text{ eq } (\omega) \cdot | E(\omega) |^2 \, d\omega,
\end{aligned}
\tag{25}
$$

where

$$
\psi \text{ eq } (\omega) = \psi(\omega) + \psi\left(\omega + \frac{2\pi}{T}\right) + \cdots + \psi\left(\omega + \frac{2N\pi}{T}\right), \tag{26}
$$

$$
\frac{(2N - 1)\pi}{T} \leqq \omega_m \leqq \frac{(2N + 1)\pi}{T}.
$$

Similarly,

$$
\begin{aligned}
y(t, \theta) &= \frac{T}{2\pi} \int_{-\omega m}^{\omega m} [S(\omega - \omega_c)T(\omega - \omega_c)R(\omega - \omega_c)e^{i\theta} \\
&\qquad + S(\omega_c + \omega)T(\omega_c + \omega)R(\omega_c + \omega)e^{-i\theta}]E(\omega)e^{i\omega t} \, d\omega \\
&= \frac{T}{2\pi} \int_{-\omega m}^{\omega m} Y(\omega, \theta)E(\omega)e^{i\omega t} \, d\omega \\
&= \frac{T}{2\pi} \left[\int_{-\omega m}^{-(\pi/T)(2N+1)} Y(\omega, \theta)E(\omega)e^{i\omega t} \, d\omega \right. \\
&\qquad + \sum_{K=-N+1}^{N-1} \int_{(\pi/T)(2K-1)}^{(\pi/T)(2K+1)} Y(\omega, \theta)E(\omega)e^{i\omega t} \, d\omega \\
&\qquad \left. + \int_{\pi/T(2N-1)}^{\omega m} Y(\omega, \theta)E(\omega)e^{i\omega t} \, d\omega \right] \\
&= \operatorname{Re} \frac{T}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega, \theta)E(\omega)e^{i\omega t} \, d\omega,
\end{aligned}
\tag{27}
$$

* $R(\omega)$ is the receiver filter transfer function.

where

$$Y \text{ eq } (\omega, \theta) = Y(\omega, \theta) + Y\left(\omega + \frac{2\pi}{T}, \theta\right) + \cdots + Y\left(\omega + \frac{2N\pi}{T}, \theta\right). \tag{28}$$

By Parseval's theorem we can write

$$\sum_n y_n^2(\theta) = \frac{T}{\pi} \int_0^{\pi/T} \mid Y \text{ eq } (\omega, \theta) \mid^2 \cdot \mid E(\omega) \mid^2 d\omega. \tag{29}$$

Therefore the normalized M.M.S.E. given by equation (4) can be rewritten as

$$[\text{M.S.E.}]_{t_0, \theta_0} = \frac{\left[\frac{T\sigma_i^2}{\pi} \int_0^{\pi/T} \psi \text{ eq } (\omega) \cdot \mid E(\omega) \mid^2 d\omega + B\right]}{y_0^2(\theta_0)}, \tag{30a}$$

where

$$B = \frac{T}{\pi} \int_0^{\pi/T} \mid Y \text{ eq } (\omega, \theta_0) \mid^2 \cdot \mid E(\omega) \mid^2 d\omega - y_0^2(\theta_0). \tag{30b}$$

Since $y_0(\theta_0)$ is fixed, minimizing $[\text{M.S.E.}]_{t_0, \theta_0}$ subject to the constraint equation (6) is equivalent to minimizing the following function,

$$V = \frac{T}{\pi} \int_0^{\pi/T} [\sigma_i^2 \psi \text{ eq } (\omega) + \mid Y \text{ eq } (\omega, \theta_0) \mid^2] \cdot \mid E(\omega) \mid^2 d\omega, \tag{31}$$

subject to the same constraint equation.

The minimization problem can be solved through a straightforward application of the method of Lagrangian multipliers.

Solving

$$\frac{\partial V}{\partial E(\omega)} = \frac{\partial}{\partial E(\omega)} \left\{ \text{Re} \frac{T}{\pi} \int_0^{\pi/T} \{ [\sigma_i^2 \psi \text{ eq } (\omega) + \mid Y \text{ eq } (\omega, \theta_0) \mid^2] \cdot \mid E(\omega) \mid^2 + \lambda Y \text{ eq } (\omega, \theta_0) E(\omega) e^{i\omega t_0} \} \, d\omega \right\} = 0, \tag{32}$$

and

$$C_1 = \text{Re} \frac{T}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega, \theta_0) E(\omega) e^{i\omega t_0} \, d\omega, \tag{33}$$

we obtain the expression for the optimum equalizer, $E_0(\omega)$, at the sampling instant, t_0, and the demodulating carrier phase, θ_0,

$$[E_0(\omega)]_{t_0,\theta_0} = \frac{\{Y \text{ eq } (\omega,\ \theta_0) e^{i\omega t_0}\}^*}{\{\sigma_i^2 \psi \text{ eq } (\omega) + |\ Y \text{ eq } (\omega,\ \theta_0)\ |^2\}} \cdot \frac{C_1}{\dfrac{T}{\pi} \displaystyle\int_0^{\pi/T} \frac{|\ Y \text{ eq } (\omega,\ \theta_0)\ |^2}{\sigma_i^2 \psi \text{ eq } (\omega) + |\ Y \text{ eq } (\omega,\ \theta_0)\ |^2}\, d\omega}, \tag{34}$$

where $\{X\}^*$ means complex conjugate of X. Substituting $[E_0(\omega)]_{t_0,\theta_0}$ into equation (4), we obtain the M.M.S.E.

$$[\text{M.M.S.E.}]_{t_0,\theta_0} = \frac{1}{\dfrac{T}{\pi} \displaystyle\int_0^{\pi/T} \frac{|\ Y \text{ eq } (\omega,\ \theta_0)\ |^2}{\sigma_i^2 \psi \text{ eq } (\omega) + |\ Y \text{ eq } (\omega,\ \theta_0)\ |^2}\, d\omega} - 1. \tag{35}$$

Equation (35) can be further minimized over all possible sampling instants and carrier phases to obtain a global minimum of mean square error.

We now consider the class IV partial-response system. The transfer function of the equivalent baseband transmitted signal and receiver filter are

$$S(\omega_c - \omega) = R(\omega_c - \omega) = \begin{cases} \sqrt{2T \sin |\ \omega\ | T}\, e^{-i(\omega/|\omega|)\pi/4} & 0 \leqq |\ \omega\ | \leqq \dfrac{\pi}{T}, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

It follows that

$$Y \text{ eq } (\omega,\ \theta) = \begin{cases} 2T \sin \omega T e^{-i(\pi/2+\theta)\omega/|\omega|} \cdot T(\omega_c - \omega) & 0 \leqq |\ \omega\ | \leqq \dfrac{\pi}{T}, \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

The constraint equations are assumed to be

$$C = y_1(\theta_0) = \text{Re} \frac{1}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega,\ \theta_0) \cdot E(\omega) \cdot e^{i\omega(T+t_0)}\, d\omega, \tag{38a}$$

and

$$C - 2 = y_{-1}(\theta_0) = \text{Re} \frac{1}{\pi} \int_0^{\pi/T} Y \text{ eq } (\omega,\ \theta_0) \cdot E(\omega) \cdot e^{-i\omega(T-t_0)}\, d\omega, \tag{38b}$$

where C is a constant.

We now wish to minimize the function

$$U = \mathrm{Re}\,\frac{1}{\pi}\int_0^{\pi/T} \{[\sigma_i^2 2T \mid \sin \omega T \mid + 4T^2 \mid \sin \omega T \mid^2 \cdot \mid T(\omega_c - \omega) \mid^2] \cdot \mid E(\omega) \mid^2 + \lambda_1 Y \text{ eq } (\omega, \theta_0) E(\omega) e^{j\omega(T+t_0)} + \lambda_2 Y \text{ eq } (\omega, \theta_0) E(\omega) e^{-j\omega(T-t_0)}\}\, d\omega \tag{39}$$

subject to the constraint equations (38a) and (38b). The expression for the optimum equalizer for a given constant C, sampling instant t_0, and carrier phases θ_0 is

$$[E_0(\omega)]_{C,t_0,\theta_0} = \frac{\left[\begin{array}{c} \dfrac{C(A_1 - A_2) + 2A_2}{A_1^2 - A_2^2}\{Y \text{ eq } (\omega, \theta_0) e^{j\omega(T+t_0)}\}^* \\ + \\ \dfrac{C(A_1 - A_2) - 2A_1}{A_1^2 - A_2^2}\{Y \text{ eq } (\omega, \theta_0) e^{-j\omega(T-t_0)}\}^* \end{array}\right]}{\sigma_i^2 \cdot 2T \mid \sin \omega T \mid + 4T^2 \sin^2 \omega T \cdot \mid T(\omega_c - \omega) \mid^2}$$

$$0 \leqq \mid \omega \mid \leqq \frac{\pi}{T}, \tag{40}$$

where

$$A_1 = \mathrm{Re}\,\frac{1}{\pi}\int_0^{\pi/T} \frac{2T \sin \omega T \cdot \mid T(\omega_c - \omega) \mid^2}{\sigma_i^2 + 2T \sin \omega T \cdot \mid T(\omega_c - \omega) \mid^2}\, d\omega, \tag{41a}$$

and

$$A_2 = \mathrm{Re}\,\frac{1}{\pi}\int_0^{\pi/T} \frac{2T \sin \omega T \cdot \mid T(\omega_c - \omega) \mid^2 \cdot e^{j\omega 2T}}{\sigma_i^2 + 2T \sin \omega T \cdot \mid T(\omega_c - \omega) \mid^2}\, d\omega. \tag{41b}$$

It follows that the M.M.S.E. is

$$[\text{M.M.S.E.}]_{C,t_0,\theta_0} = \frac{1}{\pi}\int_0^{\pi/T} [\sigma_i^2 2T \mid \sin \omega T \mid + 4T^2 \mid \sin \omega T \mid^2 \cdot \mid T(\omega_c - \omega) \mid^2] \cdot \mid [E_0(\omega)]_{C,t_0,\theta_0} \mid^2 d\omega + 2C^2 - 4C. \tag{42}$$

Since $\mid Y \text{ eq } (\omega, \theta_0) \mid^2$ is independent of t_0 and θ_0, hence $\mid [E_0(\omega)]_{C,t_0,\theta_0} \mid^2$ and $[\text{M.M.S.E.}]_{C,t_0,\theta_0}$ do not depend upon t_0 and θ_0.

Equation (42) can be further minimized with respect to C by solving

$$\frac{\partial}{\partial C}[\text{M.M.S.E.}]_{C,t_o,\theta_o} = 0. \tag{43}$$

The optimum solution $C = 1$, provides the M.M.S.E. for a class IV partial-response system.

$$\begin{aligned}[\text{M.M.S.E.}]_{C=1} &= \frac{2}{A_1 - A_2} - 2 \\ &= \min_{\text{all } C}\,[\text{M.M.S.E.}]_C\,. \end{aligned} \tag{44}$$

In the absence of channel noise, $\sigma_i = 0$, then

$$A_1 = 1, \tag{45}$$

and

$$A_2 = 0. \tag{46}$$

Hence,

$$[\text{M.M.S.E.}]_{C=1} = 0. \tag{47}$$

REFERENCES

1. Lucky, R. W., "Automatic Equalization for Digital Communication," B.S.T.J., *44,* No. 4 (April 1965), pp. 547–588.
2. Lucky, R. W., "Techniques for Adaptive Equalization of Digital Communication," B.S.T.J., *45,* No. 2 (February 1966), pp. 255–286.
3. Lucky, R. W., and Rudin, H. R., "An Automatic Equalizer for General Purpose Communication Channels," B.S.T.J . *46,* No. 9 (November 1967), pp. 2179–2208.
4. DiToro, M. J., "A New Method of High-Speed Adaptive Serial Communication Through Any Time-Variable and Dispersive Transmission Medium," Conf. Record, 1965 IEEE Annual Commun. Conv., pp. 763–767.
5. Coll, D. D. and George, D. A., "The Reception of Time-Dispersed Pulses," Conf. Record, 1965 IEEE Ann. Commun. Conv., pp. 749–752.
6. Chang, R. W., "Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication," B.S.T.J., *49,* No. 6 (July-August 1970), pp. 1069–1104.
7. Kretzmer, E. R., Generalization of a Technique for Binary Data Communication, IEEE Trans. on Commun. Tech., COM-14, No. 1 (February 1966), pp. 67–68.
8. Saltzberg, B. R., "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," IEEE Trans. Information Theory, *IT-14,* No. 4 (July 1968), pp. 563–568.

Contributors to This Issue

VÁCLAV E. BENEŠ, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, frequency modulation, traffic theory, servomechanisms, and stochastic control. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963), and of *Mathematical Theory of Connecting Networks and Telephone Traffic* (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, Mind Association, Phi Beta Kappa.

DAVID A. BERKLEY, B.S.E.E., 1961, and Ph.D. (Applied Physics), 1966, Cornell University; Research Associate in Medical Physics, Chalmers University, Gothenburg, Sweden, 1966–1968; Bell Telephone Laboratories, 1968—. Mr. Berkley is working on problems related to processing of speech for room and transmission environments, and also is involved in investigations of inner ear mechanics. Member, Acoustical Society of America, N. Y. Academy of Sciences.

PETER D. BRICKER, B.A., 1950, Bucknell University; M.A., 1952, and Ph.D., 1954, The Johns Hopkins University; Bell Telephone Laboratories, 1955—. Mr. Bricker has been concerned with human performance, perception, and judgment, particularly with regard to speech communication behavior. He is currently studying listener perception of potential electronic calling signals. Member, Acoustical Society of America, American Psychological Association, Phi Beta Kappa.

JAMES W. CARLIN, B.S.E.E., 1962, Illinois Institute of Technology; M.S.E.E., 1964, and Ph.D., 1967, University of Illinois; Bell Telephone Laboratories, 1968—. Mr. Carlin has been concerned with the electromagnetic pulse effects of nuclear bursts and with long-haul communications in millimeter waveguide.

Peter D'Agostino, B.E.E., 1965, Pratt Institute; M.S. (Electrical Engineering), 1967, New York University; Bell Telephone Laboratories, 1969—. Mr. D'Agostino's first assignment was determining the effects of the electromagnetic pulse associated with a thermonuclear blast. He is now involved in the design and evaluation of long-haul waveguide communications.

James L. Flanagan, B.S., 1948, Mississippi State University; S.M., 1950, and Sc.D., 1955, Massachusetts Institute of Technology. Faculty of Electrical Engineering, Mississippi State University, 1950–1952; Air Force Cambridge Research Center, 1954–1957. Bell Telephone Laboratories, 1957—. Mr. Flanagan has worked in speech and hearing research, computer simulation and digital encoding, and acoustics research. He is Head, Acoustics Research Department. Fellow, IEEE; Fellow, Acoustical Society of America; Tau Beta Pi; Sigma Xi; member of several government and professional society boards, including committees of the National Academy of Sciences and the National Academy of Engineering.

Richard D. Gitlin, B.E.E., 1964, City College of New York; M.S., 1965, and D.Eng.Sc., 1969, Columbia University; Bell Telephone Laboratories, 1969—. Mr. Gitlin is presently concerned with problems in data transmission. Member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

E. Y. Ho, B.S.E.E., 1964, The National Taiwan University; Ph.D., 1969, University of Pennsylvania; Bell Telephone Laboratories, 1969—. Mr. Ho has been engaged in developing and analyzing automatic equalizers for data transmission systems. Member, IEEE.

Henry J. Landau, A.B., 1953, Harvard College; A.M., 1955, and Ph.D., 1957, Harvard University; Bell Telephone Laboratories, 1957—; Institute for Advanced Study, Princeton, N. J., 1959–60, and Spring, 1967. Mr. Landau's main interest is harmonic analysis.

O. M. Mracek Mitchell, B.A., 1955, M.A., 1958, and Ph.D. (Nuclear Physics), 1962, University of Toronto; Ontario Research Foundation, 1962–63; Bell Telephone Laboratories, 1963—. Since joining Bell Laboratories, she has investigated ultrasonic loss mechanisms and interactions in materials at low temperatures, and is currently engaged in fundamental studies in acoustics and signal processing. Member, American Physical Society, Acoustical Society of America.

David C. Opferman, B.S.E.E., 1961, Pennsylvania State University; M.S.E.E., 1965, and Ph.D., 1967, University of Pittsburgh; Westinghouse Electric Corporation, 1961–1964; Bell Telephone Laboratories, 1967—. Mr. Opferman is interested in switching networks, and he is currently engaged in exploratory development of stored program processors and high-level languages for Business Communication Systems. Member, IEEE, Eta Kappa Nu, Sigma Tau, Tau Beta Pi.

Lawrence R. Rabiner, S.B. and S.M., 1964, and Ph.D. (E.E.), 1967, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1962–1964, 1967—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Since 1967, he has been engaged in research on speech communication, signal analysis, digital filtering, and techniques for waveform processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, IEEE; Fellow, Acoustical Society of America. He is secretary of the IEEE Technical Committee on Digital Signal Processing, and member of the technical committees on speech communication of both the IEEE and the Acoustical Society.

J. Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Telephone Laboratories, 1961—. Mr. Salz first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems, and is currently Supervisor of the data theory group in the data communications technology laboratory. During the academic year 1967–68 he was on leave as Professor of Electrical Engineering at the University of Florida. Member, IEEE, Sigma Xi.

Ronald W. Schafer, B.S. (E.E.), 1961, and M.S. (E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1968—. Mr. Schafer has been engaged in research on digital waveform processing techniques and speech communication. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America.

David Slepian, University of Michigan, 1941–43; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—. Mr. Slepian has been engaged in mathematical research in communication theory and noise theory, as well as in a variety of aspects of applied mathematics. During the academic year 1958–59, he was a Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley and during the Spring semesters of 1967 and 1970 he was a Visiting Professor of Electrical Engineering at the University of Hawaii. He was Editor of the Proceedings of the IEEE during 1969 and 1970. Member, AAAS, American Mathematical Society, SIAM. Fellow, IEEE, Institute of Math. Statistics.

Nelson T. Tsao-Wu, B.Sc. (Eng.), 1957, University of London; M.S., 1965, and Ph.D., 1968, Northeastern University; Bell Telephone Laboratories, 1968—. Mr. Tsao-Wu is currently engaged in applied research in switching networks, coding theory, and Customer Telephone Systems. Member, IEEE, Phi Kappa Phi, Sigma Xi.

B.S.T.J. BRIEF

A Low-Noise Metal-Semiconductor-Metal (MSM) Microwave Oscillator

By D. J. COLEMAN, JR., and S. M. SZE

(Manuscript received January 22, 1971)

I. INTRODUCTION

Low-noise microwave CW oscillations have been obtained from metal-semiconductor-metal (MSM) structures made from a 10-μm thin slice of silicon sandwiched between two PtSi Schottky barrier contacts. Microwave CW power up to 50 mW has been obtained at 5 GHz with efficiency up to 1.8 percent. The FM noise measure 1 MHz from the carrier is 22.8 dB which is considerably lower than that of a silicon avalanche oscillator. The mechanisms responsible for the microwave oscillation are (*i*) the exponential increase of the local carrier population due to injection of minority carriers at the forward-biased contact and (*ii*) the transit-time delay of injected carriers traversing the depletion region. By optimizing material and device parameters, it is believed that higher efficiency and higher power microwave oscillations can be obtained from the MSM and its related structures with the inherent low-noise characteristics.

II. DEVICE FABRICATION

Single-crystal n-type silicon wafers with 11Ω-cm resistivity (4×10^{14} cm^{-3} doping), $\langle 111 \rangle$ oriented, and with a dislocation density less than 100/cm^2 were Syton polished on both sides to a final thickness of 10 ± 2 μm. Platinum of 500 Å thickness was sputtered onto both sides of the wafer and was sintered to form approximately 1000 Å PtSi on both sides. Chromium of 300 Å was deposited on one side; this was followed by a 3000 Å layer of Au evaporation. The same depositions were then made on the other side. A standard photolithographic method was used to define circular patterns of gold dots with areas of 5×10^{-4} cm^2. The devices were separated by etching and were mounted onto V-type microwave packages.

III. DC CHARACTERISTICS

A schematic diagram of an MSM structure is shown in Fig. 1a. The band diagram at thermal equilibrium is shown in Fig. 1b for an n-type semiconductor where ϕ_{n1} and ϕ_{n2} are the barrier heights for the two metal-semiconductor contacts respectively. For the PtSi-Si-PtSi structure mentioned previously, $\phi_{n1} = \phi_{n2} = 0.85$ eV. Figure 1c shows the energy band diagram when a voltage is applied. We have electron current from the reverse-biased contact and hole current from the forward-biased contact.

The measured I-V characteristics at 300°K and 77°K of a representative device is shown in Fig. 2. The rapid increase in terminal current with applied voltage (above 30 volts) is caused by thermionic

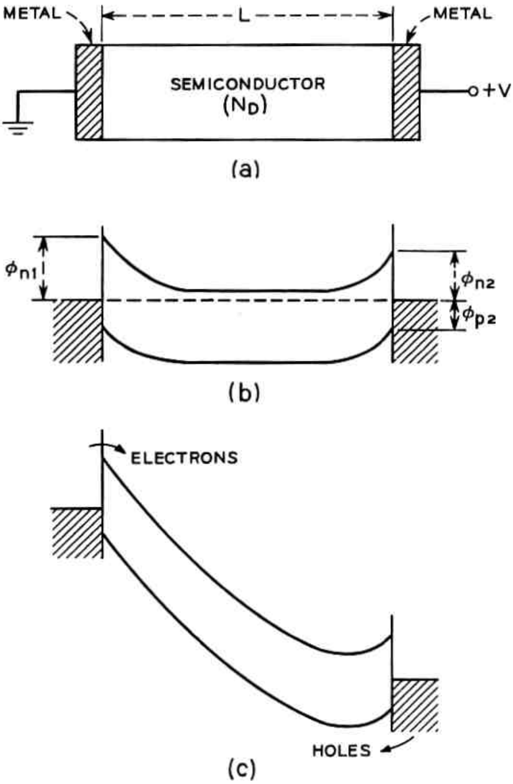

Fig. 1—(a) Schematic diagram of a metal-semiconductor-metal (MSM) structure. (b) Energy band diagram of an MSM structure in thermal equilibrium. (c) Energy band diagram of an MSM structure under biasing condition.

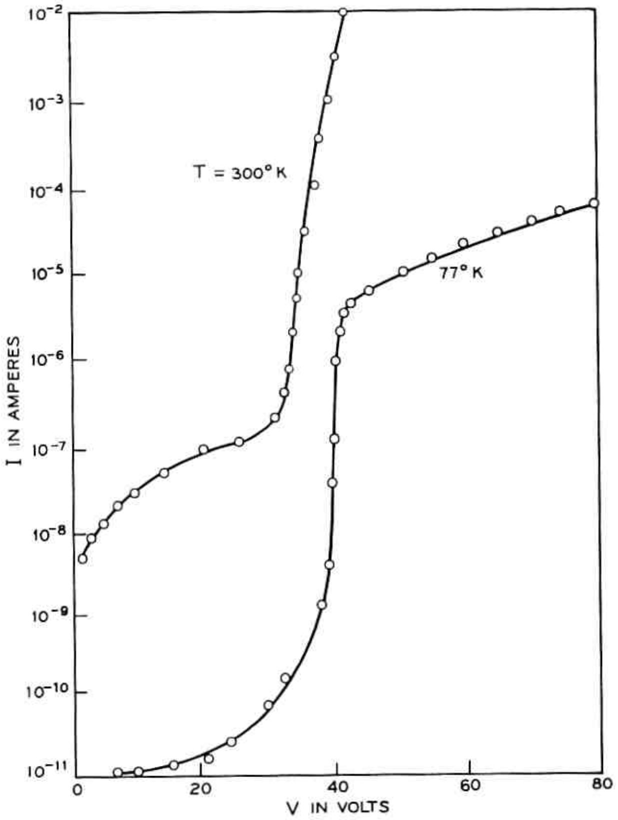

Fig. 2—Measured current vs voltage of a silicon MSM structure (PtSi-Si-PtSi) at two temperatures. The device parameters are $L = 10$ μm, $N_D = 4 \times 10^{14}$ cm^{-3}, $\phi_{n1} = \phi_{n2} = 0.85$ eV, and with an area of 5×10^{-4} cm^2.

hole injection into the semiconductor as the depletion layer of the reverse-biased contact reaches through the entire device thickness. This critical voltage is approximately given by $qNL^2/2\epsilon_s$, where N is the doping concentration, L the semiconductor thickness, and ϵ_s the dielectric permittivity.[1] The current increase is not due to avalanche multiplication as is apparent from the magnitude of the critical voltage and its negative temperature coefficient. At 77°K, the rapid increase is terminated at a current of about 10^{-5} amps. This saturated current is expected from the thermionic emission theory of hole injection[1] from the forward-biased contact with a hole barrier height (ϕ_{p2}) of about 0.15 eV.

IV. MICROWAVE PERFORMANCE

CW microwave performance of the MSM devices was measured in a coaxial Impatt circuit described by D. E. Iglesias.[2] Microwave power was obtainable over the entire C band of 4–8 GHz. The maximum power observed was 50 mW at 4.9 GHz. The maximum efficiency approached 1.8 percent. Figure 3 shows some of the measured microwave power versus current with frequency of operation indicated on each curve for three typical devices tested. The voltage indicated in parenthesis labeling each curve is the average bias voltage at the diode while oscillating. Because of the symmetry of the structure, it could be operated with either polarity of bias voltage, and similar results were obtained.

The highest-power unit was tested for FM noise when tuned to a frequency of 4.88 GHz. The FM single-sideband noise measure 1 MHz from the carrier frequency was found to be 22.8 dB at 7 mA bias

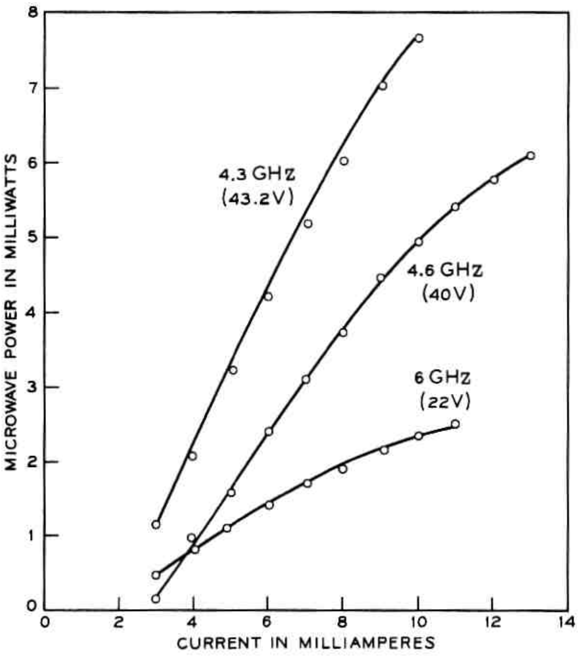

Fig. 3—CW microwave output vs input current for three Si MSM devices. Also indicated are the operating frequency and the average bias voltage while oscillating.

current. This noise measure is considerably lower than that of a silicon Impatt diode and is comparable to that of a GaAs transfer-electron oscillator.

The 6-GHz diode was used to build a stable negative conductance linear amplifier. A gain-bandwidth product of 200 MHz was obtained with 19 dB gain at 5 mA bias. The small signal noise measure was 15 ± 1 dB.

The mechanisms responsible for the microwave oscillations are believed to be (*i*) the rapid increase of carrier injection process caused by the decreasing potential barrier of the forward-biased metal-semiconductor contact and (*ii*) an apparent $3\pi/2$ transit angle of the injected carriers which traverse the semiconductor depletion region. For the 6-GHz diode, the thickness L is smaller. This results in higher frequency (since frequency is inversely proportional to L) and lower critical voltage (which is proportional to L^2). Since the main noise source for thermionic emission processes is the shot noise, one would expect a low noise measure. This is indeed observed experimentally.

If a large barrier height can be obtained for a p-type semiconductor, one can make a complementary MSM structure in the same way as described here. Since the reverse-biased metal-semiconductor contact serves mainly as a blocking contact until the reach-through voltage is obtained, it is conceivable that this contact can be replaced by a p-n junction such as p^+-n-metal structure. By optimizing the material parameters (such as doping profile, barrier heights, and semiconductor thickness) and device geometry and topology,[3] it is believed that higher efficiency and higher power microwave oscillations can be obtained from the MSM and its related structures with the inherent low-noise characteristics.

V. ACKNOWLEDGMENTS

We gratefully acknowledge the expert assistance of A. Loya who fabricated the devices. We wish also to thank A. R. Tretola for platinum deposition, J. Federico for FM noise measurements, J. M. Morris for linear amplifier noise measurements, and R. M. Ryder for many helpful discussions.

REFERENCES

1. Sze, S. M., Coleman, D. J., Jr., and Loya, A., "Current Transport in Metal-Semiconductor-Metal Structures," to be published in Solid State Electronics.
2. Iglesias, D. E., "Circuit for Testing High Efficiency Impatt Diodes," Proc. IEEE, *55* (1967), pp. 2065–67.
3. Coleman, D. J., Jr., and Sze, S. M., "A Negative Resistance Diode Circuit," U.S. Patent pending.