THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLV MAY-JUNE 1966 NUMBER 5

Electron Beam Heating of a Thin Film on a Highly Conducting Substrate

By J. A. MORRISON and S. P. MORGAN

(Manuscript received January 27, 1966)

An analysis is made of the steady-state temperature distribution in a poorly conducting plane film on a highly conducting semi-infinite substrate, owing to a time-independent heat input in a cylindrical region of the film and substrate. The problem is of interest in connection with the localized hardening of anodic oxide films on silicon by electron beam bombardment in order to produce oxide diffusion masks for the manufacture of integrated circuits. A formal solution is obtained for arbitrary dependence of the heat input on radius and depth, and a detailed study is made of a particular case in which the heat input is independent of radius across the beam, and varies in a realistic manner with depth in the film. Approximate formulas are given for the temperature in the film when the radius of the beam is large compared to the thickness of the film, and also when the conductivity of the film is small compared to the conductivity of the substrate. The approximate formulas are compared with the results of calculations based on the exact solution. Finally, a crude estimate is made of the time required to reach the steady state.

I. INTRODUCTION AND SUMMARY

Recently considerable interest has developed in the application of electron beam technology to microelectronics.[1] A number of papers have been concerned with the heat-flow problems encountered when a high-power electron beam interacts with a target. Previous investigations have considered electron heating of a uniform semi-infinite target,[2]

of a target consisting of a highly conductive metal film on a less conductive substrate,[3] and of a thin film not supported by a substrate.[4] The heating of a poorly conductive film on a highly conductive substrate does not appear to have been treated before, and forms the topic of the present investigation. It corresponds to the case of an electron beam incident upon an oxidized silicon substrate.

This analysis may find an application in the fabrication of oxide diffusion masks for integrated circuits by electron beam bombardment.[5] The etch rate of an anodic oxide film on silicon in hydrofluoric acid has been shown to decrease strongly under electron bombardment, and a proposal for producing patterns is to harden some areas of the film and then to remove the surrounding oxide with dilute HF.

It should be noted that electron beam bombardment produces radiation damage as well as thermal effects. The radiation damage alone would increase the etch rate of the film in HF, but in conjunction with high temperature it also facilitates ionic rearrangement in the SiO_2 film during irradiation, which leads to a decrease in etch rate. The latter effect predominates by far in the case of anodic SiO_2 films, so that the net result is a strong decrease in etch rate.

While the rise in temperature during irradiation is thus not the only factor contributing to the "hardening" of the oxide film, it is still the major factor, and a knowledge of the temperature distribution during irradiation is highly desirable. On the one hand one is interested in working at a high temperature in order to increase the rate of oxide "hardening"; on the other hand one must stay below the melting point of silicon (1415°C), or perhaps even lower in order not to generate excessive thermal stresses in the silicon. The edge definition of the hardened region in the oxide film is also of paramount interest for mask fabrication. Effects due to radiation damage will not be considered here, but it may be noted that radiation damage will be generated exclusively in the oxide and not in the silicon at the accelerating voltages of interest (less than 10 kv).

In this paper we consider the mathematical problem of calculating the steady-state temperature distribution due to an axially symmetric, time-independent heat input throughout a cylindrical volume of the film and substrate. The thermal properties of both materials are assumed independent of temperature, and radiation from the outer surface of the film is neglected. A formal solution of the problem is given in Section II for an arbitrary dependence of the heat input on radius and depth; but in the subsequent analysis we assume that at any given depth

the heat input is independent of radius across the beam and zero outside the beam. We also confine our attention to the temperature distribution in the film itself. The only thing we really need to know about the temperature in the substrate is that its maximum, which occurs on the axis at the film-substrate interface, is not high enough to melt the substrate.

With a fixed distribution of input heat, the normalized temperature distribution in the film depends on two dimensionless parameters, namely the ratio of beam radius to film thickness and the ratio of film conductivity to substrate conductivity. In the physical problem, the beam radius may be several times the film thickness, and the conductivity of the oxide film is between a tenth and a hundredth of the conductivity of the silicon substrate. In Section III an asymptotic approximation is given for the temperature distribution when the normalized beam radius is large. Section IV contains the solution for a perfectly conducting substrate, as well as an estimate of the first-order effect of finite but large substrate conductivity.

In order to calculate the temperature distribution numerically, it is necessary to assume a definite dependence of heat input on distance into the film ("depth-dose function"). In Section V we assume a depth-dose function which approximates the form determined empirically by Grün[6] and also employed by Wells.[7] The parameters are adjusted so that the power input is maximum at a depth equal to 40 per cent of the film thickness and zero at the bottom of the film, since, in general, one wishes to avoid direct heating of the substrate by the electron beam. Contour plots of normalized temperature have been calculated from the formulas of Section II for selected beam diameters and conductivity ratios. In addition, the exact temperature distributions along the axis and at the top and bottom of the oxide film are compared with the approximate formulas of Section III.

As a typical numerical result, we find that for an SiO_2 film of thickness 0.5 micron, bombarded by a 5 kv electron beam of diameter 20 microns with a current of 628 μa and a uniform power density of 10^6 watts/cm^2, the steady-state temperature rise on the axis is about 1800°C at the surface of the film, and about 800°C at the surface of the silicon substrate.

In Section VI a crude estimate is made of the time required to reach the steady-state temperature after the electron beam is instantaneously switched on. It appears that in an example such as the preceding, the transient time would be of the order of a few tenths of a microsecond.

II. STEADY-STATE TEMPERATURE DISTRIBUTION

The geometry of the problem to be considered is shown in Fig. 1. A plane film of thermal conductivity K_1 fills the region $0 \leqq z \leqq c$, and overlies a semi-infinite substrate of thermal conductivity K_2 which fills the region $z < 0$. We wish to find the steady-state temperature rise $T(r,z)$ under the influence of an axially symmetric, distributed heat source of strength $Q(r,z)$.

The temperature rise satisfies Poisson's equation,

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2} = \begin{cases} -Q/K_1, & 0 < z < c, \\ -Q/K_2, & z < 0, \end{cases} \tag{1}$$

and the boundary conditions are

$$\begin{aligned} T_z(r,c) &= 0, \\ T(r,0^+) &= T(r,0^-), \\ K_1 T_z(r,0^+) &= K_2 T_z(r,0^-), \\ T(r,z) \to 0 \quad &\text{as} \quad r^2 + z^2 \to \infty, \end{aligned} \tag{2}$$

where T_z denotes $\partial T/\partial z$. The first of the boundary conditions asserts that there is no heat flow across the upper boundary of the film. The method of solution which we are going to use would also allow for a linearized radiation condition at the surface, i.e., a linear relation between $T(r,c)$ and $T_z(r,c)$, if one knew the appropriate coefficients. The second and third conditions insure the continuity of temperature and heat flow across the interface between film and substrate, and the fourth condition says that the temperature rise tends to zero at great distances from the source.

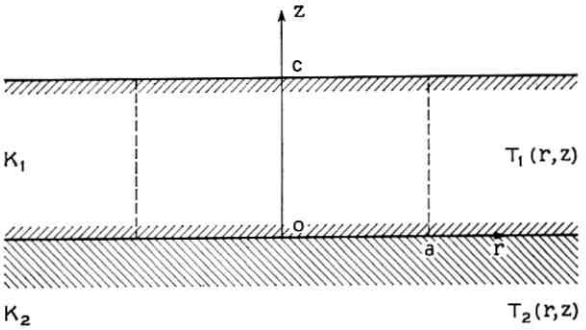

Fig. 1 — Cross section of plane film on semi-infinite substrate.

It will be convenient henceforth to work in terms of normalized, dimensionless quantities. In particular, we shall take the film thickness as the unit of length, and denote the ratio of film conductivity to substrate conductivity by ε. We also introduce a representative heat source strength Q_0 having the dimensions of power per unit volume. Thus, let

$$\begin{aligned} \xi &= r/c = \text{normalized radius} \\ \zeta &= z/c = \text{normalized depth} \\ \varepsilon &= K_1/K_2 = \text{conductivity ratio} \\ q &= Q/Q_0 = \text{normalized heat input} \\ U &= (K_1/c^2Q_0)T = \text{normalized temperature rise.} \end{aligned}$$

In terms of these normalized quantities, (1) takes the form

$$\frac{\partial^2 U}{\partial \xi^2} + \frac{1}{\xi}\frac{\partial U}{\partial \xi} + \frac{\partial^2 U}{\partial \zeta^2} = \begin{cases} -q & 0 < \zeta < 1, \\ -\varepsilon q, & \zeta < 0, \end{cases} \tag{3}$$

and the boundary conditions (2) become

$$\begin{aligned} U_\zeta(\xi,1) &= 0, \\ U(\xi,0^+) &= U(\xi,0^-), \\ \varepsilon U_\zeta(\xi,0^+) &= U_\zeta(\xi,0^-), \\ U(\xi,\zeta) \to 0 \quad &\text{as} \quad \xi^2 + \zeta^2 \to \infty. \end{aligned} \tag{4}$$

In what follows, we shall treat separately the cases of heat input to the film and heat input to the substrate. The general case follows by superposition.

2.1 *Heat Input to Film*

Assume that $q(\xi,\zeta)$ differs from zero only in the film. We denote the normalized temperature rise in the film by $U_1(\xi,\zeta)$, and seek a solution of Poisson's equation in the form

$$U_1(\xi,\zeta) = \int_0^\infty f(w,\zeta) J_0(w\xi)\, dw, \qquad 0 \leqq \zeta \leqq 1, \tag{5}$$

where $f(w,\zeta)$ is a function to be determined. Similarly, for the normalized temperature rise in the substrate, $U_2(\xi,\zeta)$, we seek a solution of Laplace's equation in the form

$$U_2(\xi,\zeta) = \int_0^\infty g(w) e^{w\zeta} J_0(w\xi)\, dw, \qquad \zeta \leqq 0, \tag{6}$$

which vanishes as $\zeta \to -\infty$.

We shall consider the case in which the normalized heat input can be written as

$$q(\xi,\zeta) = \psi(\xi)\varphi(\zeta), \tag{7}$$

that is, as the product of a function of radius times a function of depth. This will probably be justifiable if the increase in beam width with depth due to electron scattering is negligible. Furthermore, by taking $\psi(\xi)$ and $\varphi(\zeta)$ equal to δ-functions it is possible to derive the Green's function, in terms of which one can express the solution for an arbitrary axially symmetric heat input.

Substituting (5) and (7) into (3) and making use of Bessel's equation, we obtain

$$\int_0^\infty [f_{\xi\xi}(w,\zeta) - w^2 f(w,\zeta)]J_0(w\xi)dw = -\psi(\xi)\varphi(\zeta), \qquad \text{for} \quad 0 < \zeta < 1. \tag{8}$$

From the Hankel inversion formula,[8] it follows that

$$f_{\zeta\zeta}(w,\zeta) - w^2 f(w,\zeta) = -w\bar{\psi}(w)\varphi(\zeta), \tag{9}$$

where $\bar{\psi}(w)$ is the Hankel transform of $\psi(\xi)$, defined by

$$\bar{\psi}(w) = \int_0^\infty \xi\psi(\xi)J_0(w\xi)d\xi. \tag{10}$$

Extensive tables[9] of Hankel transforms are available; we note in particular the following pairs.

(*i*) Uniform beam of normalized radius α:

$$\begin{aligned} \psi(\xi) &= \begin{cases} 1, & 0 \leqq \xi < \alpha, \\ 0, & \xi > \alpha, \end{cases} \\ \bar{\psi}(w) &= (\alpha/w)J_1(\alpha w). \end{aligned} \tag{11a}$$

(*ii*) Gaussian beam:

$$\begin{aligned} \psi(\xi) &= \exp(-\xi^2/\alpha^2), \\ \bar{\psi}(w) &= (\alpha^2/2)\exp(-\alpha^2 w^2/4). \end{aligned} \tag{11b}$$

(*iii*) Infinitesimally thin, hollow beam of radius ξ_0 :

$$\begin{aligned} \psi(\xi) &= \delta(\xi - \xi_0), \\ \bar{\psi}(w) &= \xi_0 J_0(w\xi_0). \end{aligned} \tag{11c}$$

To satisfy the boundary conditions (4), we must have

$$\begin{aligned} f_\zeta(w,1) &= 0, \\ f(w,0) &= g(w), \\ \varepsilon f_\zeta(w,0) &= wg(w), \end{aligned} \tag{12}$$

from which, eliminating $g(w)$,

$$f_\zeta(w,1) = 0, \qquad \varepsilon f_\zeta(w,0) = wf(w,0). \tag{13}$$

The solution of the two-point boundary value problem for $f(w,\zeta)$ by standard methods leads to

$$f(w,\zeta) = \bar{\psi}(w)\left[\frac{\sinh w\zeta + \varepsilon \cosh w\zeta}{\cosh w + \varepsilon \sinh w}\int_0^1 \varphi(\eta)\cosh w(1-\eta)d\eta - \int_0^\zeta \varphi(\eta)\sinh w(\zeta-\eta)d\eta\right]. \tag{14}$$

From the second of (12),

$$g(w) = \frac{\varepsilon\bar{\psi}(w)}{\cosh w + \varepsilon \sinh w}\int_0^1 \varphi(\eta)\cosh w(1-\eta)d\eta. \tag{15}$$

In principle, the normalized temperature rise is completely given by (5), (6), (10), (14), and (15), provided that the heat input to the film can be represented as a product $\psi(\xi)\varphi(\zeta)$, and that there is no heat input to the substrate. A complete numerical solution for arbitrary ψ and φ would, however, involve the evaluation of five integrals, each of which depends on one or more parameters. In practice, one would try to approximate the source function in such a way that at least some of the integrations could be done analytically. An example is discussed in the following sections.

2.2 *Heat Input to Substrate*

Again we take the heat input in the product form (7), but now assume that $\varphi(\zeta)$ differs from zero only in the substrate, $\zeta < 0$. For the normalized temperature rise we assume

$$U_1(\xi,\zeta) = \int_0^\infty l(w)\cosh w(1-\zeta)J_0(w\xi)dw, \qquad 0 \leqq \zeta \leqq 1, \tag{16}$$

$$U_2(\xi,\zeta) = \int_0^\infty m(w,\zeta)J_0(w\xi)dw, \qquad \zeta \leqq 0, \tag{17}$$

in the film and substrate, respectively, where $l(w)$ and $m(w,\zeta)$ are functions to be determined. The expression for $U_1(\xi,\zeta)$ already satisfies Laplace's equation and the first of the boundary conditions (4).

As before, it is easy to show that $m(w,\zeta)$ must satisfy the differential equation

$$m_{\zeta\zeta}(w,\zeta) - w^2 m(w,\zeta) = -\varepsilon w \bar{\psi}(w)\varphi(\zeta), \tag{18}$$

and the boundary conditions

$$\begin{aligned} l(w)\cosh w &= m(w,0), \\ -\varepsilon l(w)\sinh w &= m_\zeta(w,0), \\ m(w,\zeta) &\to 0 \qquad \text{as} \qquad \zeta \to -\infty. \end{aligned} \tag{19}$$

By standard methods we find

$$m(w,\zeta) = D(w)e^{w\zeta} + \tfrac{1}{2}\varepsilon\bar{\psi}(w)\int_{-\infty}^{0} \exp\,(-w\,|\,\eta - \zeta\,|\,)\,\varphi(\eta)d\eta, \qquad \text{for} \quad \zeta \leqq 0. \tag{20}$$

It is clear that $m(w,\zeta)$ satisfies the last of the boundary conditions (19) if $w > 0$ and $\varphi(\zeta)$ vanishes for all sufficiently large negative ζ. In practical cases, it will certainly be justifiable to set the heat input identically equal to zero below some finite depth. We shall not take space to investigate the mathematical question of how slowly $\varphi(\zeta)$ could approach zero, and still have $m(w,\zeta)$ also approach zero, as $\zeta \to -\infty$.

From the first two boundary conditions (19), it is straightforward to calculate

$$l(w) = \frac{\varepsilon\bar{\psi}(w)}{\cosh w + \varepsilon \sinh w}\int_{-\infty}^{0} \exp\,(w\eta)\varphi(\eta)d\eta, \tag{21}$$

$$D(w) = \tfrac{1}{2}\varepsilon\bar{\psi}(w)\left[\frac{\cosh w - \varepsilon \sinh w}{\cosh w + \varepsilon \sinh w}\right]\int_{-\infty}^{0} \exp\,(w\eta)\varphi(\eta)d\eta. \tag{22}$$

The first of these, substituted into (16), gives the normalized temperature rise in the film; and the second, together with (17) and (20), gives the temperature rise in the substrate.

III. APPROXIMATIONS FOR A UNIFORM BEAM OF LARGE RADIUS

We shall consider henceforth only the case in which the heat input is

radially uniform out to the normalized radius $\xi = \alpha$, and zero for $\xi > \alpha$. Also we shall be interested only in the temperature distribution in the film itself. The dependence of heat input on depth will, however, still be taken as arbitrary.

In the case where heat is applied to the film by means of a radially uniform beam which does not penetrate the substrate, the normalized temperature rise in the film is given by (5), (11a), and (14). In practice, the SiO_2 film may be only half a micron thick while the beam radius is several microns. We accordingly seek an asymptotic expansion of the temperature distribution for large α. In the analysis we assume that the conductivity ratio ε is fixed with $\varepsilon \lesssim 1$. Our results will also include the physically interesting case $\varepsilon \ll 1$.

Combining (5), (11a), and (14), we may write the temperature distribution in the film in the form

$$U_1(\xi,\zeta;\varepsilon) = \int_0^\infty \left[\frac{\varepsilon\alpha}{w}\int_0^1 \varphi(\eta)d\eta + h(w,\zeta;\varepsilon)\right] J_1(\alpha w)J_0(\xi w)dw, \qquad 0 \leqq \zeta \leqq 1, \tag{23}$$

where

$$h(w,\zeta;\varepsilon) = \frac{\alpha}{w}\left[\frac{\sinh w\zeta + \varepsilon\cosh w\zeta}{\cosh w + \varepsilon\sinh w}\int_0^1 \varphi(\eta)\cosh w(1-\eta)d\eta - \varepsilon\int_0^1 \varphi(\eta)d\eta - \int_0^\zeta \varphi(\eta)\sinh w(\zeta-\eta)d\eta\right]. \tag{24}$$

It is clear that the function $h(w,\zeta;\varepsilon)$ may be expanded in a power series around $w = 0$; that is,

$$h(w,\zeta;\varepsilon) = \sum_{m=0}^{\infty} h^{(m)}(0,\zeta;\varepsilon)w^m/m!, \tag{25}$$

where the superscripts denote derivatives with respect to w. In particular,

$$h(0,\zeta;\varepsilon) = \alpha\left[(\zeta-\varepsilon^2)\int_0^1 \varphi(\eta)d\eta - \int_0^\zeta (\zeta-\eta)\varphi(\eta)d\eta\right]. \tag{26}$$

Let

$$\rho = \xi/\alpha, \qquad \xi = \alpha\rho, \tag{27}$$

so that the boundary of the heat input region is $\rho = 1$. We have[10]

$$
\begin{aligned}
P(\rho) &\equiv \int_0^\infty J_1(\alpha w) J_0(\rho\alpha w)\,\frac{dw}{w} \\
&\equiv \int_0^\infty J_1(x) J_0(\rho x)\,\frac{dx}{x} \\
&= \begin{cases} \dfrac{2}{\pi} E(\rho), & 0 \leqq \rho < 1, \\ \dfrac{2\rho}{\pi}\left[E\left(\dfrac{1}{\rho}\right) - \left(1 - \dfrac{1}{\rho^2}\right) K\left(\dfrac{1}{\rho}\right)\right], & \rho > 1, \end{cases}
\end{aligned} \tag{28}
$$

where E and K are complete elliptic integrals.

Furthermore, assuming that ρ, ζ, and ε are fixed, the following asymptotic expansions for large α are derived in the Appendix, in terms of the derivatives of $h(w,\zeta;\varepsilon)$ at $w = 0$:

$$
\begin{aligned}
&\int_0^\infty h(w,\zeta;\varepsilon) J_1(\alpha w) J_0(\rho\alpha w)\,dw \\
&\sim \begin{cases} \dfrac{h(0,\zeta;\varepsilon)}{\alpha} + \displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n \Gamma(n+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)} \\ \quad \cdot F(n+\tfrac{3}{2}, n+\tfrac{1}{2}; 1; \rho^2)\,\dfrac{h^{(2n+1)}(0,\zeta;\varepsilon)}{\alpha^{2n+2}}, & 0 \leqq \rho < 1, \\ \displaystyle\sum_{n=0}^{\infty} \frac{(-1)^{n+1}\Gamma(n+\frac{3}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)\rho^{2n+3}} \\ \quad \cdot F\left(n+\tfrac{3}{2}, n+\tfrac{3}{2}; 2; \dfrac{1}{\rho^2}\right)\dfrac{h^{(2n+1)}(0,\zeta;\varepsilon)}{\alpha^{2n+2}}, & \rho > 1, \end{cases}
\end{aligned} \tag{29}
$$

where $F(a,b;c;z)$ is the hypergeometric function. Since when ρ is near unity, we have[11]

$$
\begin{aligned}
F(n+\tfrac{3}{2}, n+\tfrac{1}{2}; 1; \rho^2) &= \mathrm{O}[(1-\rho)^{-(2n+1)}], \\
&\qquad 0 < (1-\rho) \ll 1, \\
F\left(n+\tfrac{3}{2}, n+\tfrac{3}{2}; 2; \frac{1}{\rho^2}\right) &= \mathrm{O}[(\rho-1)^{-(2n+1)}], \\
&\qquad 0 < (\rho-1) \ll 1,
\end{aligned} \tag{30}
$$

it follows that the asymptotic expansions (29) are useful only for $\alpha\,|\,1-\rho\,| \gg 1$, but not in the neighborhood of $\rho = 1$.

Combining (23), (26), (28), and (29), we obtain finally,

$$U_1(\rho\alpha,\zeta;\varepsilon) \sim \varepsilon\alpha P(\rho)\int_0^1 \varphi(\eta)d\eta + (\zeta - \varepsilon^2)\int_0^1 \varphi(\eta)d\eta - \int_0^\zeta (\zeta - \eta)\varphi(\eta)d\eta + O[\varepsilon/\alpha(1-\rho)], \qquad 0 \leqq \rho < 1; \tag{31}$$

$$U_1(\rho\alpha,\zeta;\varepsilon) \sim \varepsilon\alpha P(\rho)\int_0^1 \varphi(\eta)d\eta + O[\varepsilon/\alpha(\rho-1)\rho^2], \qquad \rho > 1,$$

where $P(\rho)$ is defined by (28). That the remainder terms are $O(\varepsilon)$ when $\varepsilon \ll 1$ may be seen from the relations

$$\begin{aligned} h(w,\zeta;\varepsilon) &= h(-w,\zeta;-\varepsilon), \\ h^{(2n+1)}(0,\zeta;\varepsilon) &= -h^{(2n+1)}(0,\zeta;-\varepsilon); \end{aligned} \tag{32}$$

i.e., the odd derivatives of h with respect to w at $w = 0$ are odd functions of ε. Note, however, that setting $\varepsilon = 0$ in the asymptotic solution (31) does not give the exact solution for a perfectly conducting substrate, inasmuch as there are exponentially small terms in α which never appear in the asymptotic solution. The exact solution for $\varepsilon = 0$ is given in Section IV.

When the product $\varepsilon\alpha$ is sufficiently large, the leading terms in the asymptotic solution (31) are proportional to $P(\rho)$; that is, they are functions of ρ $(= \xi/\alpha)$ only, and are independent of the depth ζ in the film. The function $P(\rho)$ is plotted in Fig. 2. It is continuous, with a logarithmically infinite slope, at $\rho = 1$. Numerical comparisons between the exact solution (23) and the approximate solution (31) are made in Section V.

We now look briefly at the case of heat input to the substrate by a radially uniform beam of normalized radius α and depth dependence $\varphi(\zeta)$, for $\zeta \leqq 0$. The normalized temperature rise in the film is, from (11a), (16), (21), and (27),

$$U_1(\rho\alpha,\zeta;\varepsilon) = \varepsilon\alpha\int_0^\infty \frac{\cosh w(1-\zeta)}{w(\cosh w + \varepsilon \sinh w)}\left[\int_{-\infty}^0 \exp(w\eta)\,\varphi(\eta)d\eta\right] \cdot J_1(\alpha w)J_0(\rho\alpha w)dw, \qquad 0 \leqq \zeta \leqq 1. \tag{33}$$

When ρ, ζ, and ε are fixed, and both $\alpha \gg 1$ and $\alpha\,|\,1-\rho\,| \gg 1$, an analysis entirely similar to the preceding gives

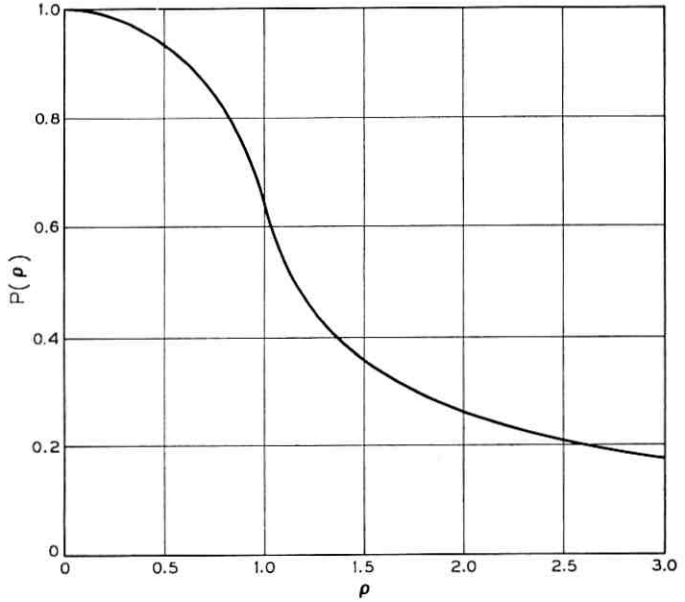

Fig. 2 — The function $P(\rho) = \int_0^\infty \frac{J_1(x)J_0(\rho x)}{x}\,dx$.

$$U_1(\rho\alpha,\zeta;\,\varepsilon) \sim \varepsilon\alpha P(\rho)\int_{-\infty}^{0}\varphi(\eta)d\eta + \varepsilon\int_{-\infty}^{0}(\eta-\varepsilon)\varphi(\eta)d\eta + O[\varepsilon/\alpha(1-\rho)], \qquad 0 \leqq \rho < 1;$$
$$U_1(\rho\alpha,\zeta;\,\varepsilon) \sim \varepsilon\alpha P(\rho)\int_{-\infty}^{0}\varphi(\eta)d\eta + O[\varepsilon/\alpha(\rho-1)\rho^2], \qquad \rho > 1. \tag{34}$$

IV. APPROXIMATIONS FOR A UNIFORM BEAM WITH LARGE SUBSTRATE CONDUCTIVITY

We now assume that $\varepsilon \ll 1$; that is, the conductivity of the substrate is large compared to the conductivity of the film. (For an SiO_2 film on silicon, ε is between 0.1 and 0.01.) Again we consider a radially uniform beam, with an arbitrary depth-dose function $\varphi(\zeta)$. No restrictions are placed on the normalized beam radius α.

For heat input to the film only, the normalized temperature rise in the film is given by (5), (11a), and (14). Referring to (14), we expand

the function $f(w,\zeta)$ in powers of ε. After a little algebra, we find that the normalized temperature rise in the film can be written as

$$U_1(\xi,\zeta;\varepsilon) = \sum_{n=0}^{\infty} \varepsilon^n U_1^{(n)}(\xi,\zeta), \tag{35}$$

where

$$U_1^{(0)}(\xi,\zeta) = \alpha \int_0^{\infty} F(w,\zeta) J_1(\alpha w) J_0(\xi w)\, dw, \tag{36}$$

with

$$F(w,\zeta) = \frac{1}{w \cosh w}\left[\sinh w\zeta \int_{\zeta}^{1} \varphi(\eta) \cosh w(1-\eta)\, d\eta + \cosh w(1-\zeta) \int_0^{\zeta} \varphi(\eta) \sinh w\eta\, d\eta\right], \tag{37}$$

and for $n \geqq 1$,

$$U_1^{(n)}(\xi,\zeta) = (-1)^{n-1}\alpha \int_0^{\infty}\left[\int_0^1 \varphi(\eta) \cosh w(1-\eta)\, d\eta\right] \times \frac{\cosh w(1-\zeta) \tanh^{n-1} w}{\cosh^2 w} J_1(\alpha w) J_0(\xi w)\, \frac{dw}{w}. \tag{38}$$

The quantity $U_1^{(0)}(\xi,\eta)$ is the normalized temperature rise in the film when the substrate is perfectly conducting. In this case, the temperature rise at the bottom of the film is zero, and in fact it is obvious from (37) that $F(w,0) \equiv 0$. The quantity $U_1^{(n)}(\xi,\zeta)$ represents the nth order correction if ε is small but not zero.

We may evaluate $U_1^{(0)}(\xi,\zeta)$ by contour integration. Let S_n denote the semicircle of radius $n\pi$ in the upper half-plane ($n = 1, 2, \cdots$), with diameter along the real axis indented at the origin. From (37) it follows that $F(w,\zeta)$ is uniformly bounded on S_n, and has simple poles within S_n at $w = (m - \frac{1}{2})\pi i$ ($m = 1, \cdots, n$). The choice of integrand depends on whether $0 \leqq \xi \leqq \alpha$, or $\xi \geqq \alpha$. For $0 \leqq \xi \leqq \alpha$ we consider

$$\int_{S_n} F(w,\zeta) J_0(\xi w) H_1^{(1)}(\alpha w)\, dw, \tag{39}$$

and for $\xi \geqq \alpha$ we consider

$$\int_{S_n} F(w,\zeta) J_1(\alpha w) H_0^{(1)}(\xi w)\, dw, \tag{40}$$

where $H_0^{(1)}$ and $H_1^{(1)}$ are Hankel functions. The integrals are evaluated

by the calculus of residues and then the limit $n \to \infty$ is taken. Since the procedure is a standard one, we omit the details and merely state the results.

We obtain, for $0 \leqq \xi \leqq \alpha$,

$$\left\{U_1^{(0)}(\xi,\zeta) - \left[\zeta \int_\zeta^1 \varphi(\eta)d\eta + \int_0^\zeta \eta\varphi(\eta)d\eta\right]\right\}$$

$$= -\frac{2\alpha}{\pi}\sum_{m=1}^{\infty} \frac{\sin[(m-\tfrac{1}{2})\pi\zeta]}{(m-\tfrac{1}{2})} I_0[(m-\tfrac{1}{2})\pi\xi]K_1[(m-\tfrac{1}{2})\pi\alpha] \qquad (41)$$

$$\cdot \int_0^1 \varphi(\eta)\sin[(m-\tfrac{1}{2})\pi\eta]d\eta,$$

and, for $\xi \geqq \alpha$,

$$U_1^{(0)}(\xi,\zeta) = \frac{2\alpha}{\pi}\sum_{m=1}^{\infty} \frac{\sin[(m-\tfrac{1}{2})\pi\zeta]}{(m-\tfrac{1}{2})} \qquad (42)$$

$$\cdot I_1[(m-\tfrac{1}{2})\pi\alpha]K_0[(m-\tfrac{1}{2})\pi\xi]\int_0^1 \varphi(\eta)\sin[(m-\tfrac{1}{2})\pi\eta]d\eta.$$

Here I_0, I_1, K_0, and K_1 are modified Bessel functions. The continuity of $\partial U_1^{(0)}/\partial\xi$ at $\xi = \alpha$ is readily verified, while that of $U_1^{(0)}$ at $\xi = \alpha$ follows from the identities

$$I_0(x)K_1(x) + I_1(x)K_0(x) = 1/x, \qquad (43)$$

and

$$\sum_{m=1}^{\infty} \frac{\cos[(m-\tfrac{1}{2})\pi\theta]}{(m-\tfrac{1}{2})^2} = \frac{\pi^2}{2}(1-\theta), \qquad 0 \leqq \theta \leqq 2. \qquad (44)$$

We remark that (41) and (42) could have been derived by separation of variables in (3). For $|\alpha - \xi| \gg 1$ the right-hand sides of (41) and (42) are exponentially small. It follows that (41) and (42) are consistent with the asymptotic expansions (31) and (32) for $\alpha \gg 1$, when $\varepsilon = 0$ and asymptotically small terms are neglected.

It does not appear possible to evaluate the first-order correction $U_1^{(1)}(\xi,\zeta)$, as given by (38), using contour integration, because the integrand has the wrong parity in w. We can, however, obtain a bound on the value of $U_1^{(1)}(0,0)$, at the "hot spot" of the film-substrate interface, where the zero-order solution $U_1^{(0)}(0,0)$ vanishes.

From (38), setting $n = 1$ and changing the order of integration, which is justified since the double integral is absolutely convergent,

$$U_1^{(1)}(0,0) = \alpha\int_0^1 \varphi(\eta)\left[\int_0^\infty \frac{\cosh w(1-\eta)}{\cosh w}\frac{J_1(\alpha w)}{w}\,dw\right]d\eta. \qquad (45)$$

When $\eta = 0$, the inner integral is equal[12] to 1. When $\eta > 0$, we transform the inner integral by substituting the integral representation[13]

$$J_1(\alpha w) = (2\alpha w/\pi) \int_0^{\pi/2} \cos(\alpha w \cos\theta) \sin^2\theta \, d\theta, \tag{46}$$

and again invoking the absolute convergence of the double integral to change the order of integration. This leads to

$$\begin{aligned}
&\int_0^\infty \frac{\cosh w(1-\eta)}{\cosh w} \frac{J_1(\alpha w)}{w} dw \\
&= \frac{2\alpha}{\pi} \int_0^{\pi/2} \left[\int_0^\infty \frac{\cosh w(1-\eta)}{\cosh w} \cos(\alpha w \cos\theta) dw \right] \sin^2\theta \, d\theta \\
&= \alpha \sin\frac{\pi\eta}{2} \int_0^{\pi/2} \frac{\cosh(\frac{1}{2}\pi\alpha\cos\theta)\sin^2\theta \, d\theta}{\sinh^2(\frac{1}{2}\pi\alpha\cos\theta) + \sin^2(\frac{1}{2}\pi\eta)} \\
&< \alpha \sin\frac{\pi\eta}{2} \int_0^{\pi/2} \frac{\cosh(\frac{1}{2}\pi\alpha\cos\theta)\sin\theta \, d\theta}{\sinh^2(\frac{1}{2}\pi\alpha\cos\theta) + \sin^2(\frac{1}{2}\pi\eta)} \\
&= \frac{2}{\pi} \tan^{-1}\left[\frac{\sinh(\frac{1}{2}\pi\alpha)}{\sin(\frac{1}{2}\pi\eta)} \right] < 1,
\end{aligned} \tag{47}$$

where the third line follows from a table of Fourier transforms.[14] Hence, finally, from (45) and (47),

$$\begin{aligned}
&U_1^{(1)}(0,0) \\
&\quad = \alpha^2 \int_0^1 \varphi(\eta) \sin\frac{\pi\eta}{2} \left[\int_0^{\pi/2} \frac{\cosh(\frac{1}{2}\pi\alpha\cos\theta)\sin^2\theta \, d\theta}{\sinh^2(\frac{1}{2}\pi\alpha\cos\theta) + \sin^2(\frac{1}{2}\pi\eta)} \right] d\eta \\
&\quad < \alpha \int_0^1 \varphi(\eta) d\eta.
\end{aligned} \tag{48}$$

We see from (31) that asymptotically, for $\alpha \gg 1$, the upper bound in (48) is attained, since $P(1) = 1$.

Now suppose that heat is put only into the substrate, so that the normalized temperature rise in the film is given by (33). Expanding in powers of ε leads to

$$U_1(\xi,\zeta;\varepsilon) = \sum_{n=1}^\infty \varepsilon^n U_1^{(n)}(\xi,\zeta), \tag{49}$$

where

$$\begin{aligned}
U_1^{(n)}(\xi,\zeta) = \alpha \int_0^\infty &\frac{(-1)^{n-1}\cosh(1-\zeta)w \tanh^{n-1} w}{w \cosh w} \\
&\times \left[\int_{-\infty}^0 \exp(wn)\varphi(\eta) d\eta \right] J_1(\alpha w) J_0(\xi w) dw.
\end{aligned} \tag{50}$$

In particular we have, on changing the order of integration,

$$U_1^{(1)}(0,0) = \alpha \int_{-\infty}^{0} \varphi(\eta) \left[\int_0^{\infty} \frac{\exp(w\eta) J_1(\alpha w)}{w} \, dw \right] d\eta$$
$$= \int_{-\infty}^{0} \varphi(\eta)[(\eta^2 + \alpha^2)^{\frac{1}{2}} + \eta] d\eta, \tag{51}$$

after substituting the known value[15] of the inner integral.

V. NUMERICAL RESULTS

In this section we give the results of some numerical computations using the exact formulas of Section II, and some comparisons with the approximations of Sections III and IV. We asume that the electron beam voltage is such that the electrons penetrate to the bottom of the oxide film (about 5 kv for an 0.5 μ film[7,16]), but do not enter the substrate. For the depth-dose function we take

$$\varphi(\zeta) = \sin \beta\zeta, \qquad \beta = 5\pi/6, \qquad 0 \leqq \zeta \leqq 1, \tag{52}$$

which is plotted in Fig. 3. The assumed depth-dose function vanishes at

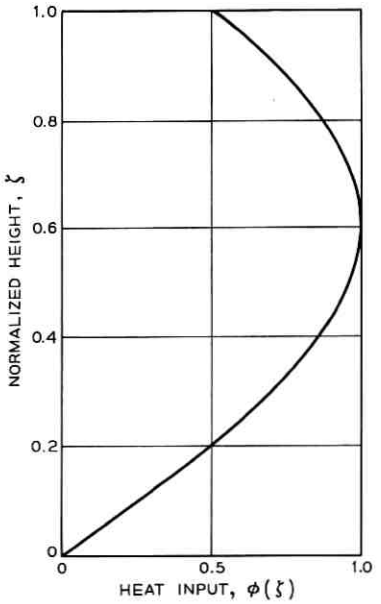

Fig. 3 — Heat input function $\varphi(\zeta) = \sin 5\pi\zeta/6$.

the bottom of the film, has a maximum at a depth equal to 40 per cent of the film thickness, and in general corresponds very closely to the empirical function used by Grün[6] and Wells.[7] The position of the maximum could be varied, of course, by changing the parameter β.

Substituting $\varphi(\zeta)$ into (5), (11a), and (14), we find after some algebra that the normalized temperature rise in the film is given by

$$U_1(\xi,\zeta) = \alpha V(\xi) \sin \beta\zeta + \alpha\beta W(\xi,\zeta), \tag{53}$$

where

$$V(\xi) = \int_0^\infty \frac{J_1(\alpha w)J_0(\xi w)}{w^2 + \beta^2}\, dw \tag{54}$$

and

$$W(\xi,\zeta) = \int_0^\infty \frac{J_1(\alpha w)J_0(\xi w)}{w(w^2 + \beta^2)} \cdot\left[\frac{\varepsilon \cosh(1 - \zeta)w - (\sinh \zeta w + \varepsilon \cosh \zeta w) \cos \beta}{\cosh w + \varepsilon \sinh w}\right] dw. \tag{55}$$

The integral on the right side of (54) can be expressed in terms of modified Bessel functions. We have[17]

$$\int_0^\infty \frac{wJ_1(\alpha w)J_1(tw)}{(w^2 + \beta^2)}\, dw = \begin{cases} I_1(\beta t)K_1(\beta\alpha), & t \leqq \alpha, \\ I_1(\beta\alpha)K_1(\beta t), & t \geqq \alpha. \end{cases} \tag{56}$$

Integrating both sides with respect to t from α to ξ and using the relationship[18]

$$\int_0^\infty \frac{J_1(\alpha w)J_0(\alpha w)}{w^2 + \beta^2}\, dw = \frac{I_1(\alpha\beta)K_0(\alpha\beta)}{\beta} \tag{57}$$

we obtain

$$V(\xi) = \begin{cases} \dfrac{1}{\beta}\left[\dfrac{1}{\beta\alpha} - K_1(\beta\alpha)I_0(\beta\xi)\right], & 0 \leqq \xi \leqq \alpha, \\ \dfrac{I_1(\beta\alpha)K_0(\beta\xi)}{\beta}, & \xi \geqq \alpha. \end{cases} \tag{58}$$

The integral for $W(\xi,\zeta)$, on the other hand, has to be evaluated numerically. The integrand is oscillatory, and falls off exponentially for large w if $0 < \zeta < 1$. If $\zeta = 0$ or $\zeta = 1$, it falls off like $1/w^4$ if $\xi \neq 0$, and like $1/w^{7/2}$ if $\xi = 0$. The numerical integration was done by Simpson's rule on an IBM 7094 computer. Combined analytic and empirical

investigations of the accuracy were made in order to guarantee that the relative error in any value of U is less (in most cases, much less) than one per cent.

Four different sets of parameters were chosen; namely, $\alpha = 2$, 10, and 20 with $\varepsilon = 1/40$, and $\alpha = 2$ with $\varepsilon = 1/4$. The normalized temperature rises at the surface of the film and at the film-substrate interface are plotted against normalized radius in Fig. 4. Note the differences in scale; in each case the edge of the beam, $\xi = \alpha$, is at the center of the plot. The temperature distribution along the vertical axis is shown for the same four cases in Fig. 5.

It is seen that the temperature distribution at the surface of the film becomes more flat-topped, and the fall-off at the edge of the beam becomes relatively (although not absolutely) mote abrupt as α increases in the first three cases of Fig. 4. Also note that the temperature levels are somewhat higher and the temperature variation through the film is less in Fig. 4(d) than in Fig. 4(a), since for the same value of α the relative conductivity of the substrate is only $\frac{1}{10}$ as large in Fig. 4(d) as in Fig. 4(a).

The dashed curves in Figs. 4 and 5 correspond to the approximate formulas (31) for large α. If $\varphi(\zeta)$ is given by (52), these approximations read:

$$U_1(\rho\alpha,\zeta;\varepsilon) \sim \varepsilon\alpha P(\rho)\frac{1-\cos\beta}{\beta} + \frac{1}{\beta}\left[\frac{\sin\beta\zeta}{\beta} - \zeta\cos\beta - \varepsilon^2(1-\cos\beta)\right], \qquad 0 \leqq \rho < 1, \tag{59}$$

$$U_1(\rho\alpha,\zeta;\varepsilon) \sim \varepsilon\alpha P(\rho)\frac{1-\cos\beta}{\beta}, \qquad \rho > 1,$$

where $\rho = \xi/\alpha$. As expected, the approximations are discontinuous at the edge of the beam, $\rho = 1$; and they are not much good when $\alpha = 2$ (worse for the larger value of ε). They are remarkably good, however, for $\alpha = 10$ and $\alpha = 20$; the dashed curves essentially coincide with the solid ones except on the surface of the film in the immediate neighborhood of the beam edge.

Contour plots for the temperature distribution in the film are given in Fig. 6 for $\alpha = 2$ and $\alpha = 10$ with $\varepsilon = 1/40$, and for $\alpha = 2$ with $\varepsilon = 1/4$. Contour plots were not made for $\alpha = 20$, because the numerical

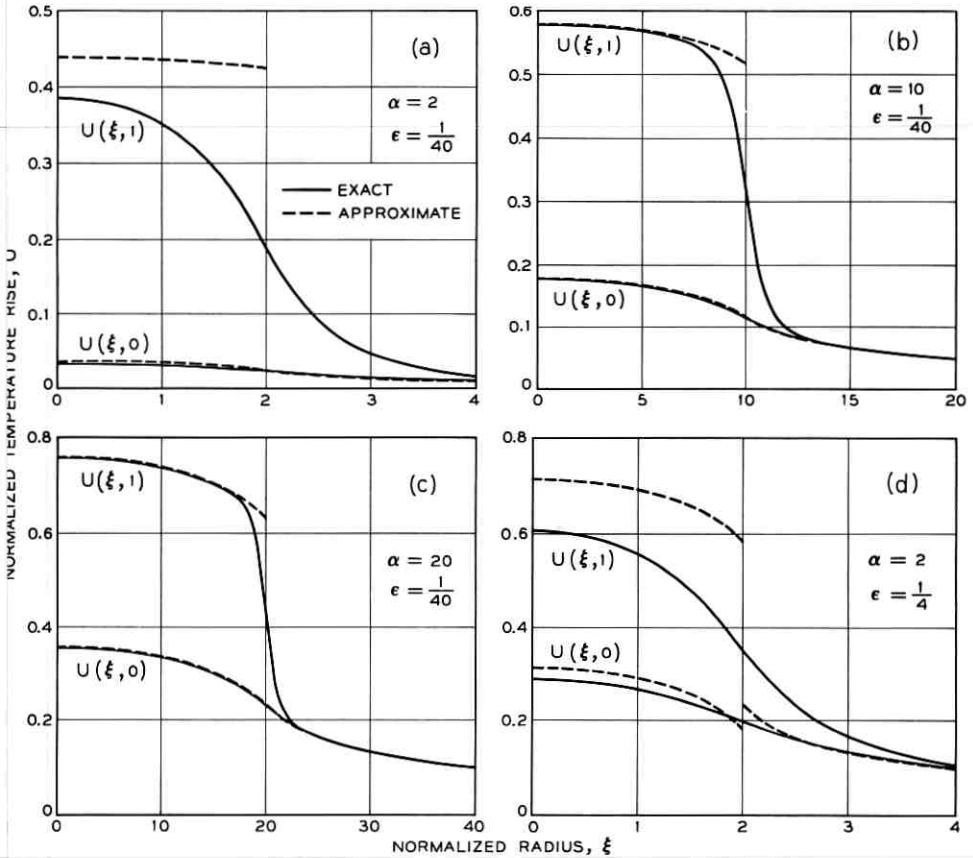

Fig. 4 — Normalized temperature rise at upper and lower surfaces of film: (a) $\alpha = 2, \varepsilon = 1/40$; (b) $\alpha = 10, \varepsilon = 1/40$; (c) $\alpha = 20, \varepsilon = 1/40$; (d) $\alpha = 2, \varepsilon = 1/4$.

integration is slow for large α (the integrands oscillate more rapidly); but it is clear that for large α the approximate formulas (59) would yield accurate contours, except very close to the beam edge.

We may also compare the bound on the first-order correction term for small ε, as given in Section IV, with the exact results. At the center of the film-substrate interface, (48) and (52) give

$$\varepsilon U_1^{(1)}(0,0) < \varepsilon\alpha(1 - \cos\beta)/\beta = 0.713\varepsilon\alpha, \qquad (60)$$

which leads to the following comparison with the exact solution $U_1(0,0)$.

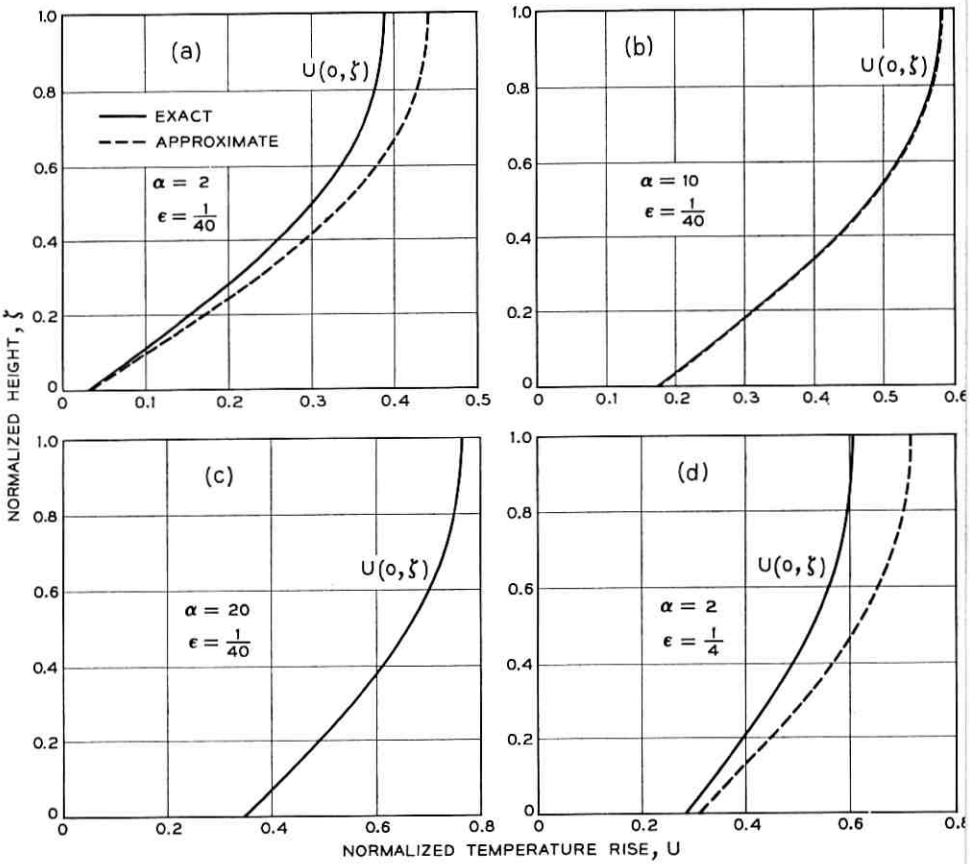

Fig. 5 — Normalized temperature rise along axis: (a) $\alpha = 2$, $\varepsilon = 1/40$; (b) $\alpha = 10$, $\varepsilon = 1/40$; (c) $\alpha = 20$, $\varepsilon = 1/40$; (d) $\alpha = 2$, $\varepsilon = 1/4$.

ε	α	$U_1(0,0)$	$0.713\varepsilon\alpha$
1/40	2	0.0314	0.0356
1/40	10	0.1770	0.1782
1/40	20	0.3556	0.3564
1/4	2	0.2874	0.3564

In order to relate the dimensionless temperature rise U_1 to the physical temperature rise T_1, it is convenient to introduce the power density (i.e., per unit area) in the incident beam. For a uniform beam of radius αc with depth-dose function $\varphi(z/c)$, the dimensional factor Q_0, which normalizes the heat input per unit volume (Section II), is related to the incident power density P_0 by

$$\pi(\alpha c)^2 P_0 = 2\pi Q_0 \int_0^c \int_0^{\alpha c} \varphi(z/c) r \, dr \, dz, \tag{61}$$

$$Q_0 = P_0 \Big/ c\int_0^1 \varphi(\zeta) d\zeta \, . \tag{62}$$

Hence, the actual temperature rise at the point (r,z) of the film is given by

$$\begin{aligned} T_1(r,z) &= (c^2 Q_0/K_1) U_1(r/c, z/c) \\ &= (1.403 c P_0/K_1) U_1(r/c, z/c), \end{aligned} \tag{63}$$

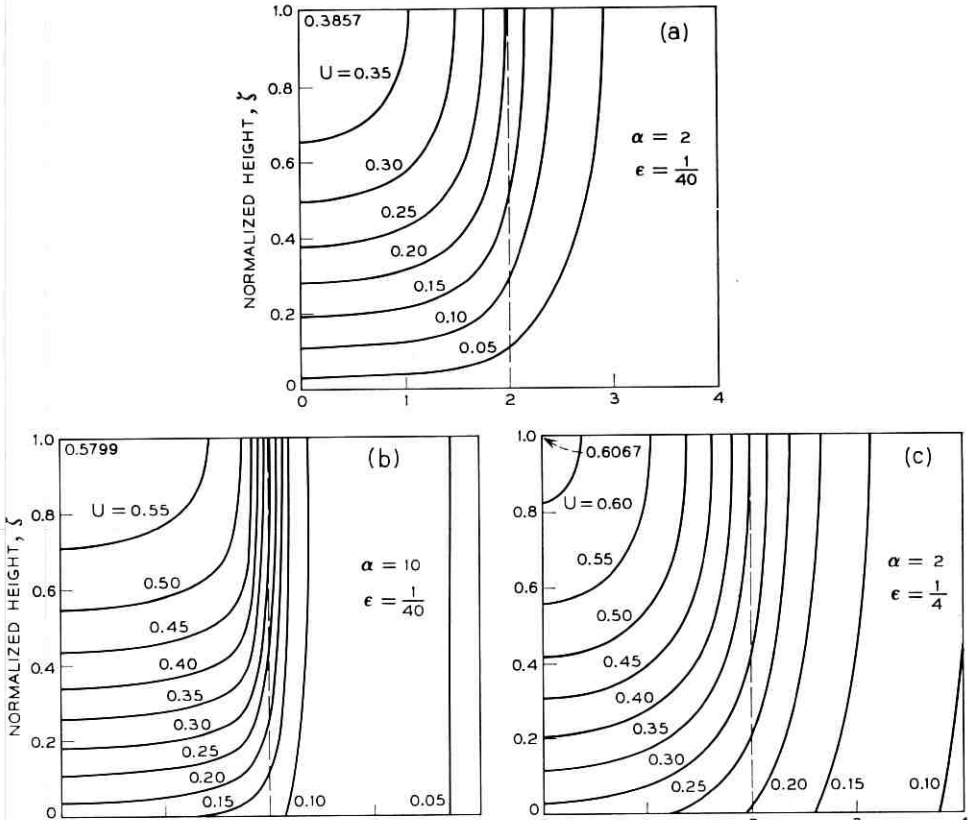

Fig. 6 — Isothermal contours. (a) $\alpha = 2$, $\varepsilon = 1/40$; (b) $\alpha = 10$, $\varepsilon = 1/40$; (c) $\alpha = 2$, $\varepsilon = 1/4$.

where the numerical coefficient corresponds to the depth-dose function (52). Consistent units for (63) are:

$$\begin{aligned} T_1 &= \text{temperature rise in °C} \\ r,z &= \text{coordinates in cm} \\ c &= \text{film thickness in cm} \\ P_0 &= \text{incident power density in watts/cm}^2 \\ K_1 &= \text{film conductivity in watt/cm °K} \\ 0.239\ K_1 &= \text{film conductivity in cal/sec cm °K.} \end{aligned}$$

We shall now look at a numerical example. It turns out that the constant-conductivities model which has been analyzed in the present paper is not a very good approximation to the real problem of a silicon dioxide film on a silicon substrate, because the thermal conductivities of both materials depend strongly on temperature. In fact, the conductivity[19] of SiO_2 *increases* from about 0.015 watt/cm °K at room temperature (various values are reported — not for thin films — which differ among themselves by as much as 2:1 depending on the crystalline orientation of the sample) to about 0.03 watt/cm °K at 900°C. For Si, on the other hand, the conductivity[20] *decreases* from about 1 watt/cm °K at room temperature to about 0.03 watt/cm °K at 900°C. For purposes of calculation we shall more or less arbitrarily assume the values

$$\begin{aligned} K_1 &= 0.03 \text{ watt/cm °K} \\ K_2 &= 1.2 \text{ watt/cm °K} \\ \varepsilon &= 1/40 \\ c &= 0.5\ \mu = 5 \times 10^{-5} \text{ cm} \\ P_0 &= 10^6 \text{ watts/cm}^2. \end{aligned} \tag{64}$$

Since these conductivities may be somewhat larger than the actual conductivities, the temperatures which we shall compute may be somewhat lower than the actual temperatures. For a 5-kv beam, the assumed power density corresponds to a current density of 200 amps/cm^2.

Table I gives the total beam current and the maximum temperature rise (i.e., on the axis) at the top and bottom of the film, for beams of diameter 2 μ, 10 μ, and 20 μ, corresponding to the previous computations with $\alpha = 2$, 10, and 20, and a conductivity ratio of 1/40. It appears, therefore, that in each case at least a part of the irradiated spot would be raised to the temperature at which the oxide hardens (900°C), but in no case would the substrate melt (1415°C).

It is probably worth repeating that the physical problem of interest is nonlinear, because of the dependence of conductivity on temperature. Bounds on the solution may be obtained from linear models, by using

TABLE I

Diameter	Current	$U_1(0,1)$	$U_1(0,0)$	$T_1(0,c)$	$T_1(0,0)$
2μ	6.28μa	0.3857	0.0314	902°	73°
10μ	157μa	0.5799	0.1771	1356°	414°
20μ	628μa	0.7589	0.3556	1775°	832°

the theorem that with a fixed heat input the steady-state temperature is not increased anywhere (usually, it is decreased everywhere) if the conductivity is increased anywhere, and vice versa. However, only a full-dress numerical treatment of the nonlinear partial differential equation, assuming that one knew the temperature dependence of the conductivity, would be likely to yield really accurate results.

VI. TRANSIENT EFFECTS

It is of interest to know how long it will take to reach the steady state if the electron beam is suddenly switched onto the film, since this gives an idea of how rapidly the beam may be scanned in laying out a mask. There have been some published analyses[7,21] of transient heating effects in electron beam machining, but we shall content ourselves with a crude estimate of the time scale in the present problem.

Consider the case of a film on a perfectly conducting substrate, with the film initially at zero temperature, and with a time-independent heat input starting at $t = 0$. Then the instantaneous temperature distribution satisfies the heat flow equation

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2} = -\frac{Q}{K_1} + \frac{C\delta}{K_1}\frac{\partial T}{\partial t}, \tag{65}$$

where K_1 is the thermal conductivity, C the heat capacity, and δ the density. The total temperature $T(r,z,t)$ may be written as the sum of a steady-state part and a transient part,

$$T(r,z,t) = T_1(r,z) + \Theta(r,z,t), \tag{66}$$

where $T_1(r,z)$ satisfies Poisson's equation (cf. Section II) and $\Theta(r,z,t)$ satisfies the homogeneous equation

$$\frac{\partial^2 \Theta}{\partial r^2} + \frac{1}{r}\frac{\partial \Theta}{\partial r} + \frac{\partial^2 \Theta}{\partial z^2} - \frac{C\delta}{K_1}\frac{\partial \Theta}{\partial t} = 0, \tag{67}$$

and vanishes as $t \to \infty$.

A sufficiently general solution of (67) may be written in the form

$$\theta(r,z,t) = \sum_{m=0}^{\infty} \int_0^{\infty} A_m(w) \exp\{-(K_1/C\delta)[w^2 + (m + \tfrac{1}{2})^2\pi^2/c^2]t\} \times J_0(wr) \sin\frac{(m + \tfrac{1}{2})\pi z}{c}\, dw. \tag{68}$$

The initial condition requires that the total temperature vanish at $t = 0$; that is,

$$T_1(r,z) + \theta(r,z,0) = 0. \tag{69}$$

Hence, from a knowledge of the steady-state temperature one can in principle use the properties of Fourier series and Fourier-Bessel integrals to determine the functions $A_m(w)$ and the transient solution $\theta(r,z,t)$.

We would like to know how fast $\theta(r,z,t)$ approaches zero with increasing time. It is clear that (68) cannot be characterized by any single exponential decay; but we observe that the most slowly decaying exponential is

$$\exp[-(K_1\pi^2/4c^2C\delta)t],$$

and it is therefore reasonable to define a crude "transient time" as

$$\tau = 4c^2C\delta/\pi^2K_1\,. \tag{70}$$

We assume the following numerical values for the SiO_2 film:

$$\begin{aligned} c &= 5 \times 10^{-5} \text{ cm} \\ C &= 1 \text{ watt sec/gm °K} \\ \delta &= 2.2 \text{ gm/cm}^3 \\ K_1 &= 0.03 \text{ watt/cm °K.} \end{aligned} \tag{71}$$

Then

$$\tau = 3.0 \times 10^{-7} \text{ sec}, \tag{72}$$

so the transient time is a fraction of a microsecond.

VII. ACKNOWLEDGMENTS

The authors are grateful to P. F. Schmidt for posing this problem and for providing several references. They wish to thank J. A. Lewis and J. McKenna for various discussions, and Miss E. Lauh for the extensive IBM computations and plots.

APPENDIX

Asymptotic Expansion of a Class of Integrals

In this appendix, the asymptotic expansion of

$$J(\rho,\alpha) \equiv \int_0^\infty h(w)J_1(\alpha w)J_0(\rho\alpha w)\,dw \tag{73}$$

is derived, where

$$h(w) = \sum_{m=0}^{\infty} h^{(m)}(0)w^m/m!, \tag{74}$$

for $\alpha \gg 1$ and $\alpha \mid 1 - \rho \mid \gg 1$, that is, for ρ not in the neighborhood of 1. It is clear that the asymptotic expansions will break down in the neighborhood of $\rho = 1$, since then the integrand will contain a term which is not rapidly oscillating.

We start from a result given by Tranter.[22] Namely, if we have the expansion

$$\int_0^\infty \exp(-\gamma w)\,\mathfrak{F}(\alpha,w)dw = \sum_{n=0}^{\infty} A_n(\alpha)\gamma^n, \tag{75}$$

then, formally,

$$\int_0^\infty h(w)\mathfrak{F}(\alpha,w)dw \sim \sum_{n=0}^{\infty} (-1)^n A_n(\alpha)h^{(n)}(0). \tag{76}$$

In the case at hand,

$$\mathfrak{F}(\alpha,w) = J_1(\alpha w)J_0(\rho\alpha w). \tag{77}$$

Assume first that ρ is fixed and $0 \leqq \rho < 1$. Then if $(\alpha^2 + \gamma^2)^{\frac{1}{2}} > \rho\alpha + \gamma$, we have[23]

$$\int_0^\infty \exp(\gamma w)\,J_1(\alpha w)J_0(\rho\alpha w)dw = \frac{1}{\alpha}\sum_{m=0}^{\infty} \frac{(-1)^m\rho^{2m}\Gamma(2m+2)}{2^{2m+1}[\Gamma(m+1)]^2} \cdot\left(\frac{\alpha^2}{\alpha^2+\gamma^2}\right)^{m+1} F\left(m+1, -m+\tfrac{1}{2}; 2; \frac{\alpha^2}{\alpha^2+\gamma^2}\right). \tag{78}$$

Using standard transformations,[24]

$$\int_0^\infty \exp(-\gamma w)\, J_1(\alpha w) J_0(\rho\alpha w)\, dw$$

$$= \frac{1}{\alpha} \sum_{m=0}^{\infty} \frac{(-1)^m \rho^{2m} \Gamma(2m+2)}{2^{2m+1}[\Gamma(m+1)]^2} \left(\frac{\alpha^2}{\gamma^2}\right)^{m+1} \times F\left(m+1, m+\tfrac{3}{2}; 2; -\frac{\alpha^2}{\gamma^2}\right)$$

$$= \frac{1}{\alpha} + \frac{\gamma}{\alpha^2} \sum_{m=0}^{\infty} \frac{(-1)^m \rho^{2m} \Gamma(2m+2)\Gamma(-\frac{1}{2})}{2^{2m+1}[\Gamma(m+1)]^3 \Gamma(\frac{1}{2}-m)} \times F\left(m+\tfrac{3}{2}, m+\tfrac{1}{2}; \tfrac{3}{2}; -\frac{\gamma^2}{\alpha^2}\right) \tag{79}$$

$$= \frac{1}{\alpha} - \frac{2\gamma}{\pi\alpha^2} \sum_{m=0}^{\infty} \frac{\rho^{2m} \Gamma(m+\frac{1}{2})\Gamma(m+\frac{3}{2})}{[\Gamma(m+1)]^2} \times F\left(m+\tfrac{3}{2}, m+\tfrac{1}{2}; \tfrac{3}{2}; -\frac{\gamma^2}{\alpha^2}\right)$$

$$= \frac{1}{\alpha} + \frac{\gamma}{\sqrt{\pi}\alpha^2} \sum_{n=0}^{\infty} \frac{(-1)^{n+1} \gamma^{2n} \Gamma(n+\frac{1}{2})}{\alpha^{2n}\Gamma(n+1)} \times F(n+\tfrac{3}{2}, n+\tfrac{1}{2}; 1; \rho^2),$$

where in the last step we have expanded the hypergeometric function in a power series and interchanged the order of summation. Comparing (79) with (75), we see that

$$A_{2n}(\alpha) = \begin{cases} 1/\alpha & \text{if } n = 0, \\ 0 & \text{if } n > 0, \end{cases} \tag{80}$$

$$A_{2n+1}(\alpha) = \frac{(-1)^{n+1}\Gamma(n+\frac{1}{2})}{\alpha^{2n+2}\Gamma(\frac{1}{2})\Gamma(n+1)} F(n+\tfrac{3}{2}, n+\tfrac{1}{2}; 1; \rho^2).$$

It follows from (76) that

$$\int_0^\infty h(w) J_1(\alpha w) J_0(\rho\alpha w)\, dw$$

$$\sim \frac{h(0)}{\alpha} + \sum_{n=0}^{\infty} \frac{(-1)^n \Gamma(n+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)} \times F(n+\tfrac{3}{2}, n+\tfrac{1}{2}; 1; \rho^2) \frac{h^{(2n+1)}(0)}{\alpha^{2n+2}} \tag{81}$$

for $0 < \rho < 1$.

An entirely similar derivation, the details of which will be omitted, leads to the expansion

$$\int_0^\infty h(w)J_1(\alpha w)J_0(\rho\alpha w)\,dw \sim \sum_{n=0}^{\infty} \frac{(-1)^{n+1}\Gamma(n+\frac{3}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)\rho^{2n+3}} \times F\left(n+\tfrac{3}{2}, n+\tfrac{3}{2}; 2; \frac{1}{\rho^2}\right) \frac{h^{(2n+1)}(0)}{\alpha^{2n+2}}, \tag{82}$$

for $\rho > 1$.

REFERENCES

1. Symposium on Electron Beam Techniques for Microelectronics, Microelectronics and Reliability, *4*, 1965, pp. 1–122.
2. Vine, J. and Einstein, P. A., Proc. IEE (London), *111*, 1964, pp. 921–930.
3. Almasi, G. S., Blair, J., Ogilvie, R. E. and Schwartz, R. J., J. Appl. Phys., *36*, 1965, pp. 1848–1854.
4. Leisegang, S., Proceedings of the Third International Conference on Electron Microscopy, London, 1954, (Royal Microscopical Society, London, 1956), pp. 176–184.
5. Schmidt, P. F., private communication.
6. Grün, A. E., Z. Naturforschung, *12a*, 1957, pp. 89–95.
7. Wells, Oliver C., IEEE Trans. Electron Devices, *ED-12*, 1965, pp. 224–231.
8. Tranter, C. J., *Integral Transforms in Mathematical Physics*, 2nd ed., Methuen, London, 1956, p. 12.
9. Erdélyi, A., *Tables of Integral Transforms*, McGraw-Hill Book Company, Inc., New York, 1954, Vol. 2, pp. 3–92.
10. Ibid., Vol. 2, p. 20, No. (27). Note that there are misprints on p. 14, No. (19).
11. *Handbook of Mathematical Functions*, ed. by M. Abramowitz and I. A. Stegun, National Bureau of Standards, Washington, D.C., 1964, p. 560, No. 15.3.12.
12. Ibid., p. 486, No. 11.4.16.
13. Ibid., p. 360, No. 9.1.20.
14. Erdélyi, op. cit., Vol. 1, p. 31, No. (12).
15. Watson, G. N., *Theory of Bessel Functions*, 2nd ed., Cambridge University Press, 1952, p. 386.
16. Kanter, H. and Sternglass, E. J., Phys. Rev. *126*, 1962, pp. 620–626.
17. Erdélyi, op. cit., Vol. 2, p. 49, No. (10).
18. Watson, op. cit., p. 430.
19. Thermophysical Properties Research Center Data Book, Purdue University, Lafayette, Indiana, 1963, Vol. 3, Chap. 1, Fig. 3017.
20. Ibid., 1961, Vol. 1, Chap. 1, Fig. 1111.
21. Pittaway, L. G., Brit. J. Appl. Phys., *15*, 1964, pp. 967–982.
22. Tranter, op. cit., pp. 63–65.
23. Watson, op. cit., p. 400.
24. *Handbook of Mathematical Functions*, op. cit., p. 559, Nos. 15.3.4 and 15.3.7.

Predictive Quantizing Systems (Differential Pulse Code Modulation) for the Transmission of Television Signals

By J. B. O'NEAL, JR.

(Manuscript received December 27 , 1965)

Differential pulse code modulation (DPCM) and predictive quantizing are two names for a technique used to encode analog signals into digital pulses suitable for transmission over binary channels. It is the purpose of this paper to determine what kind of performance can be expected from well-designed systems of this type when used to encode television signals. Systems using both previous sample and previous line feedback are considered.

A procedure is presented for the design of nonadaptive, time invariant systems which are near optimum in the sense that the resulting signal to unweighted quantizing noise ratios (S/N) are nearly maximum. Simple formulas are derived for these S/N ratios which apply to DPCM as well as standard PCM. Standard PCM is shown to be a special case of DPCM. These formulas are verified by digital computer simulation.

Any advantage of DPCM stems from removing the redundancy from the signal to be transmitted. Redundancy in a signal, however, affords a certain protection against noise introduced in the transmission medium. The penalty for removing this redundancy, through DPCM or other means, is that the transmitted signal becomes more fragile and requires a higher-quality transmission medium than would otherwise be required. This penalty is discussed in quantitative terms.

I. INTRODUCTION

In this paper, the terms predictive quantizing and differential pulse code modulation (DPCM) will be used interchangeably. They describe a special kind of predictive communications system. A predictive communications system is one in which the difference between the actual

signal and an estimate of the signal, based on its past, is transmitted. Both the transmitter and the receiver make an estimate or prediction of the signal's value based on the previously transmitted signal. The transmitter subtracts this prediction from the true value of the signal and transmits this difference. The receiver adds this prediction to the received difference signal yielding the true signal. Highly redundant signals, such as television, are well suited for predictive transmission systems because of the accuracy possible in the prediction. If the signal is sampled, and if the difference signal is quantized and encoded into PCM, then the system is a predictive quantizing or DPCM system.

A block diagram of systems of this type is shown in Fig. 1. Although delta modulation which uses the feedback principle was introduced somewhat earlier,[1] DPCM systems are based primarily on an invention by Cutler.[2] In his original patent in 1952, Cutler used one or more integrators to perform the prediction function. His invention is based on transmitting the quantized difference between successive sample values rather than the sample values themselves. The invention is a special case of a predictive quantizing system and it turned out to be a special case admirably matched to the statistics of television signals.

In the early nineteen forties Weiner[3] developed the theory of optimum linear prediction. By 1952 Oliver, Kretzmer and Harrison at the Bell Telephone Laboratories, realized the importance of linear prediction in feedback communications systems and proposed that it be used to reduce the redundancy, and, therefore, lower the required power in highly periodic signals such as television. Oliver[4] explained how linear prediction could be used to reduce the bandwidth required to transmit redundant signals. Realizing that knowledge of the statistical properties of television signals was necessary in the design of linear prediction sys-

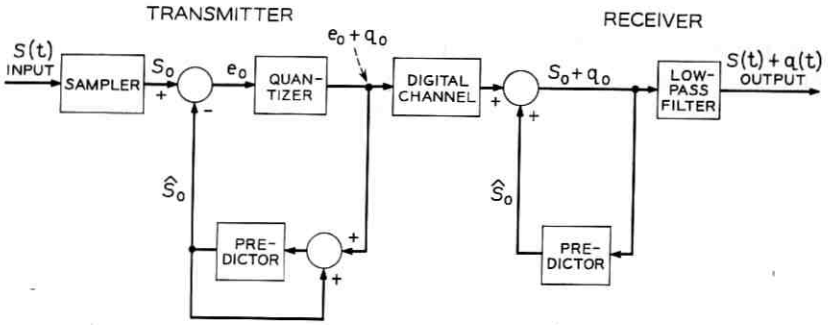

Fig. 1 — Block diagram of a DPCM system.

tems, Kretzmer[5] determined some statistics of typical television picture material. Harrison[6] actually built a signal processing system for television signals and illustrated how redundancy could be removed from these signals using linear prediction. Concurrently with this work at the Bell Telephone Laboratories, but published somewhat later, Elias[7] at MIT was developing this theory of predictive coding which explained the use of linear prediction in PCM systems.

Graham[8] recognized that the theory of prediction could be incorporated into the system described by Cutler. Since Graham's work in 1958, much effort has been expended to devise and build such a system for the transmission of television signals. Although a few experimental systems have been constructed, it is a discouraging fact that such systems have never proved to be very useful for high quality television transmission. Television signals are still being transmitted over transmission systems which do not take advantage of the signals' inherent redundancy.

It is the purpose of this paper to determine the advantages and disadvantages of well-designed DPCM systems. Such information is needed in order to establish whether or not DPCM systems are potentially useful for the transmission of television signals. To do this we present a procedure for the design of some DPCM systems which are near optimum for three television scenes and determine what kind of performance can be expected from such systems. The results obtained are verified by simulating some DPCM systems on an IBM 7094 digital computer and using, as an input, television signals derived from a flying spot scanner. Our study is restricted to nonadaptive systems using linear prediction in the feedback loop and a quantizer whose characteristics do not depend on the instantaneous value of the input signal. Both previous-sample and previous-line feedback in the prediction operation are considered in detail.

II. SUMMARY OF RESULTS AND CONCLUSIONS

Some of the more important results and conclusions about DPCM systems designed for the transmission of television signals are enumerated below. Throughout this paper the term S/N refers to the ratio of signal to quantizing noise.

(*i*) A simple formula for the S/N ratio is derived. If the sync pulses need not be transmitted, then standard PCM is shown to be a special case of DPCM and its S/N ratio is also given by this formula.

(*ii*) When the horizontal resolution is equal to the vertical resolution,

line feedback, when used in addition to previous-sample feedback, can give no more than a 1.9-db additional improvement in S/N ratio. For FCC standard monochrome entertainment television the improvement due to line feedback will be considerably less than 1.9 db.

(*iii*) Differential PCM provides more of an advantage for high resolution television systems than for low resolution systems. For monochrome entertainment television, previous-sample feedback DPCM transmission systems can provide a signal-to-quantizing noise ratio approximately 15 db higher than standard PCM. This improvement may easily vary as much as 2 or 3 db depending on picture material. A 2.8-db improvement in S/N ratio can be realized in standard PCM systems if the sync pulses can be reconstructed by the decoder and need not be transmitted. The improvement of previous-sample feedback DPCM over sync-less PCM is, therefore, only about 12 db for monochrome entertainment television. The effect of line feedback has not been included in the above numbers.

(*iv*) Since 6 db of quantizing noise is equivalent to one bit per sample, the advantage in DPCM can also be expressed in terms of bit rate. For a constant signal-to-quantizing noise ratio, a DPCM system designed for entertainment television can provide a saving of about 18 megabits (2 bits per sample) over standard PCM. This assumes a sampling rate of 9 megacycles, which is twice the bandwidth, and that the noise added by bit errors in the transmission medium is negligible. These bit rate reductions are nearly independent of the signal-to-quantizing noise ratios required.

(*v*) A signal encoded into DPCM is more vulnerable to noise in the transmission medium (bit errors) than one encoded into PCM. It is characteristic of DPCM systems that, if they decrease the quantizing noise by k db over standard PCM, then the noise in the decoded signal caused by errors in the digital transmission channel is increased by k db. This penalty means that, if DPCM is used to reduce the quantizing noise by k db, then the error rate in the digital channel required for satisfactory transmission is reduced by a factor of $(1.26)^k$. This does not imply that DPCM offers no advantage. If the limiting degradation is quantizing noise, and this is generally true for digital systems, then decreasing this quantizing noise, even at the expense of increasing the noise introduced in the transmissoin medium, is desirable. Digital transmission lines designed for PCM encoding, however, may be unsatisfactory for DPCM encoding. This result applies to DPCM systems designed for any type of signal.

(*vi*) The power spectrum of the quantizing noise is approximately

flat. The amplitude density function of the quantizing noise is found to be somewhat flatter than a Gaussian function.

(*vii*) For television input signals the amplitude density function of the quantizer input in a well-designed DPCM system is approximately Laplacian.

Television picture material which has meaning to a human observer has certain patterns which cause statistical redundancy in the resulting television signals. Differential PCM takes advantage of this statistical redundancy and the performance of DPCM systems varies with this redundancy. Conclusions (*iii*) and (*iv*) above are based on measured statistics of television signals derived from three scenes which have detail typical of television picture material.

III. PERFORMANCE CRITERION

The performance criterion used is the ordinary signal-to-quantizing noise ratio, S/N, present in the video part of the composite signal. Noise present in the sync pulses is seldom a limiting factor in television transmission. While it has often been argued that the S/N ratio is not an adequate performance criterion for television systems, a better alternative for analytical study has never been proposed. Furthermore, when used with discretion, the S/N ratio is a useful measure in determining the performance of television systems. It is especially useful in helping to decide which kinds of systems should be built and evaluated subjectively. The subjective test is the final arbiter in determining the usefulness of DPCM for the transmission of television signals.

Unless otherwise stated, the term noise used in this paper implies quantizing noise. We are concerned here with designing DPCM encoding and decoding systems which minimize the mean square difference between the decoded output signal and the analog input signal. This optimization is based on an analytical, i.e., objective, criterion, not a subjective one. Thus, the S/N ratios used are unweighted. All sampling is assumed to be at twice the bandwidth of the baseband input signal, and all the resulting quantizing noise is considered to be in-band. Systems have been proposed[9] which shape the power spectrum of the quantizing noise to make it less objectionable to the human observer. This approach, however, is complicated by the difficulty in determining the proper weighting function for noise which is not independent of the signal. In most DPCM systems the quantizing noise is highly correlated with the derivative of the signal.

IV. DESIGN PROCEDURE

The design procedure used herein is to first design the predictor ignoring the presence of the quantizer. Then the quantizer is designed to match the amplitude distribution of the signal coming from the subtractor. This procedure will result in a system which is very nearly optimum because when the number of quantizing levels is large, the inclusion of the quantizer in the circuit has very little effect on the amplitude distribution of the signal coming out of the subtractor. The predictor will be restricted to be a linear time invariant device and the theory of linear prediction will be used to optimize it. The quantizer will be designed in accordance with procedures first proposed by Panter and Dite.[10]

V. THE PREDICTOR

It is true that nonlinear prediction is superior, by the S/N ratio criterion, to linear prediction for television signals. It has never been determined, however, just how much the S/N ratio can be improved by using nonlinear prediction techniques. Graham[8] suggested one nonlinear predictor and simulated it on the computer. Fine[11] discusses the general case where both nonlinear prediction and quantization are allowed. In this paper, however, only linear prediction is used.

5.1 *Theory of Linear Prediction*

The following brief explanation of the procedure of linear prediction is based on the terse exposition of this subject given by Papoulis.[12]

Let a stationary signal $S(t)$ with mean 0 and rms value σ be sampled at the times $t_1, t_2, \cdots, t_n, \cdots$ and let the sample values be $S_1, S_2, \cdots, S_n, \cdots$, respectively.

A linear estimate of the next sample value S_0 based on the previous n sample values $S_1, S_2, \cdots, S_n$ is defined to be

$$\hat{S}_0 = a_1 S_1 + a_2 S_2 + \cdots + a_n S_n . \tag{1}$$

For simplicity, we assume here that the a's and S's are real numbers. A linear predictive encoder forms this estimate $\hat{S}_0$ and transmits the difference or error

$$e_0 = S_0 - \hat{S}_0 . \tag{2}$$

A block diagram of such a system is shown in Fig. 2. The D's represent delay elements.

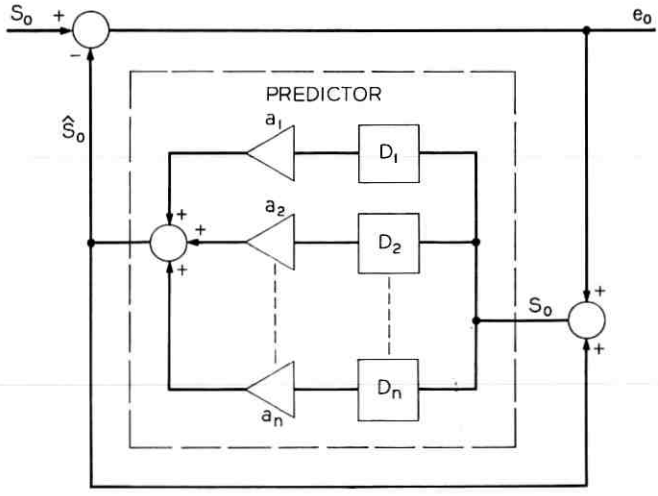

Fig. 2 — Block diagram of a linear predictive encoder.

We define the best estimate of S_0 to be that value of $\hat{S}_0$ for which the expected value of the squared error is minimum. To find the values of the a's which satisfy this condition we first take the partial derivatives of $E[(S_0 - \hat{S}_0)^2]$ with respect to each one of the a's. $E[x]$ denotes the expected value of x.

$$\frac{\delta E[(S_0 - \hat{S}_0)^2]}{\delta a_i} = \frac{\delta E[(S_0 - (a_1S_1 + a_2S_2 + \cdots + a_nS_n))^2]}{\delta a_i}$$

$$= -2E[(S_0 - (a_1S_1 + a_2S_2 + \cdots + a_nS_n))S_i]$$

$$i = 1, 2, \cdots, n.$$

To find an extremum, in this case a minimum, we set this equal to zero giving

$$E[(S_0 - (a_1S_1 + a_2S_2 + \cdots + a_nS_n))S_i] = 0$$

$$E[(S_0 - \hat{S}_0)S_i] = 0 \qquad i = 1, 2, \cdots, n. \tag{3}$$

If we represent the covariance of S_i and S_j by

$$R_{ij} = E[S_iS_j], \tag{4}$$

then from (3) we can rewrite the conditions for the best linear mean square estimate as

$$R_{0i} = a_1R_{1i} + a_2R_{2i} + \cdots + a_nR_{ni} \qquad i = 1, 2, \cdots, n. \tag{5}$$

Equation (5) defines a set of n simultaneous linear equations in the n unknowns a_i, $i = 1, 2, \cdots, n$, which can be found if the covariances R_{ij} are known. These covariances are found from the autocovariance $\psi(\tau)$ of the signal itself,

$$R_{ij} = \psi(t_i - t_j). \tag{6}$$

If $\hat{S}_0$ is the best linear mean square estimate of S_0, then the expected value of the square of the error signal e_0 is

$$\sigma_e^2 = E[(S_0 - \hat{S}_0)^2] = E[(S_0 - \hat{S}_0)S_0]$$
$$\sigma_e^2 = R_{00} - (a_1R_{01} + a_2R_{02} \cdots + a_nR_{0n}). \tag{7}$$

In (7), R_{00} is simply the variance σ^2 of the original sequence S_0, S_1, $\cdots = \{S_i\}$.

The sequence of transmitted error samples is $e_0, e_1, \cdots = \{e_i\}$ where

$$e_i = S_i - \hat{S}_i \qquad i = 0,1, \cdots, \tag{8}$$

and

$$\hat{S}_i = a_1S_{i+1} + a_2S_{i+2} + \cdots + a_nS_{i+n}.$$

The error sequence $\{e_i\}$ is less correlated and has smaller variance than the signal sequence $\{S_i\}$. The use of linear prediction has produced a sequence $\{e_i\}$ from which the sequence $\{S_i\}$ can be reconstructed. The variance σ_e^2 of the error sequence $\{e_i\}$ is less than the variance of the original sequence $\{S_i\}$ by the amount shown in the parenthesis in (7). If the number of samples n used in forming the estimate is unlimited, then the sequence of error samples can always be made completely uncorrelated. If the sequence of samples S_0, S_1, $\cdots = \{S_i\}$ is an rth order Markoff sequence, then only r samples need be used in forming the best estimate of S_0 and the resulting sequence of error samples will be uncorrelated.

As an example of particular relevance to television, consider the 1st order Markoff sequence formed by sampling a signal whose autocorrelation is the exponential function $e^{-\alpha t}$. In this case, even if all previous sample values are available, the estimate of S_0 which minimizes σ_e^2 is $\hat{S}_0 = (R_{01}/\sigma^2)S_1$ where S_1 is the most recent sample value available. It is easy to show that, in this case, the error sequence $\{e_i\}$ is completely uncorrelated, i.e.,

$$E[e_ie_j] = 0 \qquad i \neq j$$
$$= \sigma_e^2 \qquad i = j.$$

The autocorrelation function of one line of a television signal is very similar to $e^{-\alpha t}$ so in this case we expect that basing our estimate only on the previous sample value will be almost as good as using many sample values on the same line. It will be shown, however, that, if we have access to sample values on the adjacent line and/or on the previous frame, we can improve our prediction.

5.2 *Application to Television Signals*

The samples $S_1, S_2, \cdots, S_n$ used in (1) to form the estimate $\hat{S}_0$ need not be the most recently transmitted ones and they need not be in any particular order. They are simply n sample values which have been transmitted in the past. Fig. 3 illustrates 7 sample values which can be used to form a reasonably good estimate of S_0. Such an estimate would be

$$\hat{S}_0 = a_1S_1 + a_2S_2 + a_3S_3 + a_4S_4 + a_5S_5 + a_6S_6 + a_7S_7, \quad (9)$$

where the a's are chosen to satisfy (5). Also shown in the figure, are covariances between these samples.

It will be shown that there is little advantage in using samples S_3 through S_7 for, once S_1 and S_2 are used in the prediction, the other five

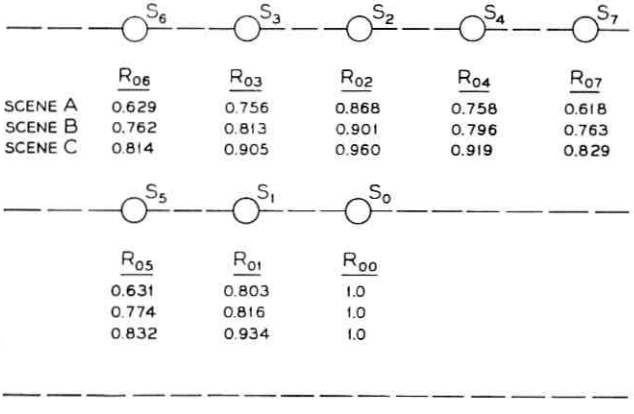

Fig. 3 — Television scan showing sample values near S_0. Covariances for the three scenes in Fig. 6 are also shown.

samples contain little additional information about S_0. Most DPCM systems built in the past use only the previous sample S_1 and form the estimate

$$\hat{S}_0 = a_1 S_1 .$$

In this simple case, it is clear from (5) that the constant a_1 should be R_{01}/σ^2, the covariance between adjacent sample points divided by the mean square value of the input sequence. DPCM systems of this type are called previous-sample feedback systems, and a block diagram of the predictor used in such a system is shown in Fig. 4.

A DPCM system which forms its estimate of S_0 by using the previous sample S_1 and the adjacent sample on the previous line S_2 will be called a line-and-sample feedback system. In this case,

$$\hat{S}_0 = a_1 S_1 + a_2 S_2 . \tag{11}$$

A block diagram of the predictor for this system is shown in Fig. 5.

This concept can easily be extended to take advantage of frame-to-frame correlation. A frame-line-and-sample feedback system would form its estimate of the next sample value by

$$\hat{S}_0 = a_1 S_1 + a_2 S_2 + a_f S_f \tag{12}$$

where S_f is the sample value which is equivalent to S_0 but on the previous frame. Frame feedback systems are not considered in detail in this paper primarily because statistics of frame-to-frame correlations are not available.

5.3 *Statistics of Television Signals*

In order to determine some statistics of television signals and to use television signals as inputs to DPCM systems simulated on the IBM 7094 digital computer, some television signals were obtained from a slow-speed flying-spot scanner.[13] These signals were sampled and encoded into 11 bit PCM and placed on a magnetic tape suitable as an

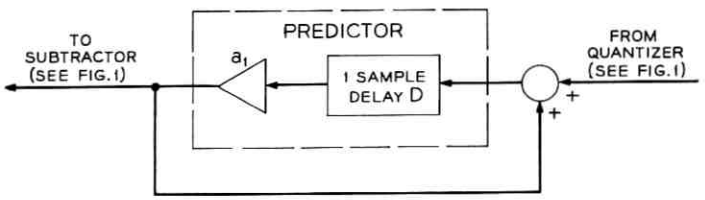

Fig. 4 — Previous-sample predictor.

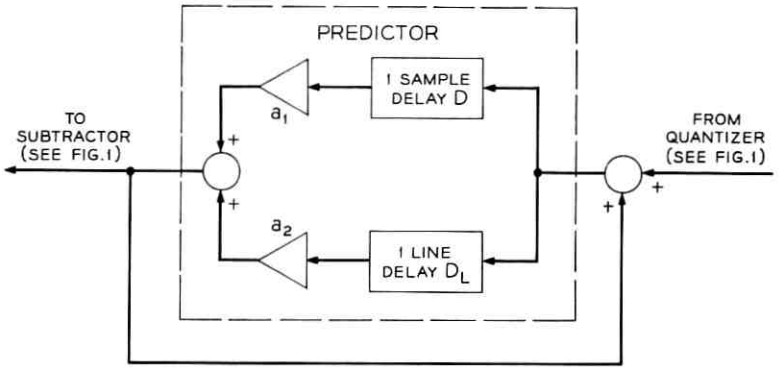

Fig. 5 — Previous line-and-sample predictor.

input to the computer. The signals were obtained by scanning the three square slides shown in Fig. 6 and represent only one frame of a television signal. In conformity with television practices, the video signal was a function of the 0.4 power of the brightness of the original scene. The standards used gave 100 lines and 100 samples per line for the visible part of the pictures and all samples taken were on a symmetric lattice or grid.

The signals on magnetic tape were composite signals, i.e., they consisted of the video signal and a train of sync pulses. Noise and distortions in the sync pulses do not govern the quality of a television signal as long as synchronization is maintained. Therefore, DPCM systems should be matched to the statistics of the video part alone. For this reason, the horizontal sync pulses were ignored and the autocovariance functions of the video part of the signals were obtained. For convenience, the signals were first normalized so that the rms value σ of the video was 1 and its mean value was 0. The autocovariance functions $\psi(\tau)$ are shown in Fig. 7. For small values of τ, these functions are very similar to exponential functions. Since we are dealing with sample values rather than with continuous signals, the autocovariance is actually a set of points at integer values of the lag τ and these points represent the values of R_{0i}, $i = 0, 1, \cdots$. Fig. 7 was constructed by finding these points and drawing lines between them. This is also true of Figs. 8 and 16. The peaks at $\tau = 100$ are due, of course, to the high correlation between adjacent lines of the television signals. Correlations between adjacent frames are not illustrated in the figure because the signals used represented only one frame of a television signal.

Fig. 3 illustrates some of the covariances between neighboring points

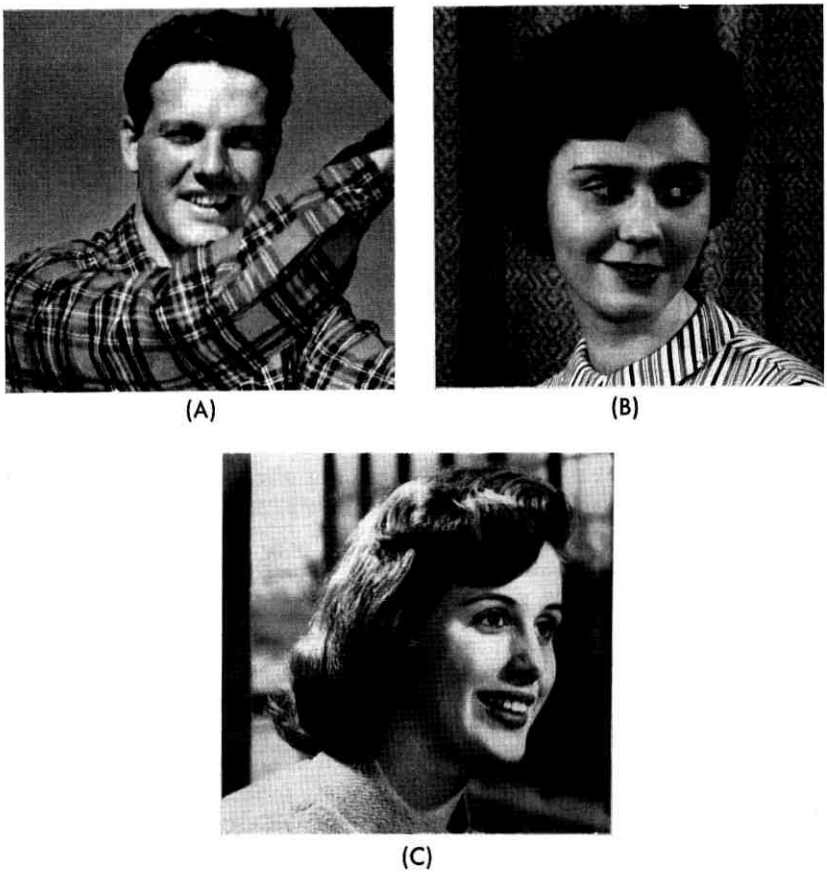

Fig. 6 — Pictures of three slides scanned to obtain television signals.

for the three scenes. For example, the covariance between points S_0 and S_4 in scene B is $R_{04} = 0.796$. The three pictures used had higher vertical than horizontal correlation.

A transmission system is useful only if it can satisfactorily transmit a vast ensemble of signals and its performance must be judged on the basis of its ability to transmit almost all members of this ensemble. The statistics we use here have been obtained from only three members of this ensemble and, since the members of this ensemble are derived from a nonergodic process, we cannot obtain the statistics of the ensemble by examination of these three members. Nevertheless, it is useful to determine the design and performance of DPCM systems when used

to transmit these members which, in some sense at least, are representative of the whole ensemble.

The autocovariance functions in Fig. 7 are averages over the time for each of the three signals used. The autocovariance function of the random process from which these three signals could be derived could not be determined here. Franks,[14] however, has proposed a model for this random process in which the autocovariance function of the picture material is exponential in both the horizontal and vertical directions. Data obtained in this study, some of which is illustrated in Fig. 7, indicates that this is a good approximation for the three scenes used here.

5.4 *Linear Predictors*

Using the data in Fig. 3, we can solve (5) and (7) for the a_i and σ_e for several practical linear predictors. Table I illustrates the optimum values of the a's and the resulting mean square error signals for 8 different predictors. The relative positions of the sample values in this table are those of Fig. 3. For example, if the prediction of S_0 is based on the three sample values S_1, S_2, and S_4 (predictor number 6) then the linear

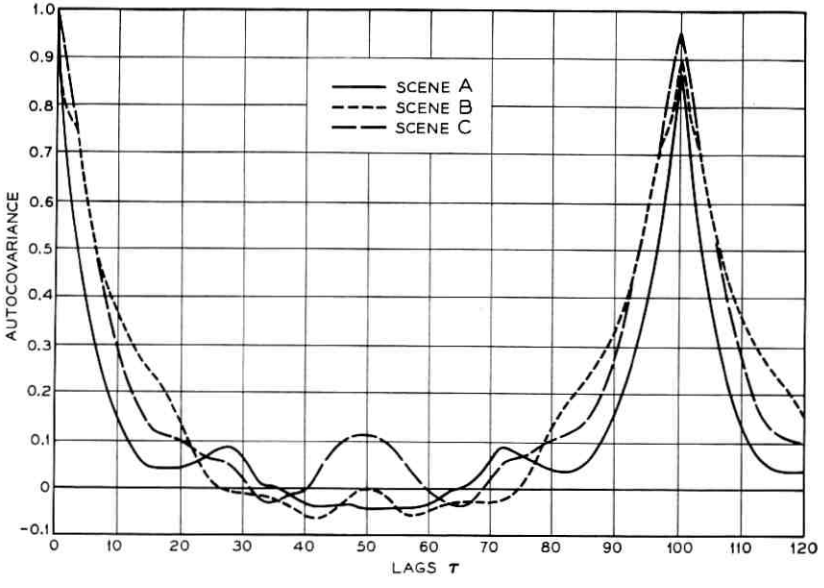

Fig. 7 — Autocovariances of the 100 line, 100 samples per line video signals obtained from the three scenes of Fig. 6. (The sync pulses were not included in computing these functions.)

predictor giving the smallest value of σ_e for scene C forms $\hat{S}_0$ by the equation

$$\hat{S}_0 = 0.383\ S_1 + 0.362\ S_2 + 0.263\ S_4\ .$$

When no quantization is present, such a system results in an error signal whose rms value σ_e is 0.230. This is 12.8 db below the rms value of the signal itself whose rms value σ is 1. The transmitted error sequence will be much less correlated than the original signal sequence. We may think of this signal processing as the removal of redundancy, in this case, 12.8 db of redundancy.

Examination of Table I reveals that once samples S_1 and S_2 have been used in the prediction, there is little advantage in using any others. This means that samples S_1 and S_2 provide almost all of the information about sample S_0 which can be obtained from the previous samples. Samples S_3 and S_4 contain almost no additional information. For a system with line feedback (a system which can store and, therefore, has access to the previous line), there is little point in using any samples but S_1 and S_2. Similarly, for a system without line feedback there is little point in using any sample but S_1 to predict S_0. Furthermore, for the pictures tested, line feedback itself provides only about a 3-db improvement in the estimate of S_0. This is somewhat disappointing especially in view of the fact that the scenes tested had higher vertical than horizontal correlation. Exactly what can be obtained from frame feedback must await the availability of frame-to-frame covariance statistics. Although the above conclusions about line feedback are based on statistics obtained from some 100 line and 100 samples per line pictures it will be shown in section VIII that they apply to television systems in general.

This study suggests that the sequence of sample values derived from one frame of a television signal (or a facsimile signal) may be approximated well by a second-order Markoff sequence.* Furthermore, studies by Deriugin[15] indicate that the sequence derived from many frames of a typical television signal may be, approximately, a third-order Markoff sequence.* In this case, the state (value) of the next sample S_0 may be statistically dependent only on S_1, S_2 and S_F, where these are the sample values adjacent to S_0 on the same line, the previous line, and the previous frame, respectively. More work is required to determine just how well Markoff sequences can represent sample values of television signals.

* This might more properly be called a distant second (or third) order Markoff sequence because, although S_1 is the previous sample value, there are many intervening samples between S_1, S_2, and S_F.

5.5 *Computer Simulation of Predictors*

In order to determine how effectively redundancy could be removed from a television signal by using prediction, the predictors number 1, 3, and 8 shown in Table I were simulated on the computer for all three scenes. The actual rms value σ_e of the errors in the prediction agreed well with those shown in Table I. The autocovariances of the error signals were found and for scene C they are illustrated in Fig. 8. The autocovariance functions shown in Fig. 8 are also representative of what was found for scenes A and B.

Figs. 9 and 10 show the amplitude distribution of the error sequence

TABLE I — VALUES OF THE AMPLIFIER GAINS AND RMS PREDICTION ERROR FOR 8 PREDICTORS MATCHED TO THE 3 PICTURES OF FIG. 6

Predictor Number	Samples Used in Prediction (see Fig. 3)	Scene	Theoretical* rms Prediction Error		a_1	a_2	a_3	a_4	a_5
			σ_e	$-20 \log \sigma_e$					
1	S_1 (see Fig. 4)	A	0.597	4.5	0.803				
		B	0.578	4.8	0.816				
		C	0.358	8.9	0.934				
2	S_2	A	0.498	6.1		0.868			
		B	0.434	7.2		0.901			
		C	0.279	10.1		0.960			
3	S_1, S_2 (see Fig. 5)	A	0.444	7.0	0.341	0.610			
		B	0.402	7.9	0.270	0.686			
		C	0.247	12.1	0.333	0.654			
4	S_1, S_5	A	0.595	4.5	0.834				−0.039
		B	0.547	5.2	0.552				0.324
		C	0.339	9.4	1.229				−0.316
5	S_1, S_4	A	0.494	6.1	0.541			0.423	
		B	0.512	5.8	0.499			0.415	
		C	0.246	12.2	0.550			0.463	
6	S_1, S_2, S_4	A	0.443	7.1	0.337	0.481		0.163	
		B	0.398	8.0	0.238	0.629		0.101	
		C	0.230	12.8	0.383	0.362		0.263	
7	S_1, S_2, S_3	A	0.439	7.2	0.432	0.660	−0.149		
		B	0.401	7.9	0.227	0.670	0.062		
		C	0.224	13.0	0.606	0.793	−0.417		
8	S_1, S_2, S_3, S_4	A	0.429	7.3	0.419	0.533	−0.134	0.155	
		B	0.398	8.0	0.210	0.620	0.047	0.097	
		C	0.214	13.4	0.598	0.544	−0.346	0.203	

* The rms value σ of the input signal is 1. This table is concerned only with prediction error and does not consider the eff ects of quantization.

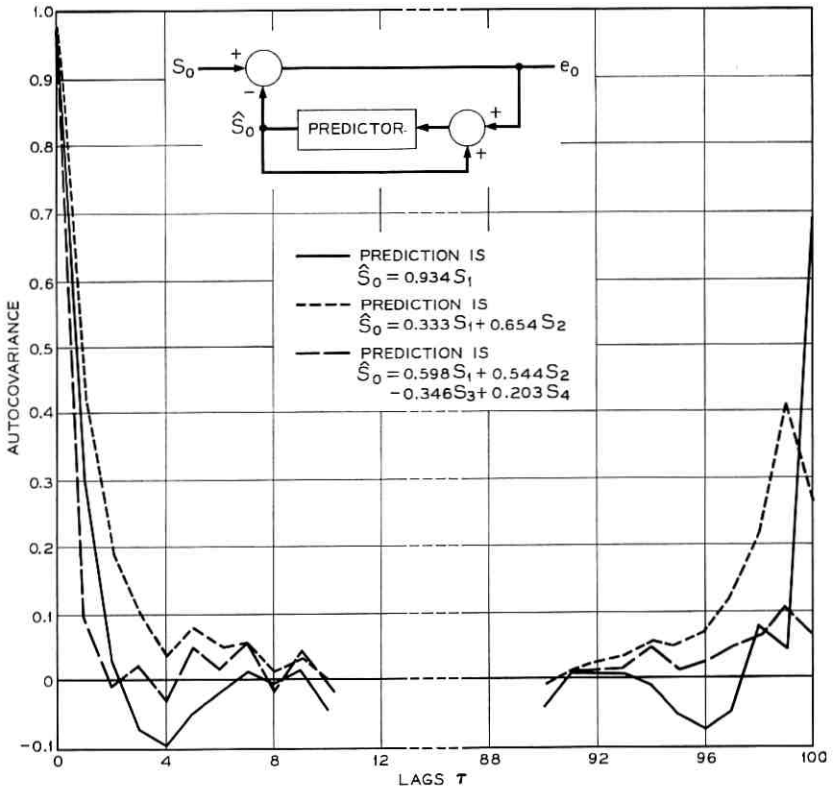

Fig. 8 — Autocovariance of error sequence $\{e_i\}$ of scene C for three linear predictors matched to this scene. (The samples S_1, S_2, S_3, and S_4 are defined in Fig. 3.)

for the three pictures using previous-sample prediction (predictor number 1 in Table I), and line-and-sample prediction (predictor number 3 in Table I), respectively. The shape of these density functions is of foremost importance in designing an optimum quantizer. In both figures the density functions can be approximated reasonably well by Laplacian functions. These amplitude density functions were found by dividing the range $\pm 4\sigma$ into 25 equal intervals and finding the number of sample values in each interval. The points so found were normalized and curves drawn between them.

VI. THE QUANTIZER

In analog systems, it is difficult to evaluate the wisdom in reducing the power by removing the redundancy from a signal. For this process

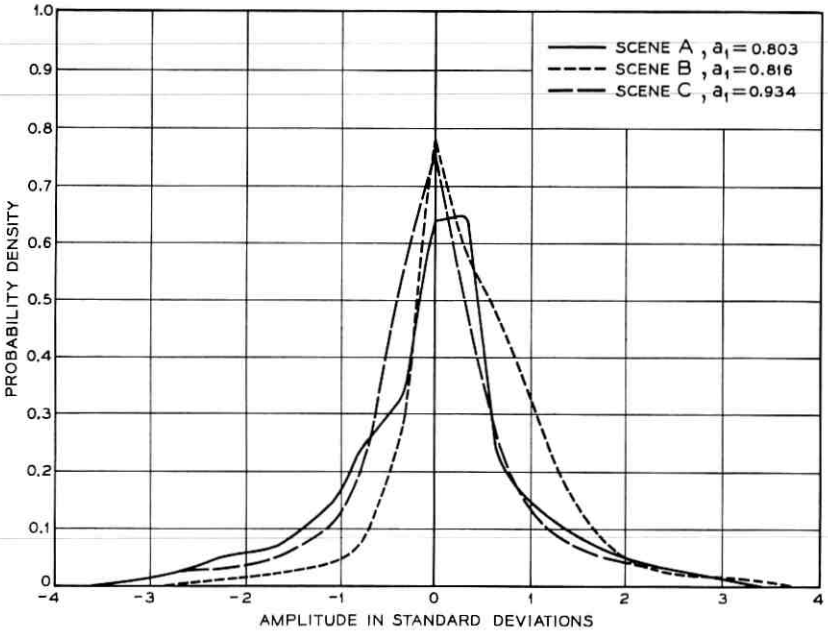

Fig. 9 — Probability density for amplitude of error sequence $\{e_i\}$ for scenes A, B, and C using previous-sample prediction, $\hat{S}_0 = a_1 s_1$. (The value of a_1 is chosen to match each scene.)

automatically makes a signal more susceptible to noise in the transmission medium. While this is still true in digital systems, as will be shown later, we are assured both by logic[16] and by experience[17] that the errors in properly designed digital systems can be made small enough to be neglected. And, if we can ignore this transmission noise, i.e., assume that the probability of error in a digital system can be made as small as we like, there is a dividend in reducing the rms value of the signal to be transmitted. In fact, it will be shown that (for the signals used here, at least) reducing the rms value of the transmitted signal from σ to σ_e decreases the rms value of the quantizing noise by a factor σ/σ_e.

If the input to the quantizer in Fig. 1 is e_0, then its output is $e_0 + q_0$ where q_0 is the quantizing noise. Since the receiver forms the decoded output by adding $e_0 + q_0$ to the estimate $\hat{S}_0$, the quantizing noise in the decoded signal is also q_0. Minimizing the quantizing noise in the decoded output, therefore, is equivalent to minimizing the rms value of the quantizing noise coming out of the quantizer. This method of minimizing the quantizing noise was recognized independently by Nitadori.[18]

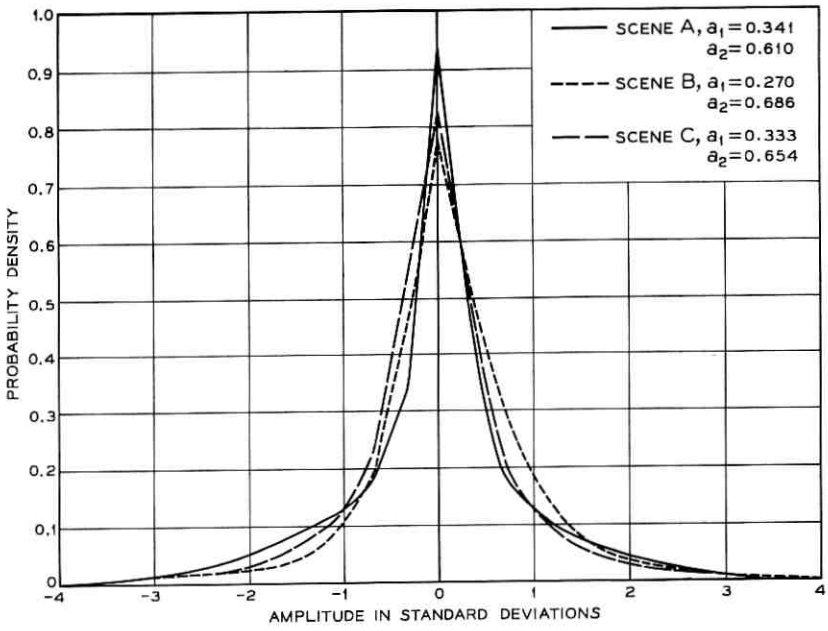

Fig. 10 — Probability density for amplitude of error sequence $\{e_i\}$ for scenes A, B, and C using previous line-and-sample prediction, $\hat{S}_0 = a_1 s_1 + a_2 s_2$. (The values of a_1 and a_2 are chosen to match each scene.)

In what follows, approximations are made which apply when the number of quantizing levels $N = 2^n$ is large. Figs. 12 and 13, to be discussed later, illustrate that the S/N formulas for DPCM systems are accurate for small N as well. It is possible, nevertheless, that inaccuracies may occur under certain conditions when N is too small. If N is less than about 8, the formulas and design procedures presented here should be used with caution.

6.1 *Optimum Quantization*

For n bit quantization, each member of the error sequence is made to assume one of $N = 2^n$ different levels. It has long been known that nonuniform quantization is generally preferable to uniform quantization in DPCM systems. Panter and Dite[10] have shown that the minimum mean square quantizing error is given by

$$\sigma_q^2 = \frac{2}{3N^2}\left[\int_0^V P^{\frac{1}{3}}(e)\, de\right]^3, \tag{13}$$

where $P(e)$ is an even function representing the probability density of the input to the quantizer and $P(e)$ is zero outside the interval $(-V,V)$ which represents the range of the quantizer input.

The curves for $P(e)$ shown in Figs. 9 and 10 may be approximated reasonably well by the Laplacian density function

$$P(e) = \frac{1}{\sqrt{2}\,\sigma_e} \exp\left(-\frac{\sqrt{2}}{\sigma_e} \mid e \mid\right), \tag{14}$$

where σ_e is the rms value of the quantizer input. Since the amplitude density function is different for each scene to be transmitted, the best we can do is to choose some representative density function and match the quantizer to it. We choose this function to be the exponential of (14) and we feel that this will give results which can be expected in practice. Although Figs. 9 and 10 are plots of the error signal without a quantizer in the circuit, computer simulations with the quantizer in the circuit showed that these amplitude density functions are effected very little by the addition of the quantizer as long as the number of levels N was greater than 4. Solving the integral in (13) for this $P(e)$ and taking the limit as V gets large gives, as an approximation for the mean square value of the quantizing noise,

$$\sigma_q^2 = \frac{9}{2N^2}\sigma_e^2. \tag{15}$$

Refining this approximation by using the actual value of V changes it very little since, for cases of interest in DPCM, V, which is the peak-to-peak value of the input signal $S(t)$, is always large compared to σ_e. For the three scenes considered here V is about 7 times the rms value σ of the signal $S(t)$, and σ_e is generally much less than σ. For small values of V the density function in (14) must be truncated. This changes and complicates the value of σ_q^2 given in (15).

From (15) we see that if the rms value in the input video signal is σ then the rms S/N ratio in the video (considering only quantizing noise) of a decoded television signal transmitted through a well-designed n digit differential PCM system is (in db)

$$S/N = 10 \log \frac{2N^2\sigma^2}{9\sigma_e^2}$$

$$S/N \cong -6.5 + 6n + 10 \log \frac{\sigma^2}{\sigma_e^2}. \tag{16}$$

Equation (16) gives the ratio of the rms video signal to rms quantizing

noise in the video. A bound on the S/N ratio to be presented later differs from (16) only by a constant and suggests that this S/N ratio is within about 5 db of that possible for any encoding system. To convert this S/N to the more useful measure, peak-to-peak composite signal to rms noise in the video, we must add a constant to (16) giving (in db)

$$\text{S/N} \cong -6.5 + C + 6n + 10 \log \frac{\sigma^2}{\sigma_e^2},$$

where C is the ratio in db of peak-to-peak composite signal to the rms value of the video. The value of C is determined by the peak value of the sync pulses. It is also dependent on picture material and upon such apparently extraneous factors as the man or electronic device which regulates the peak values of the video signal. For FCC standard monochrome entertainment television some measurements, as well as some data derived from the flying spot scanner used for these studies, indicate that the rms value of the video is about one tenth the peak value of the composite signal and the value of C is, therefore, about 20 db.* Actual measurements of the signals derived from scenes A, B, and C of Fig. 6 give 20.0, 19.8, and 18.1 db, respectively, for the value of C (assuming FCC sync standards). An approximation, then, for the ratio of peak-to-peak composite signal to rms quantizing noise in the video for a typical FCC standard television signal is (in db)

$$\text{S/N} \cong 13.5 + 6n + 10 \log \frac{\sigma^2}{\sigma_e^2}. \tag{17}$$

Bennett[20] showed that if the input signal is distributed evenly between the quantizing levels, the rms value of the quantizing noise for standard PCM is $E_0/\sqrt{12}$. E_0 is the step size of the uniform quantizer and $E_0 = V_{\text{peak}}/2^n$, where V_{peak} is the peak value of the signal to be encoded and n is the number of quantizing digits. Therefore, the peak-to-peak composite signal to rms noise ratio for standard PCM is (in db)

$$\begin{aligned} \text{S/N} &= 20 \log\sqrt{12} + 20 \log 2^n \\ &\cong 10.8 + 6n. \end{aligned} \tag{18}$$

If the sync pulses could be reconstructed by the decoder, then all the PCM levels could be used for the video and the constant in (18) would be $20 \log (\sqrt{12}/0.072) = 13.6$. In other words, if the sync pulses need not

* Some unpublished studies by J. W. Smith indicate that for systems which have automatic regulation of the peak signal excursions the constant C may be several db less than this. Such systems, in attempting to determine peak white and peak black, introduce a certain amount of clipping.

be transmitted then the ratio of peak-to-peak composite signal to rms noise in the video for standard PCM becomes (in db)

$$S/N \cong 13.6 + 6n. \tag{19}$$

Transmitting the sync lowers the S/N by 2.8 db for PCM. As one might expect, provided we neglect the sync pulses, the S/N ratios for standard PCM and differential PCM can be approximated by the same expression, namely (17). Since the constant in (17) is somewhat arbitrary, it would be easy to justify making it 13.6 to agree with (19). In standard PCM, there is no feedback loop and the estimate of the sample value S_0 based on previous sample values is simply 0, the mean value of the input sequence. In this trivial case, since $\sigma_e = \sigma$, the DPCM system becomes identical to standard PCM and (17) reduces approximately to (19). Therefore, we may consider standard PCM to be a special case of DPCM which is optimum when all the covariances R_{ij} for $i \neq j$, are zero.

When the feedback loop exists and when the amplifier gain(s) are reasonably large, then the DPCM system can adequately encode the sync pulses as well as the video. However, when the amplifier gain(s) are too small, or when the feedback loop is not provided at all, as in standard PCM, then either the decoder must be arranged to reconstruct the sync pulses, or the range of the quantizer must be increased beyond what is required for the video in order to accommodate the sync pulses.

From (5) and (7), we can express σ_e in terms of the covariances R_{ij}. For the simplest case, the previous-sample feedback system, the peak-to-peak composite signal to rms quantizing noise in the video S/N ratio can be expressed as (in db)

$$S/N \cong -6.5 + C + 6n + 10 \log \frac{\sigma^2}{\sigma^2 - R_{01}^2/\sigma^2}. \tag{20}$$

This equation illustrates that when R_{01}/σ^2 is close to 1, doubling the bandwidth and the sampling rate (this doubles the horizontal resolution), which is roughly equivalent to halving the value of $\sigma^2 - R_{01}/\sigma^2$, increases the S/N ratio by about 3 db.

6.2 *Design of the Quantizer*

One way to obtain the proper quantizer levels for minimizing the rms quantizing noise is to form a function $y(z)$ such that when z takes on uniformly spaced levels between $-V$ and V, y assumes the proper quantizing levels. Smith[21] shows that, when the probability density

of the signal to be quantized is that of (14), the function $y(z)$ is given by

$$y(z) = -\frac{V}{m} \ln \left[1 - \frac{z}{V} (1 - \exp(-m)) \right], \quad 0 \leqq z \tag{21}*$$

$$y(-z) = y(z),$$

where

$$m = \sqrt{2}\, V/3\sigma_e .$$

There are more elegant and exact ways for finding the quantizing levels,[22,23] but it is doubtful if they can be incorporated into practical systems. Furthermore, it is unlikely that these more sophisticated techniques offer a significant decrease in the quantizing noise over what can be obtained by the simple quantizers described here.

Smith studied quantizers with the characteristic of (21) in some detail for the application to standard PCM systems for speech. His rejection of this characteristic in favor of another results primarily from the wide variation of talker volumes present in speech channels. This objection does not apply to television channels whose signal levels are relatively constant.

A typical 8 level quantizer designed by using the characteristic of (21) is shown in Fig. 11. The case shown is for $V = 7$ and $m = 5.5$. The output signal always assumes the quantizing level nearest to the input signal. Overload noise, which occurs when the signal to be quantized is outside the range of the quantizer (± 2.61 in Fig. 11), is a part of the quantizing noise which is minimized here. It must be considered separately only when the range of the quantizer is so small that overload causes a significant alteration in the probability density function of (14).

VII. COMPUTER SIMULATION OF DPCM SYSTEMS

The results of computer simulations verify that systems designed by the procedures presented here do function as predicted.

By applying the principles outlined herein, the parameters for some DPCM systems were determined and these systems were simulated on the IBM 7094 digital computer The input signals used were the 100 line, 100 samples per line television pictures obtained from the scenes of Fig. 6. Section 5.3 contains a description of how these signals were obtained. The results of the simulation are shown in Figs. 12 and 13. The S/N ratios in these figures are ratios of rms video signal to rms

* This is the inversion of (A-6) of Ref. 21.

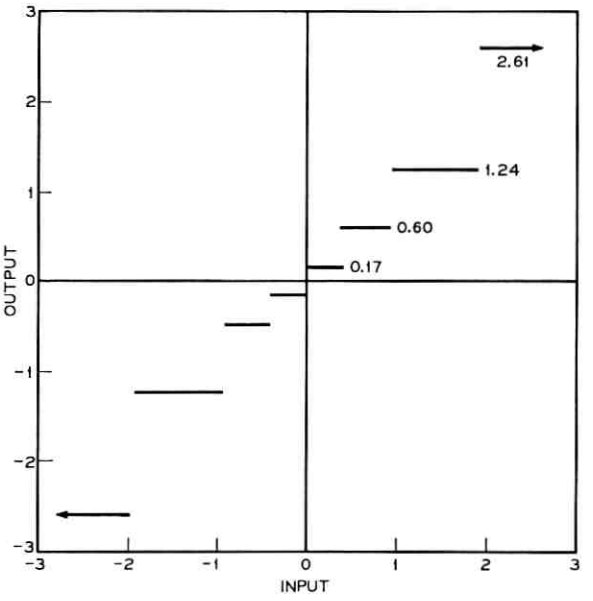

Fig. 11 — Typical 8-level exponential quantizer characteristic obtained from (21). (Case shown is for $m = 5.5$ and $V = 7$.)

noise in the video. To get the ratio of peak-to-peak composite signal to rms noise in the video from these curves, we must add C, the ratio in db of peak-to-peak composite signal to rms video. For scenes A, B, and C this value of C is 20.0, 19.8, and 18.1, respectively (assuming FCC sync standards). Also plotted in these figures are curves of S/N ratio which were predicted for these systems using (16). Fig. 12 gives the results for previous-sample feedback systems (predictor number 1 in Table I), and Fig. 13 presents results for line-and-sample feedback systems (predictor number 3 in Table I). Both figures illustrate the performance of systems whose predictors are tailored to the incoming signal. Table I illustrates that the use of more complicated predictive systems, using samples S_3, S_4, and S_5 in addition to S_1 and S_2, does not significantly lower the rms error in the prediction. Furthermore, the optimum designs of predictors 6, 7, and 8, which give only a slight decrease in the rms error, are radically different for scenes A, B, and C. A system using predictor 6, 7, or 8 designed to give good performance for scene B is likely to give poor performance for scenes A and C. This is not the case with predictors 1 and 3 which were simulated. To verify this, previous-sample feedback DPCM systems were simulated for 4, 5,

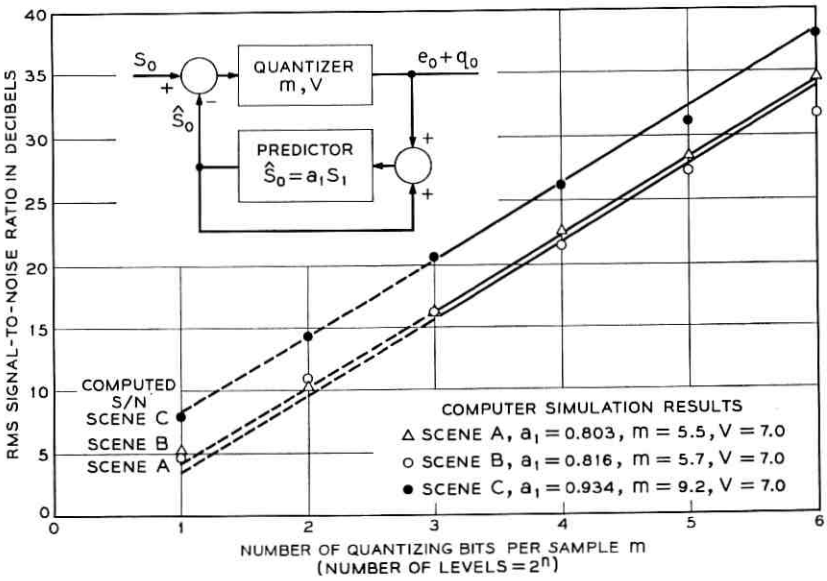

Fig. 12 — Ratio of rms video to rms quantizing noise in the video for systems matched to each scene using previous-sample feedback DPCM. (Theoretical results from (16) [straight lines] are compared with results of computer simulation.)

and 6-bit encoding with $a_1 = 0.815$, $m = 6$ and $V = 7$. The signals from all three scenes were used as inputs and the results were almost identical to those shown in Fig. 12. Similarly, line-and-sample feedback DPCM systems were simulated for 4, 5, and 6-bit encoding with $a_1 = 0.315$, $a_2 = 0.650$, $m = 8$ and $V = 7$, and the three signals all produced S/N ratios essentially the same as those in Fig. 13. The parameters in these two DPCM systems are not critical and need not be exactly matched to the picture material from which the incoming signals were obtained.

VIII. THE MARGINAL UTILITY OF LINE FEEDBACK

For the 100 by 100 matrix pictures used in this study, the use of line feedback increased the S/N ratio by only about 2 or 3 db. This increase is small because the sample values S_1 and S_2 contain substantially the same information about S_0, the sample value to be predicted. And, once S_1 has been used in the prediction, there is only a 2 or 3 db advantage in simultaneously using S_2 in the prediction.

To illustrate this point, consider a scene whose contours of equal

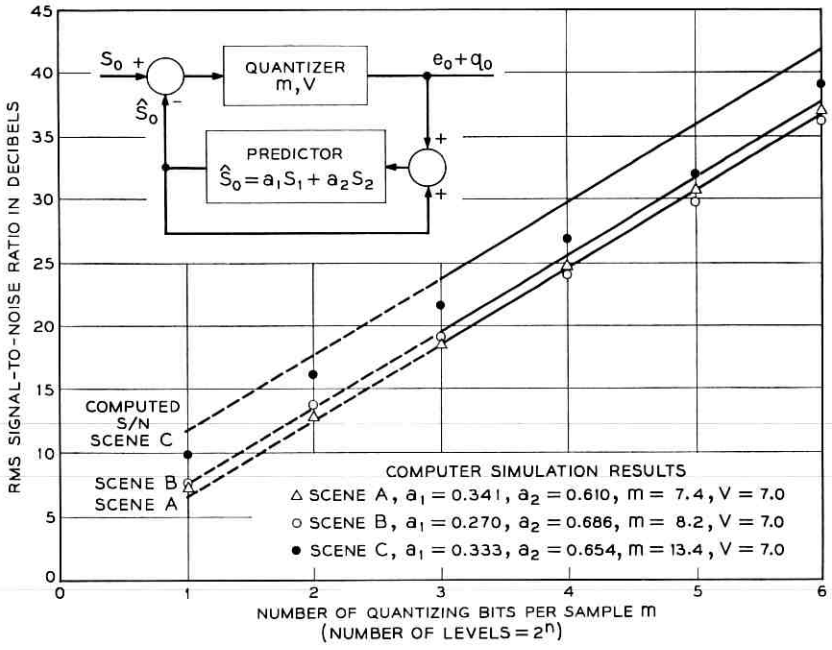

Fig. 13 — Ratio of rms video to rms quantizing noise in the video for systems matched to each scene using line-and-sample feedback DPCM. (Theoretical results from (16) [straight lines] are compared with results of computer simulation.)

autocovariance are circles. Further assume that the autocovariance between any two points, S_i and S_j, separated by a distance D can be expressed as $R_{ij} = \sigma^2 e^{-\alpha D}$. Both of these assumptions are reasonable ones for television picture material. If S_1 (the previous sample) and S_2 (the adjacent sample on the previous line) are equidistant from S_0, then $R_{01} = R_{02}$ and $R_{21} = \sigma^2(R_{01}/\sigma^2)\sqrt{2}$. From (5) the values of the coefficients a_1 and a_2 are

$$a_1 = a_2 = \frac{R_{01}}{\sigma^2 + \sigma^2(R_{01}/\sigma^2)\sqrt{2}} \tag{22}$$

$$\sigma_e^{2'} = \sigma^2\left[1 - \frac{2(R_{01}/\sigma^2)^2}{1 + (R_{01}/\sigma^2)\sqrt{2}}\right]. \tag{23}$$

Compare this with the mean square value of the prediction error when only S_1 is used in the prediction

$$\sigma_e^2 = \sigma^2 - R_{01}^2/\sigma^2. \tag{24}$$

In Fig. 14, the advantage to be gained in a DPCM system by pro-

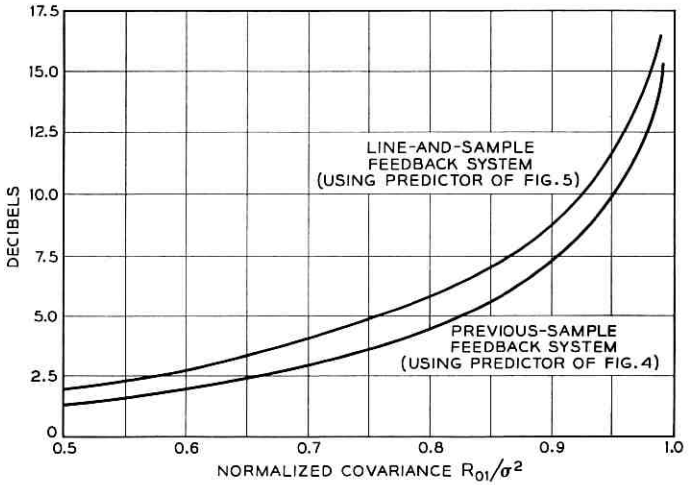

Fig. 14 — Comparison of line-and-sample feedback system and previous-sample feedback system when $R_{01} = R_{02}$. (To get actual peak composite signal to rms noise ratios in an n-digit television system, add $13.5 + 6n$ db.)

viding a line-and-sample feedback predictor is compared with that of a simple previous-sample predictor. This figure applies to sequentially scanned television systems in which the covariance between adjacent samples on the same line is equal to the covariance between neighboring samples on adjacent lines, i.e., $R_{01} = R_{02}$. In television systems using interlace, the S/N ratio improvement provided by using line feedback in addition to sample feedback will be even less. The two curves in this figure are simply plots of 10 log σ^2/σ_e^2 where the value of σ_e^2 is given by (23) for the line-and-sample feedback system, and by (24) for the previous-sample feedback system. From (17) we see that the term log σ^2/σ_e^2 represents the S/N improvement to be expected from the feedback loop. The two curves in Fig. 14 then show the value of the feedback loops for the two predictors of interest. The distance between the two curves is the improvement provided by the line-and-sample feedback system over the previous-sample feedback system. It can be shown that the maximum value of this improvement approaches about 1.9 db and this occurs as R_{01}/σ^2 approaches 1. In other words, in television signals whose samples have the same covariance in the horizontal direction as in the vertical direction, a line-and-sample feedback system can provide, at best, only 1.9-db improvement in S/N ratio over a simple previous-sample feedback system.

For the scenes used in this simulation, the line feedback loop provided S/N ratio improvements between 2 and 3 db. This was more than the

1.9 db maximum because the scenes used had higher vertical than horizontal correlation. When line-and-sample feedback DPCM is used, sequentially scanning a scene can provide as much as, but no more than, 3-db improvement in S/N ratio over 2:1 interlaced scanning with the same number of lines. This is true because the value of $1 - (R_{02}/\sigma^2)$ for the sequential scanning is about half of what it would be for interlaced scanning. Exactly how much improvement is afforded by sequential scanning depends on the values of R_{01}, R_{02}, and R_{12} which are determined by the scene scanned as well as by the television standards used.

The lower curve in Fig. 14 can also be used to predict the advantage to be gained by frame feedback DPCM. In this case, the abscissa would be R_{0F}/σ^2 where R_{0F} is the covariance between equivalent points on adjacent frames. The S/N ratio for a frame feedback system is given by (20) if R_{01} is replaced by R_{0F}. Some measurements by Kretzmer[5] and Deriugin[15] suggest that R_{0F}/σ^2 may, in general, be less than R_{01}/σ^2. This implies that frame feedback may be of little value in reducing the S/N ratio in DPCM systems.

IX. DPCM FOR MONOCHROME ENTERTAINMENT TELEVISION

In 4.5-Mc/s entertainment black and white television there is little advantage in basing the prediction on any sample values except the previous one, unless, of course, sample values from previous fields are available. For this previous-sample feedback system the approximate value of the S/N ratio to be expected is given in (20). In a 525-line picture at a frame rate of 30 per second, sampling at twice the bandwidth or 9 Mc/s means that there are about 571 samples per line. Only 83 per cent of these, or 474, occur in the video while the others occur during the horizontal and vertical sync pulses.

Using simple linear interpolation* we see that, if scene A of Fig. 6 were sampled at 9 Mc/s, the covariance between adjacent points would be $R_{01} \cong 0.958$. For scene C, $R_{01} \cong 0.986$. Using these numbers and the appropriate values of the constant C in (20), and remembering that $\sigma^2 = 1$ for these signals, the S/N ratios to be expected for transmitting these scenes over a DPCM channel can be found. The S/N ratio for scene B would fall somewhere between those of scenes A and C. These S/N ratios are compared with standard PCM and delta modulation in Fig. 15. For the lower bit rates, these curves must be used with discretion. The line representing the PCM performance is simply a plot of (18) which assumes that the sync is transmitted. The PCM S/N

* Since the aspect ratio is 4:3 we could not transmit these pictures as they are. We assume here that either the top or bottom $\frac{1}{4}$ of the pictures is not transmitted.

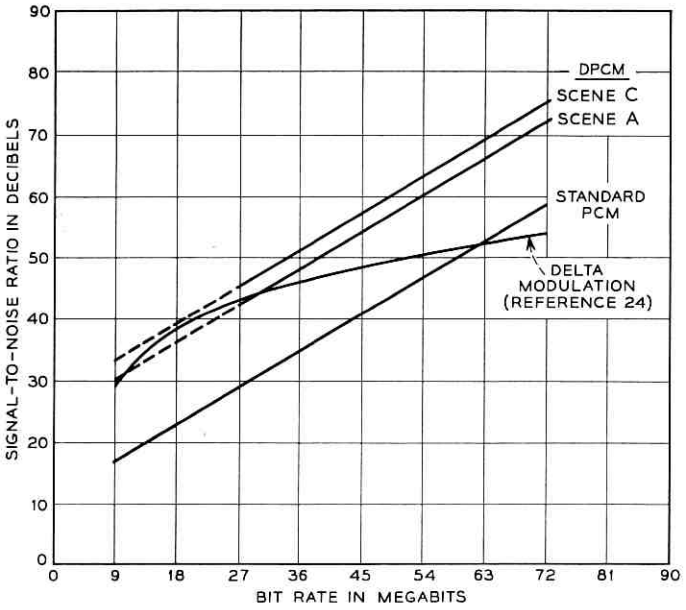

Fig. 15 — Comparison of previous-sample feedback DPCM with standard PCM and delta modulation for monochrome 4.5-Mc/s entertainment television. (Peak-to-peak composite signal to rms quantizing noise in the video is the signal-to-noise ratio shown. The sampling rate is 9 Mc/s.)

ratios can be increased by 2.8 db if the sync is not transmitted. The delta modulation S/N ratios were found by an entirely different technique[24] and it is gratifying that they are reasonably consistent with the results found here for 1 digit DPCM, which of course, is identical to delta modulation with a sampling rate of twice the bandwidth.

Fig. 15 shows that, for a fixed bit rate, DPCM would give a 14-db improvement over standard PCM for scene A and an 16.8-db improvement for scene C. The advantage of DPCM can also be expressed in terms of bit rate. For a given S/N ratio, DPCM gives a reduction in bit rate over standard PCM of about 18 megabits (2 bits/sample). Since the sampling rate for DPCM and PCM is assumed to be twice the bandwidth or 9 Mc/s, these curves in Fig. 15 are actually defined only at multiples of 9 megabits.

X. THE CHARACTER OF THE QUANTIZING NOISE

Fig. 16 illustrates the autocovariance for the quantizing noise in a previous-sample feedback system. The cases shown are for scene B but these curves are typical of all three scenes. The exact autocovariance

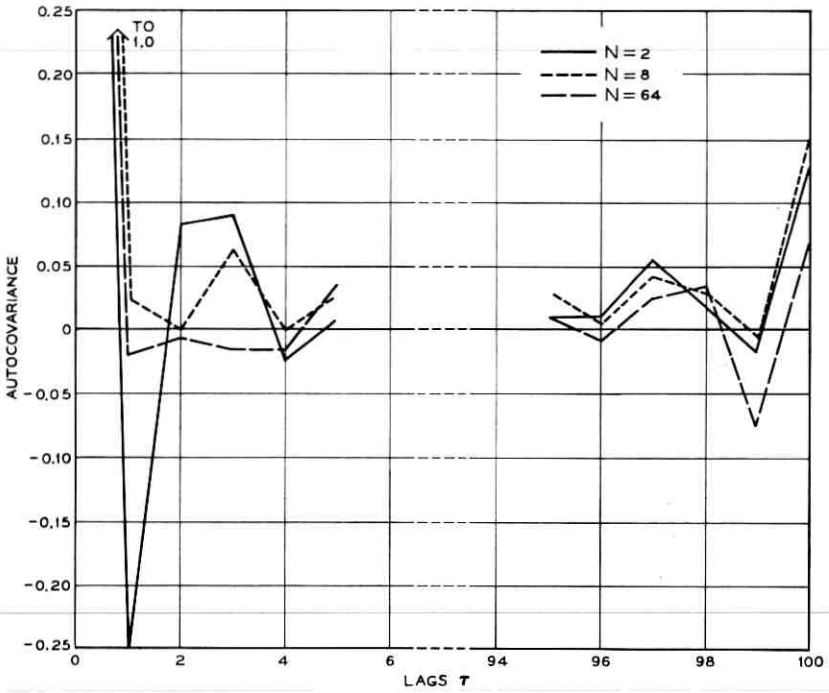

Fig. 16 — Autocovariance of quantizing noise for $N = 2$, 8, and 64 level previous-sample feedback DPCM. (Case shown is for scene B with the DPCM system optimized for this scene.)

function of the quantizing noise depends on the picture material of the incoming signal. These autocovariance functions were essentially the same when the parameters of the DPCM system did not exactly match the statistics of the incoming signal. The spectra of the quantizing noise in the three scenes is found by taking the Fourier transforms of their autocovariance functions. These spectra were found to be relatively flat for both previous-sample feedback and line-and-sample feedback systems. In both cases, there were erratic peaks and valleys at multiples of the line rate but the peaks and valleys differed from each other only by about 2 to 4 db, these differences being slightly greater for the previous-sample feedback system than for the line-and-sample feedback systems. In neither case did the spectra show any general tendency to increase or decrease for higher frequencies. Fig. 16 illustrates that the correlations between sample values is quite weak. The usual assumption of flat quantizing noise in DPCM systems is probably a good one for most purposes.

Fig. 17 shows a plot of the amplitude density function of the quantizing noise for scene B. It is typical of all the scenes that the amplitude density is relatively flat for a small number of quantizing levels N and becomes more Gaussian shaped as N gets large. In all cases, however, even when N was 64, the amplitude density function was flatter than Gaussian. For all three scenes with $N = 2$ the quantizing noise amplitude density function had a dip near zero.

XI. THE PENALTY

Removing the redundancy from the transmitted signal has the disadvantage that the signal becomes more vulnerable to noise introduced in the medium of transmission. This is true of predictive systems, in general, whether or not they are digital. A technique for reducing the redundancy in analog television signals by linear filtering has been proposed by Franks.[14] The similarity between this analog technique and DPCM is apparent. The utility of digital transmission itself is simply that it provides a desirable trade of bandwidth for noise immunity in the transmission medium. We may think of DPCM as a counter-trade. For a given amount of quantizing noise, DPCM allows transmission at a lower bit rate (and therefore bandwidth) than standard PCM. Errors in the transmission channel, however, degrade the decoded DPCM signal more than they would in standard PCM.

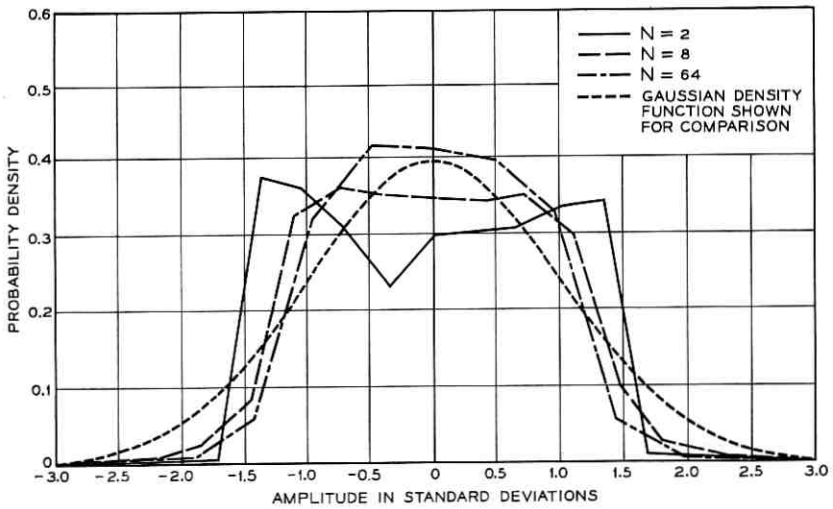

Fig. 17 — Quantizing noise amplitude density functions for $N = 2$, 8, and 64 level previous-sample feedback DPCM. (Case shown is for scene B with the DPCM system optimized for this scene.)

The decoder in a DPCM system is a linear device which operates on an incoming sequence with rms value σ_e to produce a decoded output with rms value σ. Just as the decoder increases the level of the incoming signal by $20 \log \sigma/\sigma_e = k$ db, it will also increase the level of accompanying noise (caused by digit errors in the transmission channel) by k db.

This is easily illustrated by considering the decoder of the previous-sample feedback system with the predictor shown in Fig. 4. Noise η, caused by a digit error, on a member of the incoming sequence is fed through the feedback loop and occurs on all subsequent samples. Such a noise on a single member of the incoming sequence causes the error sequence $\eta, a_1\eta, a_1^2\eta, a_1^3\eta, \ldots$, in the decoded output sequence. The noise energy in the decoded signal is, therefore

$$\eta^2 + (a_1\eta)^2 + (a_1^2\eta)^2 + \cdots = \eta^2\left(\frac{1}{1 - a_1^2}\right).$$

For the properly designed previous-sample feedback system $a_1 = R_{01}/\sigma^2$ and $1/(1 - R_{01}^2/\sigma^4) = \sigma^2/\sigma_e^2$. Therefore, a noise of energy η^2 in the transmission channel appears in the decoded signal as noise with energy $\eta^2(\sigma^2/\sigma_e^2)$, a gain of $10 \log \sigma^2/\sigma_e^2$ db.

We have already shown that DPCM provides a decrease in quantizing noise of about $10 \log \sigma^2/\sigma_e^2$ db over standard PCM (assuming the sync pulses are not transmitted). The penalty paid for this decrease in quantizing noise is that the noise in the decoded signal introduced in the transmission medium is increased by exactly that same amount. This does not mean that DPCM provides no advantage. For, in digital systems, noise introduced in the transmission medium can be made extremely small and the limiting degradation in DPCM systems is generally quantizing noise. Decreasing the quantizing noise by k db may be desirable even if the noise introduced in the channel is increased by this amount.

When the probability P of a digit error in the transmission medium is small enough so that the probability of getting two errors in the same word may be neglected, then the noise power N_t in the decoded output introduced by the transmission medium is directly proportional to P and we can express N_t (in db) as

$$N_t = K_1 + 10 \log P + 10 \log \sigma^2/\sigma_e^2. \tag{25}$$

From (17) the quantizing noise N_q can be expressed (in db) as

$$N_q = K_2 - 10 \log \sigma^2/\sigma_e^2. \tag{26}$$

The constants K_1 and K_2 are both dependent on the number of quantizing levels as well as other parameters. The term $10 \log \sigma^2/\sigma_e^2$ represents

the effect of DPCM in both equations. Reducing the quantizing noise N_q by k db through DPCM requires increasing 10 log σ^2/σ_e^2 by this amount and this increases the noise N_t introduced in the medium of transmission by k db. Whether or not DPCM can be used to advantage depends on the relative importance of N_t and N_q in limiting the performance of the system. From (25) we see that if we require N_t to remain constant while reducing the quantizing noise by k db, we must reduce the term 10 log P by k db. This requires reducing the value of P by a factor of $10^{0.1\,k} \cong (1.26)^k$.

XII. ACKNOWLEDGMENT

The author expresses his gratitude to T. V. Crater, whose insight and encouragement were essential ingredients in the formulation of the ideas presented in this paper, and to A. D. Hall and J. E. Abate who contributed many helpful suggestions.

REFERENCES

1. French Patent No. 987,238, August 10, 1951 (applied for May 23, 1949 by the N. V. Phillips' Gloeilampenfabrieken of Holland).
2. Cutler, C. C., Differential Quantization of Communication Signals, Patent No. 2,605,361, July 29, 1952 (applied for June 29, 1950).
3. Weiner, N., *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, MIT Press, Cambridge, Massachusetts, 1949.
4. Oliver, B. N., Efficient Coding, B.S.T.J., *31*, July, 1952, pp. 724–750.
5. Kretzmer, E. R., Statistics of Television Signals, B.S.T.J., *31*, July, 1952, pp. 751–763.
6. Harrison, C. W., Experiments with Linear Prediction in Television, B.S.T.J., *31*, July, 1952, pp. 764–783.
7. Elias, P., Predictive Coding, IRE Trans. Inform. Theor., *IT-1*, March, 1955, pp. 16–33.
8. Graham, R. E., Predictive Quantizing of Television Signals, IRE Wescon Convention Record, Part 4, August, 1958, pp. 147–156.
9. Kimme, E. G. and Kuo, F. F., Synthesis of optimal Filters for a Feedback Quantization System, IEEE Trans. Circuit Theor., *CT-10*, No. 3, Sept., 1963, pp. 405–413.
10. Panter, P. F. and Dite, W., Quantization Distortion in Pulse-Count Modulation with Nonuniform Spacing of Levels, Proc. IRE, *39*, January, 1951, pp. 44–48.
11. Fine, T., Properties of an Optimal Digital System and Applications, IEEE Trans. Inform. Theor., *IT-10*, October, 1964, pp. 287–296.
12. Papoulis, A., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill Book Company, Inc., New York, 1965.
13. Graham, R. E. and Kelly, J. L., Jr., A Computer Simulation Chain for Research on Picture Coding, IRE Wescon Conv. Rec., Part 4, August, 1958, pp. 41–46.
14. Franks, L. E., A Model for the Random Video Process, B.S.T.J., *45*, April, 1966, pp. 609–630.
15. Deriugin, N. G., The Power Spectrum and Correlation Function of the Television Signal, Telecommunications, *7*, 1957, pp. 1–12.
16. Oliver, B. N., Pierce, J. R., and Shannon, C. E., The Philosophy of PCM, Proc. IRE, *36*, November, 1948, pp. 1324–1331.

17. Mayo, J. S., A Bipolar Repeater for Pulse Code Modulation Signals, B.S.T.J., *41*, January, 1962, pp. 25–97.
18. Nitadori, K., Statistical Analysis of DPCM, Electronics and Communications in Japan (English Translation of The J. Inst. Elec. Commun. Eng. Jap.), *48*, No. 2, Feb., 1965.
19. O'Neal, J. B., Jr., A Bound on Signal-to-Quantizing Noise Ratios for Digital Encoding Systems, to be published.
20. Bennett, W. R., Spectra of Quantized Signals, B.S.T.J., *27*, July, 1948, pp. 446–472.
21. Smith, B., Instantaneous Companding of Quantized Signals, B.S.T.J., *36*, May, 1957, pp. 653–709.
22. Max, J., Quantizing for Minimum Distortion, IRE Trans. Inform. Theor., *IT-6*, 1960, pp. 7–12.
23. Bruce, J. D., Optimum Quantization, MIT Research Lab. of Electronics, Tech. Report No. 429, March 1, 1965.
24. O'Neal, J. B., Jr., Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Inputs, B.S.T.J., *45*, January, 1966, pp. 117–141.

Observed 50 to 60 Gc/s Attenuation for the Circular Electric Wave in Dielectric-Coated Cylindrical Waveguide Bends

By GLENN E. CONKLIN

(Manuscript received March 8, 1966)

A thin low-loss dielectric coating inside a 0.875-inch i.d. round waveguide was achieved by drawing a slightly oversize flexible tube of polyethylene into the waveguide. The plastic tube fits snugly to the guide wall. Several 14-foot lengths of waveguide were lined and loss measurements made. The lined pipes were 8 per cent lossier than the unlined. The increase in loss could be almost completely accounted for by the theory of H. Unger.

A 92° Fresnel integral bend (curvature a linear function of arc length) was constructed from a 14-foot section of the 0.875-inch i.d. lined waveguide. The insertion loss between 50 and 60 Gc/s had a value of about 0.1 db. This small value is close to the theoretical result for such bends.

I. INTRODUCTION

In a nonideal section of circular cylindrical waveguide, the circular electric wave TE_{01} couples to other modes of transmission which are permitted.[1] The strongest coupling for bends is to the TM_{11} mode since it is degenerate in a perfectly conducting straight waveguide. The degeneracy of the phase factors of the TE_{01} and TM_{11} modes in the uncoated cylindrical waveguide gives rise to the serious problem of an increase in loss upon bending. For example, in a 2-inch i.d. bare pipe at a 5.4-mm wavelength, the attenuation constant is doubled when the radius of curvature is a few miles.[1] The power transfer between the two modes can be reduced by removing the phase degeneracy. One technique for doing this is to introduce a thin layer of dielectric next to the wall of the waveguide. The electric field intensity of the TE_{01} mode vanishes at the wall but that of the TM_{11} mode has a large value there, hence, the effect on the propagation constant of the TM_{11} mode is larger than on the TE_{01}.

The solution of the characteristic equation for the normal modes of the circular cylindrical waveguide with dielectric coating obtained by Unger and Morgan[1,2] indicates there is an increase in attenuation of the TE_{01} mode attributable to dielectric loss and an added copper loss arising from the tendency of the electric field intensity to concentrate in the dielectric next to the wall. Attenuation measurements on such a straight length of cylindrical waveguide are reported in Section II and compared to theoretical predictions insofar as is possible.

A bend for which the curvature is a linear function of bend length follows the Fresnel integral curve. This is the elastic bend which is obtained by applying forces at the center and at the two ends of a rod. According to Unger, a bend in a dielectric-coated waveguide with such a curvature has a minimum of mode conversion, hence, minimum loss.[3] Measurements on a 92° bend with the linearly tapered curvature are discussed in Section III.

II. STRAIGHT DIELECTRIC-COATED WAVEGUIDE

Long straight sectious of dielectric-lined waveguide were prepared by using the technique of pulling a slightly oversize polyethylene liner into an oxygen-free copper waveguide. Only a small length of the liner was in contact with the copper wall during the pulling. The remainder was stretched enough elastically to have an outside diameter less than the inside diameter of the pipe. The liner expanded in place against the waveguide wall after the tension was removed.

The polyethylene sleeve was made by extrusion molding from a crease-resistant formulation of polyethylene and rubber.[4] The dielectric constant and loss tangent at 55.2 Gc/s of this material were measured[5] and are 2.25 and 1×10^{-4}. This means that the material is very good from the dielectric attenuation point of view. A finished extrusion tended to have small air pockets and trapped dust particles. Another difficulty existed with the extrusion. The inner die happened to be slightly eccentric with the outer during manufacture. A maximum wall thickness t_{max} of 0.0134 inch was measured on one side and a minimum t_{min} of 0.0095 inch on the other. Furthermore, the tubing had hash marks remaining from the water cooling used during extrusion.

The TE_{01} attenuation coefficient of the dielectric-coated waveguide was measured by observing the 3-db bandwidth of a cavity containing the test section. Any cavity technique for measuring the attenuation coefficient implies the use of a considerable amount of auxiliary equipment. The cavity quality test set which was used, operated between 50 and 60 Gc/s and used an M-1977 backward-wave oscillator.[5] The cavity

quality was measured by sweeping the frequency of the oscillator through the cavity resonance and then observing the frequency difference between the half-power points of the transmission curve. The apparatus was made sensitive to dispersion rather than absorption so that the measurements would be relatively insensitive to drifts in amplifier gain. The cavity assembly consisted of a two-hole coupler constructed from helix waveguide, the long section of waveguide to be studied, a 12-inch section of helix waveguide, and a piston as a shorting termination. The piston had a port for removal of air from inside the cavity. A short section of helix in the cavity was necessary in order to remove unwanted modes. Mode interference is manifest by observing markedly increased bandwidth values for some positions of resonance for the piston; thus, the minimum measured TE_{01} bandwidth was selected as the most probably correct value.

The long section of waveguide under test has a quality factor Q_2 given by the expression[6]

$$1/Q_2 = 2\alpha_2\beta_2/(\beta_{c2}^2 + \beta_2^2). \tag{1}$$

The parameters α_2 and β_2 are the real and imaginary portions of the propagation constant of the TE_{01} wave in the waveguide section under consideration. The factor β_{c2} is the free-space propagation constant corresponding to the TE_{01} cutoff frequency. The reciprocal cavity quality factor in terms of the measured 3-db frequency bandwidth Δf_m is $1/Q_m = \Delta f_m/f$ and is the sum of two portions. The first is that associated with the intrinsic attenuation of the waveguide, $\Delta f_2/f = 1/Q_2$. The second is associated with the attenuation at the ends of the cavity, b/L, where b is a constant characteristic of the ends and L the length of the cavity. The resulting expression for the measured cavity quality factor is

$$\frac{1}{Q_m} = \frac{\Delta f_m}{f} = \frac{\Delta f_2}{f} + \frac{b}{L}. \tag{2}$$

The expression for the attenuation factor of the test section becomes

$$\alpha_2 = \frac{\beta_{c2}^2 + \beta_2^2}{2\beta_2 f}\left(\Delta f_m - \frac{fb}{L}\right). \tag{3}$$

Unlined evacuated cylindrical copper waveguides were measured first in order to establish the end loss corrections and the loss characteristics of the bare copper pipe to be used later for lining and in making bends. These parameters were found by measuring the cavity bandwidth as a function of length. The intrinsic bandwidth of the test section was given directly from a plot of Δf_m versus $1/L$ by the intercept and the end

loss factor calculated from the slope. The attenuation coefficients in Fig. 3 of the bare copper pipes agree well with the results of King.[7]

The losses due to air were determined using a long section of unlined copper waveguide open to the atmosphere. These measurements were needed in order to correct the measurements taken on the dielectric-coated waveguide. The dielectric-coated waveguide could not be evacuated since an occasional small pocket of air between the liner and the copper wall caused the liner to collapse upon evacuation. The plot of the measured atmospheric losses versus frequency in Fig. 1 is in satisfactory agreement with the results of others.[8]

An initial check of the placement of the dielectric coating inside the guide was made by fabricating many sections of various lengths and then measuring the bandwidth of the cavity formed from each. The bandwidth measurements were made at a fixed frequency near 55 Gc/s (Fig. 2). Of interest are the points associated with the bare pipe. These are not scattered appreciably, which means that all of the bare pipes are quite uniform. The circled points associated with the liner show considerable scattering. Apparently the liner was not in close enough contact with the copper wall. Any blister in the liner causes the lined pipe to be lossy. The liner contact with the wall was improved by alternately cooling and warming for several cycles. A new set of measurements were made, and the decrease in attenuation indicated by the triangular points was observed. Next, a 14-foot section of the shrunken dielectric-coated guide was measured across the frequency band from 50 to 60 Gc/s. The increase in the attenuation coefficient of the waveguide due to the liner

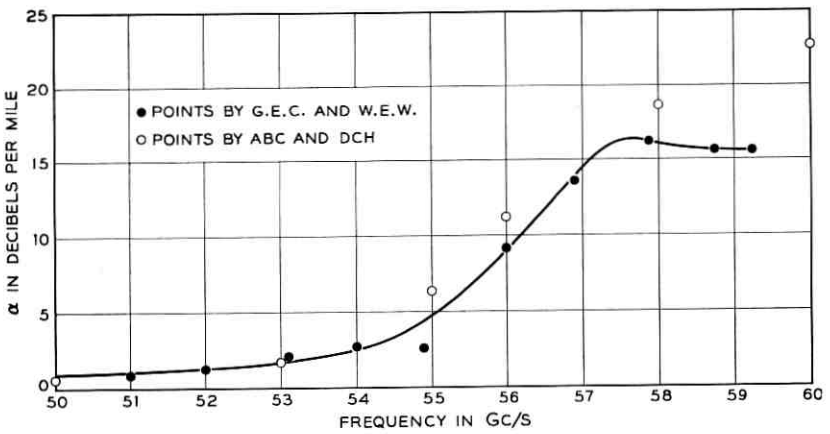

Fig. 1 — Attenuation of air in 0.875-inch i.d. waveguide.

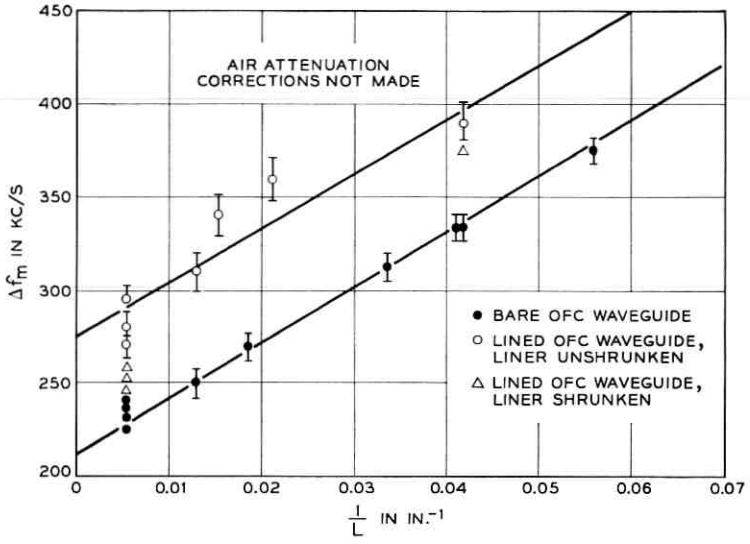

Fig. 2 — Cavity bandwidth as a function of reciprocal length.

$\Delta\alpha_l$ was calculated by subtracting the losses due to air α_a and the measured unlined waveguide α_{01} from the measured attenuation coefficient α_m at each frequency. The results for this section of coated guide are given in Fig. 3.

The increase in the TE_{01} attenuation coefficient of the dielectric-lined waveguide due to the dielectric loss in the liner is given by Unger as

$$\Delta\alpha_D = 1.835 \times 10^5 \frac{k_0^2 p_{01}'^2 t^3}{\beta_{01} a^3} \epsilon'' \text{ db/mile}, \tag{4}$$

and the increase in copper loss due to the liner being in contact with the waveguide wall as

$$\Delta\alpha_\omega = \alpha_{01}(\epsilon' - 1)k_0^2 t^2. \tag{5}$$

In these equations, $\epsilon^* = \epsilon' - j\epsilon''$ is the dielectric constant and t the thickness of the coating, a the radius of the waveguide, k_0 the phase propagation constant in unbounded vacuo, p_{01}' is the first root of the Bessel equation $J_0'(p_{01}'a) = 0$, and β_{01} is the phase propagation coefficient of the TE_{01} wave in the waveguide. Equations (4) and (5) are perturbation expressions for which β_{01} of the lined waveguide is assumed to be the same as that of the bare pipe.

An analysis of the increase in attenuation coefficient of the dielectric-

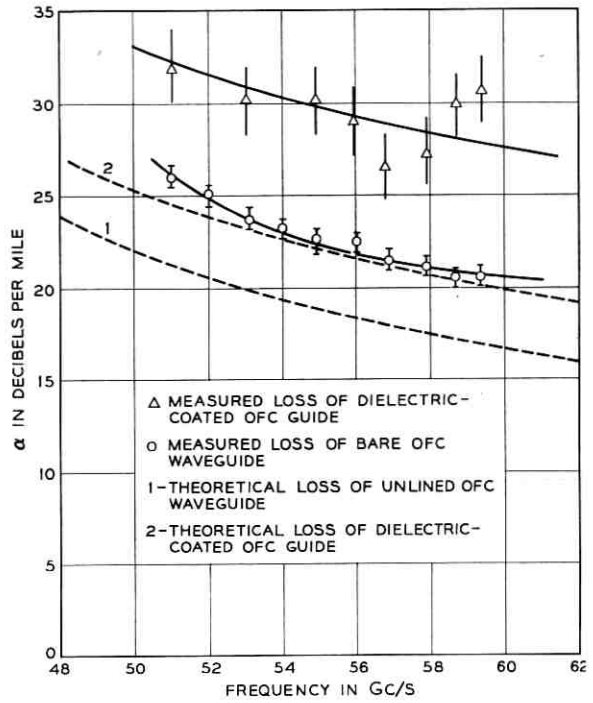

Fig. 3 — Attenuation of coated and uncoated waveguide.

coated waveguide due to the coating is dependent upon knowing the magnitude of the coating thickness. However, the inside and outside radii associated with the sleeve were slightly eccentric. A first approach to the problem of eccentricity can be made by assuming that the functions for calculating the attenuation coefficient in cylindrical geometry are changed only negligibly by the introduction of the eccentricity. In such case the average values for the quantities t, t^2, and t^3 can be used,

$$\bar{t} = \tfrac{1}{2}(t_{\max} + t_{\min}), \tag{6}$$

$$\overline{t^2} = \bar{t}^2 + \frac{c^2}{2}, \tag{7}$$

$$\overline{t^3} = \bar{t}^3 + \frac{3c^2}{2}\,\bar{t}, \tag{8}$$

where c is the distance of separation of centers

$$c = \tfrac{1}{2}(t_{\max} - t_{\min}). \tag{9}$$

Substitution of the measured values of t_{max} and t_{min} into (6), (7), (8), and (9) indicates that the arithmetic average is suitable for use in the calculations. The measured average wall thickness for the tubing used in this work was 0.0115 inch and the separation of centers was 0.0020 inch.

The values for $\Delta\alpha_D$ and $\Delta\alpha_\omega$ were calculated for several frequencies and their sum is compared in Fig. 4 to the values measured for a long section of the dielectric-coated waveguide. There is apparently some conversion of TE_{01} to other modes of propagation.

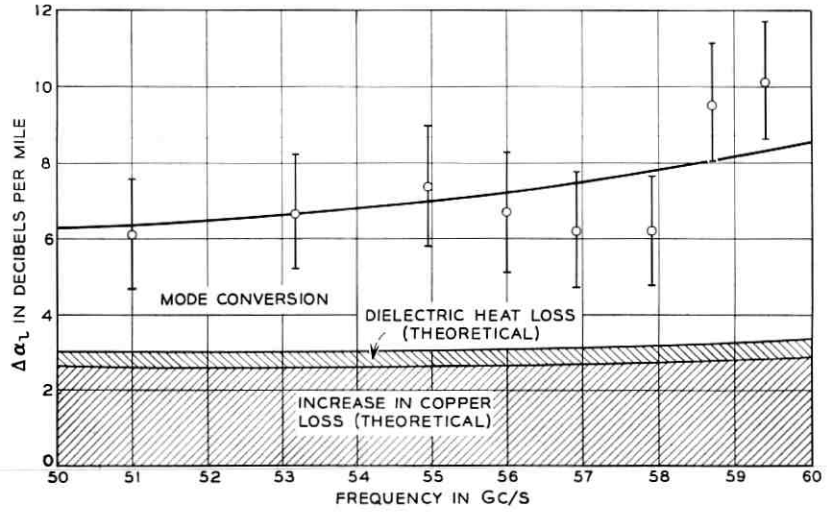

Fig. 4 — Increase in attenuation $\Delta\alpha_1$ of 0.875-inch i.d. copper waveguide due to linear.

III. DIELECTRIC-COATED WAVEGUIDE BEND

A typical bend was constructed to fit into an existing waveguide transmission system. This required that the angle of bend θ_0 be 92° and the length l_b be 14.135 feet. In order to design a linearly tapered bend with a given bend angle and length, consider a bend with the curvature being an arbitrary function of the length of arc z taken along the axis of the pipe $K = f(z)$. The curvature of the bend described in the rectangular coordinate system (Γ, Σ) of Fig. 5 is given by the expression

$$K = \left[1 + \left(\frac{d\Gamma}{d\Sigma}\right)^2\right]^{-\frac{3}{2}} \frac{d^2\Gamma}{d\Sigma^2}. \tag{10}$$

Of supplementary use is the equation for the differential length of arc,

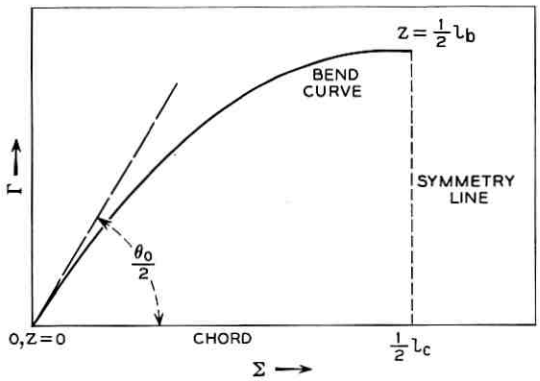

Fig. 5 — Coordinate system for half of the bend.

$$dz = \left[1 + \left(\frac{d\Gamma}{d\Sigma}\right)^2\right]^{\frac{1}{2}} dS. \tag{11}$$

Equation (10) is solvable in terms of the length of bend under the conditions that at the origin $z = 0$, $K = 0$, and $d\Gamma/d\Sigma \,\lrcorner_{z=0} = \tan \theta_0/2$. The solution is then:

$$\Sigma = \int_0^z \cos\left(\frac{\theta_0}{2} + \int_0^z K dz_1\right) dz \tag{12}$$

$$\Gamma = \int_0^z \sin\left(\frac{\theta_0}{2} + \int_0^z K dz\right) dz. \tag{13}$$

The chord length l_c of a symmetrical bend of length l_b is

$$l_c = 2\int_0^{l_b/2} \cos\left(\frac{\theta_0}{2} + \int_0^z K dz\right) dz. \tag{14}$$

When the curvature is made a linear function of length of arc, $K = -k'z$ with k' being a constant, the result is a Fresnel integral curve. The curvature parameter k' can be evaluated in terms of design parameters from the definition of the differential bend angle and the radius of curvature,

$$d\theta = \frac{1}{R} dz = K dz = -k' z dz. \tag{15}$$

The half bend angle occurs over the half-length of the pipe; thus,

$$k' = 4\theta_0/l_b^2. \tag{16}$$

Equations (12), (13), and (16) can be used to calculate the coordinate points for a Fresnel integral bend of arbitrary length and bend angle.

A centerline curve was calculated and a bend constructed to fit the desired design parameters. A convenient technique for making a bend with linearly tapered curvature is to form the waveguide around a wooden jig. The waveguide is then clamped in place. The attenuation of the bend was measured using the resonant cavity technique described previously and the results are plotted in Fig. 6. The attenuation was found to decrease slightly with increasing frequency. Theoretical values for the attenuation of the bend can be calculated from the equations of Unger and Morgan[2,3] and are shown in Fig. 6.

IV. CONCLUSIONS

The straight waveguide with liner had a slightly greater TE_{01} transmission loss than the bare pipe. Part of the loss, as suggested by Unger and Morgan, arises from dielectric heat loss and an increase in copper loss. The remaining portion of the added loss could probably be decreased because the liner was not as smooth and did not fit as snugly to the wall as desired. Extrusion of the dielectric directly inside the waveguide should yield improved performance.

The Fresnel integral bend as constructed had an insertion loss of about 0.1 db between 50 and 60 Gc/s, which is close to the theoretical

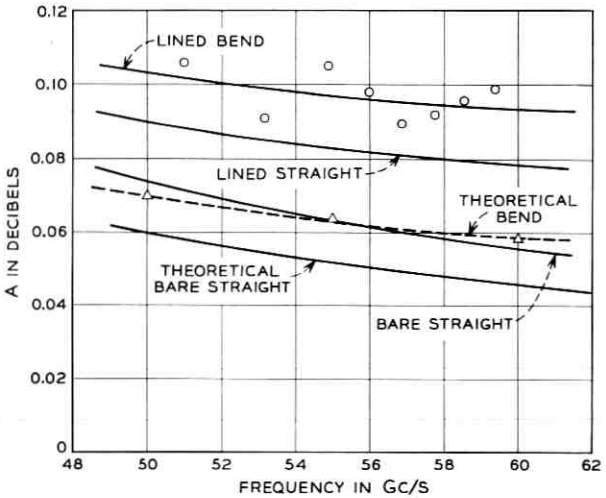

Fig. 6 — Total attenuation of the lined bend compared to straight section of equal length.

value. Lining the round waveguide removed the TE_{01} and TM_{11} mode degeneracy and most of the large loss associated with bending. This enabled the construction of a new and useful device.

The author wishes to express his appreciation to D. H. Ring for his critical review of this article and to J. W. Bell and W. E. Whitacre for aid in construction and test of the bend.

REFERENCES

1. Unger, H. G., Circular Electric Wave Transmission in Dielectric-Coated Waveguide, B.S.T.J., *36*, Sept., 1957, p. 1253.
2. Morgan, S. P., Theory of Curved Circular Waveguide Containing an Inhomogeneous Dielectric, B.S.T.J., *36*, Sept., 1957, p. 1209.
3. Unger, H. G., Normal Mode Bends for Circular Electric Waves, B.S.T.J., *36*, Sept., 1957, p. 1292.
4. Aeroflex-P, Anchor Plastics Company.
5. Conklin, G. E., Reduction of Dielectric Loss in Polyethylene, J. Applied Phys., *35*, 1964, p. 3228.
6. von Hipple, A., *Dielectrics and Waves*, John Wiley & Sons, New York, 1954.
7. King, A. P., Observed 5–6 mm Attenuation for Circular Electric Wave in Small and Medium-Sized Pipe, B.S.T.J., *35*, Sept., 1956, p. 1115.
8. Crawford, A. B. and Hogg, D. C., Measurement of Atmospheric Attenuation at Millimeter Wavelengths, B.S.T.J., *35*, July, 1956, p. 907.

Estimation of the Mean of a Stationary Random Process by Periodic Sampling

By H. T. BALCH, J. C. DALE, T. W. EDDY and R. M. LAUVER

(Manuscript received February 23, 1966)

Estimating the mean of a stationary random process from the average of equally weighted samples taken periodically in a closed interval (0, T) is considered. The variance of this estimator as a function of the number of samples in the interval is given in the form of a modified sampling theorem.

I. INTRODUCTION

This paper* considers the problem, commonly encountered in detection theory, of estimating the mean of a stationary random process from samples taken periodically in a closed interval (0, T). The samples are, in general, correlated and the estimator used is the average of equally weighted samples taken in the interval. Existing results are extended to give a clearer interpretation of the dependence of the variance on the number of samples. The dependence is obtained in terms of the power spectral density of the process in the form of a modified sampling theorem.

II. THEORY

2.1 *General*

To estimate the mean value, A, of $s(t)$ where

$$s(t) = A + n(t), \tag{1}$$

the first sample is taken at $t = 0$, and a total of $N + 1$ samples is taken in time T. $n(t)$ is a sample function from a wide-sense stationary random process with mean zero and known autocorrelation function $R(\tau)$.

* This work was supported by the U. S. Navy, Bureau of Ships under contract No. Nobsr-89401.

The estimator of A is

$$\hat{A} = [1/(N+1)] \sum_{m=0}^{N} s(mT/N). \tag{2}$$

For a fixed T, N is to be chosen to minimize the variance of $\hat{A}$.

The variance of $\hat{A}$ is given by

$$\sigma^2(\hat{A}) = [1/(N+1)] \sum_{m=-N}^{N} \left(1 - \frac{|m|}{N+1}\right) R(mT/N). \tag{3}$$

Equation (3) may be found in slightly different form in the literature.[1,2]

It is now convenient to define a weighting function, $q_{\tau_0}(\tau)$, by

$$q_{\tau_0}(\tau) = \begin{cases} 1 - \dfrac{|\tau|}{\tau_0} & |\tau| \leqq \tau_0 \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{4}$$

With this definition, (3) may be written as

$$\sigma^2(\hat{A}) = \frac{1}{(N+1)} \int_{-\infty}^{\infty} q_{[(N+1)/N]T}(\tau) R(\tau) \sum_{m=-\infty}^{\infty} \delta\left(\tau - \frac{mT}{N}\right) d\tau, \tag{5}$$

where $\delta(\tau)$ is the Dirac delta function. It is more revealing to express the variance in terms of spectral densities; thus, we make the following definitions:

$$\begin{aligned} F(\omega) &= \frac{1}{(N+1)} \int_{-\infty}^{\infty} q_{[(N+1)/N]T}(\tau) R(\tau) \exp(-j\omega\tau)\, d\tau \\ &= Q(\omega) \circledast S(\omega), \end{aligned} \tag{6}$$

where

$$\begin{aligned} Q(\omega) &= \frac{1}{(N+1)} \int_{-\infty}^{\infty} q_{[(N+1)/N]T}(\tau) \exp(-j\omega\tau)\, d\tau \\ &= \frac{T}{N} \left\{ \frac{\sin[\omega(N+1)/2N]T}{[\omega(N+1)/2N]T} \right\}^2 \end{aligned} \tag{7}$$

and $S(\omega)$ is the spectral density of $n(t)$.

Also define

$$\begin{aligned} G(\omega) &= \frac{1}{(N+1)} \int_{-\infty}^{\infty} q_{[(N+1)/N]T}(\tau) R(\tau) \\ &\qquad \exp(-j\omega\tau) \sum_{m=-\infty}^{\infty} \delta\left(\tau - \frac{mT}{N}\right) d\tau. \end{aligned} \tag{8}$$

By using Poisson's sum formula we can show that

$$G(\omega) = \frac{N}{T} \sum_{m=-\infty}^{\infty} F\left(\omega - m\frac{2\pi N}{T}\right). \tag{9}$$

Now, comparing (5), (8), and (9), we observe that

$$\sigma^2(\hat{A}) = G(0) = \frac{N}{T} \sum_{m=-\infty}^{\infty} F\left(m\frac{2\pi N}{T}\right). \tag{10}$$

Because $Q(\omega)$ is approximately zero for

$$|\omega| \geqq (2\pi/T)[N/(N+1)],$$

if $S(\omega)$ is zero for $|\omega| \geqq 2\pi B$, their convolution, $F(\omega)$, will be approximately zero for $|\omega| \leqq 2\pi[B + N/T(N+1)]$. From this result and (10) we observe that choosing

$$\frac{2\pi N}{T} \geqq 2\pi\left[B + \frac{1}{T}\frac{N}{(N+1)}\right] \tag{11}$$

makes

$$\sigma^2(\hat{A}) = G(0) \approx (N/T)F(0). \tag{12}$$

Although the restriction of (11) appears to minimize the variance of $\hat{A}$, it should be observed that $F(0)$ is also a function of N, namely

$$\frac{N}{T}F(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\left\{\frac{\sin\left[\omega\frac{(N+1)}{2N}T\right]}{\omega\frac{(N+1)}{2N}T}\right\}^2 d\omega. \tag{13}$$

If $2\pi/T$ is of the same order of magnitude as $2\pi B$, the bandwidth of $S(\omega)$, then $(N/T)F(0)$ and therefore the $\sigma^2(\hat{A})$ may be minimized by choosing the smallest value of N satisfying (11). In such cases, making N larger may actually increase the variance, as illustrated in the examples. $(N/T)F(0)$ would be independent of N if the first sample is taken at $t = T/N$ rather than $t = 0$. Solving (11) for N yields an approximate rule that

$$N \approx BT\frac{1 + \sqrt{1 + (4/BT)}}{2}, \tag{14}$$

where N is an integer, will minimize the variance of $\hat{A}$. It should be recalled that the total number of samples taken in T is $N + 1$.

The form of (9) is frequently encountered in sampling theory, where one sometimes thinks in terms of the original spectra, $F(\omega)$, shifted

by integral multiples of the sampling frequency, $2\pi N/T$. The Nyquist frequency is determined such that overlapping of the sideband spectra is small. This is too restrictive when one is interested in the variance, since then only the value of $G(\omega)$ at $\omega = 0$ is of interest. Thus, sampling can be done at a rate sufficient to prevent overlapping of sidebands at $\omega = 0$. Equation (14) may be considered as a modified sampling theorem, stating that, to minimize the variance, the sampling frequency, f_s, must satisfy

$$f_s = \frac{N}{T} = B\,\frac{1 + \sqrt{1 + (4/BT)}}{2}. \tag{15}$$

For large T, f_s is equal to one half of the Nyquist frequency required to reconstruct the time function.

2.2 *Variance for Large T*

When T is sufficiently large, the $Q(\omega)$ function approaches a delta-function, namely

$$Q(\omega) \approx 2\pi\delta(\omega)/(N + 1), \tag{16}$$

and

$$\sigma^2(\hat{A}) \approx [N/T(N + 1)] \sum_{m=-\infty}^{\infty} S(m2\pi N/T). \tag{17}$$

If, as before, $S(\omega)$ is bandlimited and

$$2\pi N/T \geqq 2\pi\left[B + \frac{1}{T}\,\frac{N}{(N + 1)}\right] \approx 2\pi B \tag{18}$$

then

$$\sigma^2(\hat{A}) \approx S(0)/T. \tag{19}$$

Notice that taking more than BT samples will not decrease the variance appreciably. Taking less than BT samples will increase the variance at a rate which depends on $S(\omega)$. The dependence of the variance on N can be easily obtained for this limiting situation from (17). In general, if $dS(\omega)/d\omega \leqq 0$ for $\omega \geqq 0$, then the variance of $\hat{A}$ will also be a monotonically decreasing function of N for $N/(N + 1) \approx 1$. On the other hand, if the spectral density of the noise is not monotonically decreasing for $\omega \geqq 0$, then the variance of $\hat{A}$ will have local minima for values of $N < BT$. These statements concerning monotonicity would be true for all T and N if $(N/T)F(0)$ were independent of N.

2.3 *Limit of Continuous Sampling*

The limit of continuous sampling has been derived elsewhere[1] and is easily obtained from (3). The result is

$$\sigma^2(\hat{A})_c = \frac{2}{T}\int_0^T \left(1 - \frac{\tau}{T}\right) R(\tau)\, d\tau = \frac{1}{T}\int_{-\infty}^{\infty} q_T(\tau)R(\tau)\, d\tau \quad (20)$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\left(\frac{\sin \omega T/2}{\omega T/2}\right)^2 d\omega. \quad (21)$$

As T becomes large,

$$\sigma^2(\hat{A})_c \approx \int_{-\infty}^{\infty} \frac{S(\omega)}{T}\,\delta(\omega)d\omega = \frac{S(0)}{T} \quad (22)$$

Thus, for large T, taking $N = BT$ samples gives the same variance as sampling continuously.

2.4 *Equivalent Independent Samples*

When T is large, one can determine the number of independent samples required to achieve the same variance as continuous sampling. The variance for N_i independent samples is

$$\sigma^2(A)_{N_i} = \frac{\frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)d\omega}{N_i}. \quad (23)$$

Equating this variance to the variance of (22) requires

$$N_i = T\,\frac{\frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)d\omega}{S(0)}. \quad (24)$$

Defining the effective bandwidth as

$$2B_e = \frac{\frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)d\omega}{S(0)}, \quad (25)$$

(24) can be written as

$$N_i = 2B_eT. \quad (26)$$

However, the variance achieved by continuous sampling may be obtained by taking $N = BT$ samples in time T. Thus,

$$N_i = (2B_e/B) \quad (27)$$

equates N for minimum variance to the number of independent samples required to achieve the same variance.

III. EXAMPLES

The variance of the sample mean as a function of number of samples $(N + 1)$ and length of record (T) has been computed for several spectral densities.

The variance of the sample mean shown on the following figures was computed using (3).

3.1 *Rectangular Spectrum*

$$S_1(\omega) = \begin{cases} \frac{1}{2}, & -2\pi < \omega < 2\pi \\ 0, & \text{elsewhere.} \end{cases} \tag{28}$$

Fig. 1 shows $\sigma^2(\hat{A})$ plotted against number of samples. Each curve of the set represents a different length of record T. The table on the figure shows the relationship of the curves to the length of record.

The most striking feature of the curves on Fig. 1 is the abrupt steps

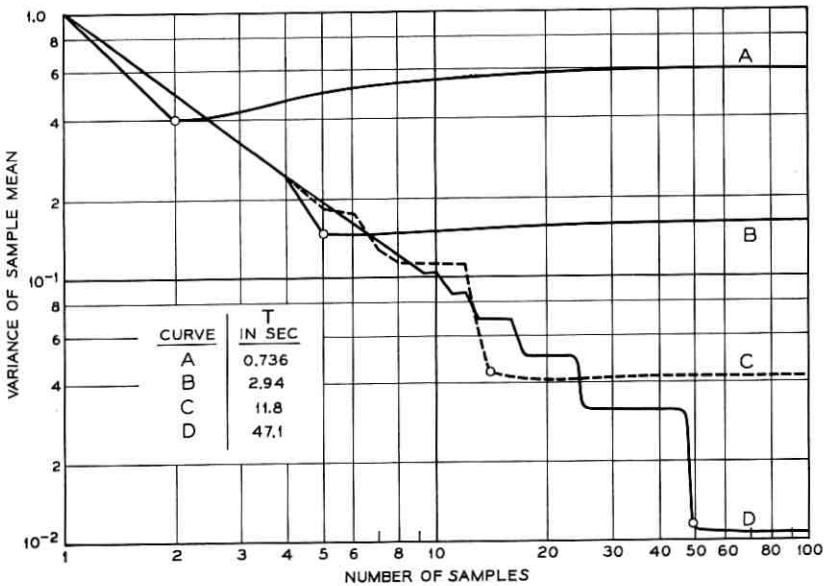

Fig. 1 — Variance of the sample mean as a function of the number of samples and length of record for a process with rectangular spectral density.

in $\sigma^2(\hat{A})$ as N is increased. This behavior is predicted by (17). Equation (14) predicts the approximate value of N for minimum variance. These values of N are shown for each of the curves by a small circle.

An interesting point to note here is that in some regions a better estimate of the mean is obtained when the same number of samples is taken for a smaller T. Also for small T, $\sigma^2(\hat{A})$ reaches a minimum and then increases as more samples are taken. This implies that for small values of T a smaller variance is obtained by taking a smaller number of samples (but including the end points) than would be obtained by continuous sampling.

3.2 *Sawtooth Spectrum*

$$S_2(\omega) = \begin{cases} \left| \dfrac{\omega}{2\pi} \right|, & -2\pi \leqq \omega \leqq 2\pi \\ 0\ , & \text{elsewhere.} \end{cases} \tag{29}$$

This is an interesting case for two reasons. First, its spectrum is not monotonically decreasing. This gives rise to local minima and maxima in $\sigma^2(\hat{A})$ as a function of N caused by the spectrum shape.

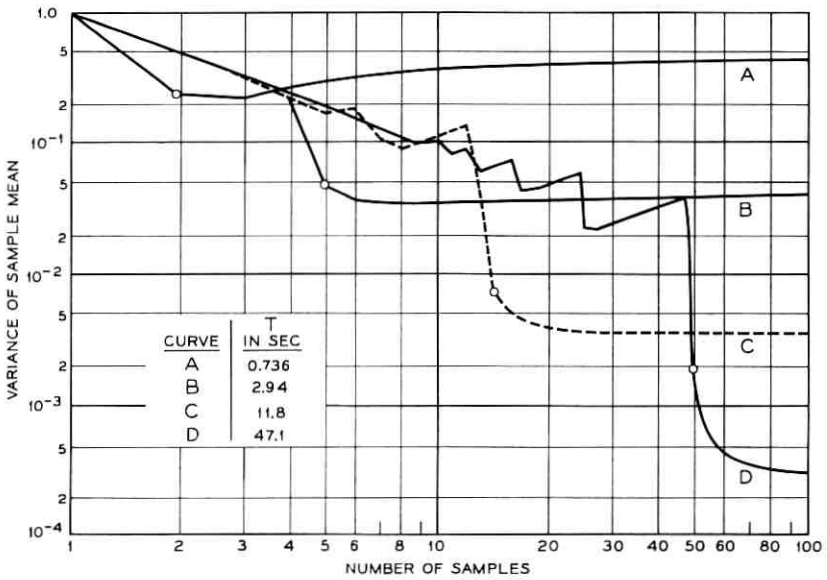

Fig. 2 — Variance of sample mean as a function of number of samples and length of record for a process with a sawtooth spectral density.

Second, its spectrum at $\omega = 0$ is 0, thus enhancing the error due to approximating $Q(\omega)$. The results are shown in Fig. 2.

3.3 *Markoff Spectrum*

$$S_3(\omega) = \frac{8}{\omega^2 + 16}. \tag{30}$$

This is an example of a nonbandlimited spectrum. The values of $\sigma^2(\hat{A})$ are shown in Fig. 3. A point worth noting here is that if the bandwidth of the process was defined as the width at the one-half power points and the time function sampled according to (14), the value of $\sigma^2(\hat{A})$ obtained would be larger by about a factor of 2 than the minimum value obtained by letting N approach infinity.

This example is also the same one treated by Fine and Johnson[3] for small values of T. Curve A on Fig. 3 agrees with their results.

IV. SUMMARY

Theory has been presented which predicts the behavior of the variance of the sample mean of periodic samples taken from a stationary random

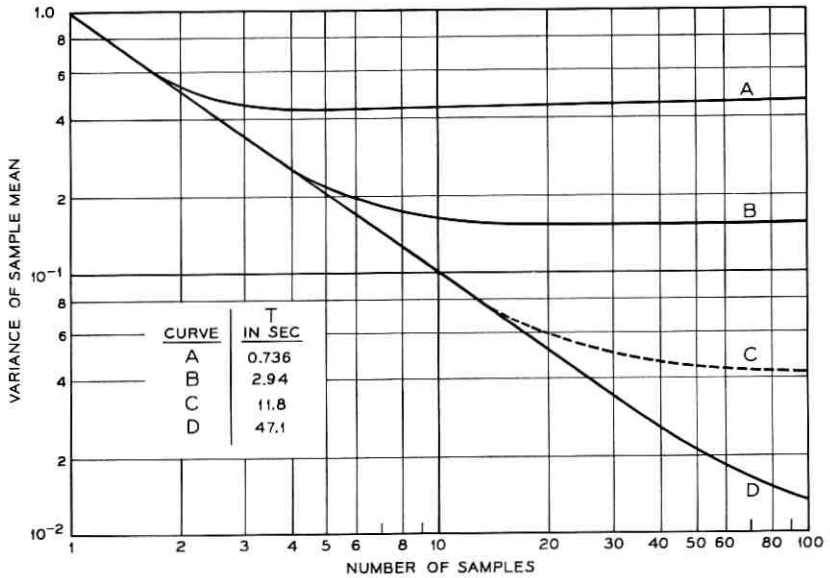

Fig. 3 — Variance of the sample mean as a function of the number of samples and length of record for a process with Markoff spectral density.

process. The variance is given in terms of the power spectrum of the sampled process. Three interesting results have been shown:

(*i*) When $BT \gg 1$, the variance of the sample mean is essentially minimized when BT samples are taken.

(*ii*) The variance of the sample mean is not necessarily monotonically decreasing as a function of the number of samples taken in a fixed record length.

(*iii*) For short record lengths, it is possible to obtain a smaller variance with a small, finite number of samples than with continuous sampling.

REFERENCES

1. Davenport, W. B. Jr. and Root, W. L., *Random Signals and Noise,* McGraw-Hill Book Company, New York, 1958.
2. Costas, J. P., Periodic Sampling of Stationary Time Series, MIT Technical Report No. 156, 16 May, 1950, p. 4, (11).
3. Fine, T. and Johnson, N., On the Estimation of the Mean of a Random Process, Proc. IEEE, Feb., 1965, pp. 187–188.

Physical Limitations on Ray Oscillation Suppressors

By D. MARCUSE

(Manuscript received January 12, 1966)

The question of whether it is possible to suppress ray oscillations in light waveguides is important for the design of light communications systems. With the help of Liouville's theorem of statistical mechanics it is shown that it is impossible to reduce simultaneously the amplitudes and the angles of ray oscillations if the ray originates in and returns to a region of low index of refraction. A reduction of both ray amplitudes and angles can be achieved only if the ray moves from a region of low to one of high index of refraction.

Liouville's theorem is used to derive a condition relating the output position and slope of a ray which traverses an optical transformer to its input position and slope. With $\mathbf{p}_i$, $\mathbf{x}_i$ *denoting the canonically conjugate variables of the output ray and* p_i, x_i *those of the input ray, the condition derived from Liouville's theorem states that the Jacobian of the transformation is one.*

$$\frac{\partial(\mathbf{p}_i, \mathbf{x}_i)}{\partial(p_i, x_i)} = 1.$$

I. INTRODUCTION

Light transmission systems can be built in various ways. A continuous dielectric medium of rotational symmetry with an index of refraction which depends on the distance r from the optical axis

$$n = n(r)$$

is capable of guiding light rays if $n(r)$ decreases monotonically with increasing r. Another example is the beam-waveguide consisting of a series of lenses which refocuses the light beam periodically counteracting diffraction.

Both of these examples have one point in common — a ray which is launched off-axis into the waveguide follows an oscillatory trajectory. However, even if a light ray travels on-axis it will be forced into an oscillatory trajectory by any imperfection of the guidance medium.[1] To

keep the ray amplitudes small requires a very high precision of alignment which might be hard to obtain for long waveguides.

It seemed natural, therefore, to consider means of suppressing these ray oscillations, and if all such efforts fail, to ask for a general physical principle which says that such ray oscillation suppressors are impossible.

The search for such a general principle is even more important as it is easy to construct models of beam waveguides which violate physical principles in subtle ways thus seeming to lead to ray oscillation suppressors. One such system is shown in Fig. 1. Assume that we deform thin lenses as indicated in the figure and assume further that these lenses behave just like plane thin lenses in that they break each ray by an amount β_n which depends only on the radius r_n of the ray but not on the input angle.

$$\tan \beta_n = -r_n/f.$$

Making the paraxial approximation, which means replacing $\tan \beta_n$ by β_n and $\tan \alpha_n$ by α_n, we obtain the ray equation

$$r_{n+1} = r_n + \alpha_n(z_{n+1} - z_n) \tag{1a}$$

$$\alpha_{n+1} = \alpha_n - \frac{r_{n+1}}{f}. \tag{1b}$$

If the lenses are warped to form parabolas,

$$z_{n+1} - z_n = d + b(r_n^2 - r_{n+1}^2). \tag{2}$$

Equations (1a) and (1b) together with (2) describe rays which, if they travel from the left to the right in Fig. 1, exhibit decreasing amplitudes. In fact, if one allows each ray to travel a sufficient distance they approach the axis arbitrarily closely.

It appears that we have invented a ray oscillation suppressor.

The object of this paper is to prove that such a device is impossible. So the question arises: What went wrong with the argument presented above? A closer examination shows that the assumption that β_n is

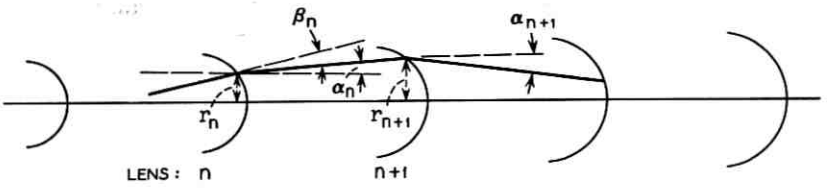

Fig. 1 — Beam-waveguide composed of warped, thin lenses.

independent of α_n violates Liouville's theorem. We will return to this question later.

The general proof of the impossibility of constructing a ray oscillation suppressor was suggested by J. R. Pierce.

II. PROOF OF THE IMPOSSIBILITY OF A RAY OSCILLATION SUPPRESSOR

The proof is based on Liouville's theorem.[2] It refers to the representation of physical systems in phase space. Phase space is the space of the canonically conjugate variables q_i and p_i describing the system. Each system is represented by one point in phase space. Many identical systems which happen to be in different states described by different values of their coordinator q_i and p_i can be described by the density of their representation points in phase space. Liouville's theorem states that the density of any given configuration of points in phase space is constant if the systems under consideration obey the canonical differential equations

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \tag{3}$$

H is the Hamiltonian function describing the system. Another version of Liouville's theorem states that the volume containing a constant number of points in phase space remains constant in time.

For Liouville's theorem to be applicable to light rays we have only to show that light rays can be described by equations of the form (3). The derivation of the Hamiltonian equations of geometric optics can be found in Ref. 5. The derivations are sketched here for the sake of convenience.

To show this we start with Fermat's principle which states that a light ray connecting two arbitrary points P_1 and P_2 in a medium of index of refraction

$$n = n(x,y,z) \tag{5}$$

follows a path such that

$$J = \frac{1}{c}\int_{P_1}^{P_2} n\, d\, s = \text{extremum}. \tag{6}$$

Here, c is the velocity of light in vacuum and s is the path length measured along the ray trajectory. Introducing coordinates x,y,z, we can rewrite (6) as

$$J = \frac{1}{c}\int_{P_1}^{P_2} n\sqrt{1 + x'^2 + y'^2}\, dz = \text{extremum} \tag{7}$$

with

$$x' = \frac{dx}{dz} \quad \text{and} \quad y' = \frac{dy}{dz}. \tag{8}$$

Equation (7) is analogous to Hamilton's principle of least action with the Lagrangian

$$L = n\sqrt{1 + x'^2 + y'^2} \tag{9}$$

and the time t being replaced by the z-coordinate.

Once the Lagrangian of a system is known the moments p_x and p_y canonically conjugate to the x and y coordinates are defined by

$$p_x = \frac{\partial L}{\partial x'} = n\frac{x'}{\sqrt{1 + x'^2 + y'^2}} \tag{10a}$$

$$p_y = \frac{\partial L}{\partial y'} = n\frac{y'}{\sqrt{1 + x'^2 + y'^2}} \tag{10b}$$

and the Hamiltonian function by

$$H = p_x x' + p_y y' - L = -\sqrt{n^2 - p_x^2 - p_y^2}. \tag{11}$$

The variational problem (7) is solved by the equations[3]

$$x' = \frac{\partial H}{\partial p_x} \qquad y' = \frac{\partial H}{\partial p_y} \tag{12a}$$

$$p_x' = -\frac{\partial H}{\partial x} \qquad p_y' = -\frac{\partial H}{\partial y}. \tag{12b}$$

Equations (12a) and (12b) are analogous to (3) which shows that the ray description can be given in terms of canonical differential equations. The equations of (12a) are satisfied identically while the equations of (12b) lead to the well-known ray equations

$$\frac{1}{\sqrt{1 + x'^2 + y'^2}}\frac{d}{dz}\left(n\frac{x'}{\sqrt{1 + x'^2 + y'^2}}\right) = \frac{\partial n}{\partial x} \tag{13a}$$

$$\frac{1}{\sqrt{1 + x'^2 + y'^2}}\frac{d}{dz}\left(n\frac{y'}{\sqrt{1 + x'^2 + y'^2}}\right) = \frac{\partial n}{\partial y}. \tag{13b}$$

Introducing

$$\frac{ds}{dz} = \sqrt{1 + x'^2 + y'^2}, \tag{14}$$

(13a) and (13b) can be written in the more familiar form[4]

$$\frac{d}{ds}\left(n\frac{dx}{ds}\right) = \frac{\partial n}{\partial x} \tag{15a}$$

$$\frac{d}{ds}\left(n\frac{dy}{ds}\right) = \frac{\partial n}{\partial y}. \tag{15b}$$

The preceding discussion of ray dynamics was sketched only to prove that Liouville's theorem applies to light rays.

Now, we are finally in a position to prove the impossibility of a ray oscillation suppressor. To simplify the discussion let us limit the problem to two dimensions, x and z. Assume that z is the axis of the system. The phase space is now two dimensional and is spanned by the coordinates x and p_x. Let us further assume that we consider an ensemble of rays whose initial conditions are such that the representation points of all these rays fill a square area centered around the origion of phase space as shown in Fig. 2. Each ray represented in this area has a certain distance x from the optical axis z and a certain slope given by (10)

$$x' = \frac{p_x}{\sqrt{n^2 - p_x^2}}. \tag{16}$$

If an oscillation suppressor were possible we would require that all the rays initially contained in the square of phase space of Fig. 2 would approach the z-axis more closely. In addition, we would require that the angles between the rays and the z-axis don't increase or perhaps even decrease. If we look at the rays initially and finally in a region of constant index of refraction n for example in vacuum, $n = 1$, we would find that the square of Fig. 2 has deformed either into the rectangle, if the angles don't shrink, or into the smaller square, if the angles as well as the amplitudes shrink, as indicated in Fig. 2 by dotted lines. In either

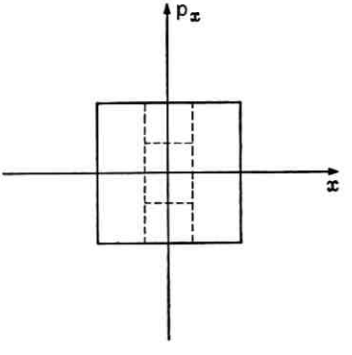

Fig. 2 — Volume in phase space occupied by light ray representation points.

case, we find that the area (volume in two dimensions) of phase space occupied by the points representing the initial ray positions has decreased. However, Liouville's theorem says that this is impossible so that we see that a ray oscillation suppressor is impossible. We can trade off amplitude at the expense of spread in p_x direction. In this case, either the tangent of the ray angles x' or the index of refraction has to increase. It is even possible to decrease both the ray amplitudes and angles by increasing n along the z-axis. However, the area in phase space has to stay constant. The initial square has to deform into a rectangle of equal area which stretches along the p_x axis. After some distance we have reached a region of high index of refraction and find both the amplitude and the ray angles decreased (but not the p_x values which have increased). For most applications to be able to make use of the effect, we have to leave the high index medium. But as soon as n drops to a low value the angles have to increase to keep the spread in p_x constant and again we have traded a decrease in the ray amplitudes for an increase in the ray angles. The ray position in most optical systems will eventually spread far apart if we allow the rays to travel far enough. This means that the volume in phase space, though its volume content remains constant, assumes a "filamentous" appearance and extends to many different parts of phase space.[2]

III. A BASIC RELATION FOR OPTICAL TRANSFORMERS

Liouville's theorem allows one to formulate a theorem which all rays passing through an optical device (optical transformer) have to obey.

Let us assume we have an arbitrary optical transformer with input rays whose positions and slopes are described by the canonically conjugate variables q_i, p_i and corresponding output ray with variables $\mathbf{q}_i$, $\mathbf{p}_i$. The output variables are related to the input variables by

$$\mathbf{q}_i = \mathbf{q}_i(q_i, p_i)$$

$$\mathbf{p}_i = \mathbf{p}_i(q_i, p_i).$$

The input rays may occupy a volume $dV = dq_1 dq_2 dp_1 dp_2$ in phase space. This volume deforms, as the rays propagate, to $d\bar{V}$. Liouville's theorem states that these volumes are identical:

$$dV = d\bar{V}. \tag{17}$$

The volume $d\bar{V}$ on the right hand side of (17) can be rewritten as

$$d\bar{V} = \frac{\partial(\mathbf{q}_i, \mathbf{p}_i)}{\partial(q_i, p_i)}, dq_1\, dq_2\, dp_1\, dp_2 \tag{18a}$$

or

$$d\bar{V} = \frac{\partial(\mathbf{q}_i, \mathbf{p}_i)}{\partial(q_i, p_i)}\, dV. \tag{18b}$$

We conclude from (17) and (18) that the Jacobian must be equal to unity

$$\frac{\partial(\mathbf{q}_i, \mathbf{p}_i)}{\partial(q_i, p_i)} = 1. \tag{19}$$

Equation (19) is stated in Ref. 6 without proof.

The derivation of (19) is based on the fact that the ray trajectory can be described by the differential equations of (13). However, there may be discontinuities in the index of refraction, n, where the ray equations can not be applied. But it is well known that rays traverse discontinuities of the index of refraction. The ray trajectory is unaltered if the discontinuity is replaced by a rapidly changing but continous transition of n. In this way we assure that the ray equations hold everywhere and that (19) is applicable even in that case.

Limiting the problem to two dimensions we can write (19) as

$$\frac{\partial \mathbf{p}}{\partial p}\frac{\partial \mathbf{x}}{\partial x} - \frac{\partial \mathbf{p}}{\partial x}\frac{\partial \mathbf{x}}{\partial p} = 1. \tag{20}$$

Equation (20) allows us to derive an interesting relation between the input and output angles of rays passing through an infinitesimally thin optical transformer (Fig. 3). If the thickness of the optical transformer shrinks to zero we have $\mathbf{x} = x$ and consequently,

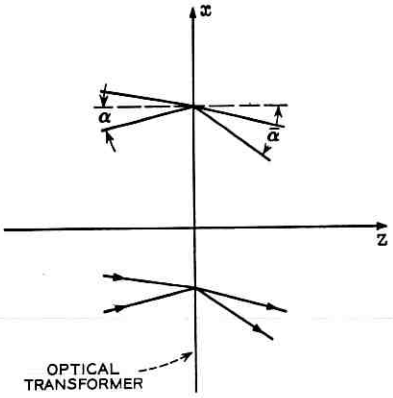

Fig. 3 — Illustrations of a thin optical transformer.

$$\frac{\partial \mathbf{x}}{\partial p} = 0 \qquad \frac{\partial \mathbf{x}}{\partial x} = 1$$

so that (20) reduces to

$$\frac{\partial \mathbf{p}}{\partial p} = 1$$

whose solution is

$$\mathbf{p} = p + f(x). \tag{21}$$

With the help of (10) we see that if

$$\frac{dx}{dz} = \tan \alpha \qquad \frac{d\mathbf{x}}{dz} = \tan \bar{\alpha},$$

it follows that

$$p = n \sin \alpha \qquad \mathbf{p} = \bar{n} \sin \bar{\alpha}$$

so that (21) can be written as

$$\bar{n} \sin \bar{\alpha} = n \sin \alpha + (\bar{n} \sin \bar{\alpha})_{\alpha=0}. \tag{22}$$

This is a fundamental relation which all rays passing through thin lenses or any other thin optical device have to obey.

If both $\alpha \ll 1$ and $\bar{\alpha} \ll 1$ and $\bar{n} = n = 1$, (22) simplifies

$$\beta = \bar{\alpha} - \alpha = (\bar{\alpha})_{\alpha=0}. \tag{23}$$

This is the relation which is used to describe the change of ray angles passing through a thin lens. Equation (22) shows that this thin lens relation holds approximately for rays which impinge nearly perpendicular to the lens. If the rays make large angles with respect to the direction normal to the lens surface (23) has to be replaced by (22). This explains the error which was made in deriving (1). If this equation is corrected by using (22) rather than (16) the ray oscillation suppressing quality of the warped thin lenses disappears.

The general expressions (19), (20), or (22) can be used to check the physical realizability of optical models.

IV. ACKNOWLEDGMENT

I am grateful to J. R. Pierce who brought Liouville's theorem to my attention, thus putting an end to attempts to invent ray oscillation suppressors.

REFERENCES

1. Marcuse, D., Statistical Treatment of Light Ray Propagation in Beam-Waveguides, B.S.T.J., *44*, November, 1965, pp. 2065–2081.
2. Tolman, R. C., *The Principles of Statistical Mechanics,* Oxford University Press, p. 48–52.
3. Ibid., p. 19–27.
4. Born, M. and Wolff, E., *Principles of Optics,* Pergamon Press, New York, 1959, p. 121.
5. R. K. Luneburg, *Mathematical Theory of Optics,* University of California Press, 1964.
6. Ibid., p. 221.

Contributors to This Issue

HUGH T. BALCH, B.S.E. (EE), 1945 and M.S.E. (EE), 1947, University of Michigan; Bell Telephone Laboratories, 1947—. Mr. Balch was first involved in the development of mobile radio systems for Bell System use. He then participated in the development of the AN/TRC-24 radio system for the U.S. Army Signal Corps. Since 1953 he has been concerned with the development of sonar systems for the U.S. Navy. He currently heads a group studying the application of new data processing techniques to future passive sonars. Licensed Professional Engineer — New Jersey.

GLENN E. CONKLIN, B.S., 1951 and M.S., 1953, University of Wichita; Ph.D., 1962, University of Kansas; Bell Telephone Laboratories, 1960–1966. Initially Mr. Conklin was engaged in investigating the millimeter microwave dielectric properties of plastics. Recently his work has been directed toward precision optical measurements. Member, American Physical Society.

JOHN C. DALE, B.E.E., 1957, University of Florida; M.E.E., 1959, New York University; Bell Telephone Laboratories, 1957—. Mr. Dale has been involved in the design of signal processors for use in anti-submarine warfare. He is presently engaged in applying digital data processing techniques to detect and recognize patterns. Member, Sigma Tau, Phi Kappa Phi, IEEE.

T. W. EDDY, B.S., 1958, University of Idaho; M.S., 1960, New York University; Bell Telephone Laboratories, 1958—. Until recently, Mr. Eddy has been engaged in the development of special purpose transmission systems for military applications. Recently, he has been concerned with problems of detecting signals in the presence of noise. Member, IEEE, Sigma Tau.

ROBERT M. LAUVER, B.S., 1956, University of Connecticut; M.E.E., 1961, New York University. At present, Mr. Lauver is working toward the Ph.D. degree in E.E. at Polytechnic Institute of Brooklyn. Bell Telephone Laboratories, 1956—. Mr. Lauver was first engaged in

development work on airborne missile guidance systems. He has since been concerned with development of sonar systems and is currently engaged in the study of signal processing techniques applied to sonar systems. Member, IEEE, AAAS, Tau Beta Pi, Eta Kapp Nu.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, and Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He is presently working on the transmission aspect of a light communications system. Member, IEEE.

SAMUEL P. MORGAN, B.S., 1943, M.S., 1944, and Ph.D., 1947, California Institute of Technology; Bell Telephone Laboratories, 1947—. A research mathematician, Mr. Morgan has been particularly concerned with the applications of electromagnetic theory to microwave and other problems. As Head, Mathematical Physics Department, he now supervises a research group in various fields of mathematical physics. Fellow, IEEE; member, American Physical Society, American Mathematical Society, SIAM, Sigma Xi, Tau Beta Pi, AAAS.

J. A. MORRISON, B.Sc. 1952, King's College, London University; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Telephone Laboratories, 1956—. A research mathematician, Mr. Morrison has worked on a variety of problems in mathematical physics. Currently, he is concentrating his interests on averaging methods, as applied to perturbed nonlinear oscillations and to problems in celestial mechanics. Member, American Mathematical Society, Society for Industrial and Applied Mathematics, Sigma Xi.

J. B. O'NEAL, JR., B.E.E., 1957, Georgia Institute of Technology; M.E.E., 1960, University of South Carolina; Ph.D., 1963, University of Florida; Bell Telephone Laboratories, 1964—. While employed at the Martin Company in 1959–60, he worked on random access communications systems. He is currently engaged in studies to determine efficient coding techniques for converting analog television signals into digital form. Member, IEEE, Eta Kappa Nu; associate member, Sigma Xi.

B.S.T.J. BRIEFS

A New Signal Format for Efficient Data Transmisson

By F. K. BECKER, E. R. KRETZMER, and J. R. SHEEHAN

(Manuscript received February 18, 1966)

I. BACKGROUND

Data communication systems in current use generally require substantially more bandwidth than the Nyquist minimum of one-half cycle per symbol. This comes about for two main reasons: first, the baseband signal spectrum has a gradual roll-off beyond the theoretical minimum;[1] second, the modulation process needed to translate the baseband spectrum to the bandpass channel generates additional side frequencies which must be preserved to permit recovery of the signal. For example, a recently described vestigial-sideband system[2] uses an extra 50 per cent of bandwidth for each of these two reasons. Consequently, such a system handles one symbol per cycle, and each symbol can convey as many levels as the signal-to-noise ratio permits — independent of all adjacent symbols. Thus, with n levels, each symbol yields $\log_2 n$ binary digits.

II. NEW TECHNIQUE

A new approach recently implemented avoids the need for excess bandwidth by using baseband shaping such that the received signal spectrum is a half-period sinusoid.[3,4] This shaping not only permits two symbols per cycle of bandwidth, but it also forces the baseband signal to be free of any dc component. This, in turn, permits single-sideband techniques for translation to any desired frequency band without increase in bandwidth.

III. UNDERLYING PRINCIPLE

The new technique is a departure from the conventional methods which are based on zero intersymbol interference.[1] Instead, it permits intersymbol interference — but in precisely prescribed amounts. This is best illustrated by examining the impulse response for each case — or, more accurately, the end-to-end response to a single symbol (e.g., a

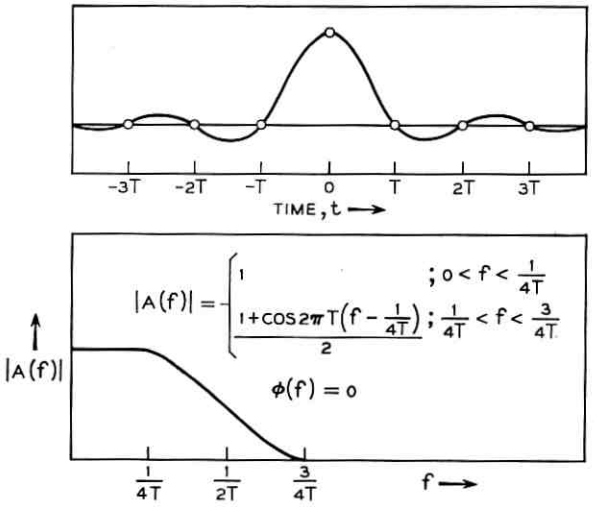

Fig. 1 — Conventional system — pulse response and frequency domain function.

"one" in a background of all zeros). Fig. 1 shows the response of a conventional system, alongside of the corresponding frequency-domain function. Fig. 2 shows the corresponding functions for the new system. In both cases, the spacing between successive symbols is T, but only in the second case is the bandwidth confined to $1/2T$.

The fact that the symbol response, as shown in Fig. 2, extends over several symbol intervals requires compensating decoding at the receiver or, advantageously, precoding at the transmitter[3] similar to "duobinary"

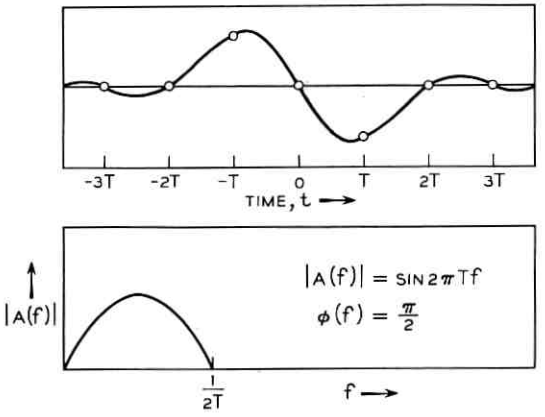

Fig. 2 — Partial response system — pulse response and frequency domain function.

or "biternary" coding.[4,5] Performance with either method is comparable to that achieved with other three-level systems.[6] The precoding used in the present implementation converts the original binary data sequence $a_1, a_2 \cdots a_n$ into a new binary sequence $b_1, b_2 \cdots b_n$, which the channel converts into the received three-level sequence $c_1, c_2 \cdots c_n$. The following relations hold

$$c_n = b_n - b_{n-2} \tag{1}$$

(by definition of the system response)

$$a_n \equiv [b_n + b_{n-2}] \bmod 2 \tag{2}$$

(by design of the precoder).

It follows that $a_n = [c_n] \bmod 2$, which means odd and even-numbered levels of c_n signify $a_n = 1$ and zero, respectively, the same as with biter-

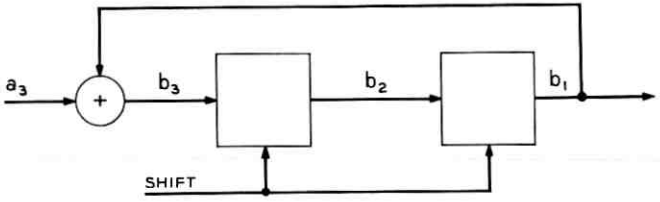

Fig. 3 — Transmitter precoding system.

nary/duobinary encoding. The precoding relation (2) is implemented with a mod-2 adder and a shift register (see Fig. 3).

IV. MODULATION PROCESS

Of all the known methods for translating a signal into a desired frequency band only single-sideband transmission preserves the signal bandwidth. This is illustrated in Fig. 4; the numbers correspond to an experimental *Data-Phone** set presently being tested on the switched telephone network at 2400 bits/sec.

In this instance the transmitted spectrum covers exactly one octave. It is generated by simply sampling the data and selecting the desired spectral band with a filter approximating the square root of the half-sinusoidal characteristic $H(f)$. A matching filter at the receiver then completes the shaping shown in Fig. 4. The carrier pilot is transmitted outside of these filters; it is recovered through a narrow-band filter (just wide enough to preserve any multiplicative noise imparted by the

* *Data-Phone* is a service mark of the Bell System.

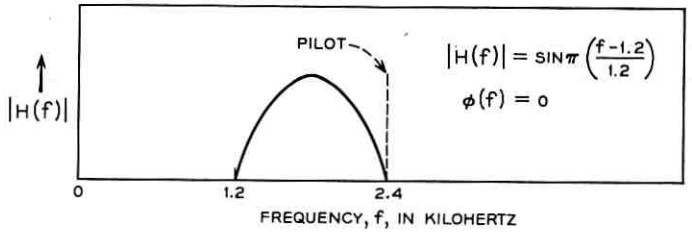

Fig. 4 — Partial response spectrum for single-sideband transmission.

channel). The carrier phase is rotated by 90 degrees before it is used for demodulation, corresponding to odd symmetry of $h(t)$ (see Fig. 2).

V. EQUALIZATION

Extensive computer simulation has shown that the system can tolerate substantial amplitude and phase distortion of various shapes. Since this tolerance is obtained at the expense of noise margin, it proved desirable to incorporate a limited amount of automatic equalization[7] for operation on the switched telephone network. This has been accomplished by formulating a new algorithm for automatic equilization of partial-response signaling formats.[8]

VI. SUMMARY

A data terminal incorporating the above principles has successfully performed in preliminary tests over a variety of cross-country dialed telephone connections. The data rate was 2400 bits/sec. In addition, a 150 bit/sec channel was operated in the reverse direction in the band below 1000 c/s. Details concerning design and performance will be published after evaluation is complete.

REFERENCES

1. Nyquist, H., Certain Topics in Telegraph Transmission Theory, AIEE Trans., *47*, April, 1928, pp. 617–644.
2. Becker, F. K., An Exploratory Multilevel Vestigial-Sideband Data Terminal for Use on High-Grade Voice Facilities, Conf. Rec. 1965 IEEE Annual Communications Convention, pp. 481–484.
3. Kretzmer, E. R., Binary Data Communication by Partial Response Transmission, Conf. Rec. 1965 IEEE Annual Communications Convention, pp. 451–455. See also, Generalization of a Technique for Binary Data Communication, IEEE Trans. Comm. Tech., Feb., 1966, pp. 67–68.
4. Lender, A., Correlative Level Coding for Binary Data Transmission, IEEE Spectrum, *3*, Feb., 1966, pp. 104–115.
5. Ringelhaan, O. E., System for Transmission of Binary Information at Twice the Normal Rate, U.S. Pat. 3,162,724; granted December, 1964.
6. Howson, R. D., An Analysis of the Capabilities of Polybinary Data Transmission, IEEE Trans. Comm. Tech., Sept., 1965, pp. 312–319.
7. Becker, F. K., Holzman, L. N., Lucky, R. W., and Port, E., Automatic Equalization for Digital Communication, Proc. IEEE, *53*, Jan., 1965, pp. 96–97.
8. Gerrish, A. M., Port, E., and Sheehan, J. R., An Automatic Equalization Technique for Partial Response Signalling Formats, to be published.

Performance of a Forward-Acting Error-Control System on the Switched Telephone Network

By E. J. WELDON, Jr.

(Manuscript received March 7, 1966)

This brief contains a summary of data taken in the course of a recent error-control experiment. In this experiment data were transmitted over switched voiceband telephone lines at 2000 bits per second using *Data-Phone** data set 201A. With the transmitting terminal located at the Holmdel, New Jersey laboratory, connections via the switched network were established to various cities (Baltimore, Cleveland, Dallas, Denver, Louisville, and St. Louis). There a return connection, again via the switched network, was established to the receiving terminal which was also located at Holmdel. Nearly all calls were made during the business day.

Errors were corrected by means of a forward-acting cyclic code which was formed by interleaving the (15,9) code generated by $x^6 + x^5 + x^4 + 1$ to degree i. As a result, $(9/15) \cdot 2000 = 1200$ information bits per second were transmitted. In the first half of the experiment, i was set to 73; in the second half, 200. Since the (15,9) code corrects all bursts of length three or less, the interleaved code can correct all bursts of length $3i$ or less. Thus, the codes are optimal burst-correctors in the sense that the equality holds in the Reiger bound,[1] i.e.,

$$b = \frac{n - k}{2}$$

where b is the guaranteed burst-correcting ability of the code, n is the code length, and k is the number of information symbols in the code.

In each case, the i subwords were decoded independently using the Peterson algorithm.[2] Decoding in this manner, rather than using the Peterson algorithm to decode the cyclic code of length $15i$ directly, enables the code to correct many error patterns which would otherwise be uncorrectable. This improves performance considerably since it enables the code to correct most error patterns of low weight (2,3,4, $\cdots$, 10, say) even though its minimum distance is only 3. It also simplifies the decoding circuitry somewhat. In this case the decoder employed approximately 210 transistors and a $14(i - 1)$-bit, limited-access, line-speed storage device.

The data are summarized in Table I. Because there does not seem to exist a single adequate measure of performance for such systems, the

* *Data-Phone* is a service mark of the Bell System.

TABLE I — SUMMARY OF DATA

Interleaving degree, i	73	200
Code length, $n = 15i$	1095	3000
Number of information symbols, $k = 9i$	657	1800
Burst-correcting ability, $b = (n - k)/2$	219	600
Number of calls	126	279
Number of hours	259	284
Number of bits transmitted	$1.9 \cdot 10^9$	$2.0 \cdot 10^9$
Line bit error rate	$5.7 \cdot 10^{-6}$	$1.2 \cdot 10^{-5}$
Number of line errors*	3613	5171
Number of delivered errors*	52	24
Improvement factor	70	215
Mean time between delivered errors* (hours)	5.0	11.9
Delivered error rate (errors* per bit)	$2.7 \cdot 10^{-8}$	$1.2 \cdot 10^{-8}$
Number of line word errors	3972	8704
Number of delivered word errors	83	59
Word improvement factor	48	150
Delivered word error rate (word errors per word)	$4.8 \cdot 10^{-5}$	$8.7 \cdot 10^{-5}$
Number of line bit errors	10607	24472
Number of delivered bit errors	2109	1703
Bit improvement factor	5	14
Delivered bit error rate (bit errors per bit)	$1.1 \cdot 10^{-6}$	$8.5 \cdot 10^{-7}$

results are presented in terms of three different types of error rate. These are based on bit errors, n-bit word errors, and errors.* The utility of the first two performance measures is apparent; however, in many situations the third is the most appropriate. For example, the mean time between errors* is the average duration of error-free intervals of useful length. This is a meaningful figure-of-merit for data users who require perfect transmission almost all of the time and for whom the cost of an error is relatively insensitive to the duration or the bit-error-density of the error. It is of interest to note that, regardless of how measured, the improvement in performance attributable to the error control system appears to increase approximately linearly with the degree of interleaving. Also the average line bit error rate encountered in this test was quite close to values reported in similar experiments on telephone facilities.

The author wishes to thank the following individuals for their co-operation throughout the course of the experiment: G. S. Robinson, who built the original error-control system (with $i = 73$); P. Mecklenburg, who changed the interleaving degree to 200 and suggested several

* An error is defined as a sequence of words all of which contain at least one bit error.

improvements in the system; and A. R. Lingenfelter, who was responsible for recording and reducing the data.

REFERENCES

1. Reiger, S. H., Codes for the Correction of "Clustered" Errors, IRE Trans., *IT6*, 1950, pp. 16–21.
2. Peterson, W. W., *Error-Correcting Codes*, MIT Press, 1961, p. 189.

A Note on a Type of Optimization Problem that Arises in Communication Theory

By I. W. SANDBERG

(Manuscript received March 16, 1966)

A problem that has arisen[1] in connection with the use of transversal filters to reduce the effect of intersymbol interference in digital communication systems is to determine a real N-vector $c \triangleq (c_1, c_2, \cdots, c_N)$ such that, with $n_0 \varepsilon \mathfrak{F} \triangleq \{1,2, \cdots, N\}$,

$$\sum_{\substack{n=-\infty \\ n \neq n_0}}^{\infty} \left| \sum_{j \varepsilon \mathfrak{F}} c_j x_{n-j} \right| \tag{1}$$

is minimized subject to the constraint

$$1 = \sum_{j \varepsilon \mathfrak{F}} c_j x_{n_0 - j} . \tag{2}$$

Here $\{x_n\}_{-\infty}^{\infty}$ denotes a set of real constants such that $| x_0 | > \sum_{n \neq 0} | x_n |$. Lucky[1] has proved the interesting theorem that the optimal choice of c coincides with the unique solution[2] of the equations

$$\begin{aligned} 1 &= \sum_{j \varepsilon \mathfrak{F}} c_j x_{n_0 - j} \\ 0 &= \sum_{j \varepsilon \mathfrak{F}} c_j x_{n-j} , \qquad n \varepsilon \mathfrak{F} - \{n_0\}. \end{aligned} \tag{3}$$

The proof of Ref. 1 consists of establishing a contradiction to the assertion that (1), with c_{n_0} eliminated with the aid of (2), is minimized for some c for which (3) is not satisfied. The reader is referred to Ref. 1 for the details.

The purpose of this note is to show that Lucky's result, and far more general results of similar type, can be directly deduced from the following proposition.

improvements in the system; and A. R. Lingenfelter, who was responsible for recording and reducing the data.

REFERENCES

1. Reiger, S. H., Codes for the Correction of "Clustered" Errors, IRE Trans., *IT6*, 1950, pp. 16–21.
2. Peterson, W. W., *Error-Correcting Codes*, MIT Press, 1961, p. 189.

A Note on a Type of Optimization Problem that Arises in Communication Theory

By I. W. SANDBERG

(Manuscript received March 16, 1966)

A problem that has arisen[1] in connection with the use of transversal filters to reduce the effect of intersymbol interference in digital communication systems is to determine a real N-vector $c \triangleq (c_1, c_2, \cdots, c_N)$ such that, with $n_0 \in \mathcal{F} \triangleq \{1,2, \cdots, N\}$,

$$\sum_{\substack{n=-\infty \\ n \neq n_0}}^{\infty} \left| \sum_{j \in \mathcal{F}} c_j x_{n-j} \right| \tag{1}$$

is minimized subject to the constraint

$$1 = \sum_{j \in \mathcal{F}} c_j x_{n_0-j} . \tag{2}$$

Here $\{x_n\}_{-\infty}^{\infty}$ denotes a set of real constants such that $|x_0| > \sum_{n \neq 0} |x_n|$. Lucky[1] has proved the interesting theorem that the optimal choice of c coincides with the unique solution[2] of the equations

$$\begin{aligned} 1 &= \sum_{j \in \mathcal{F}} c_j x_{n_0-j} \\ 0 &= \sum_{j \in \mathcal{F}} c_j x_{n-j} , \qquad n \in \mathcal{F} - \{n_0\}. \end{aligned} \tag{3}$$

The proof of Ref. 1 consists of establishing a contradiction to the assertion that (1), with c_{n_0} eliminated with the aid of (2), is minimized for some c for which (3) is not satisfied. The reader is referred to Ref. 1 for the details.

The purpose of this note is to show that Lucky's result, and far more general results of similar type, can be directly deduced from the following proposition.

Proposition: Let f^* and $\mathfrak{R}(f^*)$, respectively, denote an abstract element and a set such that $f^* \varepsilon \mathfrak{R}(f^*)$. Let $\mathcal{S} \triangleq \{f^*\} \cup \mathfrak{R}(f^*)$, and let Q denote a mapping of $\mathcal{S}$ into the set of nonnegative numbers. Let $\mathcal{S}_0$ denote a normed linear space with norm $\| \cdot \|$, and let R denote a mapping of $\mathcal{S}$ into $\mathcal{S}_0$. Let

$$\sigma(f) \triangleq Qf + \| Rf \|$$

for all $f \varepsilon \mathcal{S}$. Suppose that

(*i*) $Qf^* = 0$

(*ii*) for all $g \varepsilon \mathfrak{R}(f^*)$,

$$Qg \geqq \| Rg - Rf^* \| . \tag{4}$$

Then for all $f \epsilon \mathcal{S}$,

$$\sigma(f) \geqq \sigma(f^*) \tag{5}$$

and, if (4) holds with strict inequality for all $g \varepsilon \mathfrak{R}(f^*)$, then (5) holds with strict inequality for all $f \varepsilon \mathcal{S}$ except $f = f^*$.

Proof: Let $f \varepsilon \mathfrak{R}(f^*)$. Then

$$\begin{aligned} \sigma(f) - \sigma(f^*) &= Qf + \| Rf \| - Qf^* - \| Rf^* \| \\ &= Qf + \| Rf \| - \| Rf^* \| \\ &\geqq Qf - \| Rf - Rf^* \| , \end{aligned}$$

from which the validity of the proposition is evident.

An Application of the Proposition

For each $j \varepsilon \mathcal{F} \triangleq \{1,2, \cdots , N\}$, let $\{x_{nj}\}_{-\infty}^{\infty}$ denote a set of real numbers such that $| x_{jj} | > \sum_{n \neq j} | x_{nj} |$. Let $\mathcal{F}'$ denote a proper subset of $\mathcal{F}$ containing at least one element, and let $\{a_n \mid n \varepsilon \mathcal{F}'\}$ be a set of real numbers. Consider the problem of determining a real N-vector $c \triangleq (c_1 , c_2 , \cdots , c_N)$ such that

$$\delta(c) \triangleq \sum_{\substack{n=-\infty \\ n \notin \mathcal{F}'}}^{\infty} \left| \sum_{j \varepsilon \mathcal{F}} c_j x_{nj} \right|$$

is minimized subject to the constraints

$$a_n = \sum_{j \varepsilon \mathcal{F}} c_j x_{nj} , \qquad n \varepsilon \mathcal{F}'. \tag{6}$$

Our assumption that $| x_{jj} | > \sum_{n \neq j} | x_{nj} |$ for $j \varepsilon \mathcal{F}$ implies[2] that there

exists a unique solution c^* to the set of equations

$$a_n = \sum_{j \in \mathfrak{F}} c_j x_{nj}, \qquad n \in \mathfrak{F}'$$

$$0 = \sum_{j \in \mathfrak{F}} c_j x_{nj}, \; n \in (\mathfrak{F} - \mathfrak{F}').$$

We shall prove that if $c \neq c^*$ and c satisfies the constraints of (6), then $\delta(c) > \delta(c^*)$. For the special case in which $\mathfrak{F}'$ contains a single element, this result can be proved[3] with a modification of Lucky's technique.

Let $\mathfrak{R}(c^*)$ denote the set of all real N-vectors g, except the vector c^*, such that

$$a_n = \sum_{j \in \mathfrak{F}} g_j x_{nj}, \qquad n \in \mathfrak{F}'.$$

Let Q be the mapping of $\mathfrak{S} \triangleq \{c^*\} \cup \mathfrak{R}(c^*)$ into the set of nonnegative numbers defined by

$$Qv = \sum_{n \in (\mathfrak{F} - \mathfrak{F}')} \left| \sum_{j \in \mathfrak{F}} v_j x_{nj} \right|$$

for all $v \in \mathfrak{S}$.

Let $\mathfrak{S}_0$ denote the linear space of vectors $u = (\cdots, u_{-1}, u_0, u_{N+1}, u_{N+2}, \cdots)$ with norm

$$\|u\| = \sum_{j \in \mathfrak{F}} |u_j|,$$

and let R denote the mapping of $\mathfrak{S}$ into $\mathfrak{S}_0$ defined by

$$(Rv)_n = \sum_{j \in \mathfrak{F}} v_j x_{nj}, \qquad n \notin \mathfrak{F}$$

for all $v \in \mathfrak{S}$. Then we have

$$\delta(c) = Qc + \|Rc\|$$

for all $c \in \mathfrak{S}$. Since $Qc^* = 0$, if

$$Qg > \|Rg - Rc^*\|$$

for all $g \in \mathfrak{R}(c^*)$, that is, if

$$\rho \triangleq \sum_{n \notin (\mathfrak{F} - \mathfrak{F}')} \left| \sum_{j \in \mathfrak{F}} g_j x_{nj} \right| - \sum_{n \notin \mathfrak{F}} \left| \sum_{j \in \mathfrak{F}} (g_j - c_j^*) x_{nj} \right| > 0 \qquad (7)$$

for all $g \in \mathfrak{R}(c^*)$, then, by the proposition, $\delta(c) > \delta(c^*)$. To show that (7) is satisfied, observe that

$$\sum_{n\varepsilon(\mathfrak{F}-\mathfrak{F}')} \left| \sum_{j\varepsilon\mathfrak{F}} g_j x_{nj} \right| = \sum_{n\varepsilon(\mathfrak{F}-\mathfrak{F}')} \left| \sum_{j\varepsilon\mathfrak{F}} (g_j - c_j^*) x_{nj} \right|$$
$$= \sum_{n\varepsilon\mathfrak{F}} \left| \sum_{j\varepsilon\mathfrak{F}} (g_j - c_j^*) x_{nj} \right|$$

for $g \varepsilon \mathfrak{R}(c^*)$, and that, with $w_j \triangleq (g_j - c_j^*)$,

$$\sum_{n\varepsilon\mathfrak{F}} \left| \sum_{j\varepsilon\mathfrak{F}} w_j x_{nj} \right| \geqq 2 \sum_{n\varepsilon\mathfrak{F}} | w_n x_{nn} | - \sum_{n\varepsilon\mathfrak{F}} \sum_{j\varepsilon\mathfrak{F}} | w_j | \cdot | x_{nj} |$$

and

$$\sum_{n\notin\mathfrak{F}} \left| \sum_{j\varepsilon\mathfrak{F}} w_j x_{nj} \right| \leqq \sum_{n\notin\mathfrak{F}} \sum_{j\varepsilon\mathfrak{F}} | w_j | \cdot | x_{nj} | .$$

Therefore,

$$\begin{aligned} \rho &\geqq 2 \sum_{n\varepsilon\mathfrak{F}} | w_n x_{nn} | - \sum_{k=-\infty}^{\infty} \sum_{n\varepsilon\mathfrak{F}} | w_n | \cdot | x_{kn} | \\ &\geqq \sum_{n\varepsilon\mathfrak{F}} | w_n | \left(| x_{nn} | - \sum_{k\neq n} | x_{kn} | \right), \end{aligned} \tag{8}$$

which completes our proof, since the right side of (8) is positive for all $g\varepsilon\mathfrak{R}(c^*)$

REFERENCES

1. Lucky, R. W., Automatic Equalization for Digital Communication, B.S.T.J., *44*, April, 1965, pp. 547–588.
2. Taussky, O., A Recurring Theorem on Determinants, Am. Math. Monthly, *56*, 1949, pp. 672–676.
3. Gersho, A., Performance of Automatic Equalizers with Nonideal Delay Lines, to be published.