THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 54 November 1975 Number 9

Excitation of Parabolic-Index Fibers With Incoherent Sources

By D. MARCUSE

(Manuscript received May 5, 1975)

We study the excitation of a parabolic-index fiber by an incoherent source. The theory is based on approximating the fiber modes by Laguerre-gaussian functions. The dependence of the total light power injected into the fiber core on the separation between source and fiber, and on the transverse displacement of the source, is shown in graphic form. Also shown are far-field radiation patterns, which indicate the distribution of power versus mode number, for several launching conditions and plots of power versus azimuthal mode number for given values of the compound mode number. The study of launching efficiency versus source radius leads to prescriptions for optimizing the ratio of source to fiber core radius without a matching lens. Use of a lens for matching the image of a small source to the size of the fiber core increases the launching efficiency relative to the power consumption of the light-emitting-source diode.

I. INTRODUCTION

The excitation of step-index fibers (fibers whose cores have a constant index of refraction surrounded by a lower index cladding) has been investigated by several authors by means of ray optics.[1,2] In this paper, we study the excitation of fibers with a core whose refractive indices have a square-law dependence on the radial dimension (parabolic-index fibers). The fibers are assumed to support many modes and are excited by an incoherent source—for example, a light-emitting diode (LED). Our analysis is based on wave optics. We assume that we

may use the Laguerre-gaussian modes of the infinite square-law medium to approximate the modes of the parabolic-index fiber. Lower-order modes do not carry significant amounts of power in the region of the core boundary so that they are approximated very well by the Laguerre-gaussian modes of the infinite medium. We introduce the effect of the core boundary by considering the Laguerre-gaussian modes as being cut off when their propagation constants become equal to the plane wave propagation constant of the cladding material.

The Laguerre-gaussian modes have the advantage that they not only approximate the modes of the parabolic-index fiber, but they are also approximate solutions of beam waves in free space. The free-space modes have a beam width parameter that is a function of the z coordinate (distance from the fiber measured in the direction of its axis). At the fiber end, the free-space Laguerre-gaussian modes match the fiber modes. Thus, it is only necessary to calculate the excitation of the free-space Laguerre-gaussian beam modes by the incoherent source, if we assume that the space surrounding the fiber is matched to the fiber core (at least approximately) by means of an index-matching fluid.

We study the excitation of the modes of the parabolic-index fiber as functions of the source radius, its distance from the fiber, and as a function of its transverse displacement. This study provides information about the tolerance requirements for aligning the source with respect to the fiber. We also discuss the optimization of the source diameter with regard to the total power delivered to the fiber and with regard to the excitation efficiency relative to the electrical power required to drive the source.

We conclude that the tolerance requirements for placing the source are modest and that either the total amount of power or the power excitation efficiency of the fiber can be optimized by a suitable choice of the source diameter. It is assumed that the fiber, as well as the source, have circular cross sections.

Equation (29) states a simple law for the total amount of light power that may be injected into a parabolic-index fiber by an incoherent source of brightness B [Δ is defined by eq. (18)]. The normalization used for the power plotted in Figs. 4 through 10 is based on the expression for the total number of guided modes (28).

The efficiency of the system could be improved considerably by use of a lens or taper to match the light output of a small-area incoherent source to the core of the fiber. Use of lenses or tapers may be undesirable because such matching devices introduce added complexity into the system. However, if high overall efficiency is an important requirement, matching of the output of a small-area LED to the fiber

core with the help of an additional optical system offers a means of increasing the launching efficiency.

II. EXCITATION OF MODES BY AN INCOHERENT SOURCE

We consider a complete orthogonal set of modes obeying the orthonormality condition[3,4]

$$\frac{1}{2}\int_A (\mathbf{E}_\nu \times \mathbf{H}_\mu^*)\cdot \mathbf{e}_z dxdy = P\delta_{\nu\mu}. \tag{1}$$

The symbols $\mathbf{E}_\nu$ and $\mathbf{H}_\mu$ indicate the electric and magnetic field vectors of the modes labeled ν or μ, $\mathbf{e}_z$ is a unit vector in z direction, A is the infinite cross-sectional area of the structure, and $\delta_{\nu\mu}$ is the Kronecker delta symbol. The factor P is a normalizing constant with the dimension of power. It is assumed to be the same for all the modes.

The total electric field can be expressed as the superposition of all the modes,[4]

$$\mathbf{E} = \sum_{\nu=1}^{\infty} c_\nu \mathbf{E}_\nu. \tag{2}$$

The total power carried by the field may be expressed as[4]

$$P_t = \sum_{\nu=1}^{\infty} P\langle |c_\nu|^2\rangle. \tag{3}$$

It can be shown that a current with current density j excites each mode according to the formula,[5]

$$c_\nu = -\frac{1}{4P}\int_V \mathbf{j}\cdot\mathbf{E}_\nu^* dV. \tag{4}$$

The integral is extended over the volume in which the current density j exists; P is the normalizing parameter encountered in (1) and (3).

We are now ready to apply this formalism to the excitation of modes by an incoherent source. As a model of an incoherent source, we consider a disc of circular cross section with radius b and thickness t. The current density inside of the disc is assumed to be a random function with vanishing correlation length. The ensemble average of the absolute square magnitude of (4) is

$$\langle |c_\nu|^2\rangle = \frac{1}{16P^2}\int_V dV \int_V dV' \mathbf{E}_\nu^*\cdot\langle \mathbf{jj}^*\rangle\cdot\mathbf{E}_\nu'. \tag{5}$$

The quantity $\langle \mathbf{jj}^*\rangle$ appearing in (5) is a tensor of second rank. However, we consider that the current in the source is composed of many randomly oriented and randomly phased dipoles. This assumption allows us to assume that the off-diagonal elements of the tensor

vanish, and that all diagonal elements are equal. Thus, the tensor reduces to a multiple of the unit tensor I, and we may write

$$\langle \mathrm{jj}^* \rangle = SI\delta(\mathrm{r} - \mathrm{r}'). \tag{6}$$

Using (6) reduces (5) to the simpler form

$$\langle |c_\nu|^2 \rangle = \frac{tS}{16P^2} \int_{A_s} \mathbf{E}_\nu^* \cdot \mathbf{E}_\nu dxdy. \tag{7}$$

A_s is the area of the circular cross section of the source disc.

If we apply the formula (7) to the plane-wave modes of free space, we can calculate the amount of power ΔP that is flowing through the element of solid angle $d\Omega$ in a given direction in space. This calculation results in the expression

$$\Delta P = \frac{\omega\mu_0 St}{16\pi^2} A_s n k d\Omega, \tag{8}$$

where

$k = \omega\sqrt{\epsilon_0\mu_0}$

n = refractive index of the medium.

The derivation of (8) is sketched in the appendix.

Our model of an incoherent source behaves like a Lambert-law radiator except for a missing factor $\cos\theta$, where θ is the angle between the direction of observation and the normal direction to the surface of the disc. The factor $\cos\theta$ is missing in our theory because we treated the source as being transparent to radiation. In a partially opaque source, radiation leaving the source must originate in a volume with an effective thickness t as seen in the direction of observation. It is shown in Fig. 1 that the effective thickness of the source disc depends

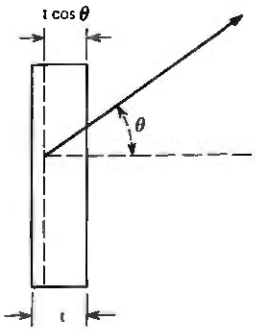


Fig. 1—Schematic of a partially opaque source disc. The effective width for waves emitted at a normal angle to the surface is t; for waves emitted at an angle θ, the effective width is $t \cos\theta$.

on the direction of observation so that we must replace t with $t \cos \theta$. Modified for the case of a partially opaque source, eq. (8) assumes the form

$$\Delta P = BA_s \cos \theta d\Omega. \tag{9}$$

The brightness of the source is defined as

$$B = \frac{\omega \mu_0 S t n k}{16\pi^2}. \tag{10}$$

Equation (10) relates the product St to the measurable quantity B. If we introduce the z component β of the propagation constant nk by the equation

$$\beta = nk \cos \theta \tag{11}$$

and modify (7) for the case of a partially opaque source, we may write

$$P\langle |c_\nu|^2 \rangle = \frac{\pi^2 B}{\omega \mu_0 n k P} \frac{\beta}{nk} \int_{A_s} \mathbf{E}_\nu^* \cdot \mathbf{E}_\nu dx dy. \tag{12}$$

The left-hand side of this expression indicates the power carried by the mode labeled ν. This power is provided by the incoherent source with brightness B. The brightness is the amount of power that the unit area of the source radiates into the unit solid angle, its dimension is W/cm^2 sr (sr = steradians).

Equation (12) is the starting point for our discussion of the excitation of the modes of the parabolic-index fiber by an incoherent source.

III. EXCITATION OF LAGUERRE-GAUSSIAN MODES

The Laguerre-gaussian beam modes of free space, normalized to carry the power P, may be expressed as follows:[6,7]

$$\mathbf{E}_{\nu,p} = \left\{ \frac{2^{\nu+3}}{e_\nu \pi w_0^2 n} \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{p!}{(p+\nu)!} P \right\}^{\frac{1}{2}} \cdot \frac{w_0}{w} \left(\frac{r}{w} \right)^\nu e^{-(r/w)^2} L_p^{(\nu)} \left(\frac{2r^2}{w^2} \right) \cos \nu\phi e^{i\psi}, \tag{13}$$

with

$$e_\nu = \begin{cases} 2 & \text{for } \nu = 0 \\ 1 & \text{for } \nu \neq 0 \end{cases}. \tag{14}$$

The beamwidth parameter is defined by the formula ($k = \omega(\epsilon_0 \mu_0)^{\frac{1}{2}}$, n is the refractive index of "free space")

$$w = w_0 \left[1 + \left(\frac{2z}{nkw_0^2} \right)^2 \right]^{\frac{1}{2}}, \tag{15}$$

and the phase function is given as

$$\psi = -nkz + (\nu + 2p + 1) \arctan\left(\frac{2z}{nkw_0^2}\right) - \frac{nkr^2}{2R}, \tag{16}$$

with the phase front radius of curvature

$$R = z\left[1 + \left(\frac{nkw_0^2}{2z}\right)^2\right]. \tag{17}$$

$\mathbf{E}_{\nu,p}$ is the x or y component of the electric field vector of the Laguerre-gaussian beam mode. Thus, the electric field is linearly polarized but is expressed in a cylindrical coordinate system with the coordinates r, ϕ, and z. We may replace the function cos $\nu\phi$ by sin $\nu\phi$ without changing any other parameter in (13). Thus, the modes are degenerate in the sense that two orthogonal polarizations, as well as both choices of the ϕ symmetry (sin $\nu\phi$ or cos $\nu\phi$), are allowed for each set of mode numbers ν, p. The function $L_p^{(\nu)}$ is a Laguerre polynomial.[8]

The origin of the z coordinate is at the narrowest point of the field distribution where we have $w = w_0$. The minimum beam width w_0 is arbitrary. However, the Laguerre-gaussian beam modes are equal to the modes of the square-law medium,[9] with refractive index distribution

$$n = n_0\left[1 - \left(\frac{r}{a}\right)^2 \Delta\right], \tag{18}$$

if we set $w = w_0$ and use[3]

$$w_0 = \left[\frac{a}{n_0 k}\sqrt{\frac{2}{\Delta}}\right]^{\frac{1}{2}}. \tag{19}$$

This choice of the beamwidth parameter w_0 ensures that, at $z = 0$, the transverse field distribution of the Laguerre-gaussian mode of free space coincides with the mode of the square-law medium (18). Both types of modes, the beam modes of free space and the modes of the square-law medium, are only approximate solutions that apply in the paraxial approximation that holds for small values of the refractive index parameter Δ and, for free space modes, for modes with sufficiently large values of w_0. For values of z other than $z = 0$, the phase function of the modes of the square-law medium must be expressed as

$$\psi = -\beta z \tag{20}$$

with the propagation constant

$$\beta = \left[n_0^2 k^2 - \frac{2n_0 k}{a}\sqrt{2\Delta}(2p + \nu + 1)\right]^{\frac{1}{2}}. \tag{21}$$

If we assume that free space consists of a medium whose refractive index n is matched to an average value of the fiber core, reflection from

the fiber end is negligibly small. We assume that the fiber core has radius a and that its refractive index is given by (18) in the region $0 \leqq r \leqq a$. In the cladding at $r > a$, the refractive index assumes the constant value

$$n_2 = n_0(1 - \Delta). \tag{22}$$

The mode fields (13) [with $w = w_0$ of (19)] approximate the modes of the parabolic-index fiber at radius $r < a$ for small values of ν and p. For large values of the mode numbers ν and p, the fields extend strongly beyond $r = a$ so that (13) (with $w = w_0$) is no longer a good approximation to the fiber modes. However, modes reaching strongly into the cladding are no longer guided by the fiber core. For this reason, we regard (13) (with $w = w_0$) as an approximation for all guided fiber modes and consider the relation

$$\beta = n_2 k = n_0(1 - \Delta)k \tag{23}$$

as a cutoff condition for the guided modes.

Because the modes of the parabolic-index fiber join smoothly with the Laguerre-gaussian beam modes of free space, we obtain the excitation of the fiber modes by determining the excitation coefficients for the Laguerre-gaussian beam modes with the help of (12). Substitution of (13) into (12) results in

$$P\langle|c_{\nu,p}|^2\rangle = \frac{8\pi B}{e_\nu w^2 (nk)^2} \frac{p!2^\nu}{(p+\nu)!} \times \int_{A_s} \left(\frac{r}{w}\right)^{2\nu} e^{-2(r/w)^2} \left[L_p^{(\nu)}\left(\frac{2r^2}{w^2}\right)\right]^2 r dr d\phi. \tag{24}$$

The ratio β/k was approximated by unity. We have added the coefficients for the modes with $\cos \nu\phi$ to those of the $\sin \nu\phi$ symmetry and used the fact that the sum of the squares of cosine and sine is unity. The integration in (24) extends over the surface of the source A_s.

If the surface of the source is larger than the area over which the mode field exists with an appreciable amplitude, the integral can be performed with the result

$$P\langle|c_{\nu,p}|^2\rangle = \frac{4\pi^2 B}{e_\nu (nk)^2}. \tag{25}$$

For $\nu = 0$, there is only one type of mode because $\sin \nu\phi = 0$. For all other values of ν, we have lumped two modes together. If we express again only the power of a single mode of a given polarization and azimuthal symmetry, we have, instead of (25),

$$P\langle|c_{\nu,p}|^2\rangle = 2\left(\frac{\pi}{nk}\right)^2 B = \frac{1}{2}\left(\frac{\lambda_0}{n}\right)^2 B. \tag{26}$$

This is an interesting formula. First of all, it shows that each mode (of

the fiber or of the free-space beam modes) receives an equal amount of power if the incoherent source is large enough. Second, comparison of eq. (26) with (9) shows that each mode acts as though it receives radiation from a square of the source surface whose sides are equal to the wavelength and as if it collects all power radiated into the solid angle $\frac{1}{2}$ *sr*. It is now easy to determine the power that is collected by all the guided modes of the fiber. We need only multiply (26) by the number of modes. The total number of guided modes is obtained from (21) and the cutoff condition (23). Combining these two equations, we obtain the following equation for the boundary in mode number space:

$$(2p + \nu + 1)_{\max} = nka\sqrt{\frac{\Delta}{2}} = \left(\frac{a}{w_0}\right)^2. \tag{27}$$

Figure 2 shows the mode number space defined by the two variables ν and p. The diagonal line (the hypotenuse of the triangle) is defined by (27). The guided modes lie inside the triangle shown in the figure. The total number of modes is approximately equal to four times the area of this triangle. The factor 4 stems from the fact that for each set of values of ν and p, we have modes with two different polarizations and two different azimuthal symmetries (except for $\nu = 0$). Thus, the total number of modes is

$$N = (nka)^2 \frac{\Delta}{2}. \tag{28}$$

The total amount of power P_f injected by a large, incoherent source into a square-law fiber is given as the product of (26) and (28),

$$P_f = \pi(\pi a^2)B\Delta. \tag{29}$$

Equation (24) can be used to calculate the power in each mode for arbitrary position of the source in relation to the fiber. In general, we assume that the fiber is separated from the source by a distance z and that it is offset by an amount d as shown in Fig. 3. The Φ integration in

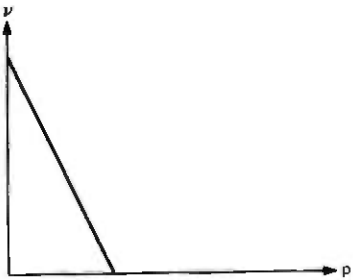


Fig. 2—Mode number plane p, ν. The diagonal line is the boundary of the guided modes that are contained inside the triangular region.

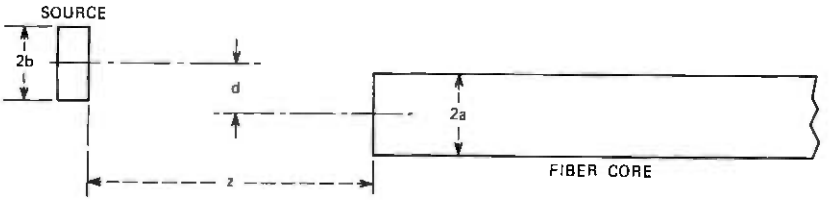


Fig. 3—Geometry of the source exciting a fiber core.

(24) can still be performed with the result

$$P\langle|c_{\nu,p}|^2\rangle = \frac{4\pi B}{(nk)^2 w^2}\frac{p!}{(p+\nu)!} \times \left\{2\pi \int_0^{|b-d|} \left(\frac{2r^2}{w^2}\right)^{\nu} e^{-2(r/w)^2}\left[L_p^{(\nu)}\left(\frac{2r^2}{w^2}\right)\right]^2 r\,dr + \int_{|b-d|}^{|b+d|} (\pi - 2\Phi_1)\left(\frac{2r^2}{w^2}\right)^{\nu} e^{-2(r/w)^2}\left[L_p^{(\nu)}\left(\frac{2r^2}{w^2}\right)\right]^2 r\,dr\right\}, \quad (30)$$

with

$$\Phi_1 = \arcsin\left(\frac{r^2 + d^2 - b^2}{2rd}\right). \quad (31)$$

Equation (30) applies again for one mode of given polarization and azimuthal symmetry. The integral in (30) must be evaluated numerically. The z dependence of the expression is hidden in the beamwidth parameter w according to (15).

IV. NUMERICAL EVALUATION AND RESULTS

In this section we show the results of numerical evaluations of eq. (30). We begin by discussing the dependence of the total power injected into the fiber on the distance between the source and the fiber end. All length variables are normalized with respect to the fiber radius a. Figure 4 shows the dependence of the normalized total power on z/a for a source whose radius is equal to the fiber radius, $b/a = 1$, for three values of Δ. The normalization of the power is apparent by comparison with (26). Since P_f indicates the total power carried by all the modes, we have divided it by the total number of modes N that is obtained from (28). This normalization results in unit normalized power at $z = 0$.

Figure 4 shows that, for $\Delta = 0.01$, the amount of power that is injected into the fiber by the incoherent source drops to approximately one-half of its maximum value for $z/a = 11$. Thus, the injection process is surprisingly insensitive to the separations between the source and the fiber end. We have checked that the curves of Fig. 4 do not change

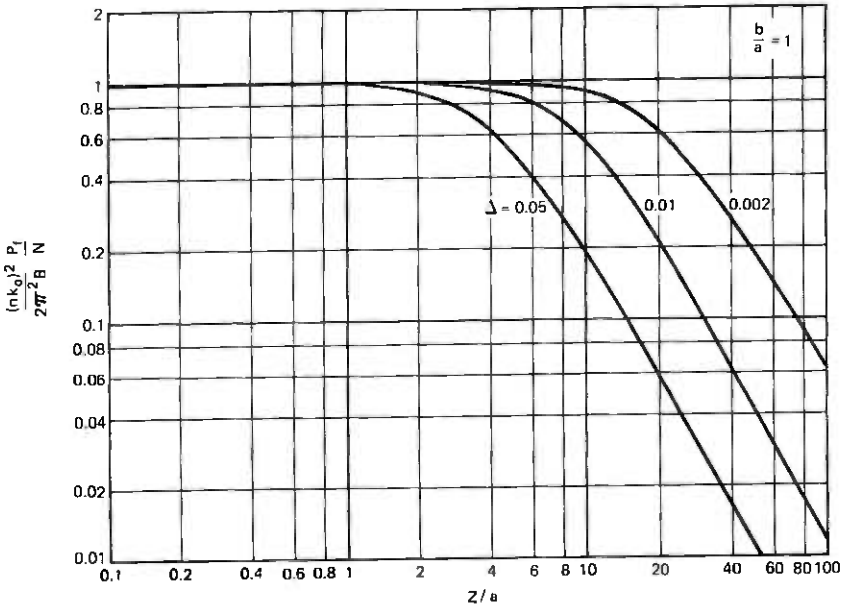


Fig. 4—Normalized total power P_f injected into the parabolic-index fiber as a function of the distance z between source and fiber. N is the total number of modes [see (28)]; B is the source brightness. This figure holds for a source-to-fiber core radius ratio $b/a = 1$.

if the value of ka is varied. Thus, the curves are universal, at least for large values of ka, that is, for fibers supporting a large number of modes. This dependence on k suggests that the curves may be derived by the methods of ray optics. The lowest part of the curves has almost reached an inverse square-law dependence on fiber-source separation. We have checked that a precise inverse square-law dependence is obtained for source-to-fiber ratios of $b/a = 0.1$. Thus, the inverse square law is reached rather slowly and will be realized by the curves of Fig. 4 for even larger values of z/a than those appearing in the figure. The curve with $\Delta = 0.01$ shows that a separation of the source from the fiber end equal to the fiber diameter, $z = 2a$, causes a drop in power coupling efficiency by only approximately 5 percent. The tolerance to source-fiber separation eases for smaller values of Δ. The curves do not reach exactly the value unity at $z = 0$ because the source is not infinitely wide. However, the slight departure from unity (0.3 percent for $N = 400$, $\Delta = 0.01$) is not apparent on the scale of the figure.

Figure 5 shows the dependence of the excitation efficiency on the amount of relative offset d/a for a source whose radius equals the fiber radius, $b/a = 1$, and which is located at three different distances from

the fiber end. The curves were computed for $\Delta = 0.01$. However, the shapes of the curves are universal and only their vertical positions depend on the value of Δ. It is immediately apparent from this figure that the relative tolerance to source displacements (offsets) becomes more liberal as the distance between the source and the fiber end is increased. A source displacement of $d/a = 1$ causes a drop in power to 40 percent of its maximum value if the source is placed directly at the fiber end, $z/a = 0$. For $z/a = 10$, the excitation efficiency has dropped from 0.55 to 0.34, that is, only 62 percent of its maximum value, for a transverse source displacement of $d/a = 1$. Like Fig. 4, Fig. 5 is independent of ka if this value is large.

So far we have concentrated on the total power injected into the fiber. It is interesting to consider the distribution of power over the various guided modes as the distance between the source and the fiber is increased. The far-field pattern emerging from the far end of the fiber (which is not facing the source) gives an indication of the mode distribution. Figure 6 shows the power density of the far-field radiation pattern as a function of angle α. The curves of this figure represent far-field radiation patterns for several values of the relative fiber-source distance z/a for $\Delta = 0.01$. The curves were obtained by taking the

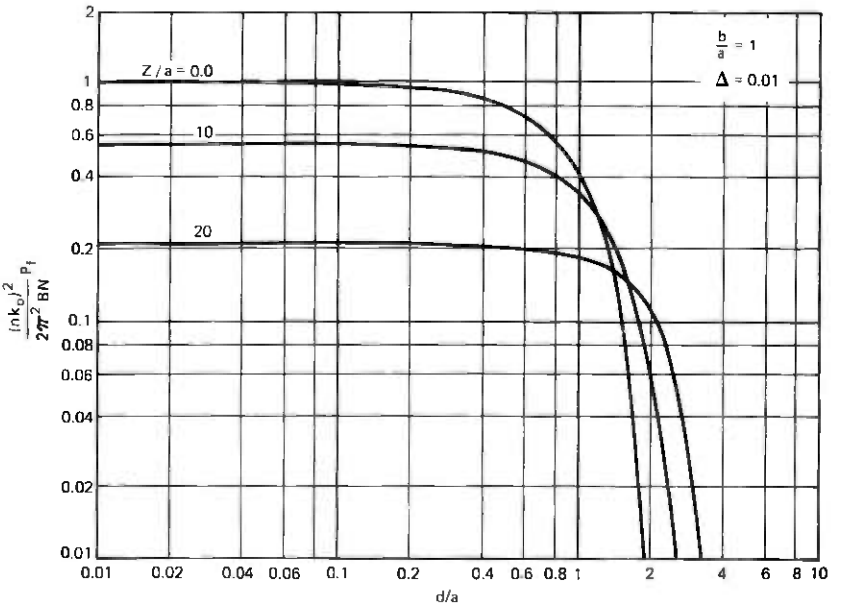


Fig. 5—Normalized total power injected into parabolic-index fiber as a function of the transverse source displacement d for several values of the distance z of the source from the fiber and for $\Delta = 0.01$.

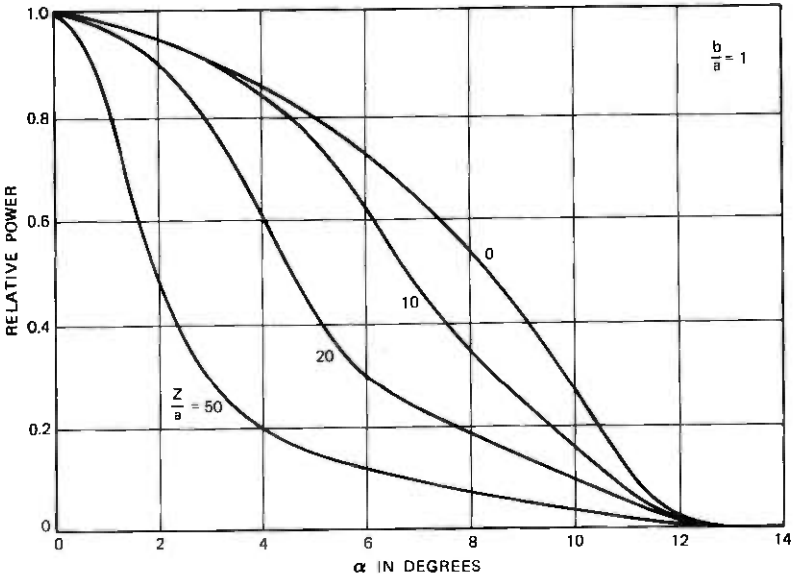


Fig. 6—Far-field radiation pattern. α is the angle (in degrees) of the direction of the observation point with respect to the fiber axis. The curves are arbitrarily normalized to unity at $\alpha = 0$. The source-to-fiber core radius ratio is $b/a = 1$, $\Delta = 0.01$.

ensemble average of the absolute square value of (2),

$$\langle |E|^2 \rangle = \sum_{\nu=1}^{N} \langle |c_\nu|^2 \rangle |\mathbf{E}_\nu|^2, \tag{32}$$

with the mode fields of (13) and the expansion coefficients of (30). All curves in Fig. 6 were computed for a source whose radius is equal to the fiber radius, $b/a = 1$, with no transverse source displacement, $d/a = 0$. The far-field pattern of a parabolic-index fiber with all modes equally excited corresponds to the curve labeled $z/a = 0$. As the source is moved away from the fiber end, the far-field radiation pattern narrows. This narrowing is caused by the fact that higher-order modes receive less power as the distance z is increased. The curves of Fig. 6 are normalized so that the power density at zero angles becomes unity.

Figures 7 through 9 give more detailed insight into the distribution of power versus mode number. Equation (21) shows that modes with equal values of the compound mode number,

$$M = 2p + \nu + 1, \tag{33}$$

have equal propagation constants. These modes lie on straight lines parallel to the diagonal line in mode-number space shown in Fig. 2. Figure 7 indicates the power distribution among the modes with

$M = 19$ that lie near the guided-mode boundary in mode-number space. If the source is placed directly in contact with the fiber, $z/a = 0$, all modes with $M = 19$ are almost equally excited, receiving almost the maximum of power. However, for $z/a = 4$, at a point where the total amount of power has dropped by only 8 percent from its maximum value, the highest-order mode group with $M = 19$ suffers a very substantial decrease in power. It is interesting that modes with higher values of ν receive more power (for constant values of M). This trend is reversed only for $z/a \geqq 8$.

Figure 8 shows the relative power in other mode groups at a fixed value of $z/a = 8$. Modes with low compound mode number, $M \leqq 6$, are still fully excited, but with increasing values of M the amount of power in the higher-order modes drops off. Figure 9 shows the same trend even more strongly for a source-fiber separation of $z/a = 20$.

So far we have studied the dependence of the excitation efficiency of the parabolic-index fiber on the source-fiber separation z/a and on the amount of offset d/a for $b/a = 1$. Figure 10 shows the normalized total amount of power for $z/a = 0$ as a function of the relative source radius b/a. This curve does not depend on the values of Δ or ka. As expected, the total amount of injected power drops off as b/a decreases. However, the decrease in total power is not proportional to the area of the source, as one might have expected, but is nearly proportional to

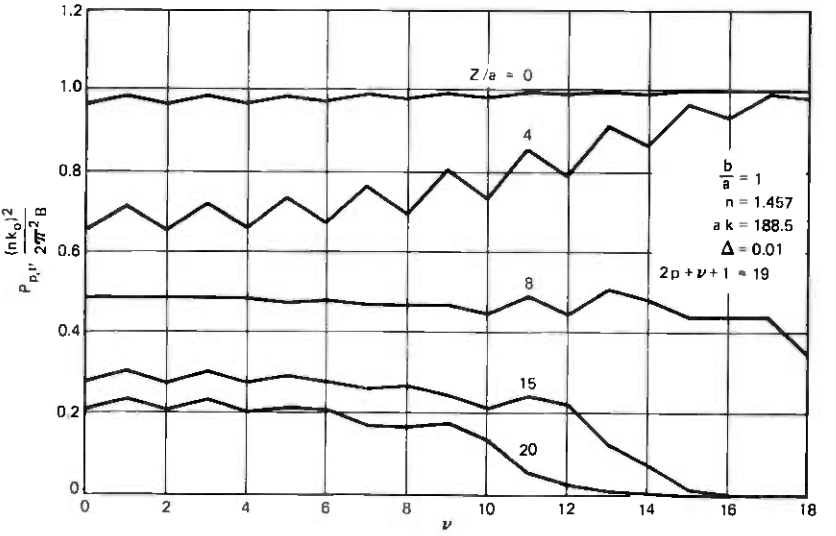


Fig. 7—Normalized power in a given mode belonging to the mode group with compound mode number $2p + \nu + 1 = 19$ as a function of the azimuthal mode number ν for several values of the distance z between source and fiber. These curves apply to the case $b/a = 1$, $ka = 188.5$, $\Delta = 0.01$, $n = 1.457$.

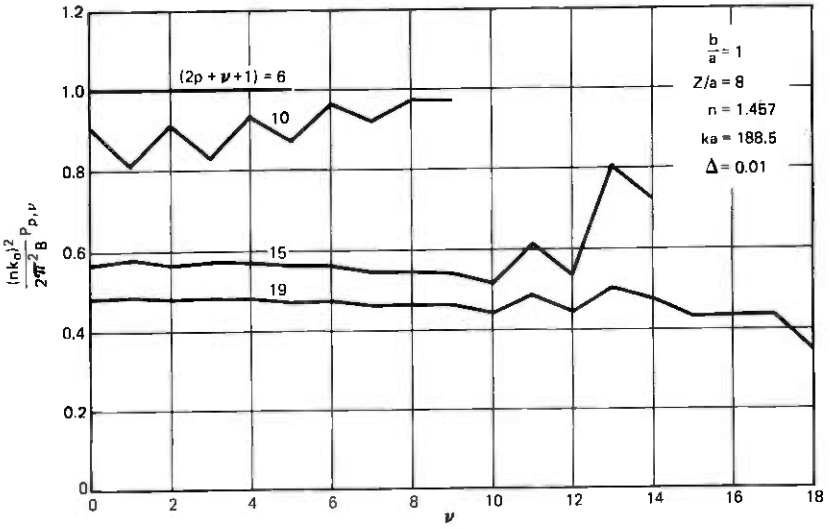


Fig. 8—Normalized power of individual modes belonging to the compound mode number $2p + \nu + 1$ (whose values are given in the figure) as a function of the azimuthal mode number ν. The source-to-fiber distance is $z/a = 8$, all other parameters are the same as in Fig. 7.

the source radius. This behavior has interesting consequences for the optimum choice of the source radius as is discussed in the next section.

Figure 11 shows far-field radiation patterns for several values of b/a. Comparison with Fig. 6 shows that the dependence of the far-field

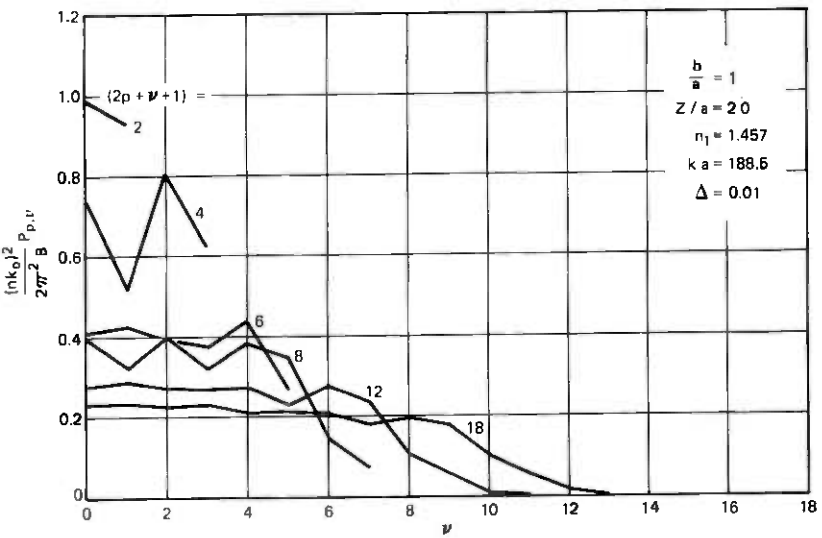


Fig. 9—This figure is similar to Fig. 8 with $z/a = 20$.

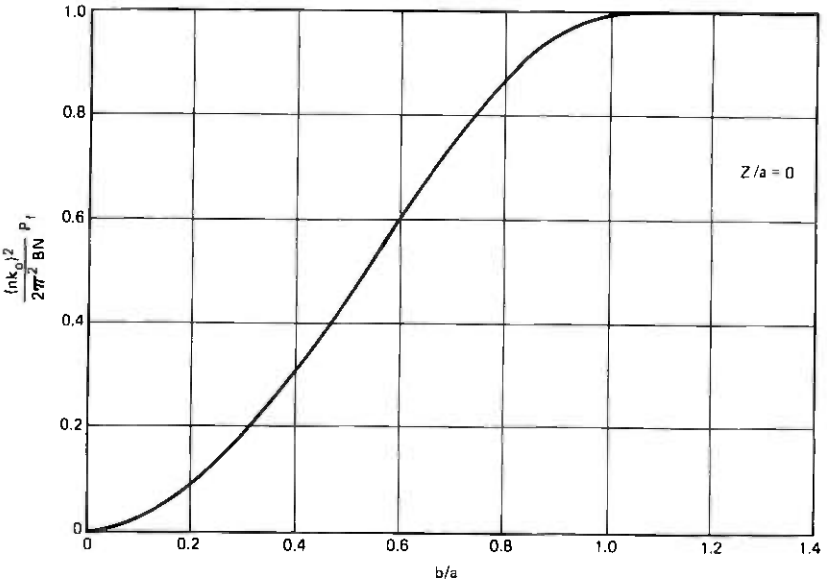


Fig. 10—Normalized total power as a function of the normalized source radius b. This figure is independent of ka and Δ.

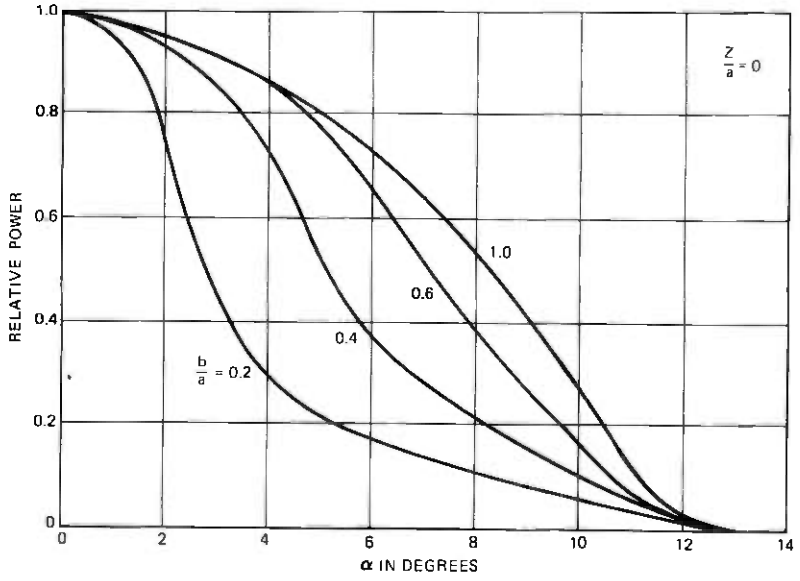


Fig. 11—Far-field radiation pattern from the end of the fiber for several values of the source-to-fiber core radius ratio b/a.

radiation patterns on the source radius bears a close resemblance to its dependence on the source-fiber separation. Higher-order modes are excited less strongly as the source radius decreases.

V. SOURCE OPTIMIZATION WITHOUT LENSES

If the source brightness B is held constant, more power is injected into the fiber as b/a increases to a value near unity. Beyond that value, no further advantage is to be gained. In fact, even though the total amount of injected power remains constant for $b/a > 1$, the overall efficiency decreases since regions of the source at $b > a$ do not contribute to the excitation of the fiber, but do waste their power. If brightness were independent of the dimensions of the source, the optimum source radius would be $b = a$. However, light-emitting diodes (LEDs) tend to be brighter if their radius decreases. C. A. Burrus has made measurements on a special type of LED operated at two-thirds of its saturation current that indicate that source brightness increases with decreasing radius.[10] This dependence is shown in Fig. 12, which was drawn from data given in Burrus' paper. The solid line is drawn

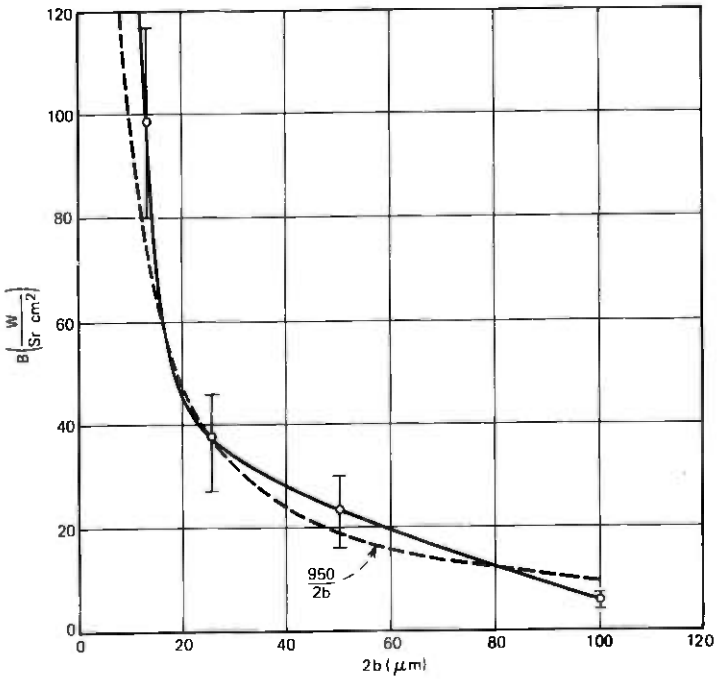


Fig. 12—Source brightness of high-intensity LEDs according to Ref. 10. The dotted line is an inverse $2b$ law used to approximate the solid curve.

through the average values, and the vertical error bars indicate the experimental uncertainty of the brightness measurements. The dotted line shown in Fig. 12 is the hyperbola

$$B = \frac{950}{2b} = \frac{W}{(b/a)} \tag{34}$$

that seems to approximate the experimental data reasonably well. To a rough approximation, the brightness of LEDs (at least of the Burrus type) seems to be inversely proportional to their radius.

To study the excitation efficiency as a function of source radius, we approximate the curve of Fig. 10 by the polynomial

$$\frac{(nk)^2}{2\pi^2}\frac{P_f}{N} = B\left[A_1\frac{b}{a} + A_2\left(\frac{b}{a}\right)^2 + A_3\left(\frac{b}{a}\right)^3\right] \tag{35}$$

with

$$\left.\begin{aligned} A_1 &= -0.06875 \\ A_2 &= 2.85 \\ A_3 &= -1.78125 \end{aligned}\right\}. \tag{36}$$

This third-order polynomial approximates the curve in Fig. 10 to within 1 percent. Substitution of (34) into (35) results in

$$\frac{(nk)^2}{2\pi^2}\frac{P_f}{WN} = A_1 + A_2\frac{b}{a} + A_3\left(\frac{b}{a}\right)^2. \tag{37}$$

This function is plotted in Fig. 13. Under the conditions prevailing in Burrus-type diodes, where the maximum attainable brightness depends on the source radius, the total power that can be injected by an LED in direct contact with a parabolic-index fiber has a maximum at a source radius of $b = 0.8a$. However, even for source radii as small as

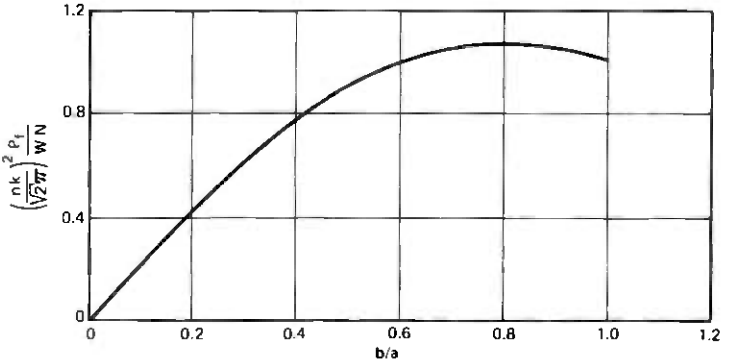


Fig. 13—Total power injected into the fiber with the high-brightness LEDs of Fig. 12 as a function of the (normalized) source radius b.

$b = 0.6a$, the total power in the fiber equals the amount that is obtained with an LED whose radius equals the fiber radius. This means that an LED with a radius only approximately half as large as that of the fiber core is still almost as effective as an LED whose radius equals the fiber core radius.

Disregarding the electrical input power into the LED, we would optimize the overall performance of the fiber system, operated without a matching lens, by choosing a source-to-fiber radius ratio of $b/a = 0.8$. However, a different optimum is obtained if we try to optimize the ratio of total power injected into the fiber to the power required to drive the diode. The power input to the LED can be estimated from the information contained in Burrus' paper[10] by multiplying the diode current with the energy gap voltage, V = 1.38 V at room temperature. This power estimate comes close to the actual power since the voltage developed across the LED's terminals varies between 1.35 and 1.6 V. Four points obtained for the LED's power consumption operated at two-thirds the saturation current are shown in Fig. 14 as a function of the diameter $2b$ of the diode. In the region between $2b = 0$ and $2b = 50$ μm, the power curve is approximately linear. According to the limited information that is available, the curve seems to turn over for larger values of $2b$. However, since only one point (at $2b = 100$ μm) does not lie on the straight line, the shape of the curve beyond $2b = 50$ μm is not known. For sufficiently small source radii, we approximate the curve in Fig. 14 by the equation

$$P_e = (8.5)(10^{-3})(2b) = W_e b/a, \tag{38}$$

keeping in mind that this linear law becomes questionable for $2b > 50$ μm. Substitution of (38) into (37) yields

$$\frac{(nk)^2}{2\pi^2}\frac{W_e}{NW}\frac{P_f}{P_e} = \frac{A_1}{b/a} + A_2 + A_3\frac{b}{a}. \tag{39}$$

This function is shown in Fig. 15. The maximum of the fiber excitation efficiency relative to the electrical drive power of the diode appears at $b/a = 0.2$, that is, at rather small source radii. It is important to remember that, even though Fig. 15 is drawn as a function of b/a, only b is allowed to vary while a must be kept constant because W_e, appearing in the normalization coefficient, is a function of a.

A good compromise between the maximum achievable total power and the desire to obtain good excitation efficiency relative to the power input to the LED may be to operate with a diode whose radius is approximately one-half of the fiber radius. In this case, $b/a = 0.5$, we lose 17 percent of the optimum operating power efficiency and work 19 percent below the maximum achievable injected power. But neither loss of efficiency is very serious and both requirements, low diode power

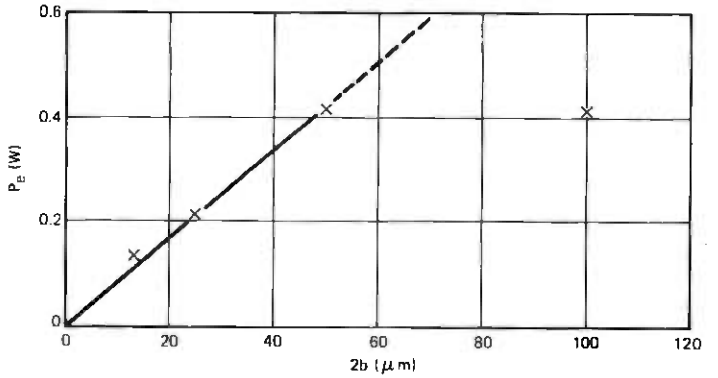


Fig. 14—Power consumption of the high-brightness LEDs as a function of source radius.

consumption and a large amount of total power launched into the fiber, are still approximately satisfied.

VI. OPERATION WITH A MATCHING LENS

Figure 10 shows that the amount of power launched into the parabolic-index fiber decreases with decreasing fiber core radius if the source brightness is held constant. The reason for this decrease in in-

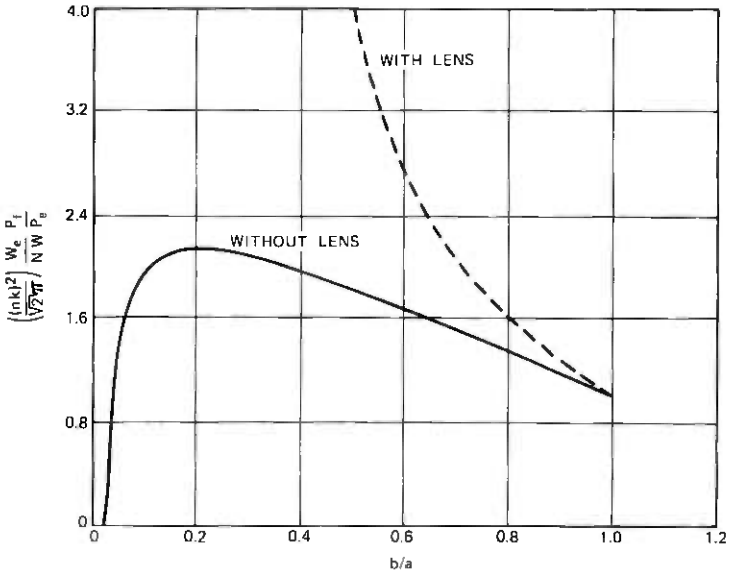


Fig. 15—Normalized ratio of total power injected into the fiber, P_f, relative to the electrical drive power P of the diode as a function of the (normalized) source radius. The solid curve applies for an LED in direct contact with the fiber; the dotted curve describes operation with a matching lens.

jected power is the loss of total source area. However, by bringing the source in direct contact with the fiber, a large amount of power is lost, because the fiber can trap only rays emitted at certain small maximum angles whose values depend on the point at which the ray is entering the fiber core. If we remove the source from the end of the fiber and focus its light onto the fiber end with a matching lens, we may increase the source image to make it coincide with the fiber core radius, but at the same time we inject all rays at a smaller angle. The loss in source brightness, caused by the magnification of its image, is thus compensated for by the fact that many of the rays, those that left the diode at angles too large to be trapped when the source was in direct contact with the fiber, are now transformed to smaller angles so that a wider cone of light leaving the source is able to be accepted by the fiber.

To investigate the beneficial effect of a matching lens, we use the transformation laws of Laguerre-gaussian beams that have been formulated by Kogelnik.[11] We assume that the source is imaged by a lens onto the end of the fiber as shown in Fig. 16. The transformation laws of gaussian beams[3,11] yield the result (in agreement with geometrical optics) that the beamwidth parameter w_0 of the fiber mode is transformed to the width w given by

$$w = \frac{b}{b'} w_0. \tag{40}$$

The source radius b has been transformed by the lens to the radius of its image b'; w_0 is given by (19). We can evaluate the effect of the matching lens by writing (30) (with $d = 0$) in the form

$$P(|c_{\nu,p}|^2) = \frac{2\pi^2 B}{(nk)^2} \frac{p!}{(p+\nu)!} \int_0^{2b^2/w^2} u^\nu e^{-u} [L_p^{(\nu)}(u)]^2 du. \tag{41}$$

Equation (41) expresses the amount of power in one mode. The total

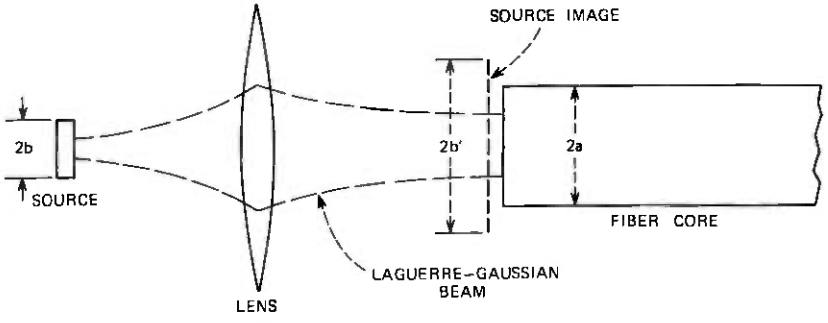


Fig. 16—Launching geometry with a matching lens.

power injected into the fiber is obtained by summing over all the guided modes. The beamwidth parameter w of the transformed Laguerre-gaussian beam enters this expression only in the upper limit of the integral. The critical term to consider is thus

$$u_1 = 2\frac{b^2}{w^2} = 2\frac{b'^2}{w_0^2}. \tag{42}$$

Equation (41) is of exactly the same form as the expression (30) for $d = 0$ in the absence of the lens. Only the beamwidth w at the position of the source is important. Choosing the lens in such a way that the image size b' of the source becomes equal to the fiber core radius ($b' = a$) results in $u_1 = 2a^2/w_0$. This means that the launching efficiency through the matching lens is identical to the launching efficiency for a source with $b = a$ in direct contact with the fiber, even though the size of the source in Fig. 16 is smaller than the fiber radius. In fact, we know from Fig. 10 that the excitation efficiency is not changed even if the source is made larger than the fiber core. The same is true for the image of the source projected onto the fiber end through the matching lens. Even if we project a source image that is larger than the fiber end, the launching efficiency is maintained if the source brightness (of the original, not the imaged source) is held constant. This seemingly contradictory result is caused by the fact that more and more rays arrive at the fiber end at smaller angles with respect to the fiber axis, making up the loss in image brightness that is caused by the larger lens magnification. This constancy of the launching efficiency, which is achievable with a matching lens, breaks down only when the lens must be so far removed from the fiber that some of the light power, which would radiate from the fiber (if we reverse the direction of all the light beams), begins to miss the lens boundary.

Thus, we have reached the conclusion that a matching lens can improve the launching efficiency of a small source to such a degree that we can obtain as much power as would be available from a large source in contact with the fiber. The total coupling efficiency with respect to the electrical power used by the diode may be considerably improved with the help of a matching lens. Instead of (39), a matching lens allows us to achieve the power ratio

$$\frac{P_f}{P_e} = \frac{2\pi^2 N W}{(nk)^2 W_e}\frac{1}{(b/a)^2}. \tag{43}$$

Equation (43) is plotted as the dotted curve in Fig. 15. This expression does not contain the effect of the finite lens aperture that must be considered for very large magnification. Equation (43) indicates that a matching lens would allow us to achieve higher launching efficiencies

than may be achieved by placing the LED directly in contact with the fiber.

Instead of a lens, a tapered dielectric waveguide of high refractive index difference between core and cladding material may be used to match a small LED to a larger fiber core. However, any optical matching device complicates the basically simple configuration of an LED in direct contact with the fiber. For many applications, it may be more advantageous to suffer the additional coupling loss that results by forgoing the procedure of matching the source to the fiber size. Incidentally, a matching lens does not increase the amount of power that may be injected from a large source.

VII. CONCLUSIONS

We have studied the excitation of parabolic-index fibers with incoherent light sources and found that a source, whose area covers the cross section of the fiber core, injects equal power into all the modes. As the source is moved away from the fiber end, it injects relatively less power (without a lens) into higher-order (as compared to low-order) modes. However, the injection mechanism is quite tolerant of source-fiber separation. At a distance of five fiber-core diameters (assuming $b/a = 1$ and $\Delta = 0.01$), the total amount of power injected into the fiber core decreases only to one-half the amount that can be achieved if the source is placed directly in contact with the diode. If the source is transversely displaced, the amount of power launched into the fiber drops to about one-half of its maximum value for a source displacement equal to the fiber radius, if the source is in contact with the fiber.

Without a matching lens, the amount of power launched into the fiber decreases with decreasing source radius (if $b/a < 1$). However, since the brightness achievable from an LED increases with decreasing radius, an optimum radius for maximum light-power injection into the fiber is obtained at a source-to-core radius ratio of $b/a = 0.8$. Relative to the electrical power requirements of the diode, the launching efficiency is optimized at $b/a = 0.2$. A compromise between these two optima may be to choose the ratio $b/a = 0.5$. These numbers are based on a special high-brightness LED developed by Burrus.[10]

Use of a matching lens allows us to inject the same amount of power from a small LED that would be available from a large source of equal brightness in direct contact with the fiber. However, because the achievable brightness increases with the decreasing radius of an LED more power can be obtained from a small LED whose light is focused into the fiber with a lens. Use of a matching lens also increases the overall efficiency of operation as shown by the dotted line in Fig. 15.

However, matching lenses or tapers complicate the basically simple launching geometry of an LED in direct contact with the fiber.

VIII. ACKNOWLEDGMENT

We profited from several valuable discussions with E. A. J. Marcatili and C. A. Burrus. Mr. Marcatili suggested the possibility of optimizing the source radius without a matching lens.

APPENDIX

We sketch the derivation of eq. (8). The plane-wave modes of free space can be expressed as[12]

$$\mathbf{E}_{\kappa,\sigma} = \frac{(2\omega\mu_0 P)^{\frac{1}{2}}}{2\pi\sqrt{\beta}}\, \mathbf{e} e^{-i(\kappa x+\sigma y+\beta z)} \tag{44}$$

with

$$\beta^2 = n^2k^2 - \kappa^2 - \sigma^2. \tag{45}$$

From (7) we obtain, by substitution of (44),

$$P\langle|c|^2\rangle = \frac{\omega\mu_0}{32\pi^2\beta} StA_s. \tag{46}$$

The total power radiated by the source is obtained from the formula:[13]

$$P_r = 2\int\int_{-\infty}^{\infty} P\langle|c|^2\rangle d\kappa d\sigma. \tag{47}$$

The factor 2 accounts for the two possible polarizations. We may express the differentials of the integral in terms of the element of solid angle $d\Omega$ into which the radiation is directed,[14]

$$d\kappa d\sigma = nk\beta d\Omega. \tag{48}$$

Thus, the fractional amount of power radiated into the element of solid angle is

$$\Delta P_r = 2P\langle|c|^2\rangle nk\beta d\Omega. \tag{49}$$

Substitution of (46) into (49) results in (8).

REFERENCES

1. K. H. Yang and J. D. Kingsley, "Calculation of Coupling Losses Between Light Emitting Diodes and Low Loss Optical Fibers," Applied Optics, *14*, No. 2 (February 1975), pp. 288–293.
2. P. Di Vita and R. Vannucci, "Geometrical Theory of Coupling Errors in Dielectric Optical Waveguides," Optics Communication, *14*, No. 1 (May 1975), pp. 139–144.
3. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold, 1972.
4. D. Marcuse, *Theory of Dielectric Optical Waveguides*, New York: Academic Press, 1974.

5. R. E. Collin, *Foundations for Microwave Engineering*, New York: McGraw-Hill, 1966, eq. (4.84a), p. 185.
6. G. Goubau and F. Schwering, "On the Guided Propagation of Electromagnetic Beam Waves," IRE Trans. on Antennas and Propagation, *AP-9*, No. 3 (May 1961), pp. 248–256.
7. H. Kogelnik and T. Li, "Laser Beams and Resonators," Applied Optics, *5*, No. 10 (October 1966), pp. 1550–1567.
8. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products*, 4th edition, New York: Academic Press, 1965, p. 1037.
9. C. N. Kurtz and W. Streifer, "Guided Waves in Inhomogeneous Focusing Media, Part I: Formulation, Solution for Quadratic Inhomogeneity," IEEE Trans. Microwave Theory and Techniques, *MTT-17*, No. 1 (January 1969), pp. 11–15.
10. C. A. Burrus, "Radiance of Small-Area High-Current-Density Electroluminescent Diodes," Proc. IEEE, *60*, No. 2 (February 1972), pp. 231–232.
11. H. Kogelnik, "Imaging of Optical Modes—Resonators with Internal Lenses," B.S.T.J., *44*, No. 3 (March 1965), pp. 455–494.
12. Ref. 3, eq. (4.7-1), p. 168.
13. Ref. 3, eq. (4.7-13), p. 169.
14. Ref. 3, eq. (4.7-15), p. 170.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

An Analysis of the Effect of Lossy Coatings on the Transmission Energy in a Multimode Optical Fiber

By A. H. CHERIN and E. J. MURPHY

(Manuscript received May 20, 1975)

Lossy plastic coatings are used as a means of providing mechanical protection for optical fibers during the optical-cable manufacturing process. A model utilizing a quasi-ray analysis has been developed in this paper to determine the effects of lossy coatings on the transmission energy in a multimoded step-index optical fiber. Cladding thickness is the dominant fiber parameter that plays a critical role in preventing transmission loss due to a lossy coating. Other parameters that significantly affect transmission loss are transmitting wavelength, the real and imaginary part of the refractive index of the lossy coating, and the fiber core diameter.

I. INTRODUCTION

A thin lossy plastic coating applied to individual optical fibers is being considered as a means of decreasing crosstalk[1] between the optical fibers and as a mechanism for protecting the fibers during the cable-manufacturing process. The effect of the lossy coating on the transmission energy in a fiber is the subject of this paper.[2,3]

A quasi-ray tracing approach is used to describe energy propagation in a multimoded step-index optical fiber with a lossy plastic coating.[4,5]

An integral expression for the power transmitted in the fiber is developed in terms of the geometry of the round fiber, the intrinsic loss of the fiber core, the reflection coefficient at the core-cladding interface, and the energy distribution at the launching end of the fiber. To calculate the reflection coefficient at the core-cladding boundary, the strategy followed is to replace the round fiber by a lossy multilayered semi-infinite slab model. A computer evaluation of the integral expressions for the transmitted and input power has been made. Included in this paper are the results of a study showing the functional relationship between power loss due to a lossy coating and cladding thickness, mode energy distribution, numerical aperture, and wave length.

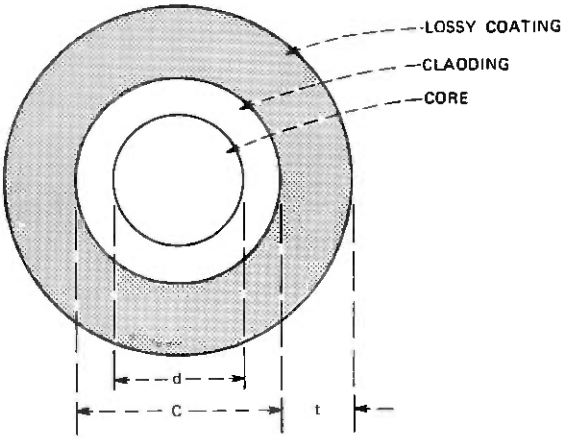


Fig. 1—Optical fiber with lossy coating.

II. DERIVATION OF TRANSMISSION AND REFLECTION LOSS FORMULAS

For the configuration shown in Fig. 1, we follow a technique developed in an earlier paper[1] and present an expression for the transmission of energy in the core of a multimode optical fiber whose cladding is surrounded by a lossy material. We assume that an optical source focuses its power on the center of the entrance end of a fiber, exciting meridional rays as shown in Fig. 2. We also assumed that:

(*i*) The input angular power distribution of the fiber is a gaussian function of the form

$$F(\theta) = F_e \epsilon^{-(\theta/\kappa\theta_c)^2}, \tag{1}$$

where κ is a parameter that is a measure of the width of the input beam and also an indication of how the power is distributed among the modes of the fiber, θ_c is the critical angle of the fiber, and F_e is a constant amplitude.

(*ii*) The propagating modes within the fiber are uncoupled, with the absorption coefficient α equal for all modes.

Under these assumptions, the fiber input and output powers are:

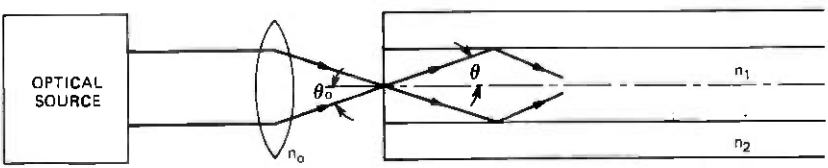


Fig. 2—Meridional ray fiber excitation.

$$P_{\text{input}} = 2\pi F_e \int_0^{\theta_{\max}} \sin\theta \epsilon^{-(\theta/\alpha\theta_c)^2} d\theta \tag{2}$$

and

$$P_{\text{output}} = 2\pi F_e \int_0^{\theta_{\max}} \sin\theta \epsilon^{-(\theta/\alpha\theta_c)^2} R^M(\theta) \epsilon^{-\alpha L \sec\theta} d\theta, \tag{3}$$

where

L is the fiber length,

M is the total number of bounces a ray makes while propagating down the fiber and is the largest integer smaller than

$$M = \left| \frac{L}{d} \tan\theta + \frac{1}{2} \right|, \tag{4}$$

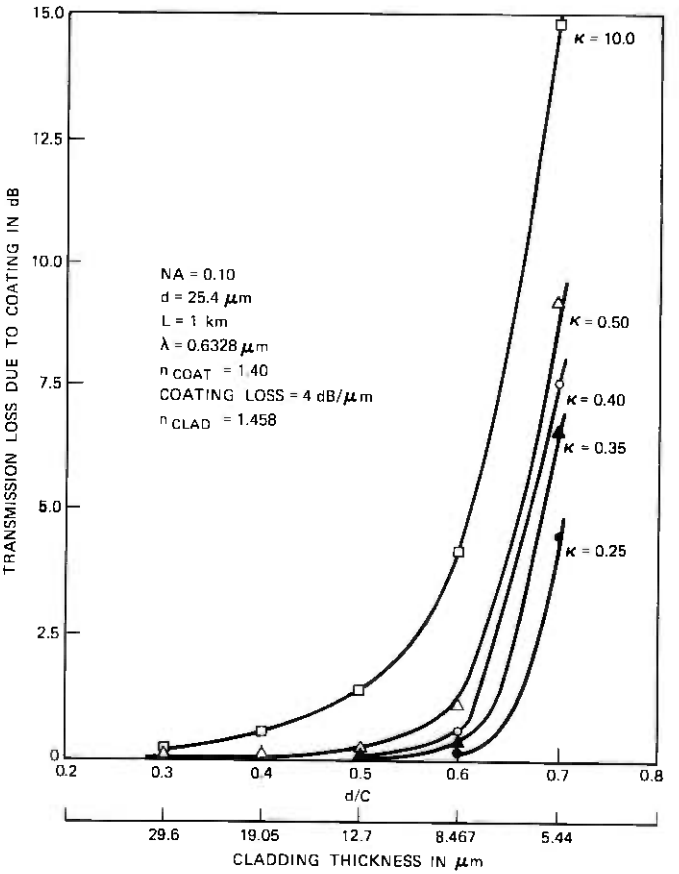


Fig. 3—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.10, d = 25.4 μm.

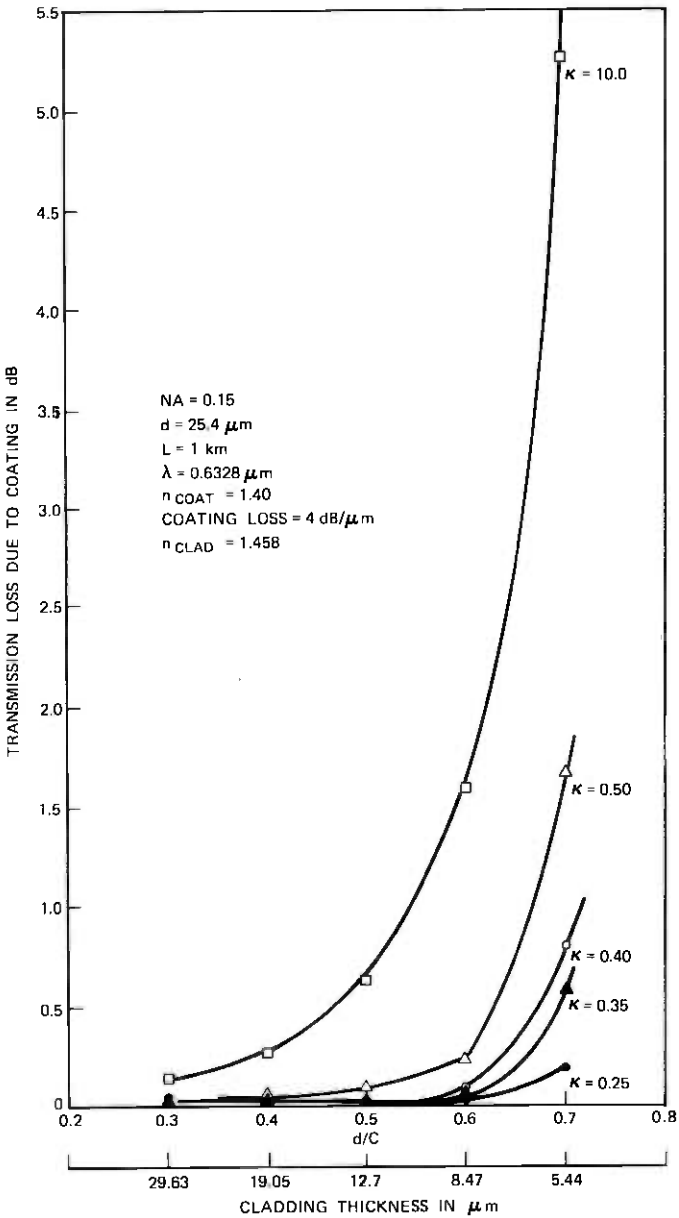


Fig. 4—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.15, d = 25.4 μm.

d is the fiber core diameter,

α is an absorption coefficient per unit length that takes into account both the fiber bulk absorption loss and scattering loss,

θ_{max} is the maximum input angle corresponding to the critical angle within the fiber, and

$R(\theta)$ is the reflection coefficient at the core-cladding boundary.

To calculate $R(\theta)$, the strategy followed is to replace the round fiber by a lossy multilayered semi-infinite slab model. The derivation of $R(\theta)$ for the slab model is shown in the appendix. The refractive indices and, in turn, the impedances of the media are complex to account for the lossy coating.

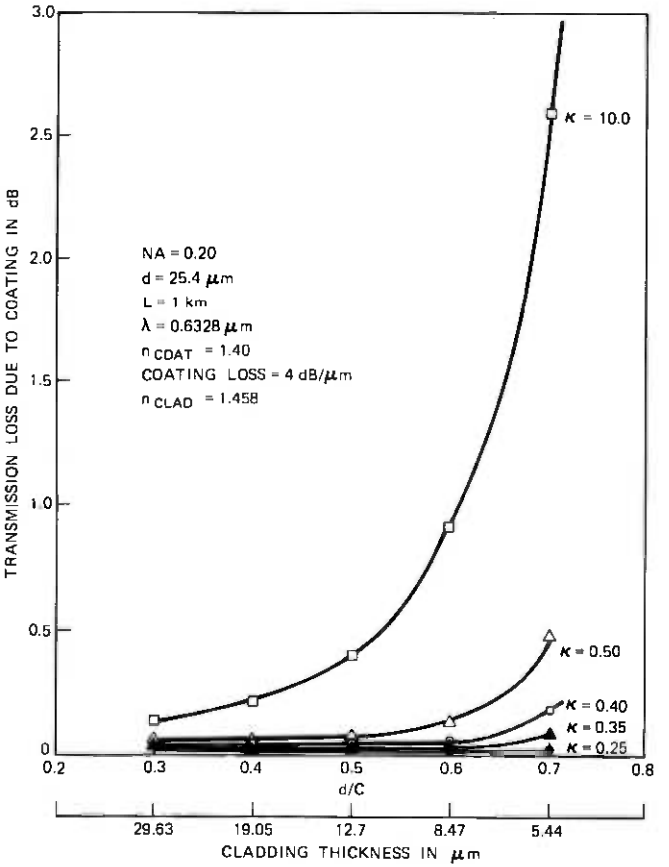


Fig. 5—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.20, d = 25.4 μm.

III. SUMMARY OF RESULTS OF THE COMPUTER STUDY

A computer program was written and the integrals (2) and (3) were evaluated for typical fiber parameters.[6] A number of studies were done to determine how transmission loss due to the coating varies as a function of cladding thickness, wavelength, core diameter, and the real and imaginary part of the coating refractive index. Figures 3 through 5 show, respectively, for fiber numerical apertures of 0.10, 0.15, and 0.20, the relationship between transmission loss due to the lossy coating and

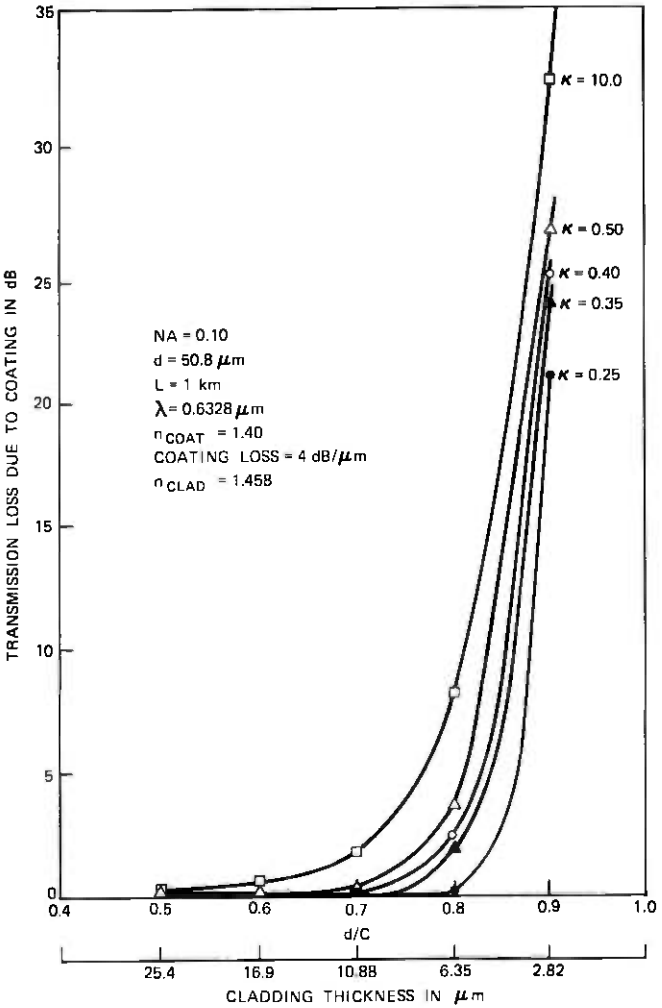


Fig. 6—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.10, d = 50.8 μm.

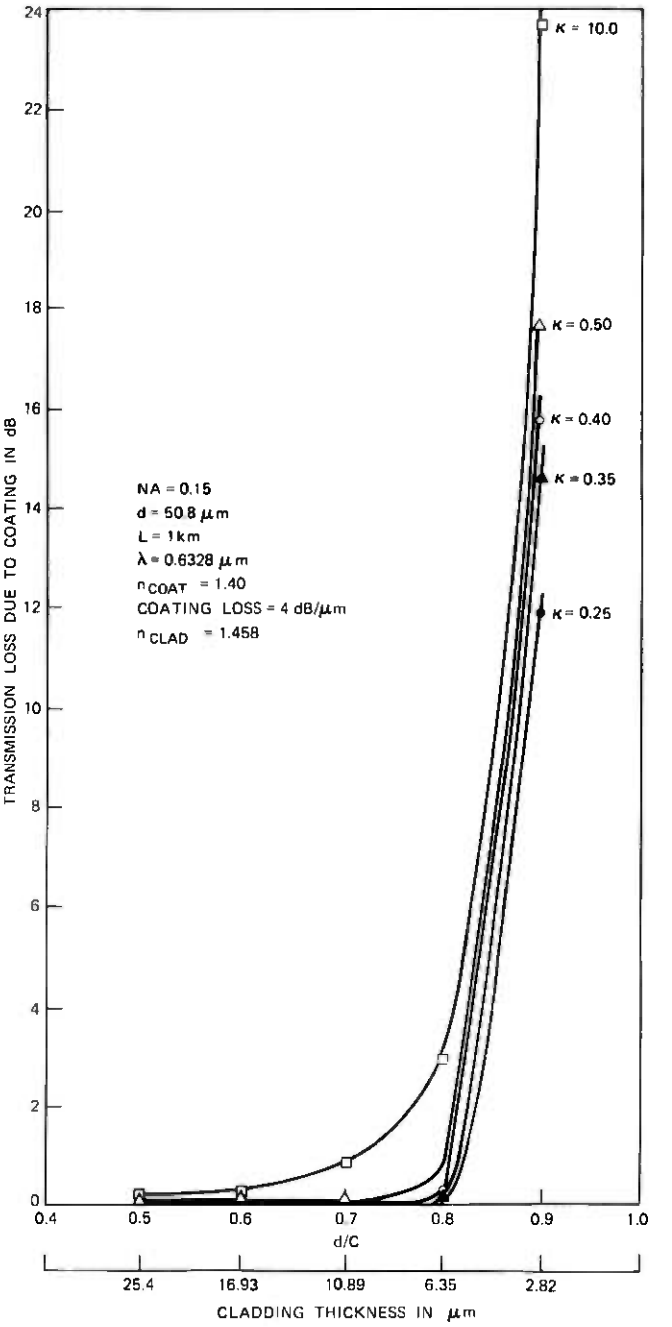


Fig. 7—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.15, d = 50.8 μm.

cladding thickness for a 25.4-μm fiber core diameter. For the same numerical apertures, Figs. 6 through 8 and Figs. 9 through 11 show this relationship for 50.8- and 75.2-μm core diameters.

For practical cladding thickness greater than 15 μm, increasing the cladding thickness will decrease the transmission loss due to the lossy coatings by approximately 0.04 dB/km per micrometer of cladding thickness for the higher-order modes ($\kappa = 10.0$). For the lower-order modes ($\kappa \leqq 0.5$), a cladding thickness of 15 μm should provide sufficient isolation to prevent transmission loss due to the presence of the lossy coating.

Calculations were made to determine the relationships between transmission loss due to a lossy coating and wavelength (λ), core

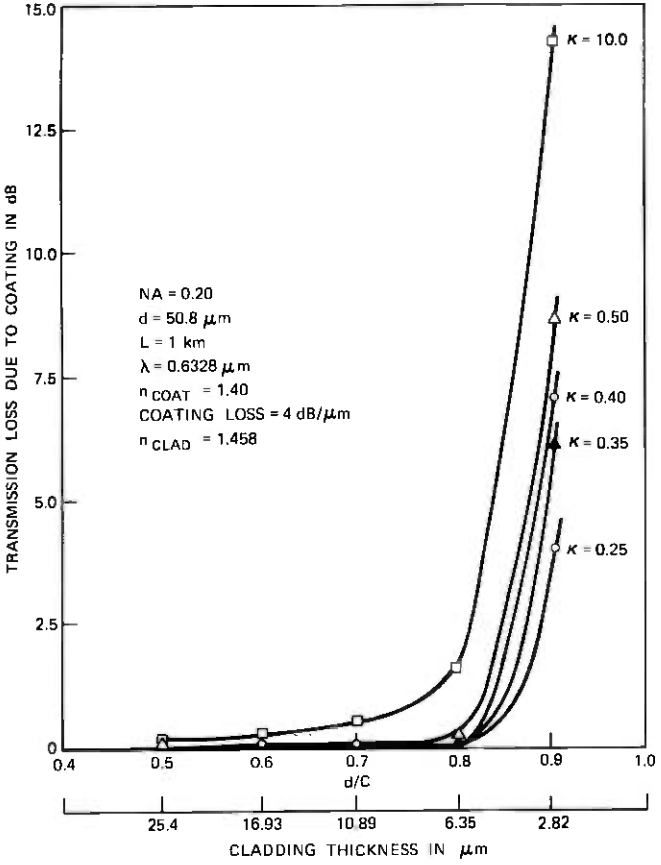


Fig. 8—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.20, d = 50.8 μm.

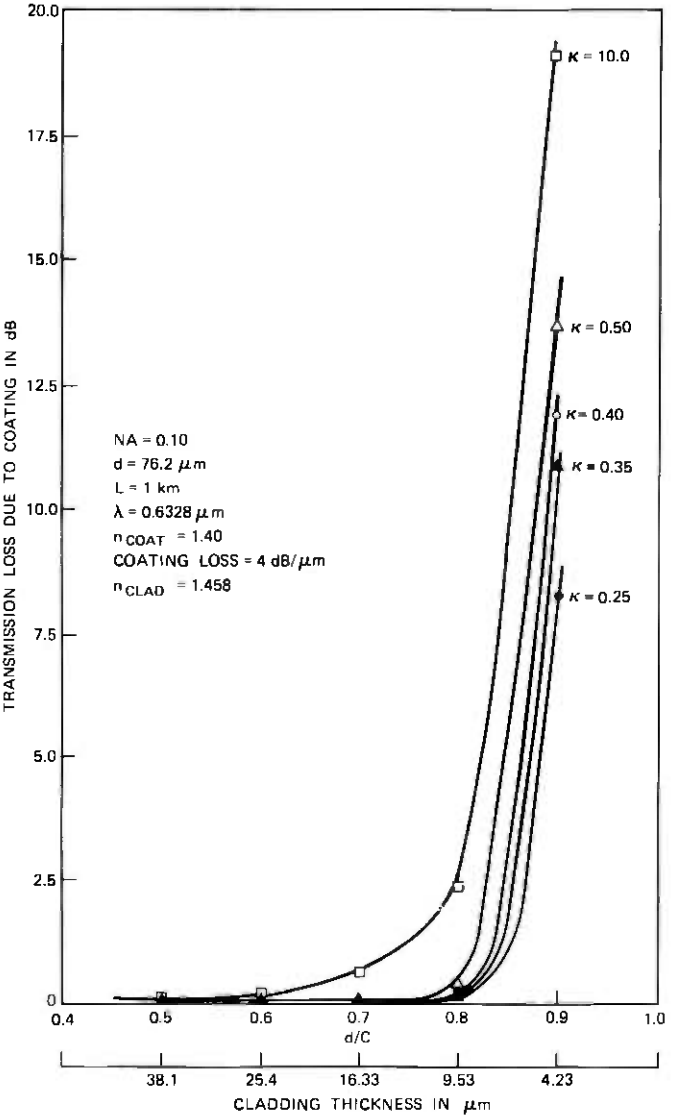


Fig. 9—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.10, d = 76.2 μm.

diameter (d), and the real and imaginary parts of the refractive index of the coating. A thin cladding of 8 μm was chosen in these calculations to easily illustrate the trends due to these parameters. This thin cladding was not intended to be a practical choice for a cladding thickness in an optical fiber. Figure 12 shows the transmission loss due to a

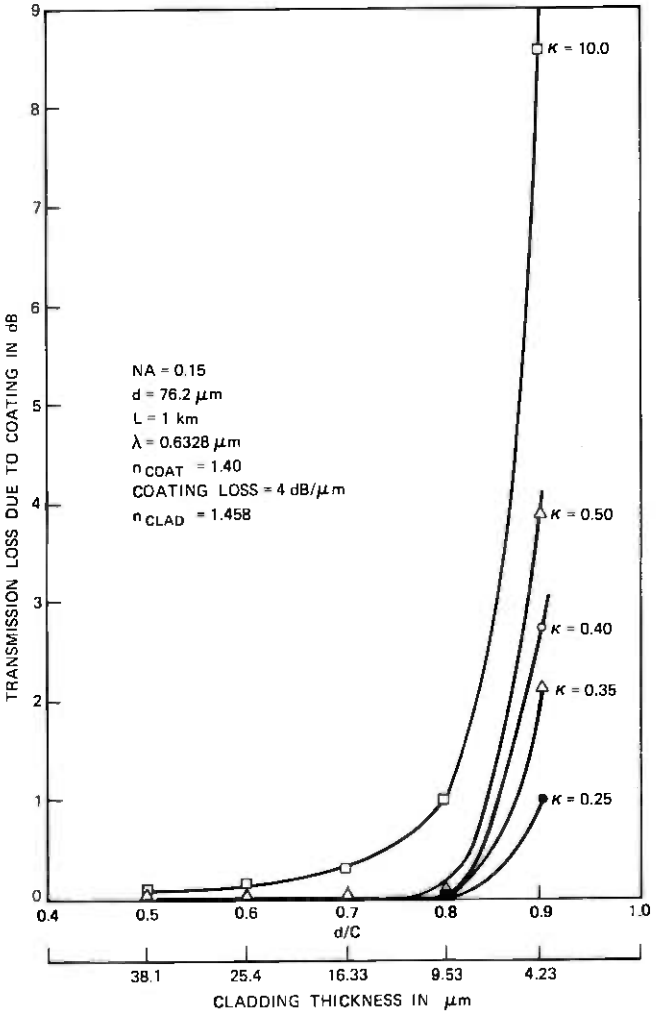


Fig. 10—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.15, d = 76.2 μm.

lossy coating of 4 dB/μm on a fiber whose numerical aperture was 0.15 and core diameter 50.8 μm. As expected, for a fixed cladding thickness, the transmission loss will increase as the wavelength increases. The increase in loss for the parameters chosen was, for the longer wavelengths (>0.8 μm), approximately 1.5 dB/km per micrometer of wavelength. Figure 13 shows the weak dependence of transmission loss on core diameter. Figures 14 and 15 show respectively the

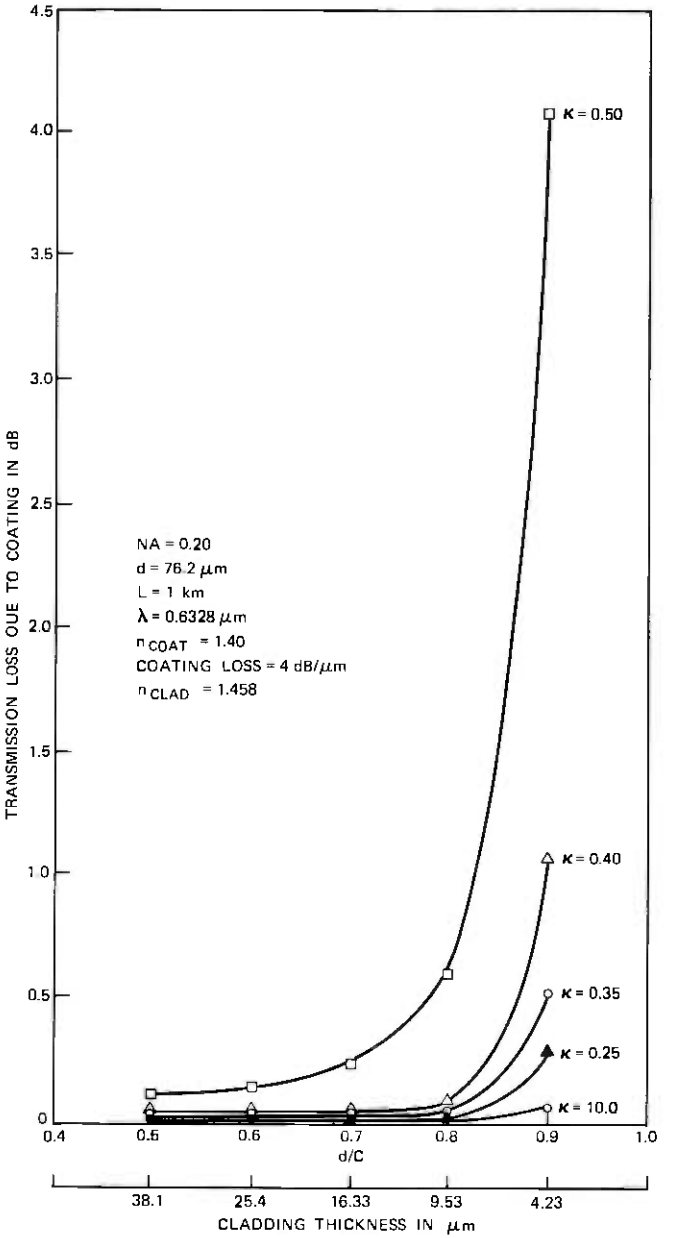


Fig. 11—Transmission loss due to coating vs d/C and vs cladding thickness; NA = 0.20, d = 76.2 μm.

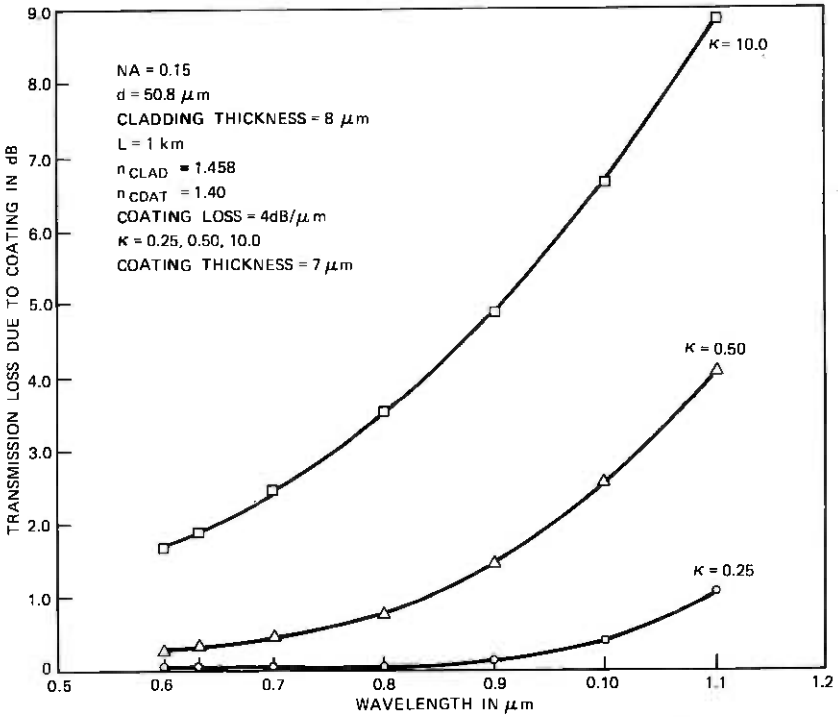


Fig. 12—Transmission loss due to coating vs wavelength.

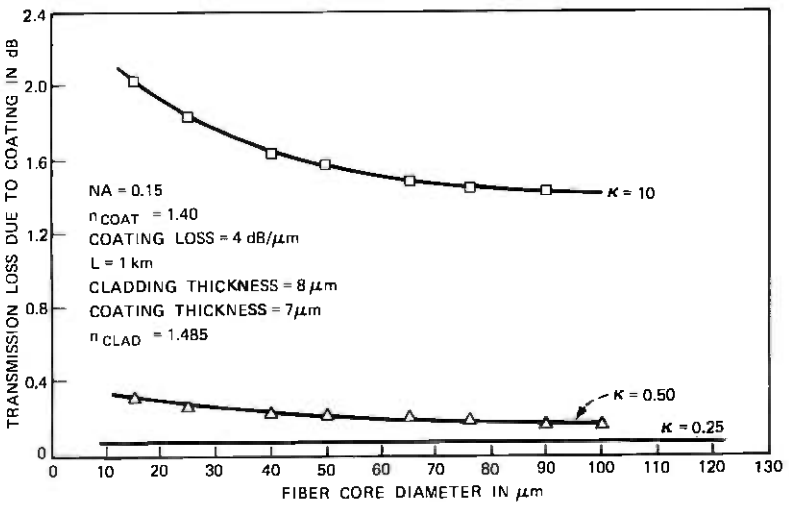


Fig. 13—Transmission loss due to coating vs core diameter.

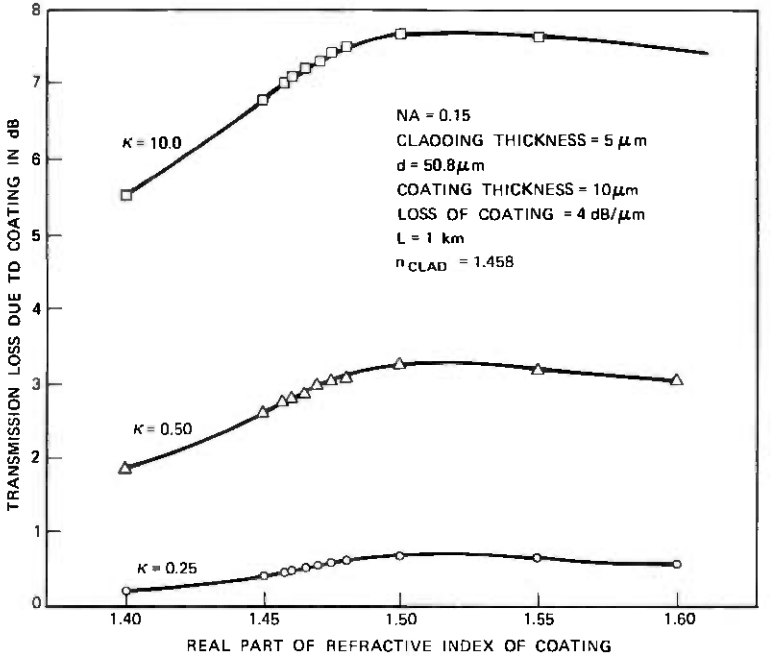


Fig. 14—Transmission loss due to coating vs real part of refractive index of coating.

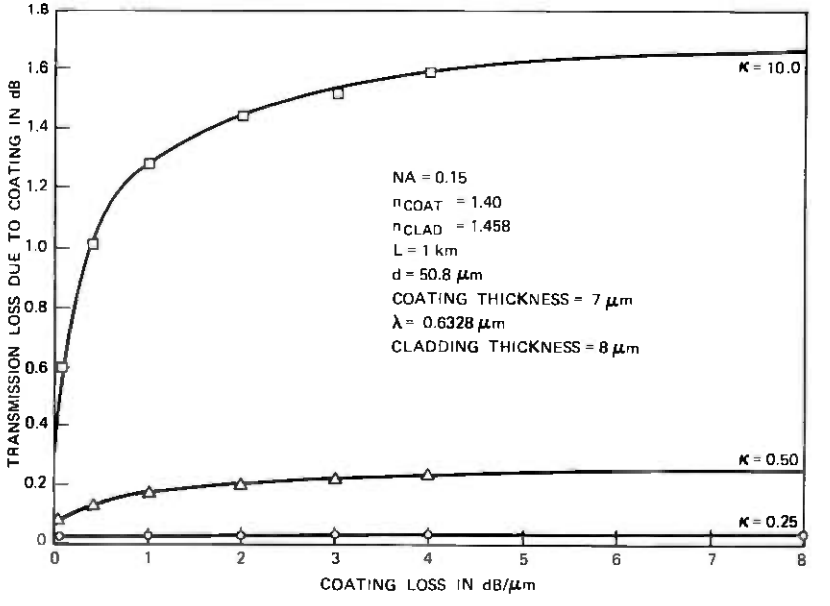


Fig. 15—Transmission loss due to coating vs imaginary part of refractive index of coating.

transmission loss as a function of the real and imaginary parts of the refractive index of the coating. Using Fig. 14, note that for a fiber with a cladding refractive index of 1.458 and coating loss of 4 dB/μm, the transmission loss increases, as the real part of the refractive index of the coating increases, up to a maximum at a value of the refractive index of approximately 1.51. As one increases the refractive index of the coating beyond 1.51, the transmission loss levels off to a constant value.

In Fig. 15, for a fixed real part of the refractive index of the coating equal to 1.40, the transmission loss was observed as a function of the coating loss (the imaginary part of the refractive index of the coating). The transmission loss increases rapidly with coating loss and reaches 80 percent of its final value for a coating loss of approximately 1 dB/μm. Further increase in the loss of the coating does not substantially affect the fiber transmission loss.

IV. CONCLUSIONS

The model described, along with an experimentally determined knowledge of the energy distribution (value of κ), can be used to choose fiber parameters that will prevent transmission loss caused by the lossy coating. Because of the meridional-ray assumption made in the analysis, a conservative estimate of the transmission loss is predicted by the model. The dominant fiber parameter that plays a critical role in preventing the transmission loss is cladding thickness. A cladding thickness of at least 20 μm is necessary to provide adequate isolation. The model also calculates the significant effect on transmission loss of transmitting wavelength, the real and imaginary part of the refractive index of the lossy coating, and the fiber core diameter.

V. ACKNOWLEDGMENT

The authors wish to thank Philip Rich for his assistance in processing the numerical data.

APPENDIX

Calculation of Input Impedance and Reflection Coefficient for a Multilayered Dielectric Medium

In this appendix, the elementary concepts of input impedance and reflection coefficient are developed for a multilayered dielectric medium. Consider the geometry shown in Fig. 16.

Let us suppose that between two semi-infinite media, denoted by 1 and $n + 1$, there are $n - 1$ layers of dielectric material denoted by 2, 3, $\cdots$, n. Let a plane wave be incident on the last layer at an angle of incidence θ_{n+1} and let the plane of incidence be the $X - Z$ plane.

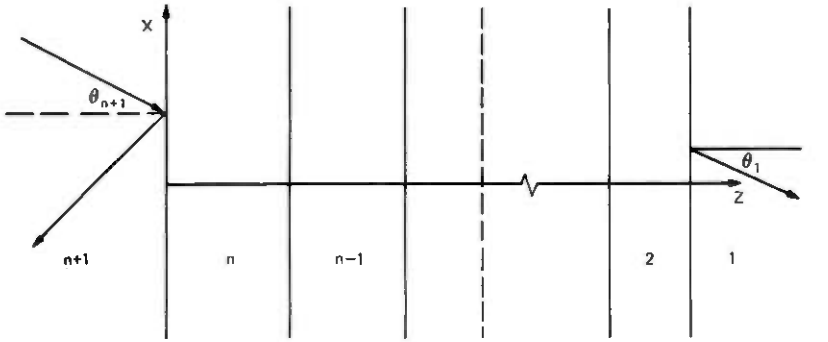


Fig. 16—Geometry used for the calculation of the reflection coefficients in a lossy multilayered dielectric medium.

As a result of multiple reflections at the boundaries of the layers, two waves will exist in each of the media with the exception of medium 1. Our problem is to determine the amplitude of the reflected wave in medium $n + 1$ and, hence, the reflection coefficient.

The following notation is used:

$d_j = z_j - z_{j-1} \equiv$ the thickness of the jth medium,

$k_j = \frac{2\pi}{\lambda} n_j \equiv$ the wave number in jth medium,

$\alpha_j = k_j \cos \theta_j \equiv z$ component of the wave vector in the jth medium,

$\phi_j = \alpha_j d_j =$ the phase change in the jth medium,

$Z_j =$ the self impedance of the jth medium,

$Z_{\text{in}}^j =$ the input impedance looking into the jth medium from the $j + 1$ medium.

The electric and magnetic fields in the jth medium are written as:

$$E_{jy} = A_j \exp\left[-i\alpha_j(z - z_{j-1})\right] + B_j \exp\left[i\alpha_j(z - z_{j-1})\right], \tag{5}$$

$$H_{jx} = \frac{1}{Z_j}\left\{A_j \exp\left[-i\alpha_j(z - z_{j-1})\right] - B_j \exp\left[i\alpha_j(z - z_{j-1})\right]\right\}. \tag{6}$$

The x and t dependency in this case is omitted for the sake of brevity but assumes the general form

$$\exp\left[i(k_{n+1}x \sin \theta_{n+1} - \omega t)\right].$$

A_j and B_j are the amplitudes of the incident and reflected waves in the jth medium ($B_1 = 0$). The amplitude A_{n+1} of the incident wave is assumed to be known. To obtain the reflection coefficient of interest:

$$\Gamma_{n+1} = B_{n+1}/A_{n+1}. \tag{7}$$

One can utilize formula (34) of Ref. 1 to express the input impedance looking into the nth medium from the $n+1$st medium:

$$Z_{\text{in}}^{(n)} = \frac{Z_{\text{in}}^{(n-1)} - iZ_n \tan \phi_n}{Z_n - iZ_{\text{in}}^{(n-1)} \tan \phi_n} \cdot Z_n. \tag{8}$$

The reflection coefficient of the incident wave can then be written in terms of input impedances as follows:

$$\Gamma = \frac{B_{n+1}}{A_{n+1}} = \frac{Z_{\text{in}}^{(n)} - Z_{n+1}}{Z_{\text{in}}^{(n)} + Z_{n+1}}. \tag{9}$$

In eq. (3), the power reflection coefficient R is used. In terms of Γ, R is defined as

$$R = |\Gamma|^2 = \left|\frac{B_{n+1}}{A_{n+1}}\right|^2. \tag{10}$$

For the examples in Section III, a four-layered medium composed of the fiber core, cladding, lossy coating, and surrounding air was used when calculating the reflection coefficient and the transmission loss due to the lossy coating. The refractive index of the lossy coating was defined as

$$N = n_{\text{coating}} - i\delta, \tag{11}$$

where δ is the imaginary or lossy part of the refractive index of the coating.

The relationship between δ and the operationally useful term, coating opacity, is[4]

$$\delta = 29.9\ \lambda\ (\text{opacity}), \tag{12}$$

where

opacity = the transmission loss of the coating in dB/μm
λ = wave length of the transmitted light in micrometers.

The term opacity is introduced here since it is an easily measurable indicator of the loss of the coating and a convenient input variable to the computer program.

REFERENCES

1. A. H. Cherin and E. J. Murphy, "Quasi-Ray Analysis of Crosstalk Between Multimode Optical Fiber," B.S.T.J., *54*, No. 1 (January 1975), pp. 17–45.
2. D. Marcuse, "Bent Optical Waveguide with Lossy Jacket," B.S.T.J., *53*, No. 6 (July–August 1974), pp. 1079–1101.
3. D. Marcuse, "The Coupling of Degenerate Modes in Two Parallel Dielectric Waveguides," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1791–1816.
4. L. M. Brekhovskikh, *Waves in Layered Media*, New York: Academic Press, 1960, pp. 56.
5. M. Born and W. Wolf, *Principles of Optics*, London: Pergamon Press, 1970.
6. B. Carnahan, H. A. Luther, and J. O. Wiles, *Applied Numerical Methods*, New York: John Wiley and Sons, 1969.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U. S. A.

A Fiber-Optic-Cable Connector

By C. M. MILLER

(Manuscript received May 30, 1975)

A technique has been developed that is potentially suitable for field-splicing an optical cable containing linear arrays of optical fibers. Linear arrays of fibers (which may reside in fiber ribbons) are placed between spacers that are grooved top and bottom to form stacked, rectangular arrays. This operation can be done without microscopes or micromanipulators. After potting, the ends of the two stacked arrays are polished to form cable terminations that are brought together in a butt joint splice. A 12 × 12 array using this technique exhibited a mean loss of 0.42 dB for 138 splices with 70 percent of the losses less than 0.5 dB. Subsequent single ribbon-to-ribbon splices had average losses less than 0.2 dB. Launching conditions can be duplicated and splice losses are repeatably low for reassembled splices; this presumably is due to polished fiber ends and accurate alignment. Experience gained thus far indicates that this mass splicing method will probably produce large array splices with a maximum loss of 0.5 dB.

I. INTRODUCTION

It is believed that splicing groups of optical fibers in the field will be necessary in fiber communication systems. Several investigators have successfully spliced individual fibers with various techniques.[1–4] Others have addressed themselves to splicing linear arrays of fibers.[5]

While some aspects of linear array or fiber-ribbon splicing appear applicable to cable splicing in the field, the operations must be performed for each linear array in a fiber-optic cable. Good ends must be obtained for each fiber and, although techniques for accomplishing this are evolving,[6] still these represent additional operations required for each fiber or fiber group during cable splicing. Another potential problem in applying individual or small-group splicing to a fiber-optic cable is in reassembly of the spliced ribbons into a compact spliced cable. The individual connectors would have to be very thin, and the splices would have to be the same length to effect compact reassembly.

The potential problems mentioned above stimulated investigation of another approach, cable splicing, by which we mean splicing all fibers of two cables by joining connector halves (terminations) formed on each cable end.

II. FIBER-OPTIC-CABLE SPLICING

As presently conceived, cable splicing involves the following operations:

(*i*) Aligning all fibers of one end of a fiber-optic cable into a uniform matrix.
(*ii*) Potting the structure to retain the geometry.
(*iii*) Grinding and polishing the ends of the potted array.
(*iv*) Joining two cable ends prepared by the previous three operations.

Each of these processes is covered in more detail in the following sections. An alternative approach is also presented which uses fiber ends prepared by controlled breaking.[6]

III. ALIGNMENT OF FIBERS

Several techniques have been attempted to align fibers in a linear array. Threading fibers through holes as opposed to laying fibers in grooves is in general a more difficult and less accurate method of fiber alignment. The "grooved" concept was adopted for this cable con-

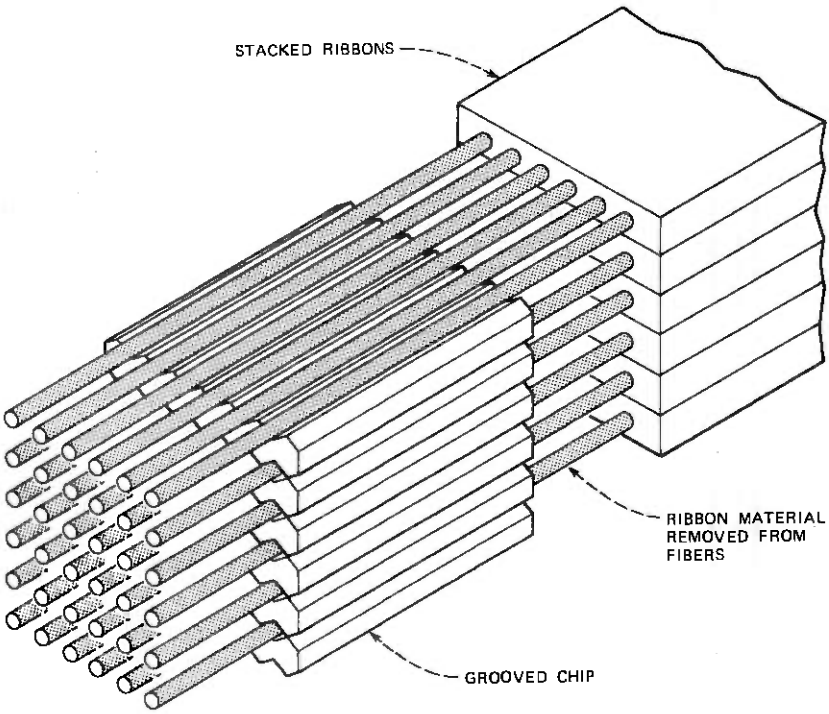


Fig. 1—Stacking fiber ribbons.

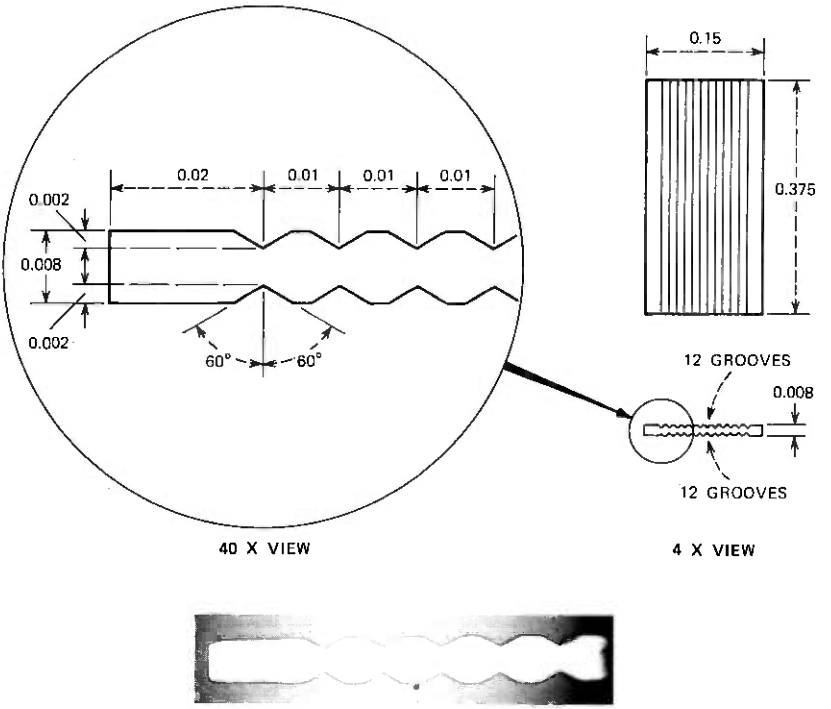


Fig. 2—Grooved chip design.

nector. Since a compact splice was desired, a thin chip design was selected which was grooved on both sides.

Figure 1 is a magnified view of how grooved chips are used to obtain fiber alignment. Ribbon ends are prepared by removing the supporting material as shown in Fig. 1. The process consists of interleaving chips and layers of fibers until all linear arrays have been stacked.

Precision grooved chips provide the primary alignment mechanism for assembly of fibers into a uniform rectangular array. These chips were produced by the Bulova Watch Company. Figure 2 is a sketch of the specifications for the chip and a photograph of the cross section of the chip. The machine work performed to make these chips was excellent. Since the chips are grooved on both sides, the top and bottom chip of the stacked array have unoccupied grooves which can be used as references for alignment during polishing and subsequently to align the two rectangular arrays in forming the butt splice.

IV. POTTING THE ARRAY

A stacked array, held by a vise, can be potted by allowing epoxy to seep through the array. Tests with close-packed arrays of fibers

showed that epoxy would seep at least $\frac{1}{2}$ inch down the length of the array before curing stops all flow. This sets the maximum length of the grooved chip; however, based on our experience a $\frac{3}{8}$-inch length appeared adequate. Approximately 15 minutes is required for epoxy to seep through a $\frac{3}{8}$-inch stacked array. Faster curing of the epoxy occurs at elevated temperatures; however, at room temperature several hours are required for the epoxy to completely cure.

A vise is used to hold the stacked array while epoxy is applied. Chips with ridges that mate with the unoccupied grooves of the stacked array will be referred to as negative chips. These negative chips are used to align the top and bottom chips of the stacked array while in the vise. Figure 3 shows the vise and the negative chips attached to the vise faces.

V. POLISHING THE ARRAY

To obtain low loss in a splice, the fiber end must be made flat and perpendicular to the fiber axis. This end preparation is usually accomplished by either controlled breaking,[6] sawing, or polishing. Sawing the epoxied arrays may yield suitable ends for splicing. Controlled breaking is not applicable to the epoxied array previously described, although polishing methods can be easily applied. Since only one polishing operation is required for each cable splice, regardless of the number of fibers or ribbons, this technique is probably the least time-consuming and the cheapest. A holding fixture for supporting and gradually advancing the epoxied array during the polishing sequence is shown in Fig. 4 with two connectors in place after polishing. The indicator at the rear of the fixture moves the inner core of the fixture (and the sample) relative to the outer cylinder. This fixture used with a grit sequence of 220×, 800×, and 0.3 μ has produced high-quality ends in approximately 15 minutes. Figure 5 shows a connector after polishing.

VI. FINAL ALIGNMENT

Several final alignment methods have been used to meet two different needs in this area. First, an alignment method has been developed for use in the laboratory in making splices that can be measured and then disassembled. Plexiglass and steel fixtures have been used for this purpose in conjunction with a grooved chip negative. These negative chips are pressed against the top and bottom of the two connectors and then placed in the final alignment fixture which further aligns the two connectors and holds them in place. This arrangement is shown in the top part of Fig. 6.

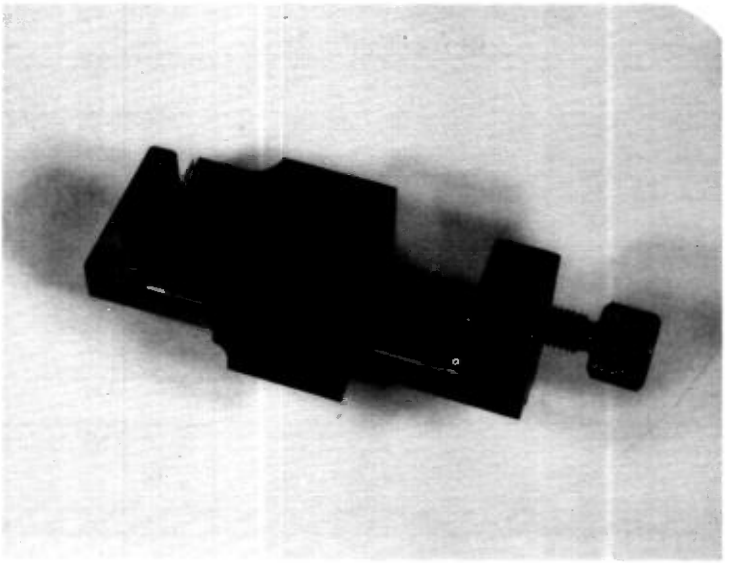


Fig. 3a—Vise used in splice fabrication.

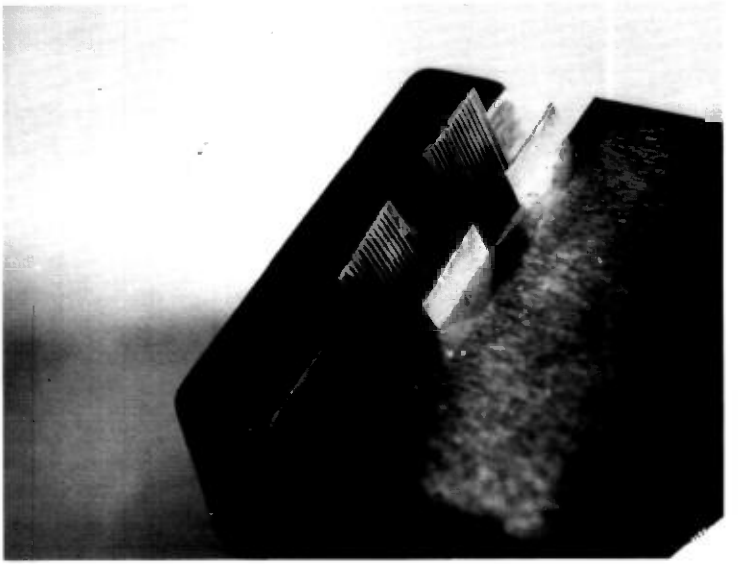


Fig. 3b—Detail showing negative chips.

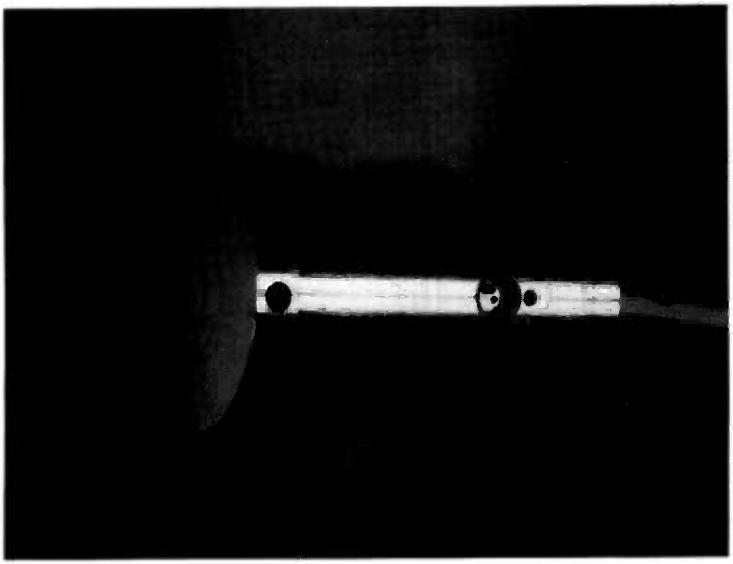


Fig. 4—Polishing fixture.

A second alignment method is used to obtain a more compact, permanent splice. For this case, the negative chips are epoxied directly onto the grooved chip connectors, again while being held and aligned

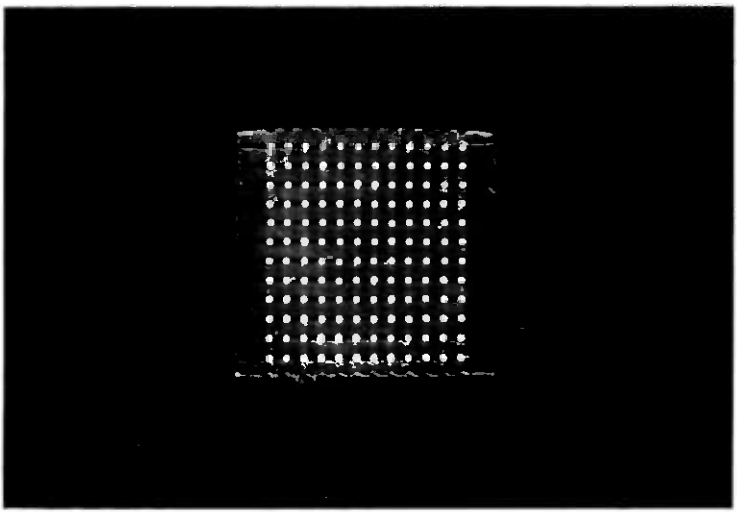


Fig. 5—End view of 12 × 12 connector after potting and polishing.

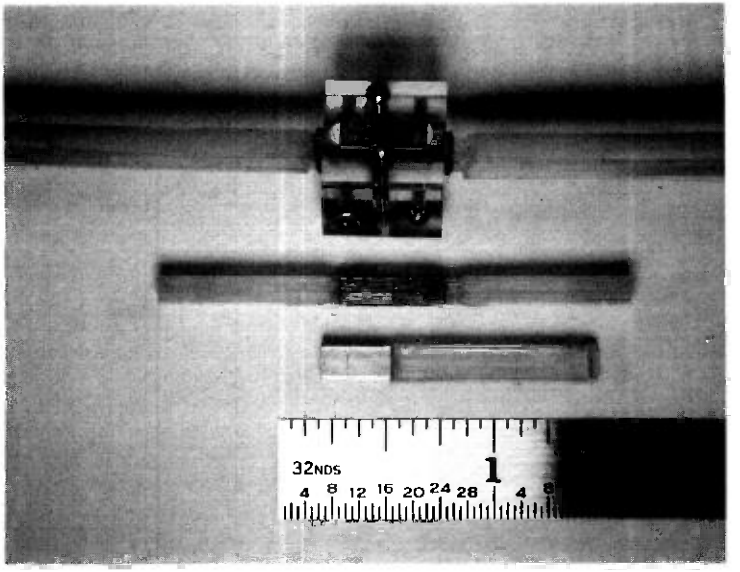


Fig. 6—Top: final alignment fixture; middle: epoxied splice; bottom: single connector.

in a suitable fixture. Figure 7 is a photograph of such a splice with index-matching epoxy used to complete splice fabrication.

VII. SPLICE-LOSS MEASUREMENTS

Using the methods and fixtures described, a 12 × 12 array splice was fabricated and measured. Six fibers were damaged or broken during the fabrication of the four connectors (connectors were placed on the input and output of the ribbon stack in addition to the splice). As shown in Fig. 8, the mean loss was 0.42 dB with a maximum loss of 1.3 dB for 138 good splices. Seventy percent of the losses were less than 0.5 dB.

Single ribbon splices have been fabricated and measured and yield lower losses than array splices. For a typical ribbon splice with index-matching, the maximum loss was below 0.5 dB and the average loss below 0.2 dB. Crosstalk coupling between adjacent splices in a typical ribbon splice is less than −65 dB.

VIII. PERMANENT CABLE SPLICE

The fiber-optic-cable connector described thus far consists of prepared cable terminations which are joined to complete the cable splice. If a final alignment fixture is used to hold the cable terminations in place, as opposed to epoxy, the splice could be disassembled. An

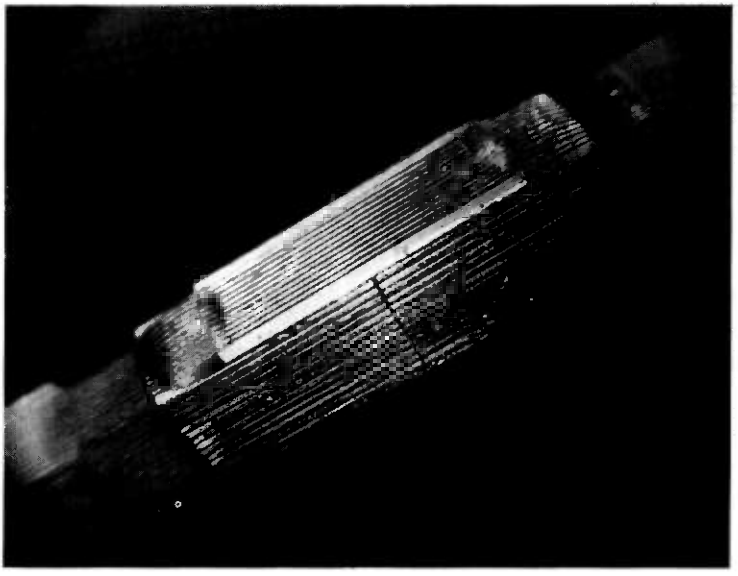


Fig. 7—Splice assembly.

alternate approach[7] can be used which, although still based on the grooved-chip concept, uses a different operational sequence and no polishing and produces a splice which cannot be disassembled. This approach is illustrated in Fig. 9. Here, the fiber ends will probably be prepared by one of the methods being developed by D. Gloge[6] et al. The stacked array is fabricated in much the same way except that two

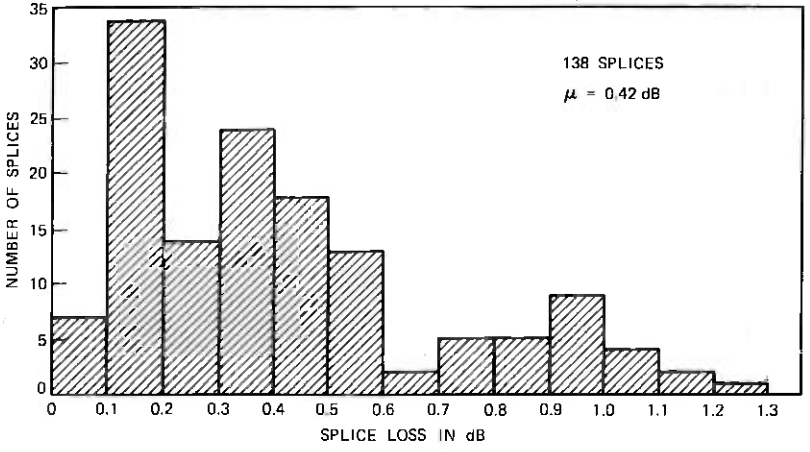


Fig. 8—Histogram of splice losses.

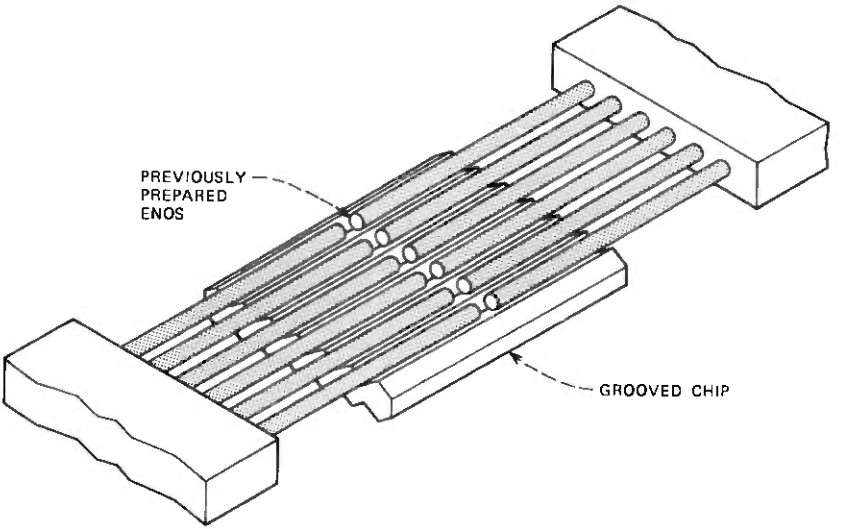


Fig. 9—Single layer of a permanent cable splice.

sets of previously prepared ends are placed in each layer. The array is then epoxied to form a permanent splice. This technique will be more difficult to assemble, and end preparation will probably be done on a ribbon-by-ribbon basis using controlled breaking. However, this permanent assembly technique should produce compact splices with losses approaching those of ribbon splices.

Some difficulties associated with ribbon splicing mentioned earlier are present in this approach. These include multiple operations required for end preparations and nearly exact ribbon length requirements. The resulting splice will, however, be small and strong, and should easily meet the goal of a large-array splice with maximum loss below 0.5 dB.

REFERENCES

1. D. L. Bisbee, "Optical Fiber Joining Technique," B.S.T.J., *50*, No. 10 (December 1971), pp. 3153–3158.
2. C. G. Someda, "Simple, Low-Loss Joints Between Single-Mode Optical Fibers," B.S.T.J., *52*, No. 4 (April 1973), pp. 583–596.
3. R. B. Dyott, J. R. Stern, and J. H. Stewart, "Fusion Junction for Glass-Fiber Waveguides," Electronics Letters, *5*, No. 11 (June 1, 1972), pp. 290–292.
4. C. M. Miller, "Loose Tube Splices for Optical Fibers," B.S.T.J., *54*, No. 7 (September 1975), pp. 1215–1225.
5. A. H. Cherin and P. J. Rich, "A Multi-Groove Embossed Plastic Splice Connector for Joining Groups of Optical Fibers," Appl. Opt., *14*, No. 12 (December 1975).
6. D. Gloge, P. W. Smith, D. L. Bisbee, and E. L. Chinnock, "Optical Fiber End Preparation for Low Loss Splices," B.S.T.J., *52*, No. 9 (November 1973), pp. 1579–1588.
7. M. I. Schwartz, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

Step-Size Transmitting Differential Coders for Mobile Telephony

By N. S. JAYANT

(Manuscript received February 7, 1975)

In using digital speech for mobile radio, we encounter the problem of severe bit-error bursts. Error clustering occurs because the bit duration is typically much smaller than that of a signal "fade," and average bit-error probabilities greater than 1 percent are not uncommon. For speech communication over such channels, this paper proposes variable step-size differential coders based on explicit (and error-protected) transmission of quantizer step size. Specifically, we discuss delta and DPCM coders to be referred to as DM-AQF and DPCM-AQF, where AQF stands for adaptive quantization with forward estimation (and transmission) of step size. (Backward estimation, based on quantized-signal history, has the nice feature that the step-size information does not have to be explicitly transmitted. Furthermore, obtaining this information does not entail any encoding delay. However, due to the dependence of step size on reconstructed signal history, backward estimation is often less reliable in the presence of bit errors than a scheme based on AQF.) The studies reported in this paper cover the problem of step-size determination in AQF, the design of time-invariant first-order predictors for DPCM-AQF, and the performances of AQF encoders with and without burst-error-protecting ploys such as redundant time-diversity coding and bit scrambling. Judging from SNR figures and informal listening tests, interesting results are obtained with the following 48-kb/s coders: three-bit DPCM-AQF with redundant error protection, and DM-AQF using bit scrambling.

I. INTRODUCTION

Recent developments in speech digitization[1] have prompted an examination of digital coding as a possibility for mobile radio telephony that conventionally employs analog techniques for speech transmission. Conceivably, much of the signaling supervision and "bookkeeping" in a mobile radio link can be digital; in this case, if the speech were handled digitally as well, it would be simple to interleave the voice bits with the control bits for transmission. Digital coding also

offers the possibilities of inexpensive coder-decoder implementation, straightforward speech encryption (by bit scrambling), and efficient signal regeneration. Perhaps the greatest incentive for the use of digital speech, however, is the thought that a properly designed digital code may be more resistant than analog systems to the multipath fading that characterizes mobile radio.

Figure 1 shows the envelope of a Rayleigh-fading signal that is typical in mobile telephony.[2] An important parameter is the fading rate, which is approximately the ratio of vehicle speed V to the carrier wavelength λ. For the example in Fig. 1, this ratio is about 15 Hz. Note also that the 5 m represent a total travel time of about 1 s at the indicated vehicle speed, and that the fading is slow or correlated in the sense that a given fade (signal strength below a specified threshold) can last for several tens of milliseconds (which will represent several hundred speech bits for the codes of this paper). The probability of a fade can be decreased by an order of magnitude by the use of diversity reception (two-branch, equal-gain or switched diversity, for example). But when a fade does occur, the signal is susceptible to noise capture as well as to co-channel interference. The end effects, with conventional analog transmissions, are impulsive "pops" and "crackles" in the

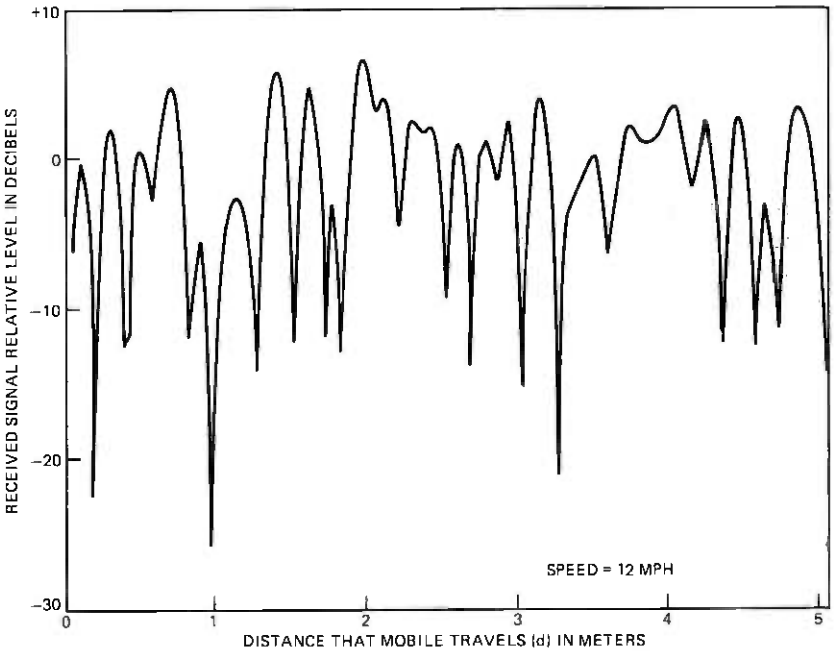


Fig. 1—Envelope of a Rayleigh fading signal (V/λ = 15.4 Hz).

received speech. Without explicit signal companding, these effects can be severe under "idle-channel" conditions, when no speech is present on the channel of interest. By contrast, if "adaptive" digital modulators are used to transmit speech, it is expected that the variable step-size mechanism in these modulators would inherently attenuate the impulsive interference signals (manifested as clustered bit errors in the digital system) during the silences of the incoming speech. This interference-squelching property is, in fact, expected to carry over to the "active-channel" condition with an ideally designed adaptive code (that exploits the statistics of signal fading or, equivalently, the time statistics of the bit-error bursts). Even if such an ideal design should be impractical, it is clear that digital coding can straightforwardly employ efficient burst-error-protection ploys such as time-diversity coding and bit-scrambling. However, these refinements (as well as the notion of forward estimation of step size) can involve significant amounts of encoding delay.

In our search for a suitable digital coder, we have used the following characteristics as guidelines. The speech bandwidth should be representative of standard telephone quality (200 to 3200 Hz); average bit-error rates higher than 1 percent are possible at times; and, finally, the overall transmission rate should not exceed a nominal 48 kb/s. When we refer to a "nominal 48-kb/s rate," we mean that additional channel capacity (in the order of 2 to 5 kb/s) may be needed for the transmission of step-size information.

A basic contention of this paper is that the "optimum" step size for a speech quantizer changes slowly enough with time for the step-size information to be transmitted reliably in a special error-protected format over a typical mobile radio channel. Thus, although the main stream of speech-carrying bits is still subject to errors, the provision of a relatively error-free step size will improve the received speech quality to a point that makes explicit step-size transmission worthwhile. We show that step-size transmitting coders are of interest for bursty as well as independent error patterns, and we include a comparison with a popular error-resistant syllabic-companded quantizer that recovers step size from the bit stream. Following Noll,[3] step-size transmitting adaptive coders will be labelled AQF (adaptive quantization with forward estimation and transmission of step size), in contradistinction to AQB (adaptive quantization with backward estimation).

The coders of this paper are differential. We discuss both DPCM (differential PCM) and DM (delta modulation). It appears from experience[1] that conventional time-invariant log-PCM quantization does not meet the error-performance requirements of mobile telephony. However, the possibility of a well-designed adaptive PCM[1,3,4] definitely

exists. A promising candidate is the technique of nearly instantaneous companding (NIC).[4]

Although our studies have included informal perceptual assessments, most performance results in this paper are objective signal-to-noise ratios termed SNRT and SNRR. These reflect, respectively, the speech quality at the output of the local and remote decoders (T and R stand for "transmitter" and "receiver"). Formal definitions appear in Fig. 2. As we shall note at appropriate points in the paper, an SNRT-maximizing encoder does not, in general, maximize SNRR, and vice versa.

Our discussions refer to computer simulations that employed band-limited (200 to 3200 Hz) speech utterances (2 s or, sometimes, longer in duration) and bit-error patterns obtained from fading simulators.[5] We believe that the main conclusions of this paper should hold for broad classes of speech and error patterns encountered in a mobile radio environment. However, our numerical results are often reflective of the specific data used in our computer simulation. To demonstrate real-world variabilities of these numerical results, we have employed variable speech data, whenever appropriate.

Section II of this paper illustrates the time characteristics of the simulated burst error channel. Section III discusses the design of a DM-AQF coder. The section also demonstrates that simple bit-protecting codes are not particularly beneficial with DM-AQF (except for the transmission of step-size information). Bit scrambling, on the other hand, provides a definite advantage. Suitable sampling rates for DM-AQF are shown to be in the order of 30 to 40 kHz. Finally, a performance comparison is made between DM-AQF and a representative DM-AQB code. Section IV describes the design of a DPCM-AQF coder, and demonstrates

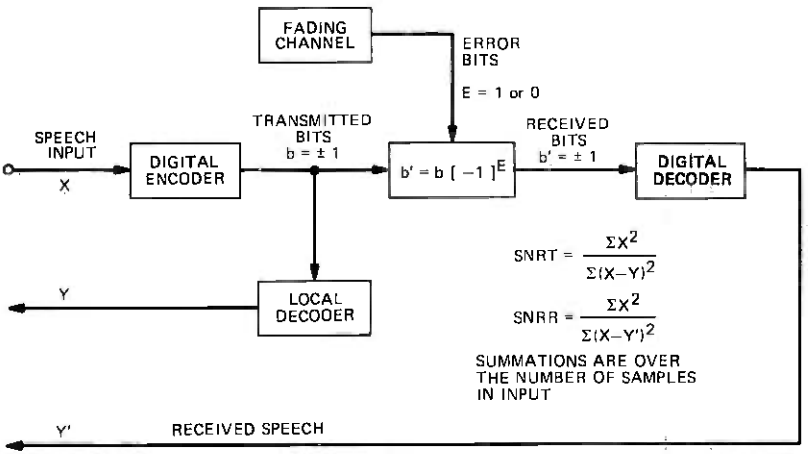


Fig. 2—Block diagram of codec simulation.

the utility of time-diversity coding for bit protection. It also indicates that a three-bit coder operating redundantly at a nominal 48 kb/s (and an 8-kHz sampling rate) is a better choice that a four-bit coder operating (also redundantly, but with less bit protection) at the same information rate. Section V provides a comparison of DPCM-AQF and DM-AQF.

II. BIT-ERROR PATTERNS

2.1 Burst errors

Two simulated-error sequences were used in this study, representing average error probabilities of 0.025 and 0.055. These numbers represent channel qualities believed to be typically "much worse than average."[5] The durations of the error sequences were long enough to simulate the transmission of all but the longest of the speech utterances being encoded. For this utterance (which was 9 s long), the bit-error sequences were used repeatedly to cover the total speech duration. Simulated bit rates ranged from 24 to 48 kb/s.

Figure 3 displays typical distribution functions for error-burst duration D and the error-free interval I. The numbers refer to a subsegment of the 0.025 error rate sequence. The error rate is denoted by $P(EB)$,

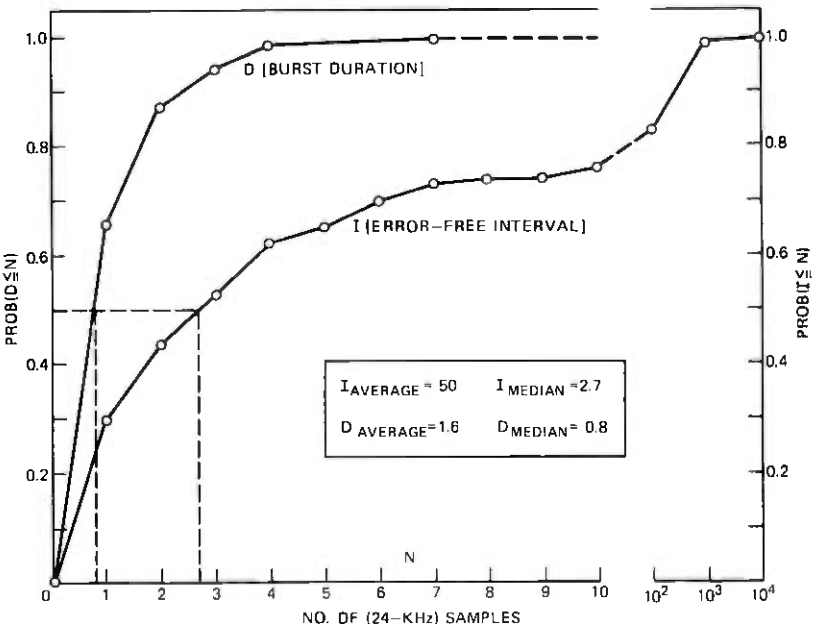


Fig. 3—Time statistics of error bursts [$P(E) = 0.025$, $V/\lambda = 36.2$ Hz].

where B refers to burst errors. An error burst in Fig. 3 is defined to have a local error probability of 1. In other words, an error burst of length D_0 implies that D_0 contiguous signal bits are in error. An isolated error, for example, is an error burst of length $D_0 = 1$. The upper rows of Table I lists the average and median values of burst duration D and error-free interval I for the subsequence examined. The ratio of (D_{average}) to $(D_{\text{average}} + I_{\text{average}})$ gives the average bit-error probability for the subsequence $(1.6/51.6 = 0.03)$. The ratio of (D_{median}) to $(D_{\text{median}} + I_{\text{median}})$, by contrast, is as high as 0.23. Of particular interest is the fact that $I_{\text{average}} \gg I_{\text{median}}$. This signifies the presence of some error-free intervals that are extremely long, together with a preponderance of intervals that are (unfortunately) very short (in fact, not much longer than D_{average}). The clustered nature of the errors is somewhat more apparent by comparison with average and median statistics that apply to an appropriate random error channel: that is, a channel where errors occur independently at every sample, but with an average error probability that is the same as that of the burst-error channel. The lower rows of Table I shows those statistics, as calculated for a random error channel whose bit-error probability is 0.03. Note that the value of D_{average} is much higher (for the same average error probability) in the case of the bursty channel, as expected.

Burst-error patterns, including that of Table I, were obtained from a fading simulator.[5] The main components of the simulation were a pseudorandom binary input, an FM transmitter-receiver, a Rayleigh fader that took into account desired ratios of vehicle speed to carrier wavelength, a noise generator, a pseudorandomly modulated carrier to approximate the effect of co-channel interference, and the option of switched-diversity reception. The numbers for the burst errors in Table I represent the impairment for a 24-kb/s signal-bit sequence (the bit duration determines the number of bits affected by a fade) when the mobile radio link is characterized by two-branch diversity reception under the following (worse than average) conditions:

$$\begin{aligned} \text{Signal-to-interference ratio} &= 9 \text{ dB} \\ \text{Signal-to-noise ratio} &= \infty \\ \frac{\text{vehicle speed}}{\text{carrier wavelength}} = \frac{V}{\lambda} = \frac{29 \text{ mi/h}}{0.353 \text{ m}} &= 36.2 \text{ Hz.} \end{aligned} \tag{1}$$

Table I — Average and median values of D and I $[P(E) = 0.03]$

Errors	I_{average}	D_{average}	I_{median}	D_{median}
Bursty	50.0	1.60	2.7	0.8
Random	32.0	1.03	23.0	0.2

2.2 Scrambled errors

Scrambled errors are of interest in sections of this paper that assume the scrambling of signal bits for error protection. The idea of bit scrambling is to expose adjacent coder bits to channel conditions that tend to be statistically independent. If the scrambling is pseudo-random, the receiver can put the received bits in proper sequence by an inverse unscrambling operation. To avoid the two operations of scrambling and unscrambling, the situation was simulated in our experiment by scrambling the known bit-error pattern and leaving the signal bits in their original sequence.

Error sequences consisted of binary entries (error bits) E that were either 0 or 1, and each entry of 1 represented a bit error in the decoding of a corresponding signal bit. The scrambling was accomplished as follows. The error-bit sequence E was handled in blocks that were M bits long, and each bit got a new position, given by a pseudo-random number (of bit intervals), as was derived from the current state of a maximal-length shift register with $\log_2 M$ stages.[6] The value of M was set at 1024, and the effect of scrambling is illustrated in Fig. 4, which is a snapshot of part of the 0.025 error-rate data. The three sections in the figure represent (contiguous) error sequences that are 1024 bits long (512 per row, two rows per block). In each of the six

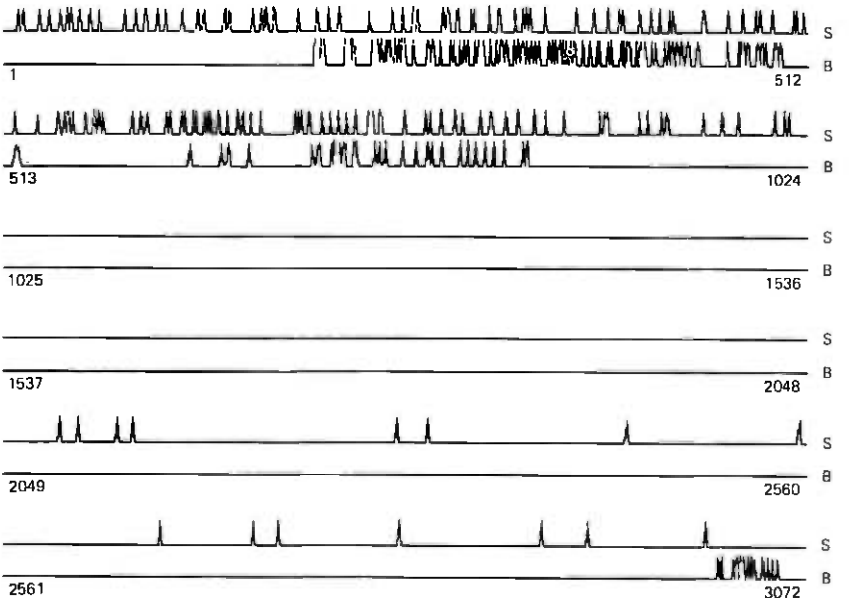


Fig. 4—Illustration of error scrambling [$P(E) = 0.025$, $V/\lambda = 36.2$ Hz].

rows, the lower sequence contains the burst errors (letter B), and the upper sequence has the scrambled errors (letter S). It is clear that a block length of $M = 1024$ is insufficient for true error randomization. Speech recordings have indicated, on the other hand, that values of M as small as 64 are sufficient to achieve useful speech encryption with signal-bit scrambling, assuming the 24-kb/s bit rate mentioned earlier.

2.3 Transmission rate and average error probability P(E)

Our simulations involved signal transmission rates of 24, 32, 40, and 48 kb/s. It is reasonable, under assumptions of constant baud rate (number of channel symbols/second), to expect higher bit-rate transmissions to be subject to correspondingly higher error rates. For example, if 24 and 48 kb/s represent two-phase and four-phase modulations of channel symbols, respectively, at a fixed 24-kHz symbol rate, the average error probability in the 48-kb/s system is expected to be typically two times* as large as that in the 24-kb/s scheme.[5] In the light of this, when we compare similar systems operating at significantly different bit rates in this paper (for example, 24 versus 48 kb/s) we assume average bit-error probabilities that are appropriately different (for example, 0.025 for 24-kb/s transmissions and 0.055 for 48-kb/s transmissions). Burst errors and scrambled errors are indicated by the notations EB and ES.

III. DM-AQF

Figure 5 illustrates the principles of variable step size delta modulation with a forward control of step size. The buffer shown in the encoder stores N input samples (typically, in linear PCM format) that are used to calculate the best step size Δ for the (future) delta modulation of the stored input block. *The step size Δ is recomputed exactly once, and explicitly transmitted to the receiver, for every block of N samples.* The rest of Fig. 5 merely represents a conventional linear delta modulator-demodulator pair.[1] The predictor is assumed to be time-invariant, and of first order. The equations describing the delta modulations are formally summarized below.

$$\begin{aligned} b_r &= \operatorname{sgn}\,(X_r - h_1 \cdot Z_{r-1}). \\ Z_r &= h_1 \cdot Z_{r-1} + \Delta \cdot b_r. \\ Z'_r &= h_1 \cdot Z'_{r-1} + \Delta \cdot b'_r. \end{aligned} \tag{2}$$

The time indices r and $r - 1$ are not shown in the figure; however, the

* Strictly speaking, this number is a function of the carrier-to-noise ratio and the modem that is employed. For example, the number can exceed two (for a typical carrier-to-noise ratio) if FSK is used as the modulation system instead of PSK (for transmitting the speech bits over the analog channel).

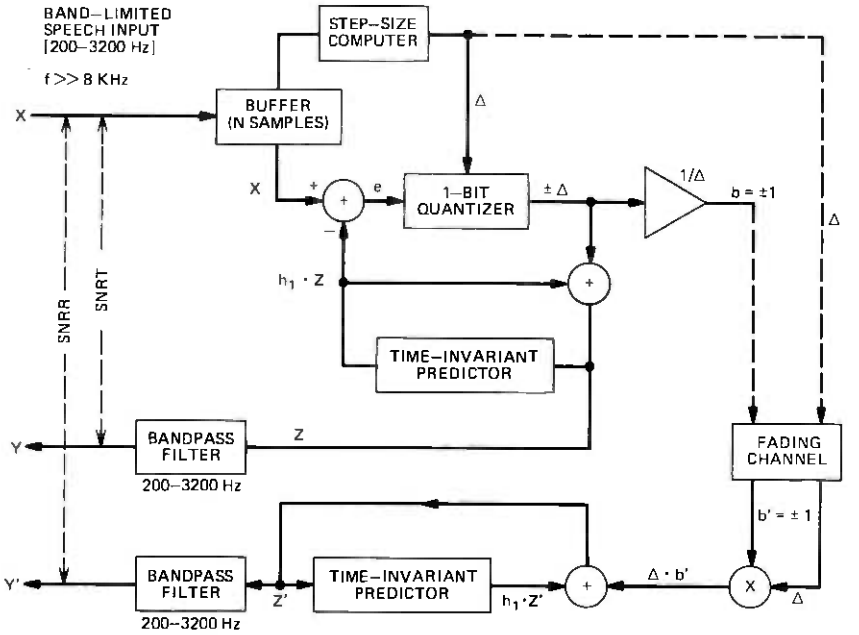


Fig. 5—Block diagram of DM-AQF codec.

implied one-sample delay occurs in the first-order predictor. Z and Z' are unfiltered staircase functions at the transmitter and receiver. The received bit b' differs from b if the error bit E is 1, and the step-size information Δ is assumed to be error-protected. For useful delta modulation, the sampling rate f should be much greater than the Nyquist frequency of the band-limited speech.

3.1 Design of Δ, N, and h_1

So that the best step-size Δ may follow the statistics of the input speech, the following algorithms were examined.

$$\Delta = K_1 \cdot \sum_{r=2}^{N} |X_r - X_{r-1}| \cdot \frac{1}{N-1}. \tag{3a}$$

$$\Delta = K_2 \cdot [\operatorname*{Max}_{2<r<N} |X_r - X_{r-1}|]. \tag{3b}$$

$$\Delta = K_3 \cdot \left[\sum_{r=2}^{N} (X_r - X_{r-1})^2 \cdot \frac{1}{N-1} \right]^{\frac{1}{2}}. \tag{3c}$$

Figure 6 plots the signal-to-noise-ratio SNRT at the encoder as a function of K_n ($n = 1, 2, 3$) for the above algorithms. The numbers refer to 24-kHz delta modulation. Each scheme exhibits an optimal K_n that

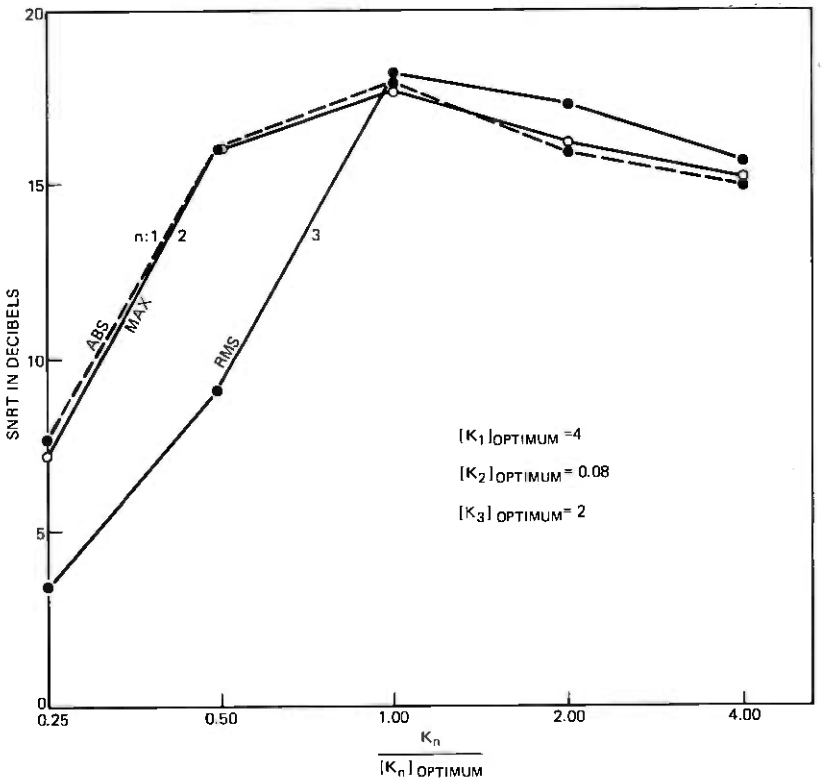


Fig. 6—Step-size computation for DM-AQF.

obviously represents the best mixture of slope-overload distortion (which predominates for $K \ll K_{OPT}$) and granular noise (which takes over for $K \gg K_{OPT}$). It is interesting that the maximum performances of the three algorithms are practically the same. This suggests that the step sizes resulting from these algorithms may not be significantly different, when optimal K values are employed. The rest of the paper will assume the use of the "average absolute slope" formula

$$\Delta = \frac{4}{N-1} \cdot \left[\sum_{r=2}^{N} |X_r - X_{r-1}| \right]. \tag{4}$$

Strictly speaking, this formula is optimal only for 24-kHz sampling and for perfect integrators ($h_1 = 1$). However, corrections for these factors were not found to be very significant for the values of f and h_1 used in our study, and formula (4) was therefore uniformly assumed for simplicity. (We may mention, however, that step-size dependencies

on h_1 and bit rate will be of interest in the design of DPCM-AQF encoders; this is seen in Section IV.)

The buffer length was set at $N = 256$. This represents a compromise among three factors: (*i*) a need to minimize the encoding delay (this suggests smaller N values), (*ii*) a need to keep down the information rate in the step-size transmitting channel (this suggests slower updating, or larger N values), and (*iii*) the need to track the changing statistics of speech with an appropriate speed. A buffer length of 5 to 10 ms turns out to be a good choice for differential coding (this is demonstrated quantitatively in the context of DPCM-AQF); and $N = 256$ does indeed correspond approximately to a 10-ms delay for $f = 24$ kHz (and a 5-ms delay for $f = 48$ kHz).

The predictor coefficient was set to be $h_1 = 0.9$. This was nearly optimal from an SNRT viewpoint for the sampling frequencies of interest. Over a noisy channel, if one uses SNRR as a performance criterion, optimal values of h_1 tend to be smaller than 0.9. This is because "leakier" integrators mitigate error propagation in the output of a differential decoder. Once again, in the interest of simplicity, a quantitative consideration of this phenomenon has been deferred to the case of multibit DPCM coding (Section IV).

3.2 Bit scrambling

Table II demonstrates how bit scrambling can provide an SNRR advantage in the presence of errors. As mentioned earlier, bit scrambling was simulated by using scrambled errors ES (in place of burst errors EB for an unscrambled bit stream). Informal listening tests indicate that the perceptual advantages of bit scrambling in DM-AQF are more significant than what the SNRR gains in Table II may suggest.

3.3 Error protection by redundant coding: EP-DM-AQF

We studied a redundant DM-AQF coder in which every pair of adjacent DM bits was protected by the transmission of a (contiguous) parity check bit. When the parity failed at the receiver, a possible bit error was detected, and the received DM bit pair were forced to form an alternating (+ − or − +) sequence. This is equivalent to the

Table II — Effect of bit scrambling in DM-AQF [$P(EB) = P(ES) = 0.055$ and entries are SNRR values in dB]

f(kHz)	Speech	Burst Errors	Scrambled Errors
32	Male	7.6	8.0
40	Female	7.8	8.8

Table III — Comparison of DM-AQF and EP-DM-AQF

Scheme	f(kHz)	Transmission rate (kb/s)	$P(ES)$	SNRR (dB)
EP-DM-AQF	32	48	0.055	8.0
DM-AQF	32	32	0.055	7.1
DM-AQF	32	32	0.025	10.0

imposition of a zero-slope segment in the speech waveform when the receiver has no confidence in the incoming bits. Table III compares the performance of this error-protected system (EP-DM-AQF) with that of an unprotected DM-AQF coder, for the example of scrambled errors. The unprotected system has a bit rate of 32 kb/s, while the EP-DM-AQF operates at $32 \times \frac{3}{2} = 48$ kb/s. We are not concerned at this point with questions like a specific baud rate. However, in view of transmission rate versus error probability relations over real channels (Section II), the interesting comparison in Table III is between rows 1 and 3 (rather than between 1 and 2). It appears that the simple parity-check-based error protection is not being useful; the advantages due to error detection at the receiver are being offset (or more than offset) by the increased error probability characteristic of the higher transmission rate in EP-DM-AQF. A similar result has been obtained in a simulation of DM-AQF with correlated errors, and also with DPCM encoders where only the most significant bit is error-protected by the use of redundancy.[3]

3.4 Unprotected DM-AQF with bit scrambling; choice of f

We have considered in some detail the specific case of unprotected (nonredundant) DM-AQF with bit scrambling. Table IV presents SNRT and SNRR values for such a system at different values of f and matched values of error probability, $P(ES)$. Some entries in Table IV are interpolated values because error sequences with the corresponding $P(ES)$ values were not available. As suggested earlier in the example of binary versus quaternary PSK, an obviously meaningful comparison is between rows 1 and 4 whose error ratios differ by a factor of two.

Table IV — DM-AQF; Effect of f

f(kHz)	$P(ES)$	SNRT (dB)	SNRR (dB)
24	0.023	17.1	8.6
32	0.032	21.2	9.6
40	0.040	23.7	10.5
48	0.048	26.0	11.2

At the transmitter end, the quantization noise is easily perceived at $f = 24$ kHz. It is barely apparent at $f = 32$ kHz, and a choice of $f = 40$ kHz is likely to be more than adequate for many situations. Notice that, as f increases, so does the difference between SNRT and SNRR; and the quantization noise has a lesser and lesser influence on the speech quality at the receiver because of the relatively greater contributions of channel noise.

3.5 A comparison with syllabic-companded DM-AQB

To demonstrate that forward step-size coding is indeed desirable for the mobile radio channel, the DM-AQF scheme was compared with a syllabic-companded delta modulator with backward-step-size control (AQB). The step-size algorithm for the DM-AQB was

$$\begin{aligned} \Delta_r &= 0.966 \cdot \Delta_{r-1} + 25 \cdot [\text{ADAPT}]_r \\ [\text{ADAPT}]_r &= 1 \text{ if } \left| \sum_{s=0}^{3} b_{r-s-p} \right| = 4, \text{ for } p = 0 \text{ or } 1 \text{ or } 2 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{5}$$

The algorithm is reminiscent of, if not identical to, the digitally controlled delta modulation (DCDM) scheme due to Greefkes,[7] which is an AQB technique well known for its error resistance. Figure 7 demonstrates that, in the presence of bit errors, the performance of DM-AQF degrades more gracefully than that of the DM-AQB defined in (5). It must be remembered, of course, that the DM-AQB system is implemented more easily and without encoding delay.[8]

3.6 The problem of step-size transmission in AQF

We have tacitly assumed so far that step-size information in DM-AQF can be very reliably transmitted, even over a fading channel, because step-size updating has to be done only infrequently. We shall now demonstrate this with some numbers.

Figure 8 illustrates a histogram of step sizes that resulted from utilizing (4) for a 32-kHz DM-AQF encoder. It was noted that the encoding was very tolerant to a maximum step-size constraint of 155, and a step-size resolution equal to 10; in other words, to a step-size dictionary of only 16 steps (5, 15, $\cdots$, 155). In practice, the maximum-to-minimum step size ratio would probably be greater than 31, in anticipation of highly nonstationary speech inputs.

The four-bit step-size information was transmitted as follows. At the beginning of each block of $N = 256$ bits, the respective four-bit word was transmitted five consecutive times. Each bit in the step-size word was decoded on the basis of a majority count over the five

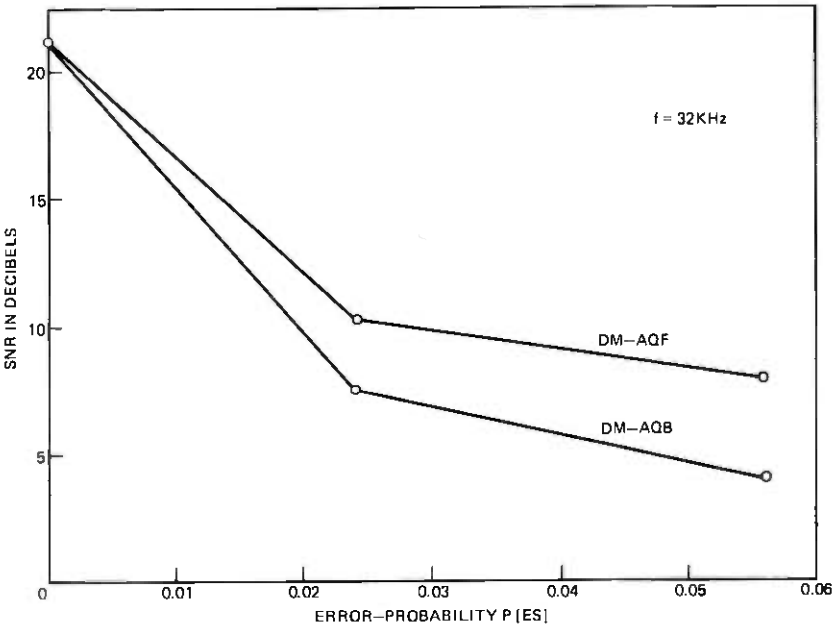


Fig. 7—Comparison of DM-AQF and DM-AQB.

received versions of the bit. The step-size transmissions increased the overall bit rate from 32 kb/s to $32[(256 + 4 \times 5)/256] = 34.5$ kb/s. For a random error rate of $P(ES) = 0.025$, the SNRR with the explicit transmission of step size, as above, was nearly identical with the value obtained in a simulation that tacitly assumed the presence of correct step-size information at the receiver. The result is not surprising; the probability of failure of a majority count of order 5 is given by

$$P(\text{M.C}; 5) = \sum_{r=3}^{5} p^r (1 - p)^{5-r} \binom{5}{r}$$
$$\sim 10p^3 \quad \text{if} \quad p \ll 1, \tag{6}$$

where p is the error probability. With $p = P(ES) = 0.025$, $P(\text{M.C}; 5) = 1.64 \times 10^{-4}$. The probability that at least one of the bits of a step-size word is wrongly decoded in our scheme is therefore no greater than 6.4 in 10,000, and there were only 250 step-size transmissions during the entire length of the (2-s) speech utterance being coded.

3.7 SNRT, SNRR, and P(E) as functions of time

We conclude our discussion of DM-AQF with an interesting demonstration of the time dependencies of SNRT, SNRR, and $P(ES)$, as measured over blocks that were $N = 256$ samples long. The sampling rate was

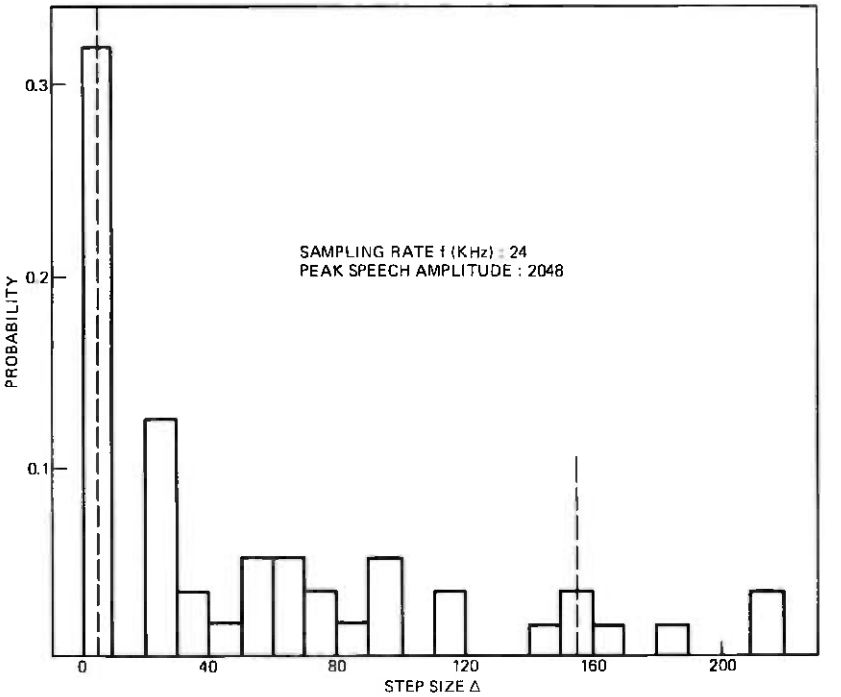


Fig. 8—Histogram of step sizes in DM-AQF (peak speech amplitude = 2048).

24 kHz, the average error probability was 0.025 [refer to Fig. 2 and eq. (1) for error characterization], and the plots on Fig. 9 used numbers taken once every 20 blocks (5120 samples). Notice the obvious negative correlation between the time functions SNRB[t] and $P(ES)[t]$. The time variation of SNRT is, of course, purely a reflection of the input speech material.

IV. DPCM-AQF

Figure 10 is a block diagram of differential PCM with forward step-size control. Differences from Fig. 5 consist in the use of a B-bit quantizer (B = 3 or 4 in this paper), and in the assumption of Nyquist-rate sampling, which obviates the need for a critical output filter. Basic DPCM notation is as follows: W is the normalized code word magnitude, e is the prediction error, and $\tilde{e}$ is the quantized value of e. The time-invariant (first-order) predictor coefficient is h_1, and r represents an instantaneous (sampled) value. The received bits b_q' (q = 1, 2, $\cdots$, B) are different from the transmitted bits b_q if a corresponding error bit E equals 1. The step size is Δ; it is assumed to be recalculated once every N samples, and successfully error-protected in transmission. The

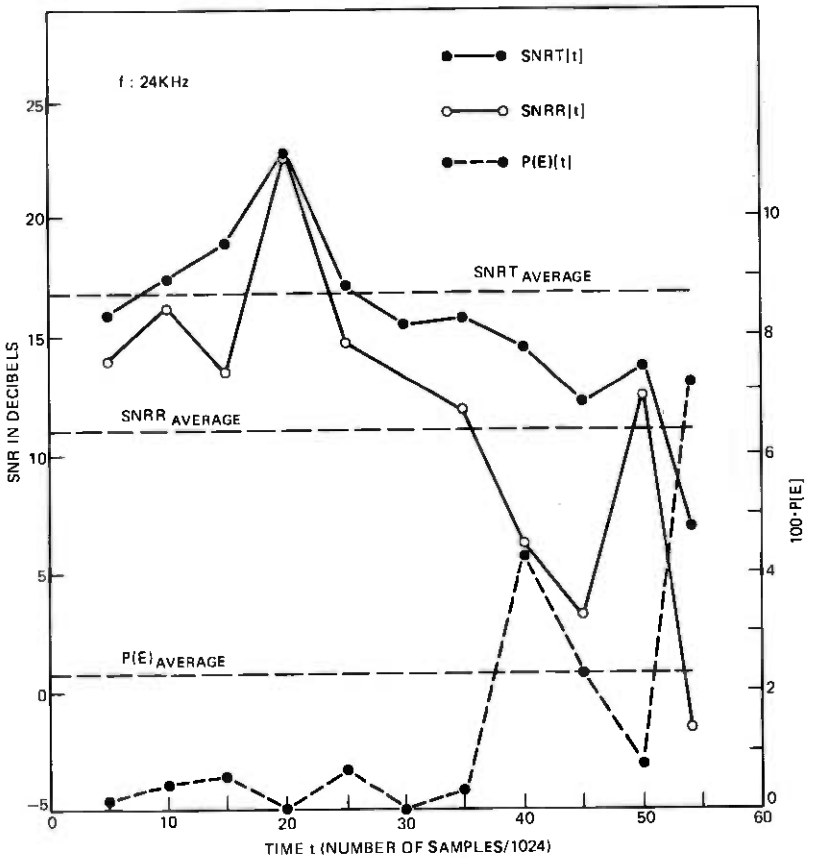


Fig. 9—Time variation of SNRT, SNRR, and $P(E)$ (V/λ = 36.2 Hz).

following are the salient DPCM equations.

$$
\begin{aligned}
b_{qr} &= \pm 1; \quad q = 1, 2, \cdots, B. \\
bb_{qr} &= 0.5 b_{qr} + 0.5 = 0 \text{ or } 1. \\
e_r &= X_r - h_1 \cdot Y_{r-1}. \\
Y_r &= h_1 \cdot Y_{r-1} + \tilde{e}_r. \\
Y'_r &= h_1 \cdot Y'_{r-1} + \tilde{e}'_r. \\
\tilde{e}_r &= W_r \cdot \Delta. \\
W_r &= \left[\sum_{q=2}^{B} 2^{B-q} \cdot bb_{qr} \right] \cdot \operatorname{sgn} b_{1r}.
\end{aligned}
\tag{7}
$$

For any sample r, the sign of the code word W is the most significant bit b_1; the least significant bit is b_B.

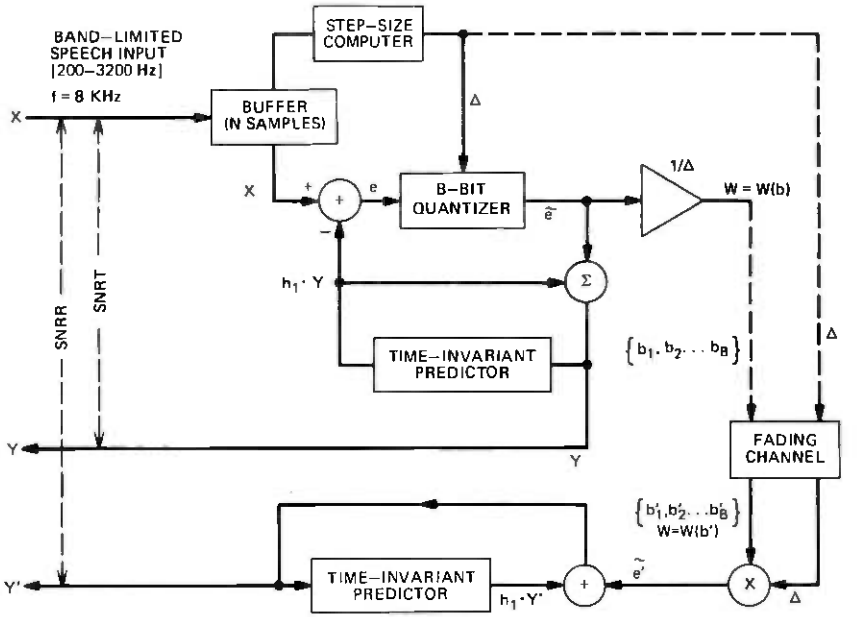


Fig. 10—Block diagram of DPCM-AQF codec.

4.1 Design of Δ, N, and h_1

AQF step sizes are derived (once for every block of N samples), using the formula

$$\Delta = K_4 \cdot \frac{1}{N-1} \cdot \sum_{r=2}^{N} |X_r - h_1 \cdot X_{r-1}|. \tag{8}$$

Figures 11, 12, and 13 illustrate typical SNRT and SNRR dependencies on the parameters K_4, N, and h_1, respectively. The curves refer to the case of $B = 4$, $P(EB) = 0.055$, and to a redundant transmission technique described in Fig. 14. It is clear that SNRT-maximizing designs are significantly different from the SNRR-maximizing values. Rather than getting bogged down in the controversial question of whether SNRT or SNRR is to be used as a performance criterion, we have elected, arbitrarily, to discuss the following SNRR-maximizing designs that were approximately good for the $P(E)$ range of 0.025 to 0.055:

$$\begin{aligned} N &= 64 \\ h_1 &= 0.6 \\ K_4 &= 0.50 \quad \text{if} \quad B = 3 \\ &= 0.25 \quad \text{if} \quad B = 4. \end{aligned} \tag{9}$$

Notice that, in Fig. 12, SNRR is maximum at $N = 128$. However, the

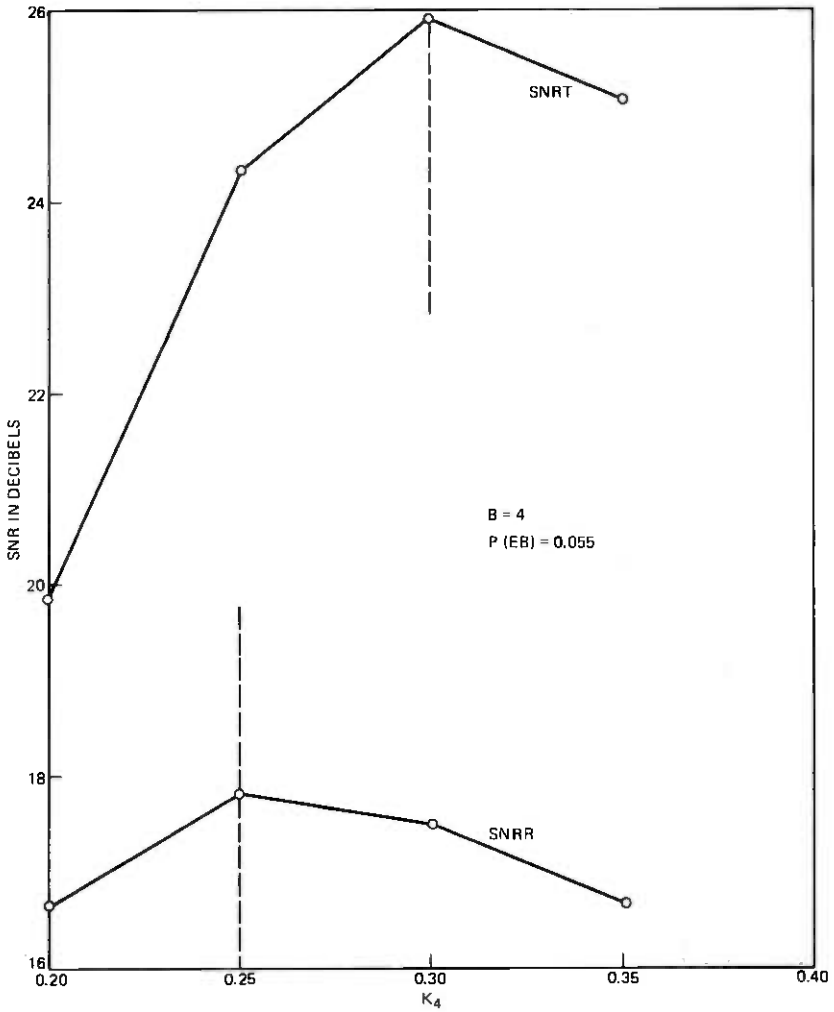


Fig. 11—Step-size computation for DPCM-AQF.

encoding delay is less objectionable (8 ms, instead of 16 ms) with $N = 64$. Note also that SNRT-maximizing designs call for higher values of both h_1 and K_4.

The maximum-to-minimum step-size ratio in the simulation was about 1000. It is possible to reduce this ratio to 100, and still provide useful coding of nonstationary speech.[1] Smaller step-size ratios enhance bit-error resistance. They also tend to simplify the problem of transmitting step-size information.

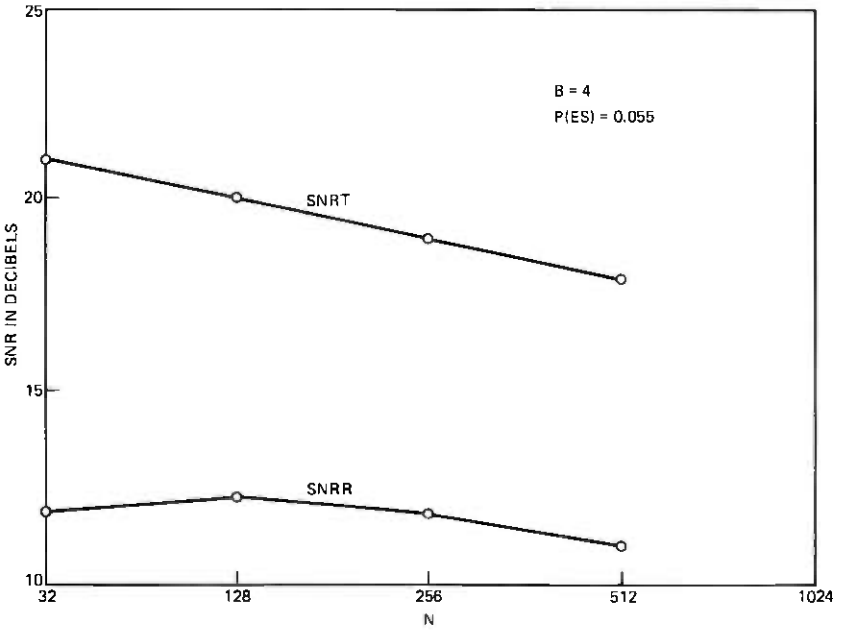


Fig. 12—Step-size updating in DPCM-AQF.

4.2 Error-protected DPCM-AQF

Figure 14 illustrates the use of time-diversity coding designed to protect DPCM bits from burst errors. The time-diversity is provided by the delay P that will be discussed presently. Figure 14a defines a three-bit EP-DPCM system where the most significant bit b_1 is transmitted three times, and the second most significant bit b_2 is sent twice. The least significant bit b_3 is transmitted only once. At the receiving end, the value of b_1 is determined on the basis of a majority count over the three received versions. In regard to the magnitude bit b_2, if the two versions of b_2 do not agree, the receiver code word is forced to its smallest magnitude (the polarity is still defined by the unequivocally decoded value of b_1). This is equivalent to forcing a "minimal-slope" segment in the decoded speech waveform when the receiver is in doubt about the code-word magnitude. Figure 14b defines a four-bit EP-DPCM system where only the most significant bit b_1 is error-protected. Once again, the decoding of b_1 at the receiver follows a majority count over the three received versions thereof. Assuming 8-kHz sampling, both EP-DPCM systems of Fig. 14 would operate at 48 kb/s. However, the three-bit system of Fig. 14a has the benefit of greater error protection.

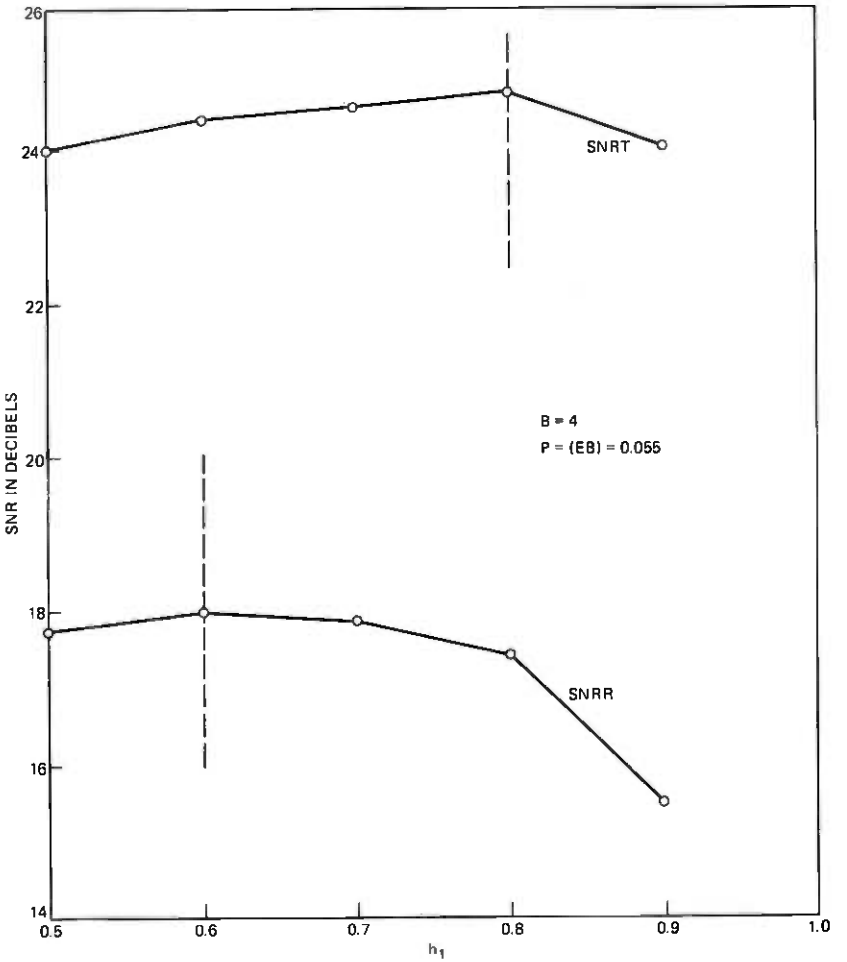


Fig. 13—Design of predictor coefficient h_1 in DPCM-AQF.

Figure 15 shows the benefits of time diversity for the example of the four-bit system of Fig. 14b. It is interesting that SNRR is still tending to increase at P values as large as 1024. It can be expected that, if $P \gg [D + I]_{\text{average}}$, successive repetitions of a given bit tend to be affected independently by the channel. D and I are the burst duration and spacing mentioned in Section II. The DPCM-AQF coders of this paper assume a uniform value of $P = 768$. For a bit rate of 48 kb/s, this implies a total encoding delay (from Fig. 13) of 2 P bits, or about 32 ms.

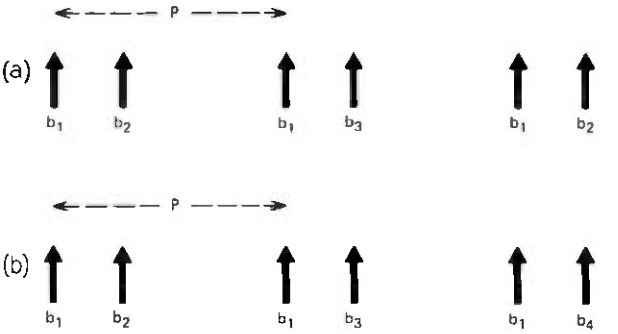


Fig. 14—Time-diversity coding for EP-DPCM.

4.3 EP-DPCM; choice of B

We now compare the two 48-kb/s systems of Fig. 14. Table V shows, for two different speech inputs, the SNRR values obtained with the three-bit and four-bit systems. The greater error protection in the three-bit system seems to make it more robust, in spite of the better quantization noise (SNRT) performance of a four-bit coder, and the better receiver-end quality of three-bit coded speech is very obvious in listening tests. The result is also mentioned by Noll.[3] It is true that four-bit coding can provide a 6-dB superiority in SNRT. It appears, on the other hand, that the subjective SNRT in DPCM is known to be considerably higher than a measured objective SNRT,[1] and the SNRT of

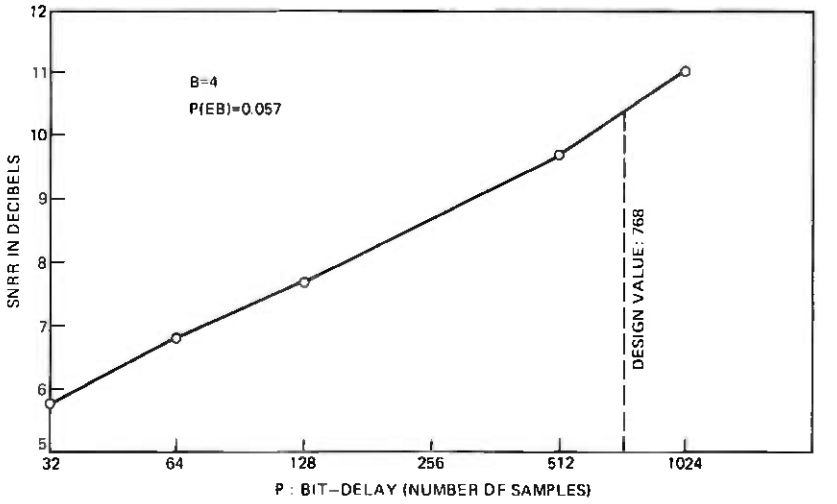


Fig. 15—Effect of time diversity on received-speech quality (V/λ = 36.2 Hz).

Table V — EP-DPCM; Comparison of three- and four-bit systems (entries are SNRR values in dB; rows 1 and 2 represent different speech inputs)

$P(EB)$	$B = 3$	$B = 4$
0.025	14.2	12.8
0.055	9.7	9.3

three-bit DPCM may prove to be subjectively adequate for some mobile links. In that case, the system of Fig. 14a would be a good configuration for error-protected DPCM.

Table VI further demonstrates the benefits of error protection for $B = 3$. In view of the transmission rate—$P(EB)$ relationships mentioned in Section II, the interesting comparison in the table is between rows 1 and 3, not 1 and 2. It is seen that EP-DPCM at 48-kb/s provides a better SNRR than unprotected 24-kb/s DPCM, in spite of the higher error probabilities that accompany the 48-kb/s transmissions. This contrasts interestingly with the results of Table II where error protection was seen to be ineffective for DM-AQF. The suitability of error protection for DPCM (and not DM) seems to be a direct consequence of the multibit quantization in DPCM: it is possible to isolate and error-protect only the more significant DPCM bits and incur an overall redundancy of 50 to 100 percent; a majority count for a 24-kHz DM would immediately result in a transmission rate of 72 kb/s (and a redundancy of 200 percent).

In a recently proposed, and not less effective, approach to EP-DPCM coding,[9] the DPCM bits are error-protected in suitably long blocks rather than on a bit-by-bit basis: The time diversity reception consists in selecting one of two time-separated blocks on the basis of an autocorrelation-type quality evaluation at the receiver.

4.4 Bit scrambling in DPCM

Informal listening tests, as well as SNRR evaluations, have shown that bit scrambling, and the resulting error-randomization, is much less effective for multibit DPCM than for DM. The reason for this is not

Table VI — Benefits of error protection for DPCM ($B = 3$)

Code	Transmission Rate (kb/s)	$P(EB)$	SNRT	SNRR
EP-DPCM	48	0.055	19.4	12.4
Unprotected DPCM	24	0.055	19.4	7.1
Unprotected DPCM	24	0.025	19.4	9.6

well understood. However, situations exist where bit scrambling can provide nominal SNRR gains even for DPCM. These have been noted by Noll.[3]

4.5 DPCM-AQB

The problem of step-size transmission for DPCM-AQF is expected to be handled through techniques not very different from those discussed in the context of DM-AQF. Following the calculation procedures of that section, it is estimated that virtually error-free transmission of DPCM step size would be possible (for the error rates considered) by the expenditure of about 5 kb/s of channel capacity.* To indicate the desirability of dedicating this kind of channel capacity for step size, we investigated two types of backward step-size control. One of these was adaptive differential quantization with a one-word memory.[1] Here, the quantized step size is modified for every sample by a factor determined solely by the magnitude of the latest code word W_r. The other adaptive scheme derived step-size information by an algorithm similar to the DPCM-AQF rule (8). The summation, however, was over the most recent N samples of quantized speech. Neither of the above backward schemes performed well enough with bit errors to merit inclusion of their results. It is conceivable, however, that, as in DM, some kind of a slowly adapting or syllabic DPCM may provide a fair result for mobile radio. It is also conjectured that the performance of such a scheme would be upper-bounded by that of DPCM-AQF in the manner of Fig. 7. At least one approach to slowly companded DPCM has been proposed to date.[10,11]

V. CONCLUSION

The object of this paper was to specify two differential coders—one from the DM family and the other from the DPCM class—that would be appropriate for digitizing speech in some types of mobile radio systems. The results of our work indeed suggest two such coders: a nonredundant 40-kHz DM-AQF coder with bit scrambling and an error-protected three-bit DPCM-AQF operating at a nominal 48 kb/s. The typical capabilities of these systems are summarized in Table VII, which is based on the example of a female utterance, "The lathe is a big tool." The transmission rates and error probabilities in Table VII are matched, albeit in a limited sense, as discussed earlier. Also, as emphasized already, the error rates in Table VII are worse-than-average numbers for many mobile radio links.

* If the overall transmission rate of the system is constrained to be 48 kb/s, it may be possible to work with a sampling rate of about 7 kHz, instead of 8 kHz, to accommodate the step-size information in the 48-kb/s channel (Ref. 4).

Table VII — Comparison of DM-AQF and EP-DPCM-AQF

Coder	B	f(kHz)	Transmission rate kb/s	Estimate of δ(kb/s)	Bit-Error Probability	SNRT (dB)	SNRR (dB)
EP-DPCM-AQF	3	8	$48 + \delta$	5	$P(EB) = 0.055$	20.5	14.5
DM-AQF	1	40	$40 + \delta$	<5	$P(ES) = 0.045$	23.7	10.1

In assessing the coders of Table VII, it may be worth noting that the DM system is more flexible. For example, the DM sampling rate can be lowered to 32 kHz with only a 2.5-dB loss in maximum speech quality SNRT (Table IV). Further, if the refinements of time-diversity coding (in DPCM) and bit scrambling (in DM) are eschewed, it is our experience that the DM system will lose less in the process.

Obviously, a common denominator in the above systems is adaptive differential quantization. Crudely speaking, adaptive quantization serves to squelch channel noise, while differential coding tends to smear it; and the combination appears to be perceptually very desirable in the context of mobile telephony.

Formal perceptual studies in this subject should appropriately include other digital techniques such as nondifferential (PCM) and backward-adaptive (AQB) coders. The studies should also include the possible effects of encoding delay. Clearly, the amount of this delay depends on what combination of refinements (forward coding, bit scrambling, and time diversity) is employed; and if the total delay gets to be long enough, the benefits of a better SNRR (due to reliable step-size information, error randomization, and redundant error protection, respectively) may be accompanied by a loss of echo performance over certain kinds of networks. The best compromise between transmitted speech quality, received speech quality, and encoding delay is very likely to be system-specific; and the nature of this compromise may influence or define a selection among analog techniques, conventional digital schemes (AQB), and step-size transmitting codes (AQF).

VI. ACKNOWLEDGMENTS

The idea of forward quantization for the fading channel was first suggested by P. Noll who, together with R. W. Schafer and M. R. Karim, provided an extensive amount of stimulation to this work. Error tapes were provided by V. H. MacDonald and G. Arredondo. Several very helpful comments on an earlier draft of this paper came from G. Arredondo and B. H. Bharucha.

REFERENCES

1. N. S. Jayant, "Digital Coding of Speech Waveforms—PCM, DPCM, and DM Quantizers," Proc. IEEE, May 1974, pp. 611–632.
2. G. A. Arredondo and J. I. Smith, "Voice and Digital Transmission in a Mobile Radio Channel at 850 MHz," Proc. Nat. Elec. Conf., *29*, 1974, pp. 74–79.
3. P. Noll, "Effects of Channel Errors on the Signal-to-Noise Performance of Speech Encoding Schemes," B.S.T.J., this issue, pp. 1615–1636.
4. D. L. Dutweiler and D. G. Messerschmidt, "Experimental Mobile Radio Digital Transmission with Time Diversity," Proc. Inter. Commun. Conf., San Francisco, June 1975.
5. V. H. MacDonald and G. Arredondo, private communication.
6. R. G. Gallager, *Information Theory and Reliable Communication*, Sec. 6.6, New York: John Wiley, 1968.
7. J. A. Greefkes, "Code Modulation with Digitally Controlled Companding for Speech Transmission," Philips Tech. Review, *31*, No. 11/12, pp. 335–353, 1970.
8. M. R. Karim, "Delta Modulation of Speech for Mobile Radio," unpublished work.
9. N. S. Jayant, "An Autocorrelation Criterion for the Time Diversity Reception of Speech Over Burst Error Channels," B.S.T.J., this issue, pp. 1583–1595.
10. S. U. H. Qureshi and G. David Forney, Jr., "A 9.6/16 KBPS Speech Digitizer," Proc. Int. Commun. Conf., San Francisco, June 1975.
11. D. J. Goodman, "A Robust Adaptive Quantizer," to be published in IEEE Trans. on Commun.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

An Autocorrelation Criterion for the Time-Diversity Reception of Speech Over Burst-Error Channels

By N. S. JAYANT

(Manuscript received March 25, 1975)

This paper proposes a new approach to signal selection in time-diversity systems. Specifically, we consider the problem of digital speech transmission over a burst-error channel using two-channel time-diversity reception.

Let every speech segment (of length W) be transmitted twice so that at least one of the transmissions escapes an error burst, with a certain useful probability. Let the received speech segments be Y_1 and Y_2. We propose an autocorrelation-maximizing signal selection procedure of the following form.

Select Y_1 (or Y_2) as the "cleaner" speech segment according as $C(Y_1, W) \geqq (\text{or} <) C(Y_2, W)$, where

$$C(Y_u, W) = \sum_{r=2}^{W} (\text{sgn } Y_{ur} \cdot \text{sgn } Y_{u(r-1)})/(W-1); u = 1, 2.$$

Y_{ur} is the speech amplitude at sample r in Y_u, W is a computational window that is typically a few milliseconds long, and sgn Y_{ur} is a polarity function that is assumed to have zero mean and unit variance.

The use of sgn Y_{ur} instead of Y_{ur} leads to a simply implemented selection procedure, and computer simulations have demonstrated its practical utility. For example, in one study of three-bit DPCM coding, autocorrelation-based burst-error detection proved to be more useful than a procedure where DPCM samples were error-protected on a bit-by-bit basis, rather than in blocks.

I. THE BURST-ERROR PROBLEM

The research reported in this paper was motivated by the problem of digital speech communication over a mobile radio channel. Signal transmissions over such a channel are characterized by multipath fading. The fading is "slow" in the sense that a given fade (signal strength below a specified threshold) can last for several tens of milli-

seconds (which will typically involve several tens or several hundred speech bits). The end effect of these "slow" fades on digital transmissions is to introduce bursts of errors in the reception of speech-carrying bits.

The time statistics of these error bursts are illustrated by the distribution functions in Fig. 1. D is the error-burst duration and I the error-free interval between successive bursts. An error burst is defined to have a local error probability of 1. In other words, a burst of length D_0 implies that D_0 contiguous speech bits are in error. An isolated error, for example, is an error burst of length $D_0 = 1$. The curves refer to a subsegment from a bit-error sequence whose average error probability was 0.06. Note that the local error probability in Fig. 1 [the ratio of D_{average} to $(D_{\text{average}} + I_{\text{average}})$] is 0.048. Notice also that $I_{\text{average}} \gg I_{\text{median}}$, suggesting a long tail in the interval distribution.

The error sequence was obtained from a fading simulator,[1] and it represents the impairment for a 24-kb/s signal-bit stream (the bit duration determines the number of bits affected by a fade) when the mobile radio link is characterized by two-branch diversity reception under the following (worse than average) conditions:

$$\begin{aligned} \text{Signal-to-interference ratio} &= 6 \text{ dB} \\ \text{Signal-to-noise ratio} &= \infty \\ \frac{\text{vehicle speed}}{\text{radio wavelength}} = \frac{V}{\lambda} = \frac{29 \text{ mi/h}}{0.353 \text{ m}} &= 36.2 \text{ Hz}. \end{aligned} \tag{1}$$

A companion paper[2] provides a somewhat more elaborate discussion of signal fades and error bursts.

II. TIME-DIVERSITY CODING

The temporal structure of clustered errors can be exploited in redundant transmission schemes where message units are repeated with an appropriate time spacing. The optimum time spacing is, in general, a function of the error statistics. For example, the spacing can be designed to minimize the probability that both of two consecutive transmissions of a given message unit are affected by an error burst or bursts. The message unit can be a block of speech-amplitude samples, or a single bit from a digital speech code, and so on.

A recent proposal discusses the use of time diversity for three-bit DPCM transmissions over mobile radio.[2] Briefly, redundancy is introduced in the form of three transmissions of the most significant (sign) bit B_1 in a DPCM word and two transmissions of the second most significant (magnitude) bit B_2. The average redundancy is therefore 100 percent. The receiver decodes the sign bit B_1 on the basis of a majority

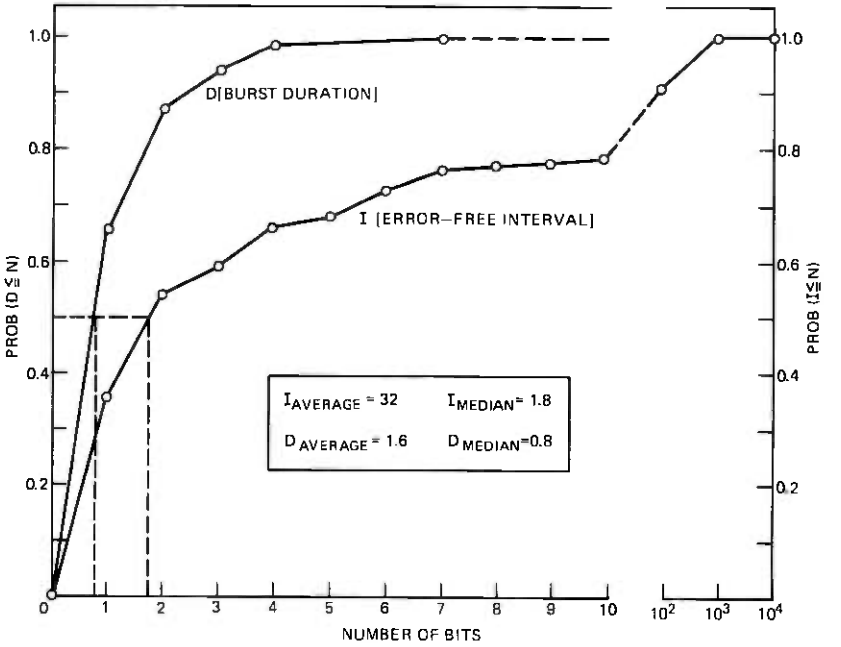


Fig. 1—Time statistics of burst errors [$P(E) = 0.06$].

count (over the three received versions). It also looks for unanimity between the two received versions of the significant magnitude bit B_2. If the unanimity does not exist, the DPCM word is forced to its minimum possible magnitude. When the spacings between the repetitions of DPCM bits are properly designed, the technique provides a significant advantage over nonredundant DPCM.[2] We comment again on this procedure at the conclusion of Section IV.

The purpose of this paper is to propose a different approach to time diversity. The method is based on error-burst detection (rather than single-error correction, as in a successful majority count); and the message unit that is error-protected is a block of contiguous speech amplitudes, rather than a basic speech-carrying DPCM bit or word. The idea of protecting message blocks using time diversity is not, in itself, claimed to be novel. What is interesting in our technique, however, is the method by which a high bit-error density is detected in a received speech segment (more strictly, in one of two segments in a diversity pair). The basis of such burst detection is a simple autocorrelation-type measurement of relative speech (or channel) quality, denoted by C. Unlike a signal-to-noise ratio (SNR), the quantity C can be evaluated over a received segment without reference to the transmitted speech.

In fact, our channel evaluations, based on C, are somewhat reminiscent of eye-pattern-based channel assessments in digital data communications.

III. AUTOCORRELATION C

The proposed measurement is the correlation

$$C(X, W) = \sum_{r=2}^{W} (\operatorname{sgn} X_r \cdot \operatorname{sgn} X_{r-1})/(W - 1), \tag{2}$$

where X_r represents a sampled speech amplitude, W is a computational window that is typically a few milliseconds long, and sgn X is a polarity function whose mean value and variance are assumed to be 0 and 1. We will also be interested in the correlations $C(XQ, W)$ and $C(Y, W)$, where the quantities XQ and Y refer to (unfiltered) staircase functions at the outputs of local and remote speech decoders (Fig. 2). $C(XQ, W)$ and $C(Y, W)$ are defined by operations similar to (2).

In simulating digital transmissions of speech over burst-error channels, we have found that clustered transmission errors tend to have the following type of effect on C: with a high probability (say, on the order of 0.9 or more), $C(Y, W) < C(X, W)$, where X and Y represent original and received speech segments. Qualitatively, the result is a consequence of increased zero-crossing activity in error-corrupted speech waveforms. Actual values of $C(Y, W)$ depend not only on the local error statistics, but also on the value of the corresponding $C(X, W)$, the nature of the quantization of X (prior to transmission) as reflected in the value of $C(XQ, W)$, and the extent of channel error propagation in the received signal (if the quantization is differential). Because of these factors, the magnitude of $C(Y, W)$ cannot be used, as such, for very reliable burst-error detection.

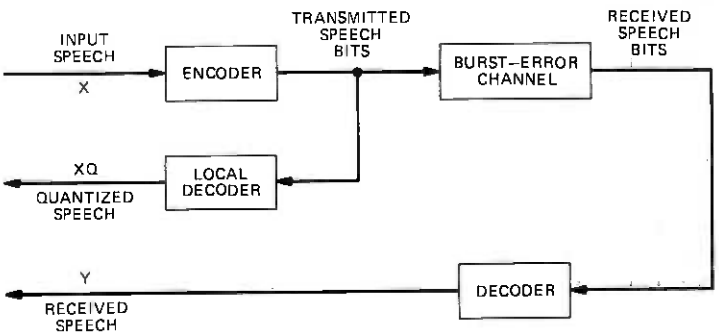


Fig. 2—Definition of X, XQ, and Y.

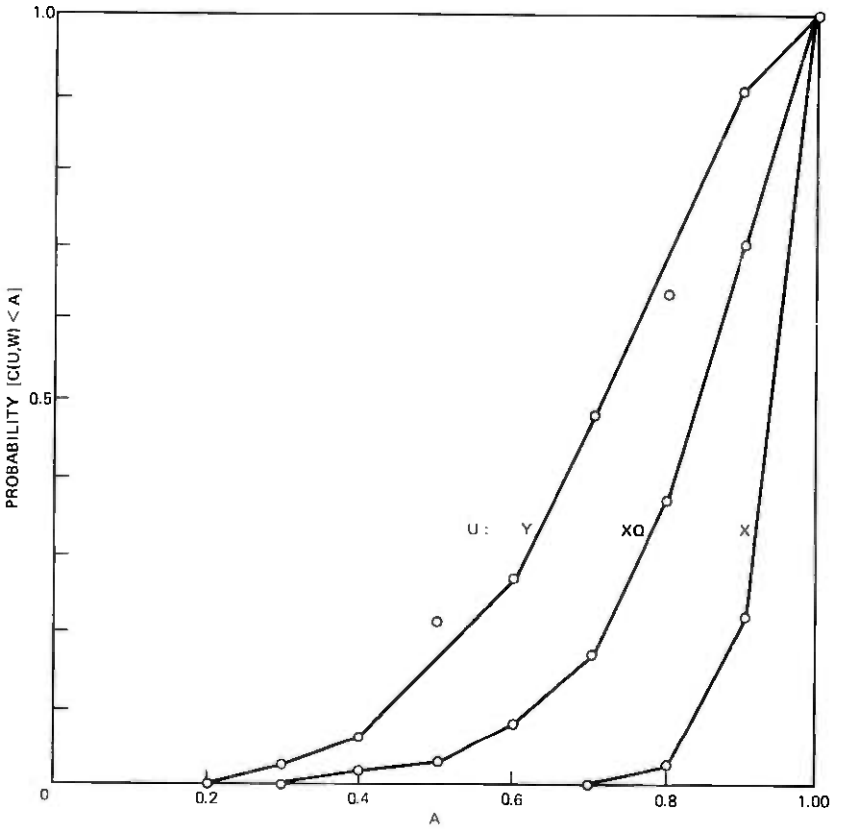


Fig. 3—Distributions of $C(X, W)$, $C(XQ, W)$, and $C(Y, W)$ in 24-kHz delta modulation $[P(E) = 0.06, W = 64]$.

These points are demonstrated by the results in Fig. 3 and Table I. These results refer to the 24-kHz delta modulation[2] of the band-limited (200 to 3200 Hz) female speech utterance, "A lathe is a big tool." The delta-modulation bits were transmitted through a simulated burst-error channel whose time statistics are shown in Fig. 1. As mentioned earlier, the average bit-error probability $P(E)$ on this channel is 0.06. Local error probabilities, as measured over blocks of W samples, will

Table I — Mean and median values of $C(X, W) - C(Y, W)$

$W = 64$	Median	Mean
$P(E, W) = 0.00$	0.11	0.14
$P(E, W) = 0.14$	0.18	0.26
ave		

be denoted by $P(E, W)$. (In delta modulation, a "sample" is synonymous with a "bit." In B-bit PCM or differential PCM, a "sample" refers to an entire B-bit word.) The windows for Fig. 3 and Table I are $W = 64$ samples long.

Figure 3 shows the distributions of $C(X, 64)$, $C(XQ, 64)$, and $C(Y, 64)$: specifically, values of the probability that $C(U, W)$ is less than A where $U = X$, XQ, or Y; $W = 64$; and $-1 \leqq A \leqq 1$. The results refer to a subset of samples characterized by nonzero values of $P(E, W)$, and an average error probability of 0.14. Notice how quantization errors, as well as transmission errors, tend to decrease the correlation C. Correlation losses due to noise and distortion are also demonstrated in Table I, which summarizes mean and median values of $[C(X, 64) - C(Y, 64)]$ for two channel conditions: the case of zero transmission errors [a subset of blocks where $P(E, W) = 0$] and the case of nonzero transmission errors [the subset of blocks where the average $P(E, W) = 0.14$]. Incidentally, both these subsets belong to the set of blocks whose average $P(E, W) = 0.06$. The top row in Table I measures the effect of quantization errors (plus, strictly speaking, the effect of error propagations in received speech), while the bottom row demonstrates the contributions of local transmission errors.

The distribution distances in Fig. 3 and the numbers in Table I both lead to the following conclusion: Although the channel quality $[P(E, W)]$ has a very clear effect on the autocorrelation C, the effect is not strong enough for $C(Y, W)$ to be employed, as such, as a reliable

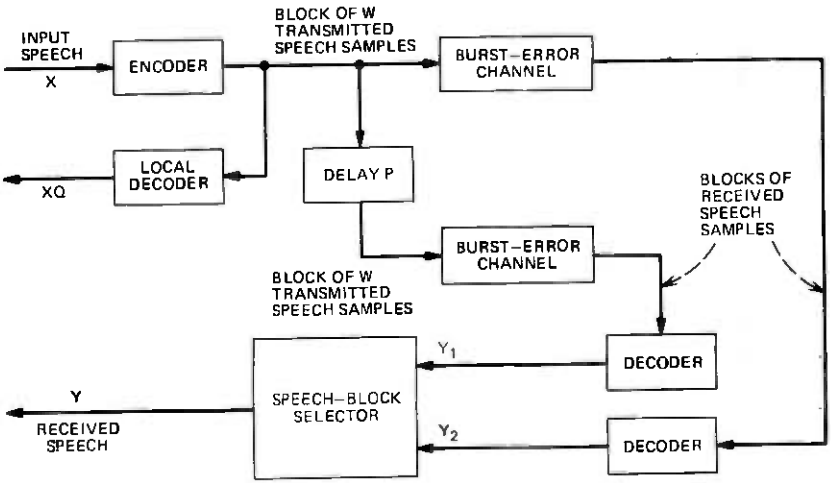


Fig. 4—Two-channel time diversity with block transmissions.

measure of speech or channel quality $[P(E)]$. To explain, a low value of $C(Y, W)$ is often indicative of a local transmission error burst. Occasionally, however, a poor autocorrelation may simply be a reflection of received waveform history and/or quantization noise, and/or an above-average high-frequency content in the local speech input.

A situation where channel information can be reliably extracted from C is in time-diversity coding. Consider, for example, the two speech segments Y_1 and Y_2 of a time-diversity pair (Fig. 4). The channel-independent factors mentioned at the end of the previous paragraph are exactly the same for both Y_1 and Y_2. Consequently, any difference between $C(Y_1, W)$ and $C(Y_2, W)$ can be safely attributed to differences in the channel conditions affecting the receptions Y_1 and Y_2.

IV. THE USE OF C IN TIME-DIVERSITY CODING

We propose that, for time-diversity reception, the autocorrelation C be used as a criterion for speech segment selection at the receiver. For example, with two-channel time diversity (Fig. 4), we suggest the

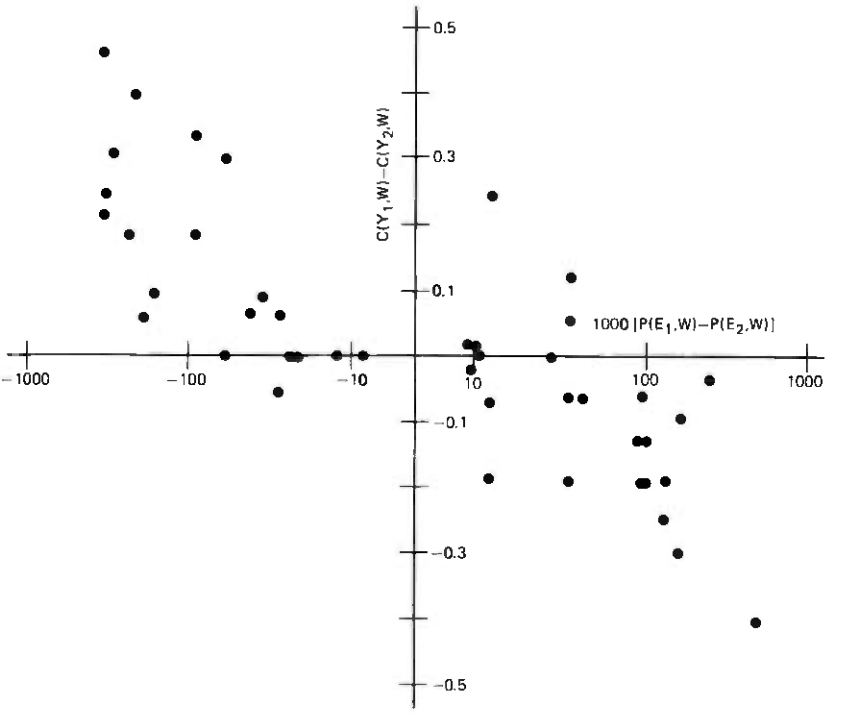


Fig. 5—Performance of $C(Y, W)$-based speech selector with three-bit DPCM $[W = P = 64; P(E) = 0.06]$.

following reception rule:

$$Y = Y_1 \text{ (or } Y_2\text{) according as } C(Y_1, W) \geqq \text{ (or } <)C(Y_2, W), \quad (3)$$

where

$$C(Y_u, W) = \sum_{r=2}^{W} (\text{sgn } Y_{ur} \cdot \text{sgn } Y_{u(r-1)})/(W - 1); \quad u = 1, 2.$$

The effect of (3) is to select the speech segment whose signum (polarity) function exhibits the higher autocorrelation. The rest of this section presents results that demonstrate the credibility of the above procedure. Specifically, we point out that very strong *negative correlations* exist between the following quantities:

$$\text{sgn } [C(Y_1, W) - C(Y_2, W)]$$

and (4)

$$\text{sgn } [P(E_1, W) - P(E_2, W)].$$

It is assumed that smaller $P(E)$ values imply better speech quality so

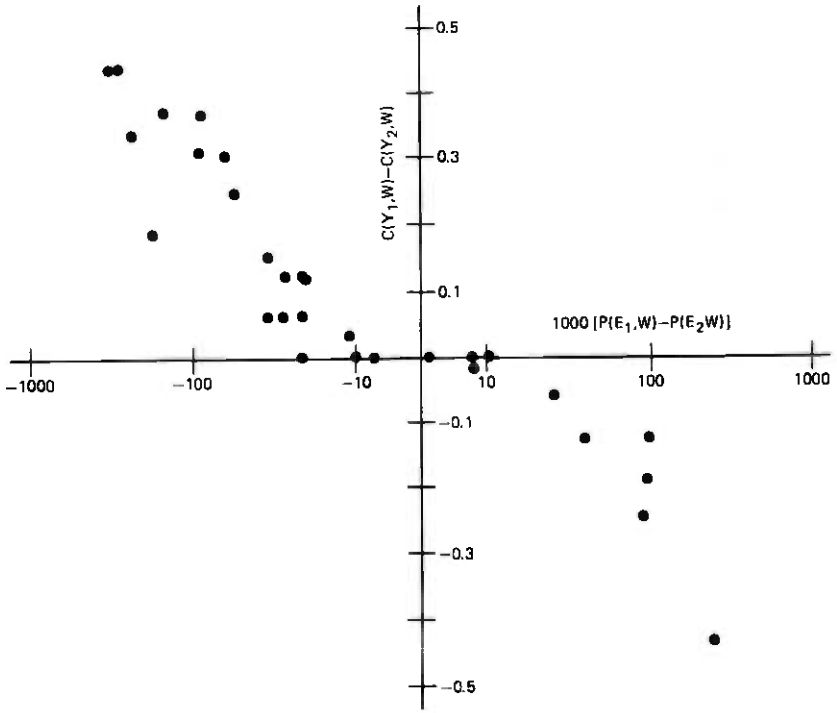


Fig. 6—Performance of $C(Y, W)$-based speech selector with three-bit PCM [$W = P = 64$; $P(E) = 0.06$].

that negative correlations between the quantities in (4) are indeed indicative of an appropriate reception rule. Most of the following discussion refers to three-bit DPCM coding. This is an example of practical interest for the time-diversity coding of speech over burst-error channels.[2]

Figures 5, 6, and 7 are scatter plots of $[C(Y_1, W) - C(Y_2, W)]$ versus $[P(E_1, W) - P(E_2, W)]$ for illustrative speech codes (PCM, DPCM) and average transmission-error rates of 0.03 and 0.06. The speech input was the same as that used in Section III, and the scatter plots represent sample subsets of simulation results. The members of the subsets were equally spaced points that spanned the total speech duration of about 1.5 seconds. Notice the negative correlation between $[C(Y_1, W) - C(Y_2, W)]$ and $[P(E_1, W) - P(E_2, W)]$ in each of Figs. 5, 6, and 7. This negative correlation reflects the fact that (for a given speech input and quantization error pattern) a higher $C(Y, W)$ value implies a lower $P(E, W)$ value, i.e., a better speech quality. The very small I- and III-quadrant occupancies reflect a low probability

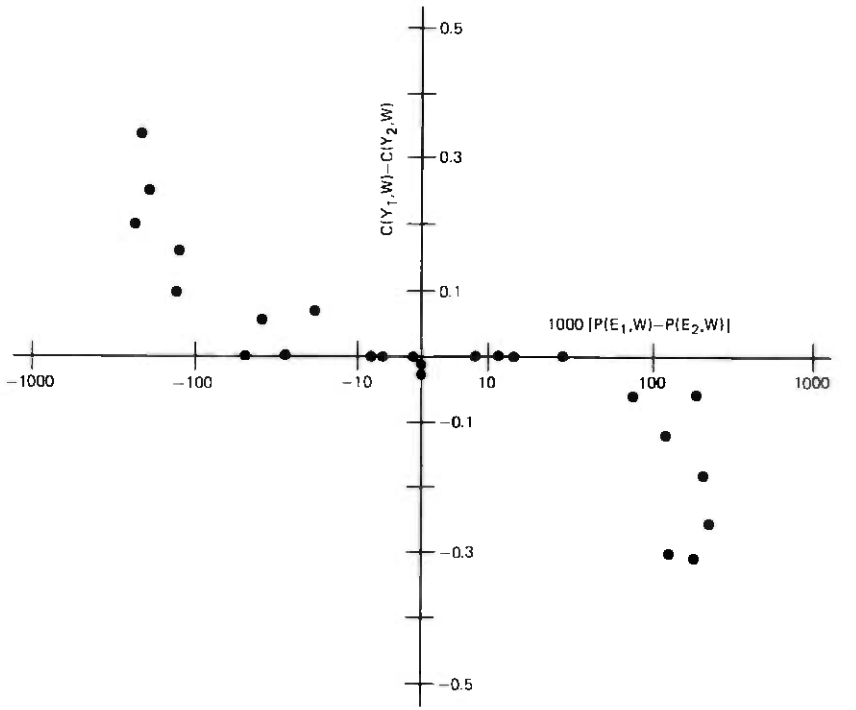


Fig. 7—Performance of $C(Y, W)$-based speech selector with three-bit DPCM $[W = P = 64; P(E) = 0.03]$.

of failure (wrong speech-segment selection for the $C(Y, W)$-based selector (3)).

We briefly discuss the effects of correlation window length W and time-diversity spacing P on received speech quality. The quantities SNRT and SNRR refer to signal-to-noise ratios measured over the duration of the entire speech utterance:

$$\begin{aligned} \text{SNRT} &= \sum X_\tau^2 / \sum (X_\tau - XQ_\tau)^2 \\ \text{SNRR} &= \sum X_\tau^2 / \sum (X_\tau - Y_\tau)^2. \end{aligned} \tag{5}$$

T and R refer to SNR values as measured at the local (transmitter-end) and remote (receiver-end) decoders (Fig. 2). We are interested in DPCM codes with a forward-adaptive quantizer: the step size is updated every 64 samples at the transmitter, and the step-size information communicated to the receiver in a special error-protected format.[2] Finally, the differential coding uses a time-invariant first-order predictor. The predictor coefficient was 0.6. This value was suggested by the need to dissipate the effects of channel errors in the reconstructed speech, as explained in the companion paper.[2]

Table II shows the effects of W and P on the received speech quality as measured by SNRR. It is seen that 100-percent redundancy, together with a good choice of W and P, can buy a more-than-4-dB improvement over unprotected DPCM. Incidentally, the overall transmission rate is approximately 48 kb/s for the time-diversity codes and 24 kb/s for the nonredundant code. The lower error rate (0.03) used for the latter is a reflection of the lower transmission rate.[1,2]

Figure 8 elaborates on the performance of the optimal ($W = 64$, $P = 256$) time-diversity code, while Table III compares its performance with that of the bit-protecting scheme[2] mentioned in Section II.

The diversity systems are formally sketched in Fig. 9. The encoding delays ($P + W$ for block protection and $2P'$ for bit protection) are

Table II — Effect of W and P on SNRR [3-bit DPCM; $P(E) = 0.06$]

W (Number of 8 kHz-samples)	P	SNRT (dB)	SNRR (dB)
64	64	20.4	14.5
128	128	20.4	12.1
256	256	20.4	13.3
64	256	20.4	15.8
Unprotected 3-bit DPCM with $P(E) = 0.03$		20.4	11.5

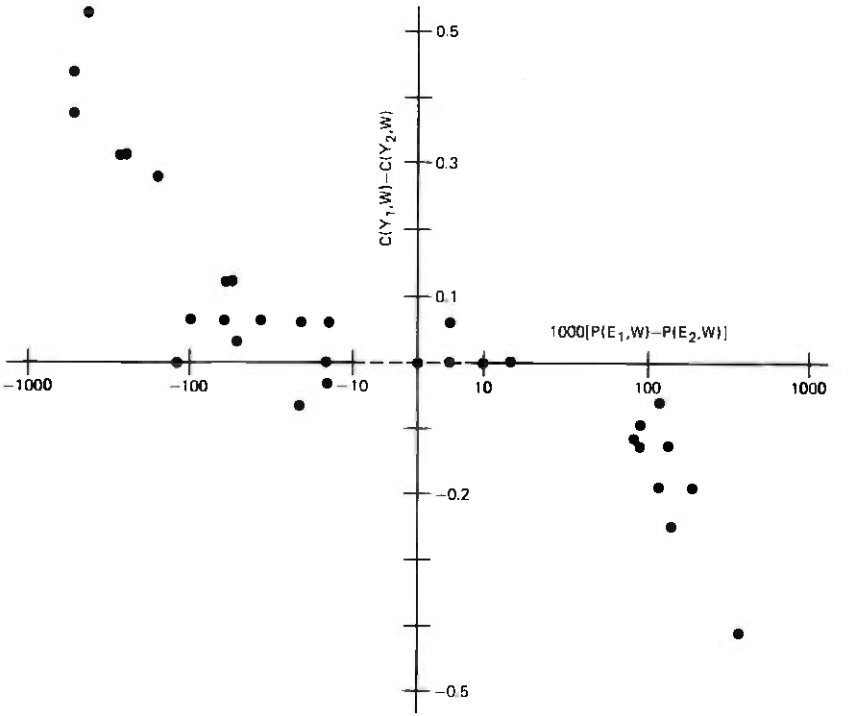


Fig. 8—Performance of $C(Y, W)$-based speech selector with three-bit DPCM [$W = 64$; $P = 256$; $P(E) = 0.06$].

chosen to be of the same order of magnitude. (Both the schemes are expected to perform slightly better with longer encoding delays.) Table III indicates a slight SNRR superiority for the block-protection technique, especially at the higher error rate. What is more significant than the SNRR advantage is a perceptual effect; the block-protected speech sounds considerably crisper. The companion paper[2] includes

Table III — SNRR values (dB) in block-protecting and bit-protecting schemes for time-diversity coding of three-bit DPCM speech

$P(E)$	Bit Protection	Block Protection
0.000	20.4	20.4
0.024	17.0	17.4
0.054	14.5	15.8

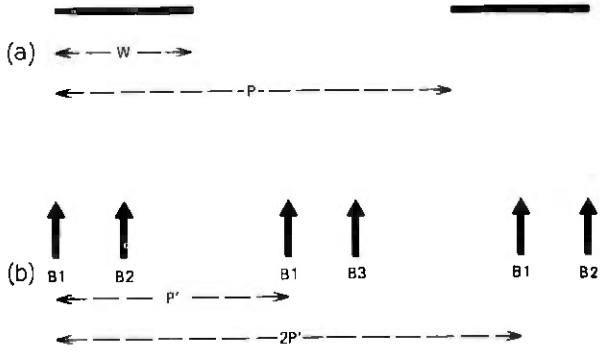


Fig. 9—Time-diversity coding based on (a) block protection (W = 8 ms, P = 32 ms); (b) bit protection ($2P'$ = 32 ms).

more observations on the speech quality resulting from error-protected DPCM.

V. CONCLUSION

This paper has demonstrated the capabilities of a new technique for signal selection in time-diversity systems. The results of Table III are a good indication of the practical utility of the new technique. We believe, however, that the contribution of this paper consists not in the specific quality improvements (over bit-protecting systems) in Table III, but in the fact that the autocorrelation of the most significant bit (polarity function) is indeed a useful measure of relative signal quality over noisy channels. This is demonstrated mainly in the scatter plots in Figs. 5, 6, 7, and 8. The use of the most significant bit in evaluating signal quality leads obviously to simple implementations.

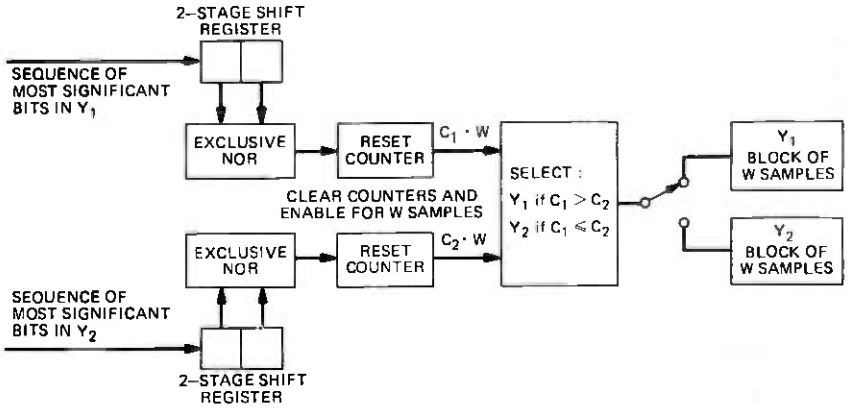


Fig. 10—Implementation of an autocorrelation-based block selector.

A possible configuration for an autocorrelation-maximizing signal selector is depicted in Fig. 10.

VI. ACKNOWLEDGMENT

The author wishes to thank P. Cummiskey for comments on an earlier version of this paper.

REFERENCES

1. V. H. McDonald and G. Arredondo, private communication.
2. N. S. Jayant, "Step-Size Transmitting Differential Coders for Mobile Telephony," B.S.T.J., this issue, pp. 1557–1581.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

A Comparative Study of Various Quantization Schemes for Speech Encoding

By P. NOLL

(Manuscript received February 7, 1975)

In this paper, the performance limits, as given by the signal-to-noise ratio (s/n), are described for different speech-encoding schemes including adaptive quantization and (linear) adaptive prediction schemes. The comparison is made on the basis of computer simulations using 8-kHz-sampled speech signals of one speaker. Different bit rates (two bits per sample—five bits per sample) have been used.

A three-bit-per-sample PCM scheme with a nonadaptive μ100 quantizer leads to an s/n value of approximately 9 dB. A maximum s/n value of approximately 25 dB has been reached using an encoding scheme including both adaptive quantization and adaptive prediction. Entropy coding of the quantizer output symbols leads to an additional gain in s/n of nearly 3 dB.

I. INTRODUCTION

Design of an efficient encoding scheme requires some knowledge of the statistics of the signal. Efforts to improve the performance of PCM systems have taken two primary directions:

(*i*) Use of quantizing schemes based on knowledge of the (one-dimensional) probability density function (PDF) of the samples to be quantized.
(*ii*) Use of quantizing schemes exploiting the correlation between successive samples.

If we had an *a priori* knowledge of the statistics of the samples, a nearly optimum quantization scheme could be used consisting of:

(*i*) A quantizer matched to the PDF of the signal to be quantized.
(*ii*) A predictor optimized for the given autocorrelation function of the signal.

The predictor lowers the variance of the signal to be quantized by

removing the correlation between successive samples. This is done by subtracting an estimation value from each incoming sample; the difference can be quantized, encoded, and transmitted (differential PCM = DPCM).

In digital speech-encoding systems, we have only a small amount of *a priori* knowledge of the statistics which, in addition, usually change with time:

(*i*) The long-period mean level differs from speaker to speaker.
(*ii*) At a given mean level, the instantaneous level changes because of variations in speech sounds.
(*iii*) The correlations between successive samples change because of variations in speech sounds.

To overcome these problems of unknown statistics, adaptive quantization and adaptive prediction schemes must be used. In these schemes, local estimates of the statistical parameters are calculated. The quantizer and/or predictor are then optimized based on these estimates.[1–3]

This paper compares different encoding schemes that include:

(*i*) Fixed quantizers.
(*ii*) Adaptive quantizers.
(*iii*) Fixed predictors.
(*iv*) Adaptive predictors.

The comparison is done on the basis of computer simulations; the signal-to-quantization noise ratio (s/n) has been used as the criterion for the comparisons. It is believed, however, that the s/n understates the subjectively perceived performance of encoders that have differential quantizers (the DPCM schemes in this paper).[1]

II. DESCRIPTION OF THE ENCODING SCHEMES

A computer program has been written that allows the simulation of encoding schemes combining the possibilities of nonadaptive or adaptive quantization and nonadaptive or adaptive prediction. The schemes that have been used are described in the following sections.

2.1 Fixed and adaptive quantizers

If the quantizer is *nonadaptive*, its characteristic is assumed to be logarithmic. Optimum, i.e., s/n-maximizing quantizers (whether uniform or nonuniform), cannot be used, not even under the assumption of a constant mean level, because the idle channel noise is higher for op-

timum quantizers than for logarithmic quantizers and results in poorer subjective performance.[3,4] The idle channel noise performance is determined by the smallest reconstruction level r_1 of the quantizer. Table I lists these values for various optimum three-bit quantizers (the term Gauss quantizer refers to a quantizer with an s/n-maximizing performance for signals with a gaussian PDF, etc.). The gamma PDF is a good model for speech amplitudes, but the smallest reconstruction level is 2.4 times higher for the corresponding optimum quantizer than for the logarithmic quantizer.

To overcome the problems of unknown mean level and the variations of the instantaneous level, adaptive quantization schemes (AQ schemes) can be used. A local estimate σ_x^2 of the variance of the input signal can be calculated; this value controls the gain of an amplifier located in front of a quantizer that is optimum for signals with unit variance. Two schemes are possible:

(*i*) *Forward estimation* (AQF): The estimation value is calculated from samples of the input signal. The input signal must be buffered, and the estimation value must be transmitted to the receiver in addition to the quantized samples.[3,5]

(*ii*) *Backward estimation* (AQB): The estimation value is calculated from quantized samples[3-6]; therefore, the state of the amplifier need not be transmitted (except for synchronizing purposes in case of channel errors).

Figure 1 shows the structures of the different PCM schemes. Note that the combination of controlled amplifier and fixed quantizer can be replaced by a quantizer with a step-size adaptation. Matching the gain of the amplifier to signal variance results in modifying the PDF of the signal to be quantized. It has been shown that different density functions can be reached by choosing an appropriate forward estimation scheme.[3] To get the best s/n performance, those quantizers can be employed that are optimum for the specific PDF.

Table I — Comparison of the smallest reconstruction levels r_1 of different optimum unit variance three-bit quantizers

Type of Quantizer	Nonuniform Quantizer r_1	Uniform Quantizer r_1
Uniform PDF quantizer	—	0.217
Gauss quantizer	0.245	0.293
Laplace quantizer	0.222	0.366
Gamma quantizer	0.149	0.398
Logarithmic μ100 quantizer	0.062	—

2.2 Types of quantizers

The following types of quantizers were used in the simulation of the speech-encoding systems:

Uniform quantizer with different loading factors.
Logarithmic quantizer with different loading factors.

Uniform optimum Gauss quantizer.
Uniform optimum Laplace quantizer.
Uniform optimum gamma quantizer.

Nonuniform optimum Gauss quantizer.
Nonuniform optimum Laplace quantizer.
Nonuniform optimum gamma quantizer.

These optimum quantizers lead to a maximum s/n for the specific probability density functions.

2.3 Algorithms of the AQ schemes

In applying adaptive quantization schemes, different possibilities of controlling the gain of the amplifier have been used. The following notation has been employed for the description of the algorithms (see also Figs. 1 and 2):

Symbol	Explanation
$x(n)$	Input sample at time instant n.
$y(n)$	Quantized sample at time instant n.
I_n	Index of quantizer step at time instant n.
G_n	Gain of the amplifier at time instant n (backward estimation).
G_N	Gain of the amplifier used in block N (forward estimation).
M	Number of quantized samples used for calculation of G_n (backward estimation).
N	Number of block.
$NSEG$	Number of input samples used for calculation of G_N (forward estimation).
NF	Number of first sample of block N. $NF = (N-1)\cdot NSEG + 1$.
ϱ_N	Vector of short-term autocorrelation coefficients calculated from the input samples of block N.
R_N	Toeplitz matrix of short-term autocorrelation coefficients calculated from the input samples.
$\alpha, \beta, \alpha_j, \epsilon$	Coefficients to be optimized for each algorithm.

In all AQ schemes, a local estimate of the quantizer input signal variance

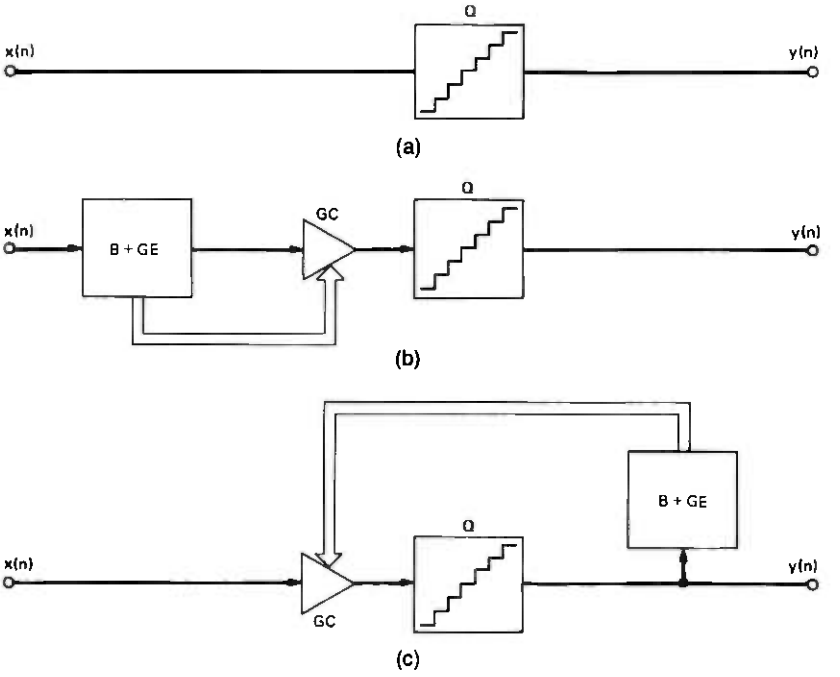


Fig. 1—PCM encoding schemes. (a) nonadaptive PCM. (b) Adaptive PCM with forward estimation (PCM-AQF). (c) Adaptive PCM with backward estimation (PCM-AQB). $x(n)$ = input signal, $y(n)$ = quantized signal, Q = quantizer, GC = gain control, B + GE = buffer and gain estimation.

is calculated; this value determines the gain of the amplifier such that the quantizer is optimal loaded.

2.3.1 Forward estimation schemes (AQF)

In the AQF schemes, the gain is only readjusted once for a new block of $NSEG$ speech samples:

$$G_N = \text{const.}; \qquad n = NF, NF + 1, \cdots, NF + NSEG - 1$$
$$NF = (N - 1) \cdot NSEG + 1.$$

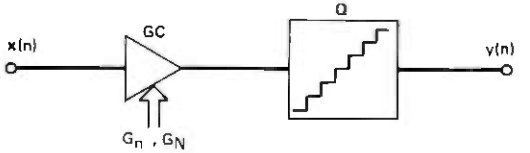


Fig. 2—AQ scheme. $x(n)$ = input signal, $y(n)$ = quantized signal, G_n = gain used at time instant n (AQB scheme), G_N = gain used in block N (AQF scheme).

The following algorithms have been used:

(*i*) PCM: *variance scheme*[3]: An unbiased estimation of the variance of the block is calculated:

$$G_N^{-2} = \alpha \cdot \frac{1}{NSEG} \sum_{j=1}^{NSEG} x^2(NF - 1 + j). \tag{1}$$

G_N is proportional to the inverse of the standard deviation estimated from the samples of the block.

(*ii*) PCM: *maximum scheme*: The maximum amplitude in the block is used:

$$G_N^{-1} = \alpha \cdot \max \{ |x(NF - 1 + j)| \}_{j=1,\cdots,NSEG}. \tag{2}$$

(*iii*) DPCM: *maximum scheme*[7]: The maximum difference between neighbored samples is used:

$$G_N^{-1} = \alpha \cdot \max \{ |x(NF - 1 + j) - x(NF - 2 + j)| \}_{j=2,\cdots,NSEG}. \tag{3}$$

This algorithm can only be used for predictors with one coefficient.

(*iv*) ADPCM: *variance scheme*[8,9]: The vector of short-term autocorrelation coefficients is used to calculate an estimation value of the variance σ_d^2 of the difference signal:

$$G_N^{-2} = \alpha \cdot \sigma_d^2 = \alpha \cdot [\sigma_x^2 - \varrho_N^T \cdot \mathbf{R}_N^{-1} \cdot \varrho_N] \tag{4}$$

2.3.2 Backward estimation schemes (AQB)

In AQB schemes, the gain of the amplifier is, in general, modified for every new input sample by a factor depending on the knowledge of the previous quantized samples or of the corresponding quantizer indices.

$$G_n = \alpha \cdot G_{n-1}.$$

The following algorithms have been used:

(*i*) *One-word memory scheme*[6]: The last gain value is multiplied by a factor that depends on the last occupied quantizer step:

$$G_n = f(|I_n|) G_{n-1}. \tag{5}$$

(*ii*) *Variance scheme*[6,10]: The last M quantized samples and (for $\beta \neq 0$) the last gain value are used to calculate a new gain value:

$$G_n^{-2} = \sum_{j=1}^{M} \alpha_j y^2(n - j) + \beta / G_{n-1}^2. \tag{6}$$

(*iii*) *Modified one-word memory scheme*[3]: The gain of the amplifier is changed if the smallest reconstruction level has been occupied α times or if the largest reconstruction level has been occupied once:

$$G_n = \begin{cases} 2.0 \cdot G_{n-1} & \text{if } |I_m| = \min \text{ for } \alpha \text{ times } (m = n, n-1, \cdots, n-\alpha+1) \\ 0.5 \cdot G_{n-1} & \text{if } |I_n| = \max \\ 1.0 \cdot G_{n-1} & \text{otherwise.} \end{cases} \tag{7}$$

2.4 Fixed and adaptive predictors

In predictive encoding systems, an estimate of each input sample is calculated and subtracted from the actual input sample; the difference is then quantized, encoded, and transmitted. The use of nonadaptive predictors (DPCM schemes) leads to a suboptimum overall performance of the encoding scheme, because the prediction is not optimum for all speakers and for all speech sounds.[1,8]

A better prediction can be reached by using adaptive algorithms (ADPCM schemes). Two schemes are possible:

(*i*) *Forward scheme*: A short-time autocorrelation function is calculated using a finite number of buffered input samples. The predictor coefficients are readjusted according to the time-variant autocorrelation function.[2,11]

(*ii*) *Backward scheme*: The predictor is optimized using the quantized information (gradient search method and Kalman filter algorithm).[1,2]

Only the forward scheme has been used in the simulations. The optimum vector $\mathbf{h}_N$ of J predictor coefficients for each block N is

$$\mathbf{h}_N = \mathbf{R}_N^{-1} \cdot \boldsymbol{\varrho}_N. \tag{8}$$

$\mathbf{R}_N$ and $\boldsymbol{\varrho}_N$ are the matrix and the vector of short-term autocorrelation coefficients calculated from the input samples of block N. The predictor coefficients have to be transmitted to the receiver, in addition to the code words of the quantized difference signal samples. An upper bound of the gain in s/n as compared to PCM is given in Section IV.

III. RESULTS

Various nonadaptive and adaptive encoding schemes have been simulated on a digital computer. The signal-to-quantization noise ratio (s/n) has been determined using 8-kHz-sampled speech samples of one speaker. The same 2.3-s utterance ("The boy was mute about his task"; female voice; bandwidth 200 to 3200 Hz) has been used in

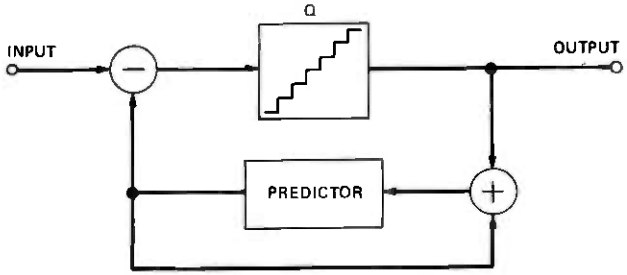


Fig. 3—DPCM scheme. Q = quantizer.

all simulations. (The simulations have not included any high-frequency emphasis of the input speech, as is characteristic of a 500-type set transmitter, for example.) The following schemes have been studied:

(*i*) Nonadaptive Quantization
PCM: see Fig. 1.
DPCM (nonadaptive prediction): see Fig. 3.
ADPCM (adaptive prediction): see Fig. 4.

(*ii*) Adaptive Quantization
PCM: see Fig. 1.
Forward scheme (PCM-AQF).
Backward scheme (PCM-AQB).
DPCM (nonadaptive prediction): see Fig. 5.
Forward scheme (DPCM-AQF).
Backward scheme (DPCM-AQB).
ADPCM (adaptive prediction): see Fig. 6.
Forward scheme (ADPCM-AQF).
Backward scheme (ADPCM-AQB).

These encoding schemes have been optimized using the s/n as criterion.

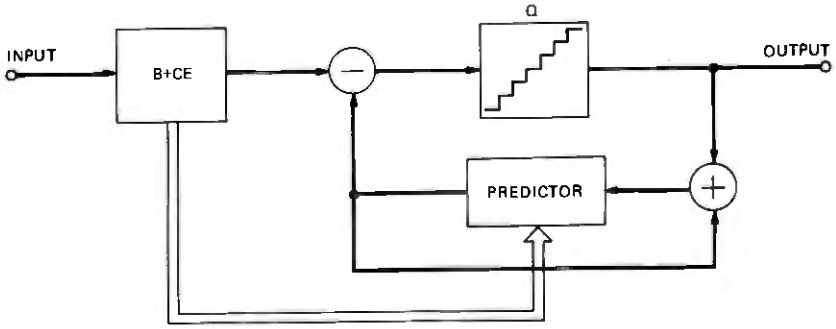


Fig. 4—ADPCM scheme. B + CE = buffer and coefficients estimator.

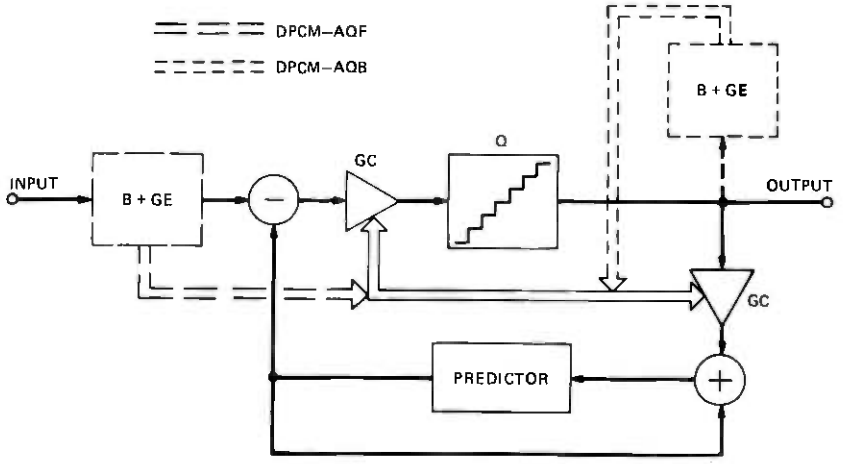


Fig. 5—DPCM-AQ schemes. B + GE = buffer and gain estimator, GC = gain control.

3.1 Optimum results: three bits/sample quantization

Figure 7 shows the optimum results reached with a three-bit quantization of the 2.3-s speech sample.

Left curves: Optimum results using a fixed quantizer.
Right curves: Optimum results using an adaptive quantizer.
Lower curves: Prediction with a first-order predictor (one coefficient).
Upper curves: Prediction with a high-order predictor.

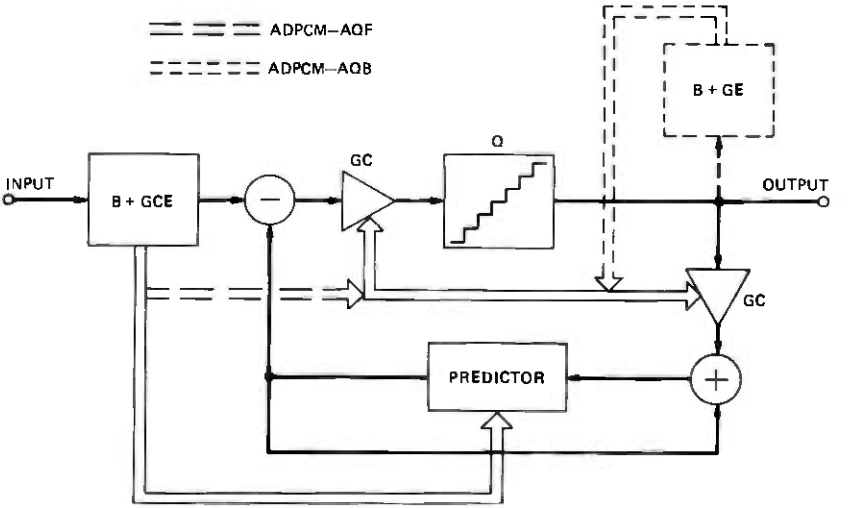


Fig. 6—ADPCM-AQ schemes. B + GCE = buffer and gain and coefficients estimator.

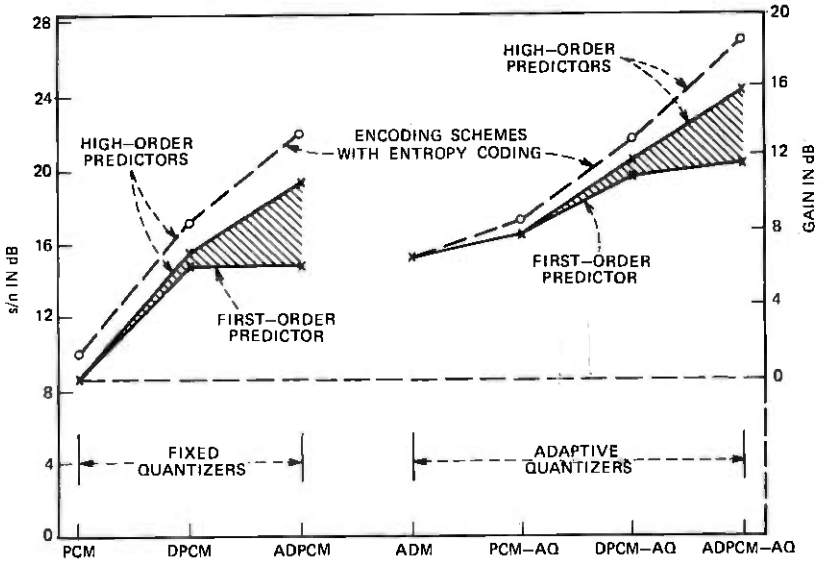


Fig. 7—Signal-to-noise ratio values and gains (over logarithmic PCM) for different three-bit speech-encoding systems.

Quantizers with a logarithmic characteristic have been used in all simulations with a fixed quantizer (curves on the left side of Fig. 7). The s/n value for a PCM scheme is

$$s/n = 8.7 \text{ dB}$$

if the quantizer has a $\mu 100$ characteristic, and if the loading is $4\sigma_x$ (σ_x is the standard deviation of the signal to be quantized). As compared to this s/n value of 8.7 dB, the following maximum gains can be reached with prediction schemes using the same type of quantizer (G^* is the gain in s/n over PCM):

$$\text{Fixed predictor, fixed quantizer: } G^* \approx 7 \text{ dB}$$
$$\text{Adaptive predictor, fixed quantizer: } G^* \approx 11 \text{ dB}.$$

Adaptive quantization (PCM-AQ) not only has the advantage of increasing the dynamic range that the quantizer can handle, but it also allows the application of quantizers that are optimum for the probability density function of the signal to be quantized. The following gain over the 8.7-dB value of nonadaptive PCM has been reached:

$$\text{Adaptive quantization: } G^* \approx 7 \text{ dB}.$$

Using predictors, the gains over PCM are now

$$\text{Fixed predictor, adaptive quantizer: } G^* \approx 12 \text{ dB}$$
$$\text{Adaptive predictor, adaptive quantizer: } G^* \approx 16 \text{ dB}.$$

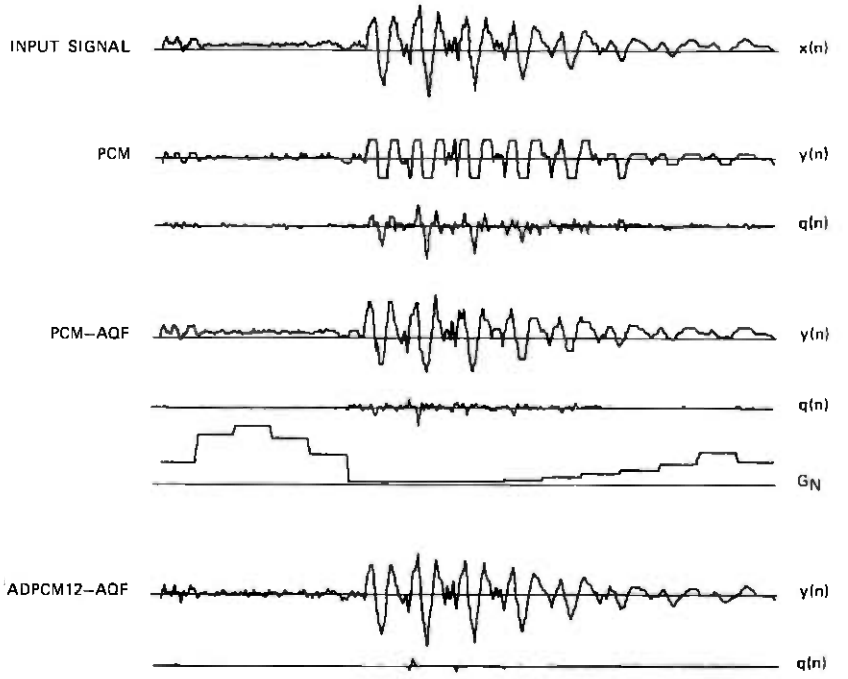


Fig. 8—Comparison of waveforms. $x(n)$ = sequence of input samples (512 samples), $y(n)$ = sequence of decoded samples, $q(n)$ = sequence of quantization errors, G_N = sequence of amplifier gains.

Figure 8 shows the waveforms of the reconstructed signal and of the quantization error for a 64-ms segment of speech. Three examples are shown:

(*i*) PCM, nonadaptive, $\mu 100$ characteristic. Only eight different levels can be used for the reconstruction (decoding) of the signal.

(*ii*) PCM-AQF, optimum Gauss quantizer, $NSEG = 32$. The number of levels is limited to eight for each segment of $NSEG$ samples. Different levels can be used for each segment.

(*iii*) ADPCM12-AQF, optimum Laplace quantizer, $NSEG = 128$. The predictive encoding with a 12th-order predictor leads to a very high s/n. For each segment, the number of levels of the *difference* signal is limited to eight, but the reconstructed signal does not suffer this limitation.

3.2 Adaptive delta modulation

To determine whether the quantization schemes represent an improvement over existing adaptive delta modulation (ADM) schemes, the s/n value of Jayant's ADM-scheme[12] has been determined at a bit rate

of 24 kb/s. The s/n value is approximately 15 dB.[7] Therefore, the gain over nonadaptive three bits/sample PCM is

$$\text{Adaptive delta modulation: } G^* \approx 6 \text{ dB}.$$

3.3 Entropy coding

Entropy coding is a variable-length coding procedure that assigns short code words to highly probable symbols and longer code words to less probable symbols. The average word length is approximately equal to the entropy of the quantizer output signal. The entropy coding technique leads to an additional gain in s/n for a given average bit rate. The number of quantizer steps can be increased without exceeding an average bit rate of three bits per sample. The dashed lines in Fig. 7 show the s/n values that can be reached by using an entropy coding technique. In this case, uniform quantizers with a large number of steps have been employed; the step sizes have been adjusted to give a quantizer output entropy of three bits. It should be noted that a buffer is needed so that the variable-length coded signal can be transmitted over a channel at a uniform bit rate.

3.4 Optimum results: two bits/sample up to five bits/sample quantization

Figure 9 shows the s/n values for quantizations with two bits/sample up to five bits/sample (corresponding to bit rates from 16 kb/s up to 40 kb/s). The following encoding schemes have been compared:

PCM	μ100 characteristic, $8\sigma_x$ loading.
PCM-AQF	$NSEG = 32$, optimum Gauss quantizer.
DPCM1-AQB	1 predictor-coefficient, fixed; optimum Gauss quantizer.
ADPCM1-AQF	1 predictor-coefficient, adaptive; optimum Gauss quantizer; $NSEG = 32$.
ADPCM4-AQF	4 predictor-coefficients, adaptive; optimum Laplace quantizer; $NSEG = 128$.
ADPCM12-AQF	12 predictor-coefficients, adaptive; optimum gamma quantizer; $NSEG = 256$.

3.5 Parameter transmission in adaptive encoding schemes

In all forward schemes, channel capacity is needed for transmission of the adaptive parameters. The problems and techniques of quantizing these parameters are not considered in this paper. It is known that the parameters tolerate coarse quantization and slow updating. If necessary, redundancy-reducing schemes can be used to lower the number of bits that have to be transmitted in addition to the encoded speech

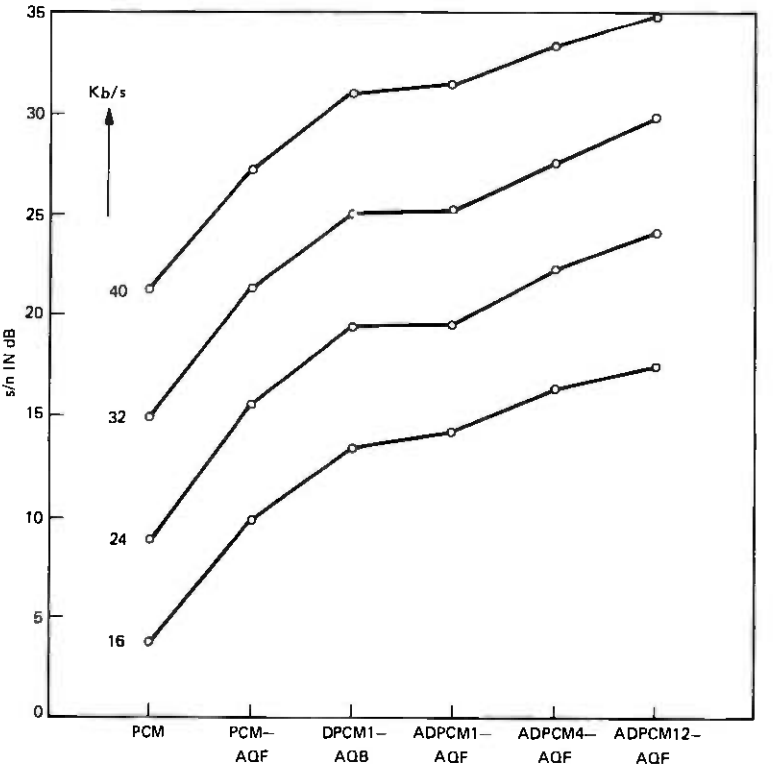


Fig. 9—Signal-to-noise ratio values for quantization with two bits per sample (16 kb/s) up to five bits per sample (40 kb/s).

samples. The needed channel capacity can be approximately transformed into an equivalent loss in s/n. If each parameter of the adaptive scheme has to be encoded with $NADD$ bits/segment, and if $NSEG$ is the number of samples/segment, then we get an equivalent reduction in s/n performance:

$$\Delta_{s/n} = 6.02 \frac{NADD \text{ (bits/segment)}}{NSEG \text{ (samples/segment)}} \text{ (dB)}. \tag{9}$$

This loss is due to the reduction of the number of quantizer steps in order not to exceed the maximum allowed bit rate.

Example:

$NSEG = 128$ (16 ms)
$NADD = 4$ bit.

The loss is 0.2 dB for each coefficient to be transmitted.

IV. UPPER BOUNDS FOR PREDICTION

The linear dependencies between the amplitudes of the speech sample being used in all simulations have been calculated to get a measure of the maximum gain that can be reached with linear prediction. Note that these upper bounds of the prediction gain cannot be reached with predictive encoding systems (especially if the quantizer has only a low number of quantization levels), because prediction is done then with decoded speech samples. These samples include a quantization error.

4.1 Nonadaptive prediction

The long-term autocorrelation function of the speech signal has been measured. Figure 10 shows the first 19 time lags of the normalized autocorrelation function $\rho(n)$. Using these data, a predictor can be optimized such that the variance of the difference signal is minimum.

The prediction gain is the ratio of the variances of the input signal and the difference signal:

$$G_P = 10 \log_{10} \frac{E[x^2(n)]}{E[d^2(n)]} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_d^2}. \tag{10}$$

G_P can be calculated directly from the normalized autocorrelation function $\rho(n)$. Figure 11 shows this gain versus the number of coefficients being used for the prediction. The maximum prediction gain is approximately 10.5 dB. This value is an upper bound of the additional gain in s/n over PCM by using nonadaptive differential encoding schemes. This gain cannot be reached if the DPCM encoder has to handle speech samples of different speakers. In this case, suboptimum predictor coefficients have to be chosen such that the DPCM encoder has

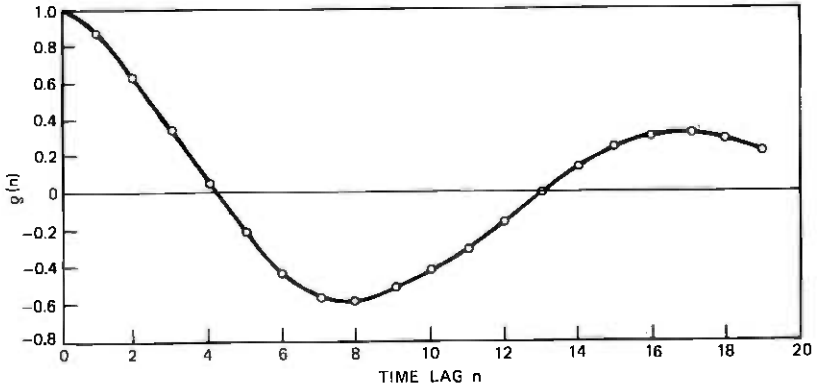


Fig. 10—Normalized autocorrelation function (female voice; 200 to 3200 Hz).

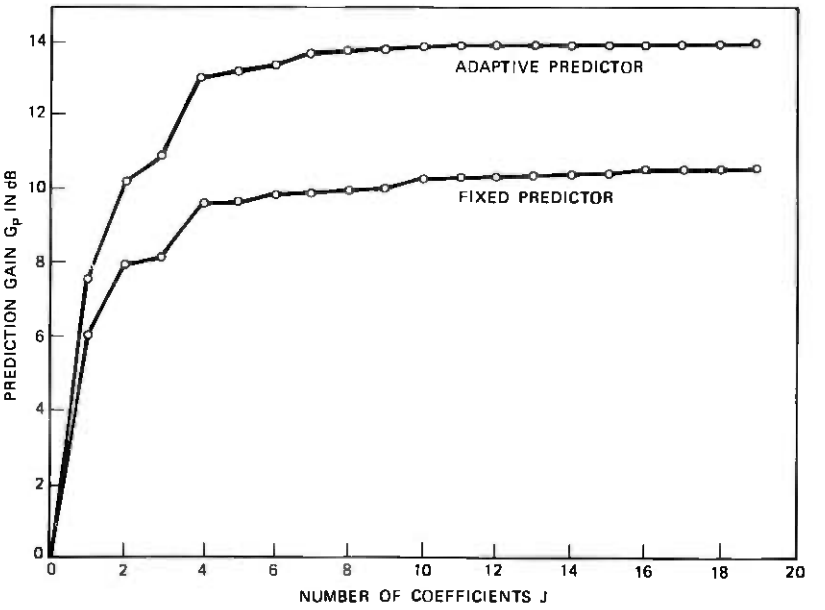


Fig. 11—Prediction gains vs number of predictor coefficients.

a good performance for all speakers. This demand can only be fulfilled with predictors of low order (up to three coefficients). It may be relevant to mention that such suboptimum predictor coefficients have been used in the simulation of the DPCM schemes.

Knowing the long-term autocorrelation function $\rho(n)$, it is possible to calculate an approximation of the power density function. This is done by calculating the power transfer function of a recursive filter, the coefficients of which are equivalent to the coefficients of the optimum predictor (maximum-entropy method[13]). Figure 12 shows the power density spectrum calculated in this way from 16 coefficients of the autocorrelation function $\rho(n)$.

4.2 Adaptive prediction

NSEG samples of the input samples are buffered, and the short-term autocorrelation function of this segment is calculated. For each segment of *NSEG* samples, the variance of the difference signal can be calculated directly from this short-term autocorrelation function [see eq. (4)]. Using these variances, a prediction gain can be determined for different numbers of predictor coefficients and for different values of *NSEG*. Figure 11 shows the optimum prediction gain for an adaptive prediction scheme versus the number of predictor coefficients. In each

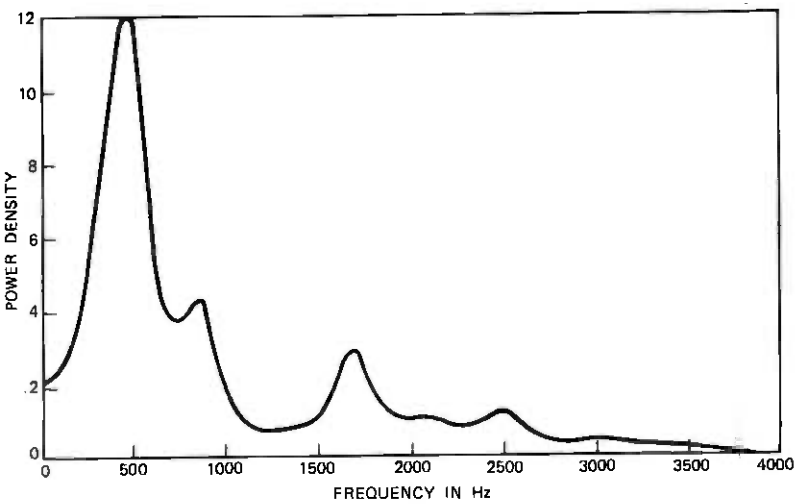


Fig. 12—Power density of the speech signal.

case, the optimum value *NSEG* has been used. The maximum prediction gain is approximately 14 dB. This value is an upper bound of the additional gain in s/n over PCM by using adaptive differential encoding schemes.

V. UPPER BOUNDS FOR QUANTIZATION

It is possible to design quantizers such that the signal-to-quantizing-noise ratio is a maximum; this is done by choosing the quantizer step sizes according to the probability density function of the signal. It is known that these optimum quantizers cannot be used for the quantization of speech signals: the s/n improvement is offset by the greater idle channel noise and smaller dynamic range (Ref. 4; see also Section 2.1 above). Optimum quantization is practical, however, if used in an adaptive quantization scheme; it gives us an s/n advantage over logarithmic quantization, and it allows a further increase in s/n by using entropy coding techniques (variable length coding). The adaptive quantization technique changes the PDF of the signal to be quantized; it has been shown[3] that different density functions can be reached with the forward estimation scheme (AQF scheme). Table II shows the s/n values for three-bit quantizers without and with entropy coding. The values of the first two columns are taken from Max[14] and Paez and Glisson.[15] In the case of entropy coding, the quantizers have been optimized so that the s/n is maximum for the given average bit rate of three bits per sample.[16] It is not possible to get higher s/n values with any encoding scheme based on memoryless single-letter quantization.

Table II — Maximum s/n values of various three-bit quantizers

	Quantizer Without Entropy Coding		Quantizer With Entropy Coding
	Optimum Uniform Quantizer	Optimum Nonuniform Quantizer	
	s/n(dB)	s/n(dB)	s/n(dB)
Uniform PDF	18.06	(18.06)	18.06
Gaussian PDF	14.27	14.62	16.53
Laplace PDF	11.44	12.61	17.09
Gamma PDF	8.78	11.47	18.78

VI. CONCLUSIONS

Comparisons of various nonadaptive and adaptive three-bit speech-encoding systems via simulation with speech inputs show that a wide range of signal-to-quantization-noise ratios can be reached starting with 9 dB (logarithmic PCM) and increasing up to 27 dB (adaptive predictive coding with adaptive quantization and entropy coding). Adaptive quantization has an s/n advantage of 7 or 5 db over logarithmic PCM when used in encoding schemes without and with prediction, respectively. Nonadaptive prediction leads to a 7-dB increase in s/n, and 11 dB can be gained using adaptive prediction techniques. Entropy coding gives an additional 2 to 3 dB improvement; such a coding technique is difficult to implement if a constant bit rate has to be achieved, but it may be of interest for asynchronous data networks. Furthermore, subjectively, DPCM gains over logarithmic PCM are believed to be greater than what the s/n gains suggest.[1]

Informal listening tests have shown that all predictive encoding schemes give a very good speech quality when used in connection with adaptive quantization (DPCM-AQ or ADPCM-AQ). Differences between the original speech and the decoded speech are not audible with adaptive prediction schemes when a high-order predictor is used (for example, ADPCM4-AQF).

The upper bounds that have been determined separately for the prediction gains and the quantizer s/n performances cannot be reached in practical predictive encoding systems. This fact is attributed to the predictor-quantizer interaction; that is, the input to the predictor is a noisy version of the input signal, and the input to the quantizer is a noisy prediction error. This interaction is not negligible when three-bit quantizers are used.

It is important to realize that all results are based on a single speech record of one speaker. Computer simulations using other speech

material show basically similar results; the main differences appear in the exact prediction gains that can be reached. In many instances, these gains are higher than those mentioned in this paper.

One object of this paper was to quantify the (relative and absolute) capabilities of a wide range of nonpitch-tracking speech coders. The coders studied have a variety of potential applications that call for different specifications of speech quality and coder complexity. A second purpose of this paper was to study the capabilities of three-bit encoding in some detail, as motivated by mobile telephone studies.[17,18]

REFERENCES

1. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," Proc. IEEE, *62*, No. 5 (May 1974), pp. 611–632.
2. P. Noll, "Untersuchungen zur Sprachcodierung mit adaptiven Prädiktionsverfahren," Nachrichtentechnische Zeitschrift, *27*, H. 2 (1974), pp. 67–72.
3. P. Noll, "Adaptive Quantizing in Speech Coding Systems," Proc. of the International Zürich Seminar on Digital Communications, 1974, pp. B3(1)–(6).
4. R. W. Stroh and M. D. Paez, "A Comparison of Optimum and Logarithmic Quantization for Speech PCM and DPCM Systems," IEEE Trans. on Commun., *COM-21*, No. 6 (June 1973), pp. 752–757.
5. R. Elsner, W. K. Endres, H. Mangold, P. Noll, E. Paulus, and D. Wolf, "Recent Progress in Digital Processing of Speech," IEEE Trans. on Commun., *COM-22*, No. 9 (1974), pp. 1168–1171.
6. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
7. N. S. Jayant, private communication.
8. P. Noll, "Nonadaptive and Adaptive Differential Pulse Code Modulation of Speech Signals," Polytech. Tijdschr., Ed. Elektrotech. Elektron. No. 19 (1972), pp. 623–629.
9. P. Noll, "Sprachübertragung mit adaptiven DPCM-Verfahren," Heinrich-Hertz-Institut, Berlin, Germany, Tech. Bericht 164, 1973.
10. R. W. Stroh, "Optimum and Adaptive Differential Pulse Code Modulation," Ph.D. thesis, Polytechnic Institute of Brooklyn, 1970.
11. B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," B.S.T.J., *49*, No. 8 (October 1970), pp. 1973–1986.
12. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–343.
13. J. P. Burg, "Maximum Entropy Spectral Analysis," Proc. NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede (Netherlands), 1968.
14. J. Max, "Quantizing for Minimum Distortion," IRE Trans. Inform. Theory, *IT-6*, No. 1 (March 1960), pp. 7–12.
15. M. D. Paez and T. H. Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems," IEEE Trans. on Commun. Tech., *COM-20*, No. 2 (April 1972), pp. 224–230.
16. P. Noll and R. Zelinski, "Ein Beitrag zur Quantisierung gedächtnisfreier Modellquellen," Heinrich-Hertz-Institut Berlin-Charlottenburg, Tech. Report No. 170, 1974.
17. P. Noll, "Effects of Channel Errors on the Signal-to-Noise Performance of Speech-Encoding Systems," B.S.T.J., this issue, pp. 1615–1636.
18. N. S. Jayant, "Step-Size Transmitting Differential Coders for Mobile Telephony," B.S.T.J., this issue, pp. 1557–1581.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

Effects of Channel Errors on the Signal-to-Noise Performance of Speech-Encoding Systems

By P. NOLL

(Manuscript received February 7, 1975)

The signal-to-noise ratios of different speech-encoding schemes have been measured in the case where the channel contains errors. Those types and probabilities of errors have been considered that are of interest for mobile telephone applications. Most encoding schemes use an adaptive three-bit quantizer with an explicit transmission of the step-size information. A scheme with an adaptive prediction algorithm has also been studied. It has been assumed in all cases that the side information about the quantizer step size and the predictor coefficients is transmitted in an error-protected format.

Measurements were made by simulating the coding schemes and the noisy channel on a digital computer. The results include upper bounds of the improvements that can be reached with error protection of the most significant bits.

I. INTRODUCTION

The suitability of digital coding and transmitting speech signals for mobile telephone systems is a question of current interest. In UHF systems, Rayleigh fading causes the carrier-to-interference ratio and the carrier-to-noise ratio to be low in frequent intervals. This leads to high bit-error probabilities in the transmission of the coded signal; the errors occur in bursts.

This paper compares the effects of channel transmission errors on the objective signal-to-noise ratio (s/n) for different encoding schemes. Most of these schemes use an adaptive quantizer with time-varying step sizes or, equivalently, a time-varying gain control of an amplifier in front of a quantizer with fixed step sizes. The side information about the step size (or about the amplifier gain) is derived from stored samples of the input signal and has to be transmitted together with the message block of coded samples (adaptive quantization with forward estimation = AQF). These AQF schemes have an excellent idle channel performance, even in the presence of channel errors, if the side informa-

tion can be transmitted in an error-protected format. Previous studies[1–3] always assumed a nonadaptive μ-logarithmic quantizer with its specific problems of allowable peak clipping and changing performance caused by different mean levels of the speech signal. The purpose of the present study is to show the s/n performance of some adaptive speech-encoding schemes suitable for mobile telephone applications. Our measurements were made by simulating the coding schemes and the noisy channels on a digital computer. The measurements include upper bounds for the s/n that can be reached by using error-protection schemes to reduce the effective channel error probability. We have studied independent, as well as clustered, channel error patterns. No attempt was made to study practical error-detection or error-correction schemes. However, the results allow us to predict the overall performance that can be reached with nonideal error-correction schemes. The s/n values of the coding systems with independently distributed channel errors were measured because it is possible to produce a nearly equivalent error pattern by scrambling the bit stream at the output of the transmitter.[6] Additionally, independent errors will be the primary impairment if the interference is low.

The organization of this paper is as follows. Section II discusses the dependence of the total s/n on the bit-error rate P, assuming gaussian-distributed quantizer input data. It has been shown[4] that the gaussian probability density function is a good approximation for signals occurring in speech AQF schemes. Both the natural binary code (NBC) and the folded (symmetrical) binary code (FBC) are considered. It is shown that the FBC code has a better s/n performance if channel transmission errors cannot be ignored. Section III discusses those speech-encoding systems used in this study, and Section IV considers the types of errors on the channel and gives the main results obtained by simulating the encoding schemes and the noisy channels on a digital computer. Comparison is made on the basis of the objective overall signal-to-noise ratio. Tape-recorded examples were used to compare the subjective and perceptual effects of signal-quantization and channel errors. Some conclusions are given in Section V.

A detailed comparison of various speech-encoding schemes on the assumption of an error-free transmission is described in another paper.[5] A companion paper by Jayant[6] gives numerous examples of the performance of practical error-protecting schemes in the presence of channel errors.

II. DERIVATION OF THE SIGNAL-TO-NOISE RATIO FOR GAUSSIAN SIGNALS

We calculate the overall signal-to-noise ratio of a three-bit PCM-quantization scheme on the assumption that a speech signal can be

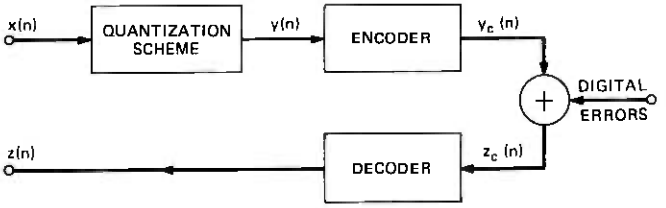


Fig. 1—Digital transmission system.

represented by a gaussian probability density function (PDF) at the quantizer input. This assumption is approximately valid if adaptive quantization with forward estimation is applied to the speech samples and if the message blocks are not too long.[4] Figure 1 shows a block diagram of the system under consideration. The quantizer with $M = 8$ steps (and $m = \log_2 M$ bits per code word) maps each input sample $x(n)$ into one of a set of eight rational numbers $y(n) \epsilon \{v_k\}_{k=1,2,\cdots,M}$. The representation level v_i is chosen if $u_{i+1} \geqq x(n) > u_i$, as illustrated in Fig. 2. The index i of the input symbol v_i of the transmission system is transmitted to the receiver in a binary format [binary code word $y_c(n)$]; the received code word $z_c(n)$ is interpreted as one of the eight output symbols $z(n) \epsilon \{w_k\}_{k=1,2,\cdots,M}$. We obtain a change $\delta_{ij} = |v_i - w_j|$ in amplitude if the transmitted quantizer index i is changed to j because of channel errors (Fig. 3). The total mean-squared error is

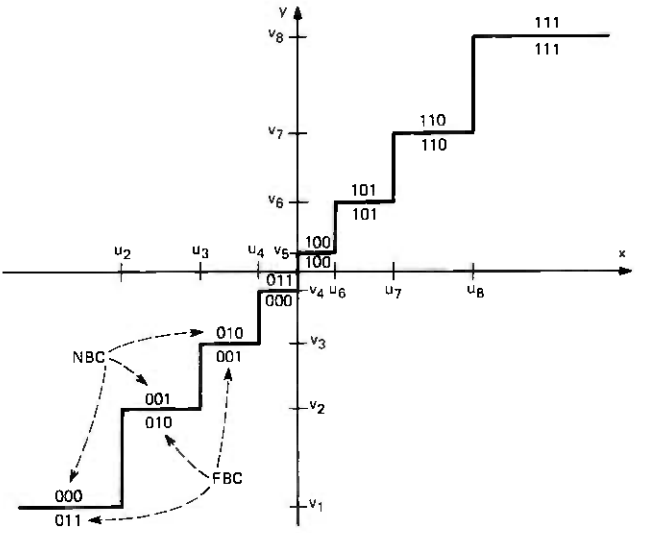


Fig. 2—Symmetric nonuniform quantizer. $m = 3$ bits, $M = 8$ steps, NBC = natural binary code, FBC = folded binary code.

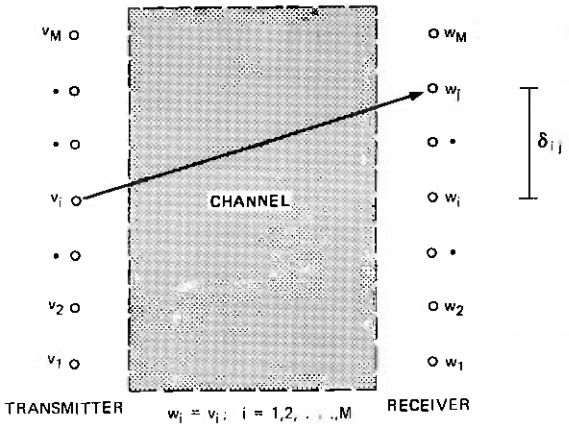


Fig. 3—Channel transmission errors.

given by

$$\epsilon_t^2 = E[x(n) - z(n)]^2 = \epsilon_q^2 + \epsilon_c^2, \tag{1}$$

where the quantization error ϵ_q^2 and the channel error ϵ_c^2 are given by

$$\epsilon_q^2 = E[x(n) - y(n)]^2 \tag{2}$$

$$\epsilon_c^2 = E[y(n) - z(n)]^2. \tag{3}$$

Equation (1) is only true on the assumption of a vanishing correlation between quantization error and channel error;[7] this is the case if the quantizer structure is that of Max[8] (these quantizers lead to a maximum s/n performance for a given PDF).

The mean-squared error caused by digital line errors is

$$\epsilon_c^2 = \sum_{i=1}^{M} \sum_{j=1}^{M} P(v_i, w_j) \cdot \delta_{ij}^2. \tag{4}$$

With

$$P(v_i, w_j) = P(v_i) \cdot P(w_j/v_i), \tag{5}$$

we have

$$\epsilon_c^2 = \sum_{i=1}^{M} \sum_{j=1}^{M} P(v_i) \cdot P(w_j/v_i) \cdot \delta_{ij}^2 = \text{trace } \{\mathbf{P}_v \cdot \mathbf{P}_c \cdot \boldsymbol{\delta}^2\}, \tag{6}$$

where

$P(v_i, w_j)$ is the joint probability of an input symbol v_i at the transmitter and an output symbol w_j at the receiver,

$\delta_{ij} = \lvert v_i - w_j \rvert$	is the amplitude of the error occurring if the input symbol v_i has been chosen at the transmitter and if the output symbol w_j has been interpreted at the receiver,
$P(v_i)$	is the probability of input symbol v_i occurring,
$P(w_j/v_i)$	is the conditional probability that the output symbol w_j will be received if the input symbol v_i is sent,
$\mathbf{P}_v$	is a diagonal matrix with elements $P(v_i)$,
$\mathbf{P}_c$	is the channel transition matrix with elements $P(w_j/v_i)$, and
$\boldsymbol{\delta}^2$	is the matrix of squared error amplitudes with elements δ_{ij}^2.

The conditional probabilities $P(w_j/v_i)$ can be calculated easily if the bit errors on the channel are distributed independently. The values depend on the bit-error probability P and on the code. The probability $P(w_j/v_i)$ that w_j will be received if v_i is sent is just the probability that digital errors will occur in the D places where they differ and that no errors will occur in the $m - D$ remaining places,

$$P(w_j/v_i) = P^D(1 - P)^{m-D}, \tag{7}$$

where P is the bit-error probability on the channel, D is the Hamming distance between the code words representing the symbols v_i and w_j, and m is the number of bits per code word. The Hamming distance D depends on the code; two codes have been considered (see Fig. 2): (*i*) the natural binary code (NBC), (*ii*) the folded binary code (FBC). For this code, the most significant bit gives polarity information; the remaining bits represent the signal magnitude in natural binary code. For example, the transition from input symbol v_1 to output symbol $w_8 = v_8$ causes an error δ_{18} in amplitude. The Hamming distances are $D = 3$ and $D = 1$ for the NBC code and FBC code, respectively (see Fig. 2). Using (7) we get $P(w_8/v_1) = P^3$ with the NBC code and $P(w_8/v_1) = P(1 - P)^2$ with the FBC code.

The mean-squared error caused by digital line errors can also be calculated if (ideal) error-protection schemes are applied:

Scheme EP1: The most significant bit of each code word is perfectly error-protected.

Scheme EP2: The two most significant bits are perfectly error-protected.

The channel transition matrix $\mathbf{P}_c$ has to be modified in these cases because some elements of the matrix are zero then. These necessary modifications are not described in this paper. Using (1) and (6), we

obtain the signal-to-noise ratio

$$\text{s/n} = 10 \cdot \log_{10} \frac{\sigma_x^2}{\epsilon_l^2} = 10 \cdot \log_{10} \frac{\sigma_x^2}{\epsilon_q^2 + \text{trace}\,\{\mathbf{P}_v \cdot \mathbf{P}_c \cdot \boldsymbol{\delta}^2\}}, \tag{8}$$

where

$$\sigma_x^2 = E[x^2(n)] \tag{9}$$

is the mean-squared power of the input signal. The normalized quantization noise variance ϵ_q^2/σ_x^2 has a value of 0.03451 if the structure of the three-bit quantizer is that of Max.[8] Using this value, the s/n performance has been calculated as a function of the bit-error rate P (Fig. 4). The two lower curves show that the FBC code outperforms the NBC code. The folded binary code has therefore been used in all simulations. The upper curves demonstrate the advantage of an (ideal) error protection of the most significant bit (EP1) and of the two most significant bits (EP2). It may be relevant to mention that this error protection leads to a reduction of the effective bit-error rate. To confirm the theoretical results, we have measured some s/n values with the simulation program that has been used for the study of the speech-

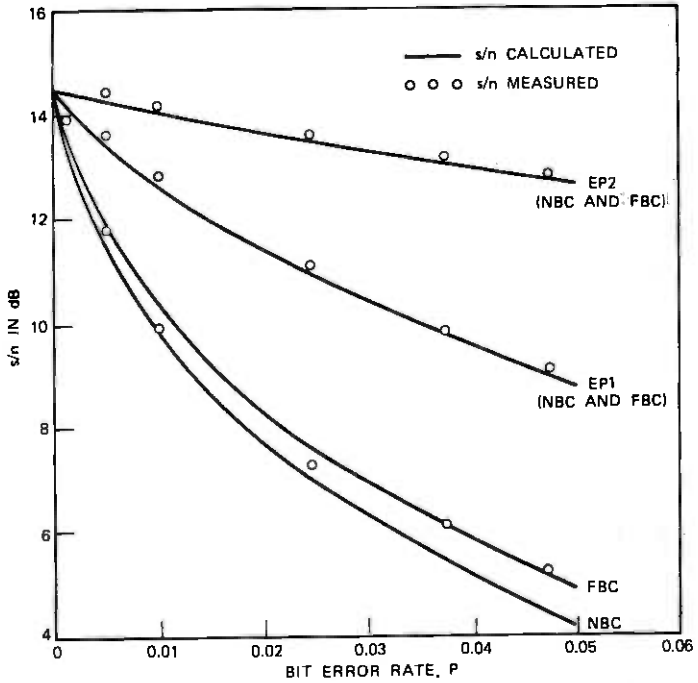


Fig. 4—Signal-to-noise ratio performance of a gaussian data transmission system. NBC = natural binary code, FBC = folded binary code, EP1 = error protection of most significant bit, EP2 = error protection of two most significant bits.

Table I — Channel coefficients of optimum three-bit Gauss quantizers. NBC = natural binary code, FBC = folded binary code.

	α_1	α_2	α_3
Optimum nonuniform quantizer			
NBC	6.9124	−2.7373	−0.3133
FBC	5.7445	−0.4015	−0.3133
Optimum uniform quantizer			
NBC	7.2113	−3.3611	0.
FBC	5.5672	−0.0128	0.

encoding systems. The results obtained with gaussian input data and the FBC code compare favorably with those determined using (8) (see Fig. 4).

The channel error variance ϵ_c^2 as given in (6) can be transformed into

$$\epsilon_c^2 = \sum_{j=1}^{m} \alpha_j \cdot P^j. \tag{10}$$

The coefficients α_j contain the total information about the effects of channel errors on the performance of an encoding-decoding scheme. Thus, different schemes can be compared easily. Table I lists the α_j coefficients of optimum uniform and nonuniform gaussian three-bit quantizers. Note that there are only very small differences between nonuniform and uniform quantizers. Note also that the FBC code should be chosen; it has a nearly 1-dB advantage over the NBC code if the bit-error rate is very high (see also Fig. 4).

In the discussions so far, a quantizer has been assumed that is optimum (in the sense of a maximum s/n) if the channel is error-free. A higher overall s/n performance can be reached, however, if the quantizer is reoptimized for a given channel transition matrix $\mathbf{P}_c$.[9] Figure 5 shows an example with a three-bit quantizer optimum for a bit-error rate of 0.025 and the NBC code. In fact, we get a better s/n performance if the bit-error rates P are high, but the shortcoming is the decrease in s/n for low P values; this reduction is approximately 2.3 dB in the error-free case ($P = 0$). The reoptimization of the quantizer can therefore only be of interest if the channel noise statistics are time-invariant, or if the s/n decrease for small error rates can be accepted. This will be the case if a quantizer with a higher number of step sizes is chosen.

The principal aim of this section was to show the calculation of the s/n of a (nonadaptive) PCM scheme if channel errors are present. These calculations can be extended to DPCM systems and to burst error

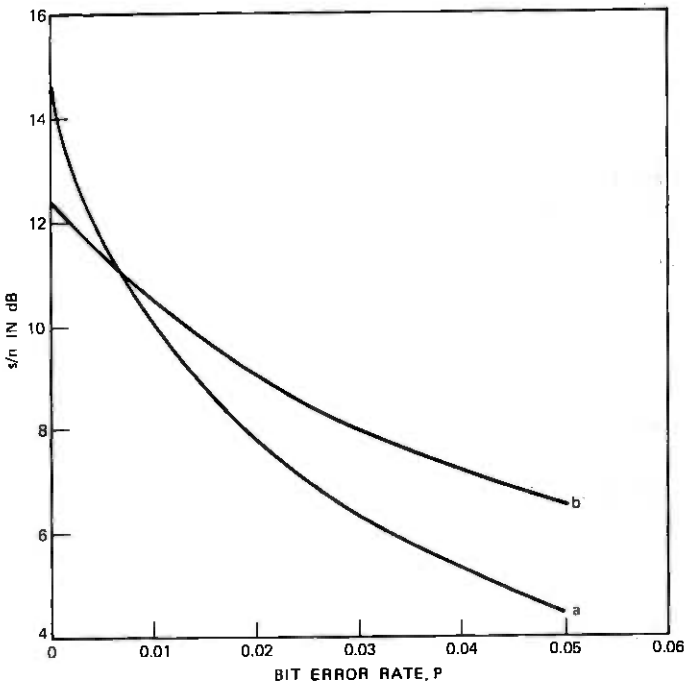


Fig. 5—Signal-to-noise ratio performance of two nonuniform gaussian three-bit quantizers. (a) Quantizer is optimized for an error-free transmission. (b) Quantizer is optimized for independent channel errors of rate $P = 0.025$. NBC code has been used in both cases.

channels. But it seems very difficult to find theoretical solutions for coding schemes that apply adaptive quantization and adaptive prediction strategies. Instead of looking for such solutions, we simulated different encoding schemes and channels on a digital computer and measured the overall s/n. We shall find some similarities to the results we obtained in this section. Particularly, we shall use the channel coefficients α_j to compare the performance of the coding schemes.

III. DESCRIPTION OF THE CODING SYSTEMS

To get a good quality of the coded speech with low bit rates, we have used PCM and differential PCM schemes that employ adaptive quantizers. Both nonadaptive prediction (DPCM) and adaptive prediction (ADPCM) have been applied. The advantage of adaptive quantization is that the quantizer is always adjusted to the highly variable variance of the speech signals. Thus, a better s/n performance is achieved.[4,10] DPCM and ADPCM schemes provide an additional s/n gain over PCM; this is especially true if the predictor responds to changes in the short-term

spectrum of speech (ADPCM).[10,11] A nonadaptive logarithmic companded PCM has been included in our study because it very often serves as a standard reference in coder comparisons. A great number of speech-encoding schemes have been compared in a companion paper[5] on the basis of s/n as performance measure using the same speech signal employed in this paper. The effect of channel errors on s/n performance has been studied using the following encoding schemes:

Scheme 1: *PCM, nonadaptive* (*Fig. 6a*). The quantizer has a $\mu 100$ characteristic,[12] and the loading is four times the standard deviation of the speech signal to be quantized.

Scheme 2: *PCM-AQF* (*Fig. 6b*). Thirty-two samples of the input signal are buffered, and the maximum value of this block determines the gain of the amplifier in front of the quantizer (adaptive quantization

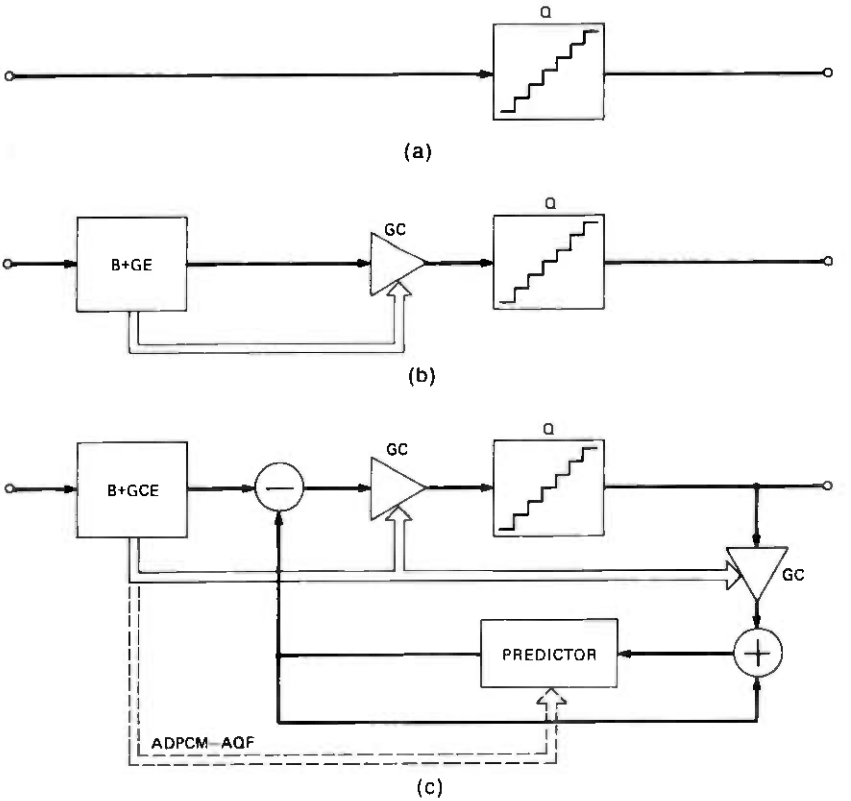


Fig. 6—Speech-encoding schemes. (a) Nonadaptive PCM. (b) PCM with adaptive quantization (PCM-AQF). (c) DPCM and adaptive DPCM (ADPCM) with adaptive quantization (DPCM-AQF and ADPCM-AQF). Q = quantizer, GC = gain control, B + GCE = buffer and gain and coefficients estimation.

with forward estimation = AQF). The characteristic of the quantizer is optimum for signals with a gaussian probability density.

Scheme 3: *DPCM1-AQF* (*Fig. 6c*). A predictor with one time-invariant coefficient is being used in connection with adaptive (forward estimation) quantization (AQF). Thirty-two samples of the input signal are buffered, and the maximum difference between neighbored samples determines the gain of the amplifier in front of the quantizer. The characteristic of the quantizer is optimum for signals with a gaussian probability density. The predictor coefficient that is optimum for the speech signal being used is $h_1 = 0.85$. Lower values lead to a better performance of the DPCM scheme in the case of high error probabilities; this will be shown in Section V.

Scheme 4: *ADPCM1-AQF* (*Fig. 6c*). In this adaptive prediction scheme, 32 samples of the input signal are buffered, the normalized short-term correlation coefficient between neighbored samples of this block is calculated, and the predictor coefficient is set to this correlation coefficient. The gain of the amplifier in front of the quantizer is determined by calculating an estimation value of the standard deviation of the difference signal; the amplifier gain is set to the inverse of this estimation value. The characteristic of the quantizer is optimum for signals with a gaussian probability density.

Scheme 5: *ADPCM4-AQF* (*Fig. 6c*). In this adaptive prediction scheme, four optimum predictor coefficients are calculated for each segment of 32 samples from the first values of the short-term autocorrelation function; see the description of Scheme 4 for further details.

The folded binary code (FBC) was used in all cases. It should be mentioned that the combination of controlled amplifier and fixed quantizer in the adaptive quantization schemes is equivalent to a quantizer with a step-size adaptation. Some adaptive quantization schemes that use the transmitted code words for the control of the amplifier gain (adaptive backward estimation) have also been studied. The simulations have shown that these schemes cannot be used for channels with high bit-error probabilities; the overall s/n turned out to be less than 0 dB in most cases.

IV. ERROR PERFORMANCE: RESULTS AND DISCUSSION

4.1 Simulation system and types of errors

The dependence of the overall signal-to-noise ratios of five speech coding schemes (see Section III) on the average bit-error probability P has been determined for different types of noisy channels. The s/n is given by

$$\text{s/n}(P) = 10 \log_{10} \frac{\sigma_x^2}{\epsilon_t^2}, \tag{11}$$

where ϵ_i^2 and σ_x^2 are defined in (1) and (9), respectively. The s/n values have been measured for bit-error rates of 0, 0.001, 0.0125, 0.025, and 0.05. The measurements were made by simulating the coding schemes and the noisy channels on a digital computer. Channels with independent, as well as correlated, error patterns have been studied. The statistically independent errors have been generated by using a pseudo-random noise generator program. Tape recordings with error patterns of actual fading signals have been used in the channel simulation of burst errors. The error patterns are typical for UHF mobile radio transmission. The statistics of these errors are described in Ref. 6. In all simulations, it has been assumed that it is possible to transmit the information about the gain of the amplifier (AQF scheme) and/or about the predictor coefficients (adaptive prediction) without any error. Increased signal-to-noise ratios can be reached for a given P value using error-detection and error-correction schemes. In studying these error-protected cases, it has been assumed that all errors are corrected. Practical schemes will not always be able to correct all errors. Therefore, the s/n values given in this paper for the error-protected case represent an upper bound on the performances of error-protecting techniques. Two types of error correction have been studied:

EP1: Protection of the most significant bit; this bit is the sign bit.
EP2: Protection of the two most significant bits. Only changes to neighbored output symbols are possible in this case (if the quantizer has eight step sizes).

A 2.3-second utterance ("the boy was mute about his task"; female voice; bandwidth 200 to 3200 Hz; sampling rate 8 kHz) has been used in all simulations.

4.2 Results

The s/n performances of the coding schemes that have been described in Section III have been measured using three-bit quantizers and the folded binary code. Figures 7 to 11 show the measured dependence of the s/n on the average bit-error rate P in the case of burst errors. The lower, middle, and upper curves refer to the unprotected transmission and to the EP1 and EP2 error-protection schemes; note that the effective bit-error rate is reduced then. We show this using Fig. 9 as an example. The s/n value for $P = 0$ is due to the quantization noise only. The lower curve shows a considerable loss in s/n for high bit-error rates P. An increase in s/n can be obtained by protecting the most significant bit (EP1; middle curve) or the two most significant bits (EP2; upper curve). For example, if $P = 0.025$, the s/n value without error protection is 10.4 dB. A value of 14 dB is obtained with

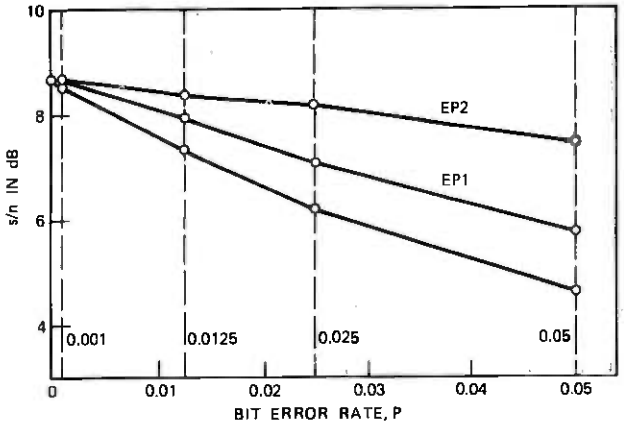


Fig. 7—Signal-to-noise ratio performance of a three-bit log PCM scheme in the presence of correlated errors.

the protection of the most significant bit (EP1). The effective bit-error rate is reduced to $\frac{2}{3} \times 0.025 = 0.0167$ in this case because $\frac{1}{3}$ of the errors are assumed to be corrected at the receiver. The 14-dB value of the EP1 curve is 2 dB higher than the s/n value we get for $P = 0.0167$

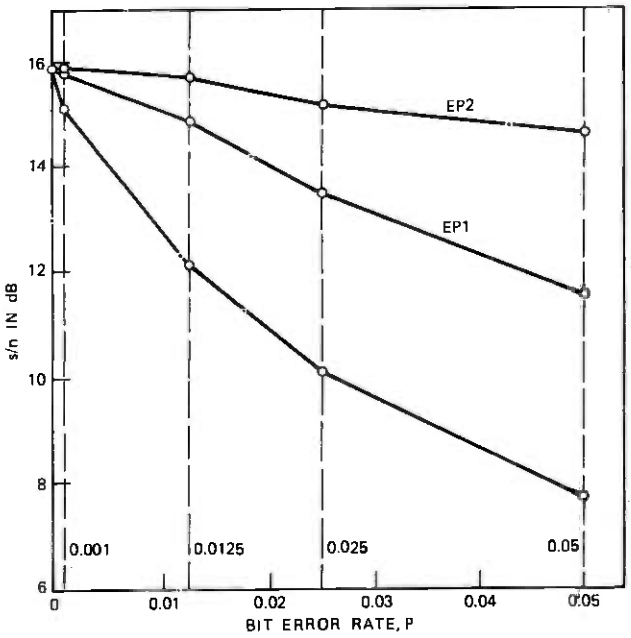


Fig. 8—Signal-to-noise ratio performance of a three-bit PCM-AQF scheme in the presence of correlated errors.

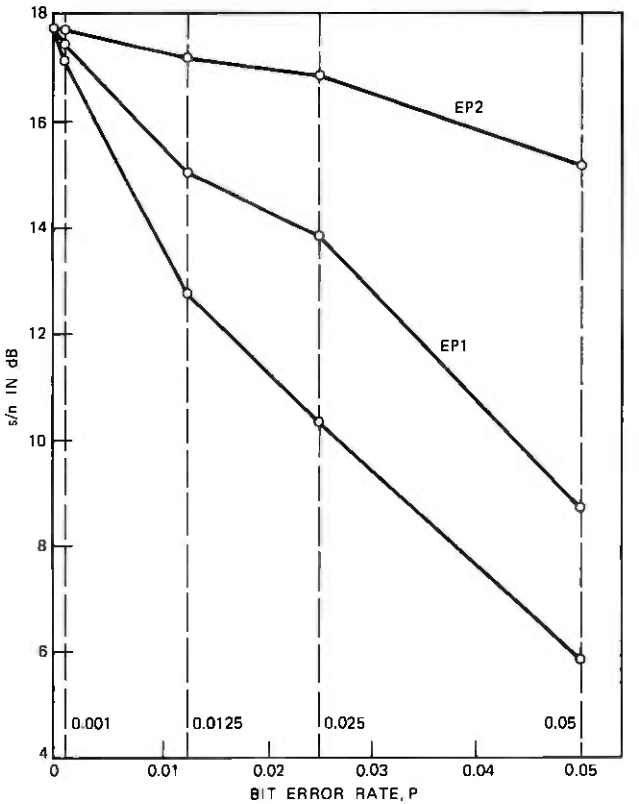


Fig. 9—Signal-to-noise ratio performance of a three-bit DPCM1-AQF scheme in the presence of correlated errors. The predictor coefficient has a value $h_1 = 0.6$.

on the lower curve (no error protection). We expect this result because no errors occur on the most significant bit for the effective 0.0167 bit-error rate of the EP1 curve. An s/n value of 16.9 dB is obtained with the protection of the two most significant bits (EP2). The s/n value for the effective bit-error rate of $\frac{1}{3} \times 0.025 = 0.083$ is 14.3 dB if the errors occur on all bits of the code words (lower curve); therefore, a 2.6-dB increase in s/n is due to the fact that only the least significant bits are affected.

Error protection, however, is only possible by inserting redundancy into the code words. Let us assume that it is possible to obtain an error protection of the two most significant bits by using three redundant bits for each three-bit code word. The total bit rate is now 6 bits per sample. Let us further assume that doubling the transmission rate causes doubling the bit-error rate (this is true for phase-modulation systems). Again using Fig. 9, we find an s/n value of 15.2 dB for

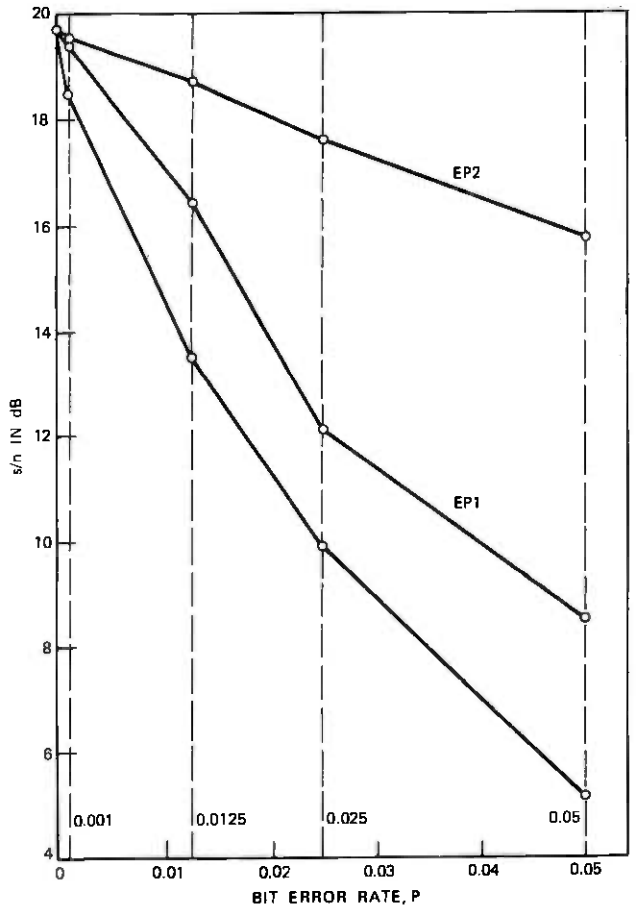


Fig. 10—Signal-to-noise ratio performance of a three-bit ADPCM1-AQF scheme in the presence of correlated errors.

$P = 0.05$ (EP2 curve). On the other hand, the s/n value without error protection is 10.4 dB for $P = 0.025$. Therefore, an improvement of nearly 5 dB over the transmission without error protection has been obtained. A similar discussion using the EP1 values shows that we get only a small s/n advantage then: an error protection of the sign bit is not sufficient for improving the overall performance.

To better compare the performances of the coding schemes in the presence of errors, we have plotted the s/n values of these schemes with P as a parameter (Fig. 12). The s/n values for $P = 0$ are due to the quantization errors only; the increase in s/n as compared to log-PCM starts with 7 dB (PCM-AQF) and goes up to 14 dB (ADPCM4-AQF). At

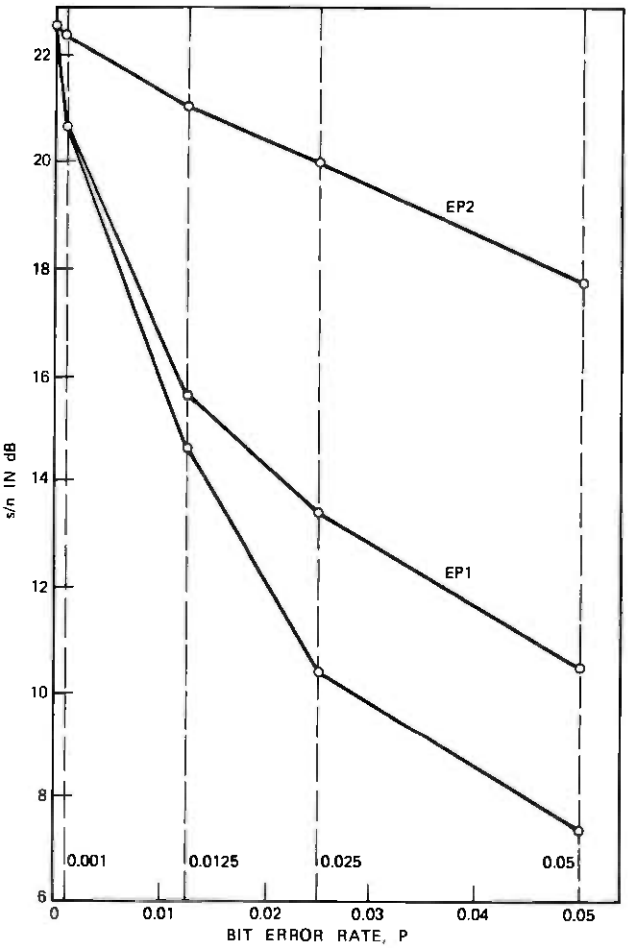


Fig. 11—Signal-to-noise ratio performance of a three-bit ADPCM4-AQF scheme in the presence of correlated errors.

higher bit-error rates, the different coding schemes have approximately identical performance because the output noise due to channel errors predominates over the quantization noise due to the quantizer.

Recall from (1) that the total error variance can be expressed as the sum of the quantization error variance ϵ_q^2 and the channel error variance ϵ_c^2 if the mutual error is neglected. The term ϵ_q^2 can be determined from the s/n for $P = 0$; hence, we can separate the values ϵ_c^2 for the four bit-error rates P that have been used in the simulations; these are the values $P = 0.001$, 0.0125, 0.025, and 0.05. The channel error can be

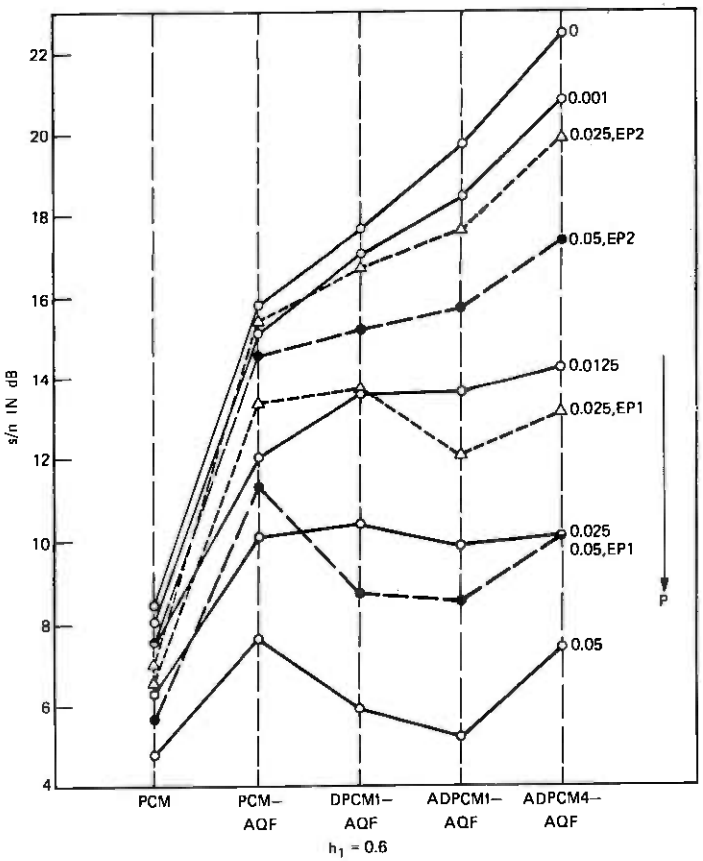


Fig. 12—A comparison of the s/n performance of three-bit encoding schemes at different bit-error rates P (correlated errors).

expressed approximately as

$$\epsilon_c^2 = \alpha_1 P + \alpha_2 P^2 \tag{12}$$

because the third term of (10) can be neglected.

The coefficients α_1 and α_2 have been determined using the measured data by searching for the minimum of the mean-squared differences between measured and calculated s/n values. These coefficients α_1 and α_2 describe the effect of channel errors on the performance of the coding schemes. The total channel error variance is mainly determined by the channel error coefficient α_1. Figure 13 shows that the α_1 values of those encoding schemes that use the same gaussian quantizer (all AQF schemes) are not very different. From this, we conclude that transmission errors are no more serious for DPCM and ADPCM schemes than

for PCM; this has already been mentioned for DPCM in Ref. 2. The channel error performance is better for burst errors than for independently distributed errors. This has partly to do with the fact that some bursts appear in low-level parts of the speech sample. On the other hand, we cannot neglect the α_2 term in the case of burst errors; α_2 is the coefficient of the P^2 term in (12); this term mainly represents the channel error contribution caused by two bit errors in a code word.

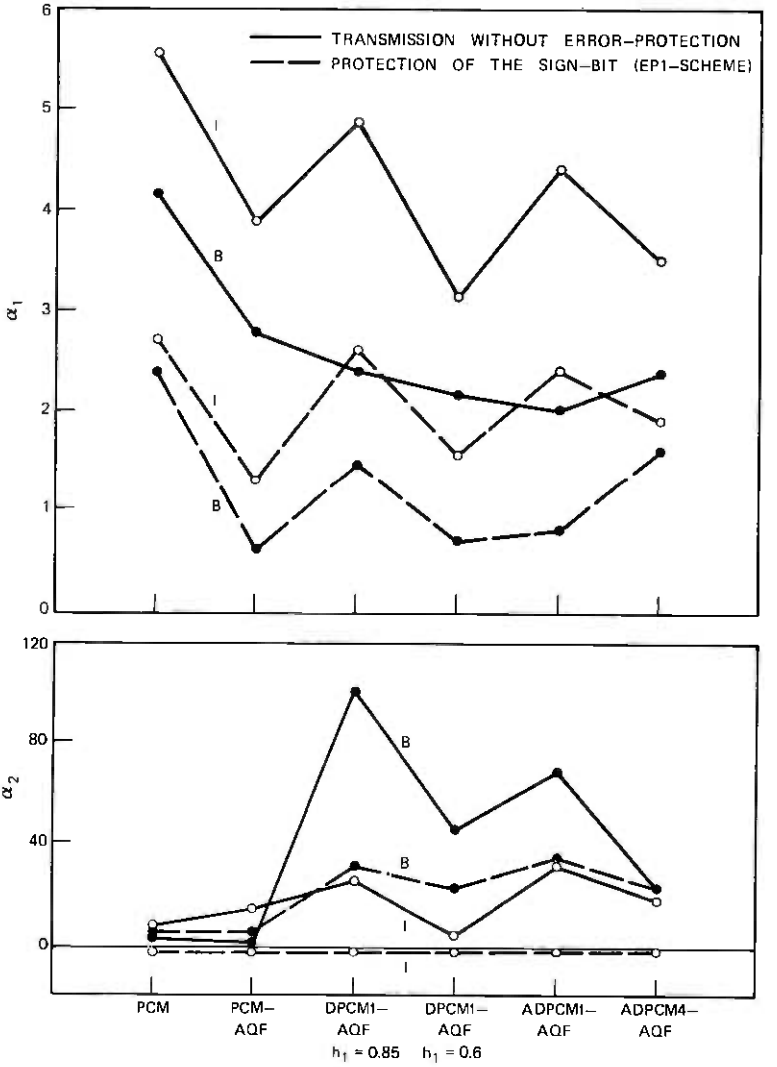


Fig. 13—Channel coefficients α_1 and α_2. I = independent errors, B = burst errors.

This contribution is higher in the case of burst errors; it causes a stronger decrease in signal-to-noise ratio as can be seen in Fig. 14 showing the s/n performance of a DPCM1-AQF scheme both for independent and correlated errors. We find the same tendency if we protect the most significant bit (EP1; see Fig. 13) or the two most significant bits. We have used the average α_1 values to calculate the s/n increase if we apply (perfect) error protection: the increases are approximately 3 and 11 dB for the EP1 scheme and EP2 scheme, respectively. Note, from Fig. 13, that the α_2 term can be neglected in the case of PCM schemes; we therefore have a slower decrease in s/n at high bit-error rates. This can also be seen from Fig. 15, where we compare the s/n performance of a PCM-AQF scheme with DPCM1-AQF schemes that have different values of the predictor coefficient. Lowering this value, we obtain a higher channel error resistance, but PCM-AQF outperforms the

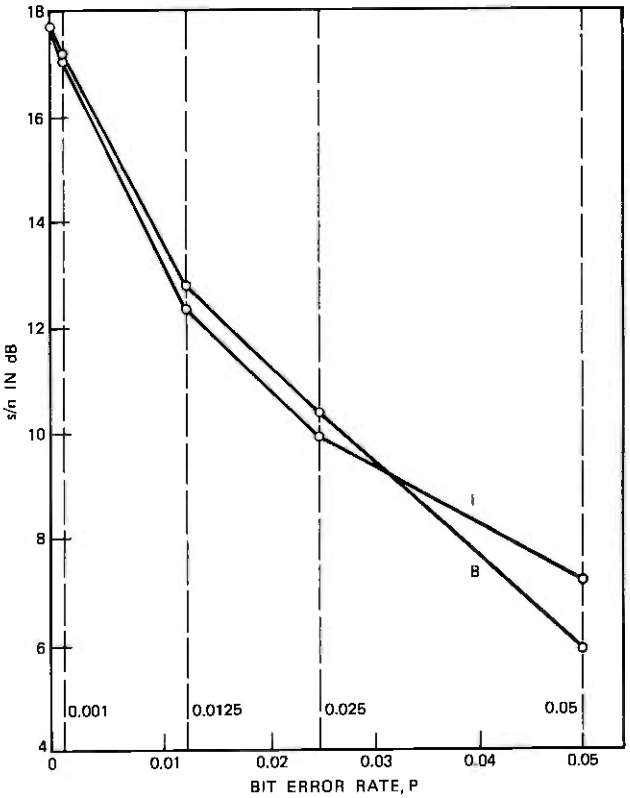


Fig. 14—Signal-to-noise ratio performance of a three-bit DPCM1-AQF scheme in the presence of independent errors (I) and burst errors (B). The value of the predictor coefficient is $h_1 = 0.6$.

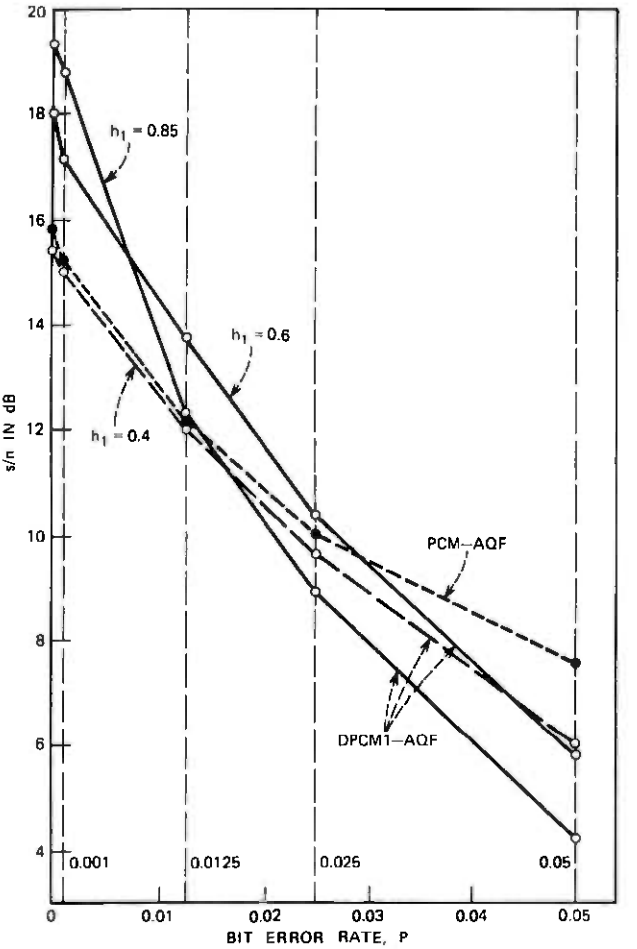


Fig. 15—Comparison of the s/n performance of DPCM1-AQF schemes with different values h_1 of the predictor coefficient with a PCM-AQF scheme.

DPCM1-AQF scheme if the predictor coefficient is too low (note the different slopes of the curves at high bit-error rates).

Our simulations involved not only three-bit quantization but also quantization schemes with a greater number of step sizes. Figure 16 illustrates a typical example: the signal-to-noise ratios of the four-bit quantization schemes are nearly 6 dB higher than the signal-to-noise ratios of the corresponding three-bit quantization schemes if the channel is error-free. But this increase is lost in the presence of high channel error rates: all three-bit and four-bit systems have a similar s/n performance for high bit-error rates. Higher s/n values can only

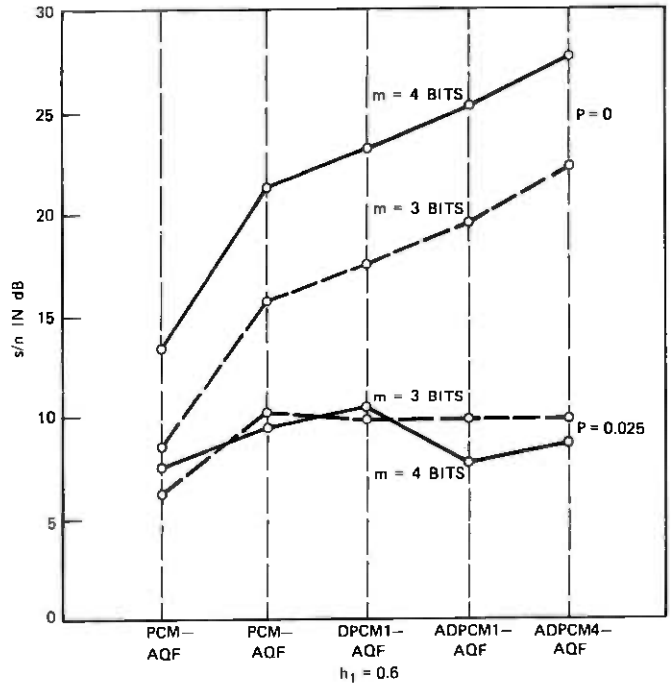


Fig. 16—Comparison of three- and four-bit encoding schemes.

be reached by error protection, that is, by reducing the effective bit-error rate.

V. CONCLUSIONS

In this paper, we have determined the s/n performance of various speech-encoding schemes in the presence of high bit-error rates (up to 5 percent); both independent and correlated error patterns have been used. Some conclusions can be drawn from the quantitative results that have been obtained and from informal subjective listening tests.

(*i*) It is possible, without pitch tracking, to quantize speech signals with three bits per sample such that the decoded signal is nearly indistinguishable from the original signal (adaptive prediction schemes in connection with adaptive quantization). A simple scheme with a fixed predictor (one coefficient) and an adaptive quantization can be chosen for a bit rate of four bits per sample.

(*ii*) Adaptive quantization lowers the idle channel noise and increases the s/n (the three-bit quantizer with a logarithmic characteristic has a very poor performance).

(*iii*) Only the adaptive quantization schemes (AQF schemes) with an explicit error-protected transmission of the step-size information can be used in the case of high channel error probabilities.

(*iv*) The folded binary code (FBC) outperforms the natural binary code (NBC) for large bit-error rates.

(*v*) Burst errors cause a stronger decrease in signal-to-noise ratio for large bit-error rates than independent errors.

(*vi*) Transmission errors are no more serious for DPCM and ADPCM schemes than for PCM.

(*vii*) All coding schemes show approximately the same s/n performance for high bit-error rates, because the contribution of the noisy channel to the total error is much higher than the contribution of the quantizer. Therefore, a better s/n performance can only be reached by using error-protection schemes, not by increasing the number of quantizer steps.

(*viii*) A high-quality decoded signal can be obtained with a protection of the two most significant bits. An improvement in decoded signal quality can be realized even if a doubling of the bit-error rate (caused by the higher transmission rate) has to be tolerated.

(*ix*) Significant-bit-packed codes that provide only protection of the sign bit (EP1 scheme) are not very efficient.

(*x*) Adaptive quantization schemes suppress the idle channel noise; therefore, channel errors produce decoded noise only with very small amplitudes in silent intervals. This fact makes the decoded speech perceptually more pleasing.

(*xi*) Nonadaptive and adaptive prediction schemes have a better perceptual quality than PCM schemes when bit errors occur on the channel. The power density spectrum of the error sequence is shaped in the DPCM or ADPCM feedback loop such that the main contribution of the error is in the low-frequency range. This error spectrum is perceptually less objectionable.

It is important to realize that the numerical results of this paper are based on a single speech record of one speaker. However, we expect the broad conclusions of this paper, as summarized above, to be true of a wide range of input speech material.

VI. ACKNOWLEDGMENTS

This work has been carried out in the Department of Acoustics Research during a summer visit. The author would like to express his very sincere thanks to many members of this department for their help, and especially to J. L. Flanagan as the head of this department for his support of this work. Special mention must be made of the benefit of stimulating discussions with N. S. Jayant and R. B. Kuc.

REFERENCES

1. I. Dostis, "The Effects of Digital Errors on PCM Transmission of Compandored Speech," B.S.T.J., *44*, No. 10 (December 1965), pp. 2227–2243.
2. K. Chang and R. W. Donaldson, "Analysis, Optimization, and Sensitivity Study of Differential PCM Systems Operating on Noisy Communication Channels," IEEE Trans. on Communications, *COM-20*, No. 3 (1972), pp. 338–350.
3. I. Yan and R. W. Donaldson, "Subjective Effects of Channel Transmission Errors on PCM and DPCM Voice Communication Systems," IEEE Trans. on Communications, *COM-20*, No. 3 (1972), pp. 281–290.
4. P. Noll, "Adaptive Quantizing in Speech Coding Systems," Proc. of the International Zürich Seminar on Digital Communications, 1974, pp. B3(1)–(6).
5. P. Noll, "A Comparative Study of Various Quantization Schemes for Speech Encoding," B.S.T.J., this issue, pp. 1597–1614.
6. N. S. Jayant, "Step-Size Transmitting Differential Coders for Mobile Telephony," B.S.T.J., this issue, pp. 1557–1581.
7. R. E. Totty and G. C. Clark, "Reconstruction Error in Waveform Transmission," IEEE Trans. Information Theory (Correspondence), *IT-13*, 1967, pp. 336–338.
8. J. Max, "Quantizing for Minimum Distortion," IRE Trans. Information Theory, *IT-6*, 1960, pp. 7–12.
9. A. J. Kurtenbach and P. A. Wintz, "Quantizing for Noisy Channels," IEEE Trans. on Communications, *COM-17*, No. 2 (1969), pp. 291–302.
10. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," Proceedings of the IEEE, *62*, No. 5 (1974), pp. 611–632.
11. P. Noll, "Untersuchungen zur Sprachcodierung mit adaptiven Prädiktionsverfahren," Nachrichtentechnische Zeitschrift, *27*, H. 2 (1974), pp. 67–72.
12. B. Smith, "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 653–709.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

A Diffusion Model Approximation for the GI/G/1 Queue in Heavy Traffic

By D. P. HEYMAN

(Manuscript received April 10, 1975)

For the single-server queue with renewal input, we obtain heavy traffic approximations for the time-dependent distributions of queue length and virtual delay by constructing approximating diffusion processes. These approximations are shown to agree with known limiting cases, and a comparison is made with results from a computer simulation.

I. INTRODUCTION

We consider a single-server queuing system where the interarrival times are independent and identically distributed (i.i.d.) random variables, customers are served in order of arrival, the service times of the various customers are i.i.d. random variables, and the interarrival and service times form independent sequences. Lindley[1] obtained a recursive equation for the delay-time of the nth arriving customer, an integral equation for the delay time of a customer in the steady-state, and conditions for the latter to have a nondegenerate limit.

Lindley's equations have not yielded to conveniently used analytical solutions, except in some special cases, stimulating a search for approximations to the distributions of general interest. In this paper, we approximate the queue length and delay processes by appropriately chosen diffusion processes. This method of approximation appears to have been introduced by Gaver[2] and Newell.[3] Gaver and Newell considered the M/G/1 queue; we extend their approximate models to the GI/G/1 queue. Other methods for obtaining diffusion approximations for queuing processes involve applying the theory of weak convergence to sequences of approximating processes. Whitt[4] is a survey of these methods and contains an extensive bibliography. We show that the diffusion models developed in this paper agree with the limiting processes obtained by weak convergence methods.

Important features of this paper are the use of the M/M/1 queue to motivate a diffusion process approximation for the single-server queue and the use of elementary renewal theory results to obtain the param-

eters of the process. This approach provides an intuitive explanation for the limit theorems.

II. PRELIMINARIES AND NOTATION

Let T_i be time between the arrival epochs of the $(i-1)$st and ith customers, $i = 1, 2, \cdots$, assume these random variables are i.i.d., and let $\lambda^{-1} = E(T_1)$ and $\sigma_A^2 = \text{Var}\ (T_1)$. Assume the customers are served in order of arrival, let S_i be the service time of the ith customer, $\mu^{-1} = E(S_1)$, $\sigma_B^2 = \text{Var}\ (S_1)$, and assume $S_1, S_2, \cdots$ are i.i.d. random variables. We define the traffic intensity by $\rho = \lambda/\mu$ and will always assume $\rho < 1$. We seek approximations for the queue size and virtual delay at time t, and obtain these approximations from suitably chosen diffusion processes. For any function F, let $F_x = \partial F/\partial x$, $F_{xx} = \partial^2 F/\partial x^2$, $F_y = \partial F/\partial y$, etc., and unambiguous arguments will be suppressed.

Let $\{X(t), t \geqq 0\}$ be a homogeneous and additive diffusion process, $F(t, x; x_0) = \Pr\{X(t) \leqq x | X(0) = x_0\}$, and a and b be the infinitesimal mean and variance of the process, respectively. Then F satisfies the forward Kolmogorov (Fokker-Planck) equation

$$F_t = -aF_x + \frac{b}{2} F_{xx}, \tag{1}$$

with initial condition

$$F(0, x; x_0) = \begin{cases} 0 & \text{if} \quad x < x_0 \\ 1 & \text{if} \quad x \geqq x_0 \end{cases}. \tag{2}$$

If the range of $X(t)$ is $[0, \infty)$, then (1) is subject to the boundary condition

$$F(t, 0; x_0) = 0, \qquad t > 0. \tag{3}$$

The solution to (1) subject to (2) and (3) is given in Newell[5] as

$$F(x, t; x_0) = \Phi\left(\frac{x - x_0 - at}{\sqrt{bt}}\right) - e^{2xa/b}\Phi\left(\frac{-x - x_0 - at}{\sqrt{bt}}\right), \tag{4}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

is the normal distribution function (d.f.). If $a < 0$, then

$$F(x) = \lim_{t \to \infty} F(x, t; x_0) = 1 - e^{-2x(-a)/b}, \tag{5}$$

which is the negative-exponential d.f. and is independent of x_0.

Returning to the GI/G/1 queue, let W_i denote the delay of the ith customer. Lindley showed that when $\rho < 1$, $W = \lim_{i \to \infty} W_i$ is non-degenerate. Kingman[6] showed that if ρ is slightly less than unity, the

d.f. of W is approximately negative-exponential with mean

$$\tfrac{1}{2}(\sigma_A^2 + \sigma_B^2)/(\lambda^{-1} - \mu^{-1}). \tag{6}$$

III. AN APPROXIMATION FOR THE M/M/1 QUEUE

To motivate the diffusion model employed in approximating the GI/G/1 queue, and to indicate its efficacy for the M/M/1 queue in particular, we first develop an approximation for the M/M/1 queue. A scheme suggested by Bailey[7] for approximating stochastic processes is used.

In the M/M/1 queue, customers arrive according to a Poisson process with rate λ, and the service-time distribution is negative-exponential with mean μ^{-1}. Let $N(t)$ denote the number of customers in the queue at time t, and $\pi(t, n; n_0) = \Pr\{N(t) = n | N(0) = n_0\}$. For $t > 0$ and $n = 1, 2, \cdots$, $\pi(t, n; n_0)$ satisfies

$$\frac{d}{dt}\pi(t, n; n_0) = \lambda\pi(t, n-1; n_0) + \mu\pi(t, n+1; n_0) - (\lambda + \mu)\pi(t, n; n_0), \tag{7a}$$

and

$$\frac{d}{dt}\pi(t, 0; n_0) = \mu\pi(t, 1; n_0) - \lambda\pi(t, 0; n_0). \tag{7b}$$

The initial condition is

$$\pi(0, n; n_0) \begin{cases} 1 & \text{if } n = n_0 \\ 0 & \text{if } n \neq n_0 \end{cases}, \tag{8}$$

and the boundary condition is

$$\pi(t, n; n_0) = 0 \quad n < 0, \qquad t \geq 0. \tag{9}$$

The idea of the approximation is to replace (7) by a partial differential equation that is easier to solve. We do this by replacing the discrete variable n by the continuous variable x, and $\pi(t, n; n_0)$ by $p(t, x; x_0)$ in (7). Expanding in a Taylor's series about the point $(t, x; x_0)$ and keeping only first- and second-order terms, we obtain

$$p_t = -(\lambda - \mu)p_x + \frac{\lambda + \mu}{2}p_{xx}, \qquad x, t > 0. \tag{10}$$

If we define

$$\mathrm{P}(t, x; x_0) = \int_{-\infty}^{x} p(t, y; x_0)dy,$$

it can be easily shown that P also satisfies (10). We take

$$\mathrm{P}(0, x; x_0) = \begin{cases} 0 & \text{if } x < x_0 \\ 1 & \text{if } x \geq x_0 \end{cases} \tag{11}$$

and

$$\mathrm{P}(t, 0; x_0) = 0, \qquad t > 0, \tag{12}$$

as natural replacements for (8) and (9). This system of equations is identical in form to (1), (2), and (3), so $\mathrm{P}(t, x; x_0)$ is given by the right-side of (4) with $a = \lambda - \mu$ and $b = \lambda + \mu$.

Consider now the asymptotic behavior of π and P: From the theory of the M/M/1 queue, we have

$$\pi_n = \lim_{t \to \infty} \pi(t, n; n_0) = (1 - \rho)\rho^{n-1} \quad \text{for} \quad n > 0$$

and from (5) we obtain

$$\mathrm{P}(t) = \lim_{t \to \infty} \mathrm{P}(t, x; x_0) = 1 - \exp\left[-2(1 - \rho)/(1 + \rho)\right].$$

If for $n > 0$ we approximate π_n by

$$\bar{\pi}_n = \int_{n-1}^{n} d\mathrm{P}(t),$$

we obtain

$$\begin{aligned} \bar{\pi}_n &= \left[1 - e^{-2(1-\rho)/(1+\rho)}\right] e^{-2(n-1)(1-\rho)/(1+\rho)} \\ &= (1 - \alpha)\alpha^{n-1}, \qquad n > 0, \end{aligned}$$

where $\alpha = \exp\left[-2(1 - \rho)/(1 + \rho)\right]$, so $\bar{\pi}_n$ has the same form as π_n. If ρ is close to unity, $\alpha \doteq \rho$ and hence

$$\bar{\pi}_n \doteq \pi_n, \qquad \rho \doteq 1,$$

so $\bar{\pi}_n$ is a good approximation of π_n when ρ is slightly less than one.

IV. THE GI/G/1 QUEUE—APPROXIMATE QUEUE LENGTH

Let us first consider a heuristic manner of obtaining (10) for the M/M/1 queue. During the time interval $(t, t + \Delta t]$, the number of customers in the system changes by the number of arrivals minus the number of service completions, and when $N(t) = n > 0$, this change has expectation $(\lambda - \mu)\Delta t + o(\Delta t)$ and variance $(\lambda + \mu)\Delta t + o(\Delta t)$. To approximate $\{N(t), t \geqq 0\}$ by a diffusion process with the same infinitesimal mean and variance, set $a = \lambda - \mu$ and $b = \lambda + \mu$ in (1), which yields (10). This suggests that an appropriate choice of a and b will yield a good approximation for the queue length of the GI/G/1 queue.

For the GI/G/1 queue, let $N(t)$ be the queue size at time t, $A(t)$ and $D(t)$ the number of arrivals and departures in $(0, t]$, respectively; then

$$N(t) = N(0) + A(t) - D(t), \qquad t > 0. \tag{13}$$

For any renewal process $\{M(t), t \geqq 0\}$ where the interevent times have mean m and variance V, for large values of t

$$E[M(t)] \approx t/m, \tag{14}$$

and

$$\text{Var}\,[M(t)] \approx tV/m^3, \tag{15}$$

(Ref. 8). By hypothesis $\{A(t), t \geqq 0\}$ is a renewal process, so from (14) and (15) we obtain

$$E[A(t)] \approx \lambda t, \qquad \text{Var}\,[A(t)] \approx \lambda^3\sigma_A^2 t. \tag{16}$$

The process $\{D(t), t \geqq 0\}$ is not a renewal process, but, in heavy traffic (ρ close to 1), the server will be occupied most of the time, so we approximate $D(t)$ by $\tilde{D}(t)$, where

$$E[\tilde{D}(t)] \approx \mu t, \qquad \text{Var}\,[\tilde{D}(t)] \approx \mu^3\sigma_B^2 t. \tag{17}$$

Cox and Smith[9] use (13), (16), and (17) to study the GI/G/1 queue for small values of t without using a diffusion model. Substituting $\tilde{D}(t)$ for $D(t)$ in (13) and using (16) and (17), we obtain the approximate results

$$\lim_{t\to\infty} E[N(t)]/t \doteq \lambda - \mu \tag{18}$$

and

$$\lim_{t\to\infty} \text{Var}\,[N(t)]/t \doteq \lambda^3\sigma_A^2 + \mu^3\sigma_B^2, \tag{19}$$

which suggest that we approximate $N(t)$ by a diffusion process $\tilde{N}(t)$, say, with infinitesimal mean and variance given by

$$a = \lambda - \mu \tag{20}$$

and

$$b = \lambda^3\sigma_A^2 + \mu^3\sigma_B^2, \tag{21}$$

respectively. If we let $F(t, x; x_0) = \Pr\{\tilde{N}(t) \leqq x \mid \tilde{N}(0) = x_0\}$, then F satisfies (1), (2), (3), and hence is given by (4), with a and b as above.

As a partial check on the efficacy of this approximation, let us define $\tilde{N} = \lim_{t\to\infty} \tilde{N}(t)$. For $\rho < 1$, $\tilde{N}$ is a proper random variable, and from (5), (20), and (21),

$$E(\tilde{N}) = \frac{\mu}{2}\,\frac{\sigma_A^2\rho^2 + \sigma_B^2\rho^{-1}}{\lambda^{-1} - \mu^{-1}}, \tag{22}$$

which together with the queuing formula $L = \lambda W$ (see Ref. 10) and $\rho = 1$ yields the heavy traffic approximation given by (6).

V. THE GI/G/1 QUEUE—APPROXIMATE VIRTUAL DELAY

The virtual delay at time t is the delay in queue a customer would experience if it arrived at time t; an exact formulation is given by

Beneš.[11] Toward developing an approximation of the virtual-delay process, define

$$L(t) = S_1 + S_2 + \cdots + S_{A(t)}, \qquad t > 0, \tag{23}$$

so $L(t)$ represents the amount of work time brought to the server in $(0, t]$. Since $S_1, S_2, \cdots, S_{A(t)}$ are i.i.d., using (17) we obtain

$$E[L(t)] = E(S_1)E[A(t)] \approx \rho t \tag{24}$$

for large values of t. Using the conditional variance relationship $\text{Var}[L(t)] = E\{\text{Var}[L(t)|A(t)]\} + \text{Var}\{E[L(t)|A(t)]\}$ and (17), we obtain for large values of t,

$$\text{Var}[L(t)] \approx \lambda(\sigma_B^2 + \rho^2\sigma_A^2)t. \tag{25}$$

The sample paths of the virtual delay process are sawtooth functions with a jump of size S_i at the arrival epoch of the ith customer followed by a decline of slope -1; the process has an impenetrable boundary at the axis of abscissas. Assume that (24) and (25) hold for all t, so that

$$\alpha = \lim_{\Delta t \to 0} \{E[L(t + \Delta t)] - E[L(t)]\}/\Delta t = \rho \tag{26}$$

and

$$\begin{aligned} \beta &= \lim_{\Delta t \to 0} \{\text{Var}[L(t + \Delta t)] - \text{Var}[L(t)]\}/\Delta t \\ &= \lambda(\sigma_B^2 + \rho^2\sigma_A^2). \end{aligned} \tag{27}$$

Following Gaver,[2] we approximate the virtual-delay process by a diffusion process $\{\tilde{V}(t), t \geqq 0\}$, say, with infinitesimal mean and variance given by

$$a = \rho - 1, \qquad b = \lambda(\sigma_B^2 + \rho^2\sigma_A^2), \tag{28}$$

respectively. Hence, the time-dependent d.f. of $\tilde{V}(t)$ is given by (4) with a and b given by (28).

Turning now to asymptotic results, when $\rho < 1$, $\tilde{V} = \lim_{t \to \infty} \tilde{V}(t)$ exits and is proper, and from (5) has a negative-exponential distribution with mean

$$E(\tilde{V}) = \tfrac{1}{2}(\sigma_B^2 + \rho^2\sigma_A^2)/(\lambda^{-1} - \mu^{-1}). \tag{29}$$

Hooke[12] showed that, if the interarrival times are nonlattice, then as $\rho \uparrow 1$, $\Pr\{W \leqq x\} \to \Pr\{V \leqq x\}$, where V is the virtual delay in the steady state. This result and (29) together show that, as $\rho \uparrow 1$, $E(\tilde{V})/E(W) \uparrow 1$, and the d.f. of $\tilde{V}$ agrees with (6). From (22) and (29) we observe that $E(\tilde{V}) < E(\tilde{N})/\mu$ when $\rho < 1$, but $E(V) \geqq E(N)/\mu$ is known to hold.

We note that once a diffusion model for the virtual-delay process is at hand, the method given in Heyman[13] for approximating the busy-period distribution can be applied.

VI. A NUMERICAL EXAMPLE

In this section, we compare $E[N(t)]$ to the results of a single computer simulation. We consider a single-server queue that is empty at time zero, where the interarrival times are uniformly distributed from 0 to 20 minutes, and the service times are uniformly distributed from 0 to 19 minutes. Thus,

$$\lambda^{-1} = 10, \qquad \sigma_A^2 = 100/3, \qquad \mu^{-1} = 9.5, \qquad \sigma_B^2 = (9.5)^2/3; \tag{30}$$

hence, $\rho = 0.95$ and from (22) we obtain $E(\tilde{N}) = 6.50$. Since the d.f. of $\tilde{N}$ is negative-exponential, the standard deviation of $\tilde{N}$ is also 6.50.

Let $N_j(t)$ denote the sample of $N(t)$ produced by the jth simulation run, N_j be the estimate of N produced by the jth run, and $\hat{N} = n^{-1} \sum_{j=1}^{n} N_j$. Assuming the simulator produces independent sample paths and approximating the mean and variance of N by those of $\tilde{N}$, standard sampling theory indicates that, for reasonably large values of n, $\hat{N}$ has a normal distribution with mean 6.5 and standard deviation $6.5/\sqrt{n}$. Hence, given $c > 0$, the number of runs required to have

$$\Pr\{|\hat{N} - E(N)| \leqq cE(N)\} \geqq 0.99$$

is the smallest integer $\geqq (2.575)^2/c$; choosing $c = 0.05$ yields $n = 134$.

These considerations lead us to simulate 150 sample paths, and indicate one of the potential uses of even a crude diffusion approximation.

Gaver[2] and Newell[5] observe that if one makes the change of variables $\xi = -(a/b)x$ and $\tau = (a^2/b)t$, (1) becomes

$$F_\tau = -F_\xi + F_{\xi\xi}/2, \tag{31}$$

which can be solved once and for all; the solutions for any a and b can be recovered by scaling. For our example, $-(b/a) = 13.00$ and $b/a^2 = 2472.2$ minutes. Table 2-2 in Gaver[2] gives the values of

$$\int_0^\infty \xi d_\xi F(\tau, \xi; 0)$$

for $\tau = 0.1, 0.2, \cdots, 1.0$. These were used to construct Table I below.

The approximations shown in the table are not as accurate as the diffusion approximations for M/G/1 shown in Table I of Gaver[2]; these comparisons share with Gaver's table the property that the diffusion process consistently overestimates the mean queue-size.

Table I — Comparison of diffusion approximation and simulation results

Time	τ									
$E[\tilde{N}(t)]$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Diffusion	2.68	3.50	4.02	4.40	4.68	4.91	5.11	5.27	5.41	5.52
Simulation	2.25	2.91	3.40	3.71	4.16	4.39	4.61	5.05	5.18	5.13

VII. COMPARISON WITH THE RESULTS OF IGLEHART AND WHITT

Iglehart and Whitt[14,15] use the theory of weak convergence of stochastic processes to obtain heavy-traffic-limit theorems for large classes of queuing processes. In particular, these theorems hold for the GI/G/1 queue. Their results are obtained by considering a sequence of GI/G/1 queuing systems, indexed by n ($n = 1, 2, \cdots$). For the nth system, let λ_n and μ_n denote the arrival and service rates, respectively, and $N_n(t)$ and $V_n(t)$ denote the queue size and virtual delay at time t, respectively. For the above quantities, let the absence of a subscript denote the limit with respect to n, e.g., $\lambda = \lim_{n\to\infty} \lambda_n$. For each n, let σ_A^2 and σ_B^2 be as before, and assume $N_n(0) = V_n(0) = 0$.

Section three of Iglehart and Whitt[14] shows that if $\lim_{n\to\infty} (\lambda_n - \mu_n)\sqrt{n} = c$, where c is some finite constant, then

$$\lim_{n\to\infty} \frac{N_n(nt')}{\gamma\sqrt{n}} \Rightarrow B(t', -c/\gamma), \tag{32}$$

where $\gamma = \lambda^3\sigma_A^2 + \mu^3\sigma_B^2$, $B(t', -c/\gamma)$ is the Wiener process with negative drift c/γ together with an impenetrable barrier at the origin and $\Rightarrow$ denotes weak convergence. From (32) it follows that for large values of n, i.e., when $\rho_n = \lambda_n/\mu_n$ is close to 1,

$$\Pr\{N_n(nt')/\gamma\sqrt{n} \leqq d\} = \Phi\left(\frac{d - ct'/\gamma}{\sqrt{t'}}\right) - e^{2cd/\gamma}\Phi\left(\frac{-d - ct'/\gamma}{\sqrt{t'}}\right). \tag{33}$$

Since $N_n(t) \doteq N(t)$, $c \doteq (\lambda - \mu)\sqrt{n}$ for large n, upon making the change of variables $x = d\gamma\sqrt{n}$ and $t = nt'$, we obtain

$$\Pr\{N(t) \leqq x\} \doteq \Phi\left(\frac{x - (\lambda - \mu)t}{\gamma\sqrt{t}}\right) - e^{2(\lambda-\mu)x/\gamma^2}\Phi\left(\frac{-x - (\lambda - \mu)t}{\gamma\sqrt{t}}\right) \tag{34}$$

from (33). Observe that the right-hand side of (34) agrees exactly with

the time-dependent distribution (when $x_0 = 0$) of $\tilde{N}(t)$ obtained in Section IV.

Let $\alpha_n = \mu_n^{-1} - \lambda_n^{-1}$ and $\sigma^2 = \sigma_A^2 + \sigma_B^2$. From theorem 6.1 of Iglehart and Whitt,[15] it is possible to derive the following result, which appears in theorem 4 of Whitt.[16] If $\alpha_n \sigma^{-1}\sqrt{n} \to -k$, $-\infty < k < \infty$, then

$$\frac{V_n(nt')}{\sigma\sqrt{n\lambda_n}} \Rightarrow B(t', -k). \tag{35}$$

For large values of n, $\alpha_n \doteq (\rho - 1)\lambda^{-1}$, $-k \doteq (\rho - 1)\lambda^{-1}\sigma^{-1}\sqrt{n}$, $V_n(t) \doteq V(t)$, and

$$\Pr\{V(nt') \leqq d\sigma\sqrt{\lambda n}\} \doteq \Phi\left(\frac{d + kt'\sqrt{\lambda}}{\sqrt{t'}}\right) - e^{-2kd\sqrt{\lambda}}\Phi\left(\frac{-d + kt'\sqrt{\lambda}}{\sqrt{t'}}\right). \tag{36}$$

Letting $t = nt'$ and $x = d\sigma\sqrt{\lambda n}$ in (35), we obtain

$$\Pr\{V(t) \leqq x\} \doteq \Phi\left(\frac{x - (\rho - 1)t}{\sigma\sqrt{\lambda t}}\right) - e^{2(\rho-1)x/\lambda\sigma^2}\Phi\left(\frac{-x - (\rho - 1)t}{\sigma\sqrt{\lambda t}}\right). \tag{37}$$

The right-hand side of (36) represents the time-dependent distribution (4) with $x_0 = 0$ and

$$a = \rho - 1, \qquad b = \lambda(\sigma_A^2 + \sigma_B^2), \tag{38}$$

which is the same value of a and almost the same value of b given in (28), where the difference between the values of b given in (28) and (38) vanishes as $\rho \to 1$.

From the results of this section, we can conclude that the heuristically constructed diffusion models for $\tilde{N}(t)$ and $\tilde{V}(t)$ given in this paper can be regarded as the limit (in the sense of weak convergence) of $N(t)$ and $V(t)$, respectively, with suitable normalization.

VIII. SUMMARY

We have obtained an approximation for the time-dependent distributions of queue length and virtual delay in a GI/G/1 queue using a diffusion model. The diffusion model for the queue length process was obtained by using the mean and variance of the asymptotic rate of change of an approximation of the queue-length process as the infinitesimal mean and variance for a diffusion process. The diffusion approximation was compared to a simulation of a particular queueing process, and reasonably close agreement for mean queue lengths was obtained. The approximation for the virtual-delay process was gen-

erated in a manner suggested in Gaver[2] and the limiting results it provided were shown to agree with a theorem of Kingman[6] for the delay process. The time-dependent distributions for the approximate queue length and virtual-delay processes agree, as the traffic intensity approaches one, with the limiting results of Iglehart and Whitt.[14,15]

IX. ACKNOWLEDGMENT

John Rath of Bell Laboratories suggested Section VII and made many helpful comments.

REFERENCES

1. D. V. Lindley, "The Theory of Queues with a Single Server," Proc. Comb. Phil. Soc., *48*, Part 2 (April 1952), pp. 277–289.
2. D. P. Gaver, "Diffusion Approximations and Models for Certain Congestion Problems," J. Appl. Prob., *5*, No. 3 (December 1968), pp. 607–623.
3. G. F. Newell, "Queues with Time-Dependent Arrival Rates I—The Transition Through Saturation," J. Appl. Prob., *5*, No. 2 (August 1968), pp. 436–451.
4. W. Whitt, "Heavy Traffic Limit Theorems for Queues: A Survey," *Mathematical Methods in Queueing Theory*, New York: Springer-Verlag, 1974, pp. 307–350.
5. G. F. Newell, *Applications of Queueing Theory*, London: Chapman and Hall, 1971.
6. J. F. C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues," *Proceedings of the Symposium on Congestion Theory*, W. L. Smith and W. E. Wilkinson eds., Chapel Hill: University of North Carolina Press, 1965, pp. 137–159.
7. N. T. J. Bailey, *The Elements of Stochastic Processes*, New York: John Wiley, 1964.
8. D. R. Cox, *Renewal Theory*. London: Methuen, 1962.
9. D. R. Cox and W. L. Smith, *Queues*. London: Methuen, 1961.
10. J. D. C. Little, "A Proof of the Queueing Formula $L = \lambda W$," Opns. Res., *9*, No. 3 (May–June 1961), pp. 383–387.
11. V. E. Beneš, *General Stochastic Processes in the Theory of Queues*, Reading, Massachusetts: Addison–Wesley, 1963.
12. J. A. Hooke, "Some Limit Theorems for Priority Queues," Tech. Report 91, Ithaca: Cornell University, 1969.
13. D. P. Heyman, "An Approximation for the Busy-Period of the M/G/1 Queue Using a Diffusion Model," J. Appl. Prob., *11*, No. 1 (March 1974), pp. 159–169.
14. D. Iglehart and W. Whitt, "Multiple Channel Queues in Heavy Traffic, I," Adv. Appl. Prob., *2*, No. 1 (Spring 1970), pp. 150–177.
15. D. Iglehart and W. Whitt," Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks, and Batches," Adv. Appl. Prob., *2*, No. 2 (Autumn 1970), pp. 355–369.
16. W. Whitt, "Heavy Traffic Approximations for Stable Queues," Technical Report, Department of Administrative Sciences, New Haven: Yale University, August 1971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 9, November 1975
Printed in U.S.A.

Optimal Rearrangeable Graphs

By F. R. K. CHUNG

(Manuscript received April 3, 1975)

Many important properties of switching networks can be effectively studied in the more general context of graph theory. In particular, the various rearrangeability properties of a network fall into this category. If G is a graph with vertex set $V = I \bigcup \Omega$, we say G is rearrangeable if, for all choices of distinct vertices, $i_1, i_2, \cdots, i_t$ in I and $j_1, j_2, \cdots, j_t$ in Ω, there exist vertex disjoint paths between i_k and j_k for all k. In this paper, we determine the minimum number of edges any rearrangeable graph may have for all choices of I and Ω. We also discuss generalizations in which V is strictly greater than $I \bigcup \Omega$ and/or t is bounded by a predetermined value. The minimal rearrangeable graphs we construct can be used to form efficient rearrangeable (and nearly rearrangeable) switching networks of arbitrary size.

I. INTRODUCTION

Let G be a finite graph with vertex set $V(G)$ and edge set $E(G)$.* Let I and Ω be nonempty subsets of $V(G)$ (not necessarily disjoint) and let S denote the set

$$V(G) \setminus (I \bigcup \Omega) = \{v \in V(G) \mid v \notin I \bigcup \Omega\}.$$

We use the following terminology:

(*i*) A *request* is an ordered pair (x, y) with $x \in I$, $y \in \Omega$, and $x \neq y$.

(*ii*) A set of requests is called an *assignment* if each vertex in the set occurs once at most.

(*iii*) An assignment A is called *realizable in* G (or we say G *satisfies* A) if we can find a set of vertex-disjoint paths connecting x and y for each pair (x, y) in A.

(*iv*) A graph G is said to be *rearrangeable* if G satisfies *any* assignment.

The problem we consider, first suggested by F. K. Hwang,[2] is to find

* I.e., $E(G)$ consists of a prescribed set of unordered pairs of distinct elements of some finite set $V(G)$. Generally, we follow the terminology of Harary.[1]

rearrangeable graphs with given I and with Ω having the least possible number of edges.

In this paper, we derive lower bounds on the minimum number of edges that a rearrangeable graph can have (see Theorem 1 in Section II). In addition, we also construct rearrangeable graphs which meet these bounds so that these graphs are optimal by this measure. Finally, we consider a generalization of rearrangeability, called k-rearrangeability, and we solve the corresponding problems in this case as well.

This study was motivated by questions of rearrangeability in switching networks (see Ref. 3). The sets I and Ω correspond to the sets of inlets and outlets, respectively; an edge $\{x, y\}$ of G corresponds to a crosspoint between x and y. The optimal rearrangeable graphs we construct can consequently be used to form efficient rearrangeable (and nearly rearrangeable) switching networks of arbitrary size.

II. BASIC PROPERTIES OF REARRANGEABLE GRAPHS

Let G be a rearrangeable graph with distinguished subsets I and Ω, where we assume without loss of generality that $|I| = n \leqq m = |\Omega|$. For the bulk of the paper, we shall restrict ourselves to the special case that S is empty, i.e., $V(G) = I \cup \Omega$.

If $\{x, y\}$ is an edge of G, we say that x and y are *adjacent* and we write $x \sim y$. Similarly, for $T \subseteq V(G)$, the notation $x \sim T$ will denote that $x \sim t$ for some $t \in T$. By the *degree* of $v \in V(G)$, written deg (v), we mean the number of edges of G containing v. More generally, if $X \subseteq V(G)$, then $\deg_X(v)$ denotes

$$|\{\{v, x\} \mid \{v, x\} \in E(G) \text{ and } x \in X\}|.$$

Suppose there is a vertex $v \in I$ with deg $(v) = k < n$, and let $v_1, \cdots, v_k$ denote the vertices that are adjacent to v.

Now consider an assignment A in which all the v_i, $1 \leqq i \leqq k$, occur as well as the pair (v, v'), where $v' \in \Omega$ is not adjacent to v. But this A is not realizable in G, which contradicts the hypothesis that G is rearrangeable. Hence, for all $v \in I$, we must have $\deg(v) \geqq n$. By a similar argument, it can be shown that deg $(v') \geqq n$ for all $v' \in \Omega$. Thus, for any rearrangeable graph G we must have:

Fact 1: For all $v \in V(G)$, deg $(v) \geqq n$.

Let us now state several more elementary facts about rearrangeable graphs G which can be proved in much the same way as Fact 1.

Fact 2: For all $v \in I$, $\deg_\Omega(v) \geqq n$.
Fact 3: For all $v \in V(G)$, $\max[\deg_I(v), \deg_\Omega(v)] \geqq n$.
Fact 4: If $v \sim I$ and $v \sim \Omega$ and $|I \cap \Omega| = 0$, then deg $(v) \geqq n + 1$.

Lemma 1: *In a rearrangeable graph G with vertex set $V(G) = I \cup \Omega$, $|I| = n < m = |\Omega|$, and $|I \cap \Omega| = 0$, there are at least n vertices with degree greater than n.*

Proof: If all $v \in I$ satisfy $\deg_\Omega(v) \geqq n + 1$, then we are done. Suppose there is an element $v \in I$ satisfy $\deg_\Omega(v) = n$, say, v is adjacent to $v_1, v_2, \cdots, v_n$. If any v_i, say, v_1, is not adjacent to Ω, let us consider an assignment in which all the v_i, $2 \leqq i \leqq n$, occur as well as the ordered pair (v, v'), where v' is a vertex in Ω different from any v_i. However, since $|I| = n$, it is impossible for G to satisfy this assignment. Hence, all the v_i are adjacent to both I and Ω. By Fact 4, $\deg v_i \geqq n + 1$ for $i = 1, \cdots, n$, which proves Lemma 1.

Lemma 2: *If G is any rearrangeable graph with vertex set $I \cup \Omega$, which has $|\Omega| > |I| = n$ and $|I \cap \Omega| = 0$, then G has at least $n(p + 1)/2$ edges, where $p = |V(G)|$.*

Proof: The number of edges in G satisfies the following inequality:

$$\begin{aligned} |E(G)| = e(G) &= \tfrac{1}{2} \sum_{v \in G} \deg(v) \\ &\geqq \tfrac{1}{2}[(p - n)n + (n + 1)n] \\ &= \tfrac{1}{2}n(p + 1). \end{aligned}$$

Lemma 3: *In a rearrangeable graph G with vertex set $V(G) = I \cup \Omega$, which has $|I| = n < m = |\Omega| < 2n$ and $|I \cap \Omega| = 0$, we have*

$$e(G) \geqq nm - \tfrac{1}{2}(m - n - 1)(m - n).$$

Proof: Denote the vertices in I by $i_1, \cdots, i_n$. Suppose the vertex i_j is adjacent to d_j vertices in Ω, where we may assume $d_1 \leqq d_2 \leqq \cdots \leqq d_n$. Let Ω_j be the union of $\{i_j\}$ and the d_j elements in Ω, which are adjacent to i_j. By Fact 3, each element in $\Omega \setminus \Omega_j$ is then adjacent to at least $n - (m - d_j) + 1$ elements in Ω_j. Hence, by counting the total number of edges $e(G)$ and using the fact that $d_1 \leqq d_i$ for all i, it follows that

$$e(G) \geqq nd_1 + (m - d_1)(n - m + d_1 + 1) + \tfrac{1}{2}(m - d_1)(m - d_1 - 1).$$

But the right-hand side is minimized by choosing d_1 as small as possible. Thus, since $m < 2n$, then by Fact 3, we have

$$e(G) \geqq mn - \tfrac{1}{2}(m - n)(m - n - 1),$$

which proves the lemma.

The preceding inequalities are summarized in the following result.

Theorem 1: *In a rearrangeable graph with vertex set $V(G) = I \cup \Omega$, and $|I| = n \leqq m = |\Omega|$, $|\Omega \cap I| = 0$, the number of edges $e(G)$ satisfies*

$$e(G) \geqq \begin{cases} \left\lceil \dfrac{n(m+n+1)}{2} \right\rceil & \text{if } m \geqq 2n, \\ \lceil mn - \frac{1}{2}(m-n)(m-n-1) \rceil & \text{if } 2n > m \geqq n, \end{cases}$$

where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x.

The proof follows at once from Lemma 2 and Lemma 3.

III. OPTIMAL REARRANGEABLE GRAPHS—MANHATTAN GRAPHS

In this section, we give a construction for a class of optimal rearrangeable graphs. The number of edges in these graphs will meet the lower bound in Theorem 1. These graphs will be called *Manhattan* graphs because they resemble a number of bridges connecting a high-density metropolitan area and low-density suburban areas.

A Manhattan graph with vertex set $V(G) = I \cup \Omega$ will be denoted by $M(I, \Omega)$. If $|I| = n$, $|\Omega| = m$, and $|I \cap \Omega| = 0$, $M(I, \Omega)$ is also denoted by $M(n, m)$.

In this section, we give the construction of $M(n, m)$ for any n and m by considering the following cases.

Case 1, $n = m$: The Manhattan graph $M(n, n)$ is the complete bipartite graph $K_{n,n}$, i.e., there is an edge between every pair of vertices (u, v), $v \in I$, $v \in \Omega$.

Case 2, $n < m < 2n$: We shall specify the edges of $M(n, m)$ by giving the subgraph spanned by various subsets of vertices of $M(n, m)$. The *spanning subgraph* of a set $S \subseteq V(G)$ is the subgraph of G with edge set $\{\{x, y\} \mid \{x, y\} \in E(G) \text{ and } x, y \in S\}$.

Let

$$I = \{i_1, i_2, \cdots, i_n\},$$
$$\Omega = \{x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_t\},$$

where

$$t = m - n.$$

$M(n, m)$ will be constructed as follows:

(*i*) The spanning subgraph of the vertices $I \cup \{x_i \mid 1 \leqq i \leqq n\}$ in $M(n, m)$ is a complete bipartite graph $K_{n,n}$;
(*ii*) y_j is adjacent to $x_j, x_{j+1}, \cdots, x_{j+(n-t)}$, for $j = 1, 2, \cdots, t$;
(*iii*) The spanning subgraph of the vertices $\{y_1, y_2, \cdots, y_t\}$ in $M(n, m)$ is a complete graph K_t.

The graph $M(n, m)$ is clearly rearrangeable. As an example of this construction, we illustrate $M(3, 5)$ in Fig. 1.

Case 3, $2n \leqq m < 3n$: The construction scheme for $M(n, m)$ in this case may be described as follows:

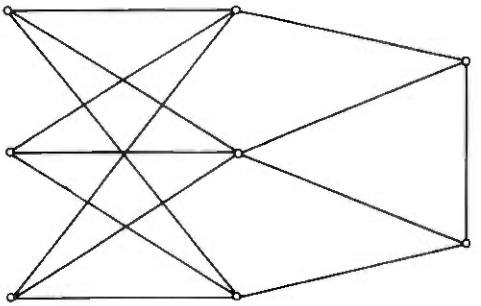


Fig. 1—Graph $M(3, 5)$.

Let

$$I = \{i_1, i_2, \cdots, i_n\},$$
$$\Omega = \{x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n, z_1, z_2, \cdots, z_t\},$$

where

$$t = m - 2n.$$

(*i*) The spanning subgraph of the vertices $I \cup \{x_1, x_2, \cdots, x_n\}$ in $M(n, m)$ is a complete bipartite graph $K_{n,n}$;

(*ii*) x_j is adjacent to y_j for $j = 1, 2, \cdots, n$;

(*iii*) z_j is adjacent to $y_1, y_2, \cdots, y_n$ for $j = 1, 2, \cdots, t$;

(*iv*) The spanning subgraph of vertices $\{y_1, y_2, \cdots, y_n\}$ in $M(n, m)$ is any graph with degree sequence

$$\underbrace{\{n - t - 1, n - t - 1, \cdots, n - t - 1, w\}}_{n - 1 \text{ times},}$$

where

$$w = \begin{cases} n - t - 1 & \text{if } n(n - t + 1) \text{ is even,} \\ n - t & \text{otherwise.} \end{cases}$$

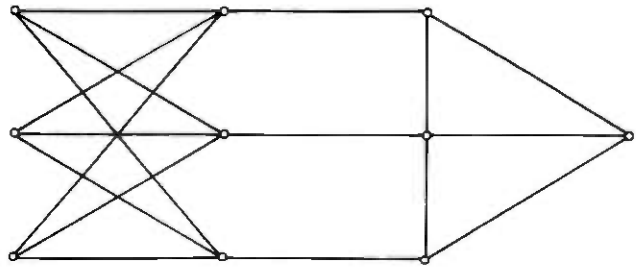
From a well-known theorem of Erdös and Gallai,[4] a graph with this degree sequence can always be constructed. The graphs $M(3, 7)$ and $M(3, 8)$ are shown in Fig. 2 as examples of this construction.

We want to show this graph is rearrangeable. Given an assignment A involving vertices $x_{a_1}, x_{a_2}, \cdots, x_{a_{n_1}}, y_{b_1}, y_{b_2}, \cdots, y_{b_{n_2}}, z_{c_1}, z_{c_2}, \cdots, z_{c_{n_3}}$, it is clear that $n \geqq n_1 + n_2 + n_3$.

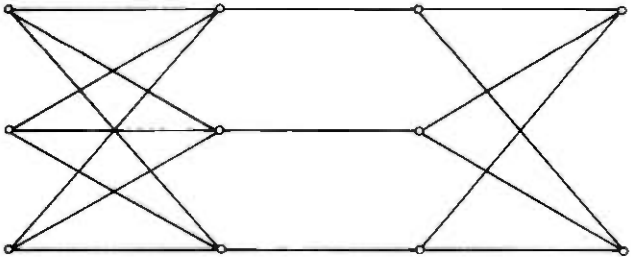
We may assume $n_2 = n_2' + n_2''$, where both x_{a_i} and y_{b_j}, $j = 1, 2, \cdots, n_2'$, appear in A.

If $t - n_3 \geqq n_2'$, it is easy to see this graph is rearrangeable.

Suppose $t - n_3 < n_2'$. Let us consider that the set $S_j = \{y_{b_i} | 1 \leqq i \leqq n_2, y_{b_i}$ is adjacent to y_{b_j}, and both y_{b_i} and x_{a_i} do not occur in $A\}$.



(a)



(b)

Fig. 2—(a) Graph $M(3, 7)$. (b) Graph $M(3, 8)$.

Since $|S_j| \geqq n - t - n_1 - n_2''$ for all $1 \leqq j \leqq n$, we know that at least $n - t - n_3 - n_2''$ of the n_2' requests involving $y_{b_1}, y_{b_2}, \cdots, y_{b_{n_2}}$ can be connected.

If

$$n_2' \geqq n - t - n_1 - n_2'',$$

then

$$\begin{aligned} n_2' - (n - t - n_1 - n_2'') &= t - (n - n_1 - n_2' - n_2'') \\ &\leqq t - n_3. \end{aligned}$$

After the remaining $n_2' - (n - t - n_1 - n_2'')$ requests are connected by a path passing through some of the $t - n_3$ z_i's, which do not occur in A, the requests involving the z_{b_i}'s or the x_{a_i}'s can be easily connected. Thus, we have proved that the graph $M(n, m)$ is rearrangeable.

For the case $m = 3n - 1$, there is another type of Manhattan graph which is a special case of the following class of graphs.

Case 4, $m = h(n - 1) + 2n$, $h \geqq 1$:
Let

$$I = \{i_1, i_2, \cdots, i_n\},$$
$$\Omega = \{x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n, z_1, z_2, \cdots, z_{(n-1)h}\}.$$

The graph $M(n, m)$ is constructed as follows:

(*i*) The spanning subgraph of vertices $I \cup \{x_i | 1 \leqq i \leqq n\}$ is the complete bipartite graph $K_{n,n}$;

(*ii*) x_j is adjacent to y_j, $j = 1, 2, \cdots, n$;

(*iii*) There is a cycle with vertices $y_1, z_1, z_2, \cdots, z_h, y_2, z_{h+1}, \cdots, z_{2h}, y_3, \cdots, y_n, y_1$;

(*iv*) There is a complete graph K_{n-1} with vertex set $\{z_{i+jh} | j = 0, 1, \cdots, n-2\}$ for $i = 1, 2, \cdots, h$.

(*v*) y_i is adjacent to $y_1, y_2, \cdots, y_{i-2}, y_{i+2}, \cdots, y_n$ for $i = 2, \cdots, n-1$.

y_1 is adjacent to $y_3, y_4, \cdots, y_{n-1}$.

y_n is adjacent to $y_2, y_3, \cdots, y_{n-2}$.

As an example of this construction, we illustrate $M(4, 14)$ in Fig. 3.

To see that this graph is rearrangeable, let us consider an assignment in which $x_{a_1}, x_{a_2}, \cdots, x_{a_{n_1}}, y_{b_1}, y_{b_2}, \cdots, y_{b_{n_2}}$, and $z_{c_1}, z_{c_2}, \cdots, z_{c_{n_3}}$ occur. Because of the structure of this graph, any request involving the x_{a_i} or y_{b_j} can easily be connected after the n_3 requests involving the z_{c_i}'s are connected.

It is clear that $n \geqq n_1 + n_2 + n_3$. First, let us consider the special case $n_1 = n_2 = 0$. Let $P_{i,j}$, $i < j$, denote the path $y_i, z_{(i-1)h+1}, z_{(i-1)h+2}, \cdots, z_{ih}, y_{i+1}, \cdots, y_j$. If each of $P_{1,2}, P_{2,3}, \cdots, P_{n-1,n}$ contains one of the z_{c_i} except for one $P_{i,i+1}$, then the assignment can be satisfied. If more than one of $P_{1,2}, P_{2,3}, \cdots, P_{n-1,n}$ contains more than one z_{c_i}'s, say $P_{1,2}$ contains z_{c_1}, z_{c_2}, and $P_{3,4}$ contains z_{c_3}, z_{c_4}, we know that at least one of the $P_{i,i+1}$'s does not contain any z_{c_i}, say $P_{t,t+1}$. Instead of considering the assignment A, it suffices to consider the assignment involving $\{z_{c_1+(t-1)h}\} \cup \{z_{c_i} | i = 2, 3, \cdots, n_3\}$. Continuing this argument, it is enough to consider an assignment satisfying the property that all the z_{c_i} involved appear in distinct $P_{i,i+1}$'s except for two of them and, therefore, this assignment is realizable.

Now, for arbitrary n_1 and n_2, let $S = \{a_1, a_2, \cdots, a_{n_1}, b_1, b_2, \cdots, b_{n_2}\}$. Relabel S by $S = \{s_1 < s_2 < \cdots < s_{n'}\}$, $n' \leqq n_1 + n_2$, and consider

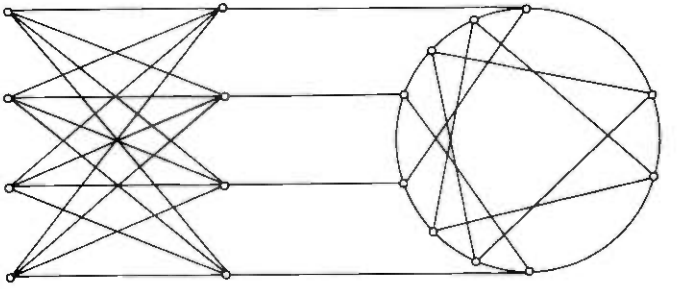


Fig. 3—Graph $M(4, 14)$.

the set $T = \{P_{i,j} | i < j,\ i, j \notin S$ and $i + 1, i + 2, \cdots, j - 1 \in S\}$. Then $|T| = n - |S| - 1$.

Let

$$R = \{z_i | i = c_1, c_2, \cdots, c_{n_3}\} \cup \{y_j | j \in \{b_1, b_2, \cdots, b_{n_2}\} \setminus \{a_1, \cdots, a_{n_1}\}\}.$$

If all paths in T contain one element of R except one path which contains two elements of R, then the assignment is clearly realizable. Otherwise, we may use an argument similar to the one above to establish the rearrangeability of $M(n, m)$.

Case 5, $m \geqq 3n$ and $m = 2n + h(n - 1) + t$, $0 < t < n - 1$: In this case, the graph is a combination of a graph of Case 3 and a graph of Case 4 except for minor modifications. Let

$$I = \{i_1, i_2, \cdots, i_n\},$$
$$\Omega = \{x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n, z_1, z_2, \cdots, z_{(n-1)h}, w_1, w_2, \cdots, w_t\}.$$

(*i*) If $t \neq n - 2$, let us delete all the edges of the form $\{y_i, y_j\}$, $1 \leqq i, j \leqq n$ in $M(n, 2n + t)$ in Case 3. We then construct a cycle of vertices $y_1, z_1, z_2, \cdots, z_{n-1}, y_2, \cdots, y_n, y_1$. The spanning subgraph of vertices $\{z_1, z_2, \cdots, z_{(n-1)h}\}$ are the same as that in $M(n, 2n + (n - 1)h)$ in Case 4. The spanning subgraph

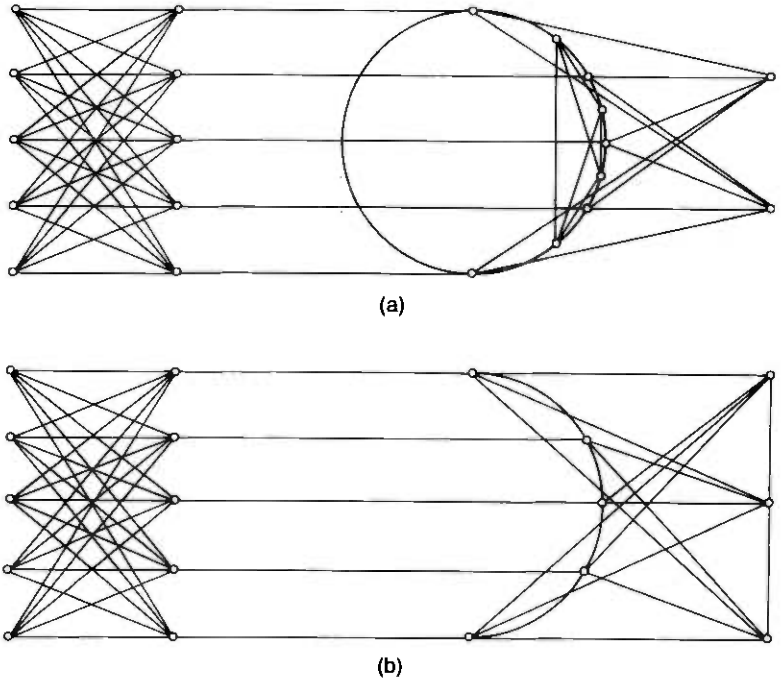


Fig. 4—(a) Graph $M(5, 16)$. (b) Graph $M(5, 17)$.

of vertices $\{y_1, y_2, \cdots, y_n\}$ with the exception of the edge y_1y_n is any graph with degree sequence

$$\underbrace{(n-t-3, n-t-3, \cdots, n-t-3, w)}_{n-1 \text{ times}},$$

where

$$w = \begin{cases} n-t-3 & \text{if } n(n-t-3) \text{ is even,} \\ n-t-2 & \text{otherwise,} \end{cases}$$

and

$$y_i \nsim y_{i+1},\ i = 1, 2, \cdots, n-1.$$

(*ii*) If $t = n-2$, the construction scheme is as follows:

(*a*) The spanning subgraph of vertices $I \cup \{x_1, x_2, \cdots, x_n\}$ is $K_{n,n}$;

(*b*) x_j is adjacent to y_j, $j = 1, 2, \cdots, n$;

(*c*) There is a path $y_1, z_1, \cdots, z_h, y_2, z_{h+1}, \cdots, z_{2h}, y_3, \cdots, y_n$;

(*d*) There is a complete graph k_{n-1} with vertex set $\{z_{i+jh} | j = 0, 1, \cdots, n-2\}$ for $j = 1, 2, \cdots, h$.

(*e*) When $n = 3$, w_1 is connected to any y_i. If $n \neq 3$, we have the following:

w_i, $i = 1, 2, \cdots, n-2$, is adjacent to all y_j except y_{i+1};
If n is even, then $w_{2i-1} \sim w_{2i}$, $i = 1, 2, \cdots, [n/2] - 1$.
If n is odd, then $w_{2i-1} \sim w_{2i}$, $i = 1, 2, \cdots, [n/2] - 1$, $w_{n-3} \sim w_{n-2}$.

It is an easy exercise to show that the Manhattan graph $M(n, m)$ thus constructed is rearrangeable. As examples of this construction, we illustrate $M(5, 16)$, $M(5, 17)$ in Fig. 4.

By a direct calculation, it is easy to verify that all the Manhattan graphs we constructed in Cases 1 through 5 achieve the lower bounds on all rearrangeable graphs for given I and Ω with $|I \cap \Omega| = 0$. From this, the following result is immediate.

Theorem 2: The Manhattan graphs $M(n, m)$ are optimal rearrangeable graphs.

We note that a complete bipartite graph $K_{n,m}$ has nm edges. Thus, by Theorem 1, a Manhattan graph has precisely $[\frac{1}{2}(m-n-1) \times \min(n, m-n)]$ fewer edges than $K_{n,m}$. When m is large compared to n, this is approximately $\frac{1}{2}nm$.

IV. MANHATTAN GRAPHS FOR THE CASE OF $|I \cap \Omega| \neq 0$

Let us now consider an optimal rearrangeable graph with vertex set $I \cup \Omega$ and $|I \cap \Omega| \neq 0$.

If $|\Omega \setminus I| \geq |I|$, then the Manhattan graph $M(I, \Omega)$ will be taken to be the same as $M(I, \Omega \setminus I)$. To prove that the Manhattan graph $M(I, \Omega)$ is an optimal rearrangeable graph for given I, Ω, and $|\Omega \setminus I| \geq |I|$, we need only show that $M(I, \Omega)$ is rearrangeable. Any request $(x, y) \in A$, $x \in I$, $y \in \Omega \setminus I$, can be connected in $M(I, \Omega \setminus I)$ as well as in $M(I, \Omega)$. If the assignment contains some request (x, y), where both x and y are in I, they can be successfully joined by a path of length 2 via some vertex in $\Omega \setminus I$.

When $|\Omega \setminus I| \leq |I| - 1$, the following construction suggested by F. K. Hwang[2] suffices for $M(I, \Omega)$.

Let $M(I, \Omega)$ be the union of a complete graph K_l and a three-partite graph $K_{n,m,l}$ as shown in Fig. 5b, where $n = |I \setminus \Omega|$, $m = |\Omega \setminus I|$, $l = |\Omega \cap I|$. To illustrate this construction more clearly, we denote the graph I_n to be the graph of n vertices without any edge. If two graphs G and H are joined by two thick lines, as shown in Fig. 5a, there is an edge connecting any vertex in G to any vertex in H.

We note that *no* edge in the above graph can be deleted without destroying the rearrangeability of the graph for the given I, Ω. Thus, we can state the following result.

Theorem 3: *The Manhattan graph $M(I, \Omega)$ is an optimal rearrangeable graph for any given I, Ω.*

V. k-REARRANGEABLE GRAPHS

A graph is said to be *rearrangeable of capacity k* or *k-rearrangeable* if it satisfies any assignment A of size at most k, i.e.,

$$A = \{(x_1, y_1), \cdots, (x_t, y_t)\}, \; t \leq k.$$

A rearrangeable graph is easily seen to be a special case of a k-rearrangeable graph with $k = |I|$.

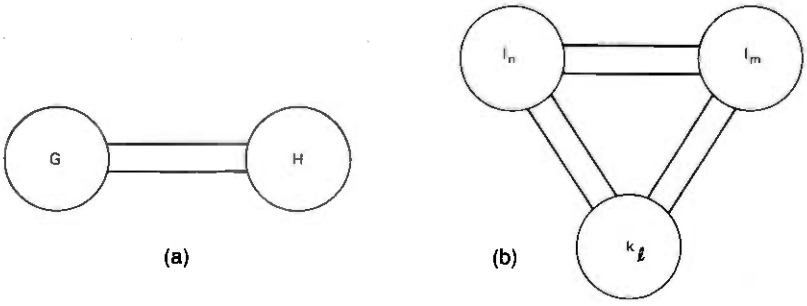


Fig. 5—(a) Complete connection between two graphs. (b) Graph $M(I,\Omega)$ with $n = |I \setminus \Omega|$, $m = |\Omega \setminus I|$, $l = |\Omega \cap I|$.

Assuming $|\Omega| \geqq |I| \geqq k$, a k-rearrangeable graph G with vertex set $V(G) = I \bigcup \Omega$ has the following properties (which are similar to those of a rearrangeable graph).

Fact 1': For all $v \in V(G)$, $\deg(v) \geqq k$.
Fact 2': If $v \in I$, $\deg_\Omega(v) \geqq k$.
Fact 3': For any $v \in V(G)$, $\max[\deg_I(v), \deg_\Omega(v)] \geqq k$.
Fact 4': If $v \sim I$, $v \sim \Omega$, then $\deg(v) \geqq k+1$.

Lemma 1': *The number of edges in a k-rearrangeable graph with vertex set $V(G) = I \bigcup \Omega$, $|I| > k$, $|\Omega \setminus I| > k$, $p = |V(G)|$, satisfies*

$$e(G) \geqq \left\lceil \frac{k(p+2)}{2} \right\rceil.$$

Proof: If there is an $x \in I$ which is not adjacent to I, then the spanning subgraph G' of the vertex set $V(G) \setminus \{v\}$ in G must be k-rearrangeable. If $|I| = k+1$, then G' has at least $\frac{1}{2}kp$ edges. If $|I| > k+1$, then G' has more than $\frac{1}{2}kp$ edges (by induction). In any case, G has at least $\frac{1}{2}k(p+2)$ edges. Similarly, G must have $\geqq \frac{1}{2}k(p+2)$ edges if there is an $y \in \Omega$ which is not adjacent to Ω.

If all vertices in I are adjacent to I and all vertices in Ω are adjacent to Ω, consider the sets

$$I' = \{x \in I \mid x \sim \Omega, x \sim I\},$$
$$\Omega' = \{y \in \Omega \mid y \sim I, y \sim \Omega\}.$$

Any element in I' or Ω' has degree $\geqq k+1$ and also $|I'| \geqq k$, $|\Omega'| \geqq k$. Hence,

$$e(G) \geqq \left\lceil \frac{k(p+2)}{2} \right\rceil.$$

Similar to Theorem 1, we have Theorem 4.

Theorem 4: *The number of edges in a k-rearrangeable graph with vertex set $V(G) = I \bigcup \Omega$, $|I| = n$, $|\Omega| = m$, $k < n \leqq m$, $|I \bigcap \Omega| = 0$, satisfies*

$$e(G) \geqq \begin{cases} \lceil \frac{1}{2}k(m+n+2) \rceil & \text{if } 2k \leqq n \leqq m, \\ \lceil \frac{1}{2}\{k(m+n+1) + t(k-t+1)\} \rceil & \\ \qquad \text{if } k < n < 2k \leqq m, \quad n = k+t, & \\ \lceil \frac{1}{2}\{k(m+n) + t(k-t+1) + t'(k-t'+1)\} \rceil & \\ \qquad \text{if } k < n \leqq m < 2k, \quad n = k+t, \quad m = k+t'. & \end{cases}$$

If I and Ω are disjoint, an optimal k-rearrangeable graph can be constructed by combining two optimal rearrangeable graphs $M(k, n)$, $M(k, m)$ by overlapping $K_{k,k}$ as shown in Fig. 6. These are called

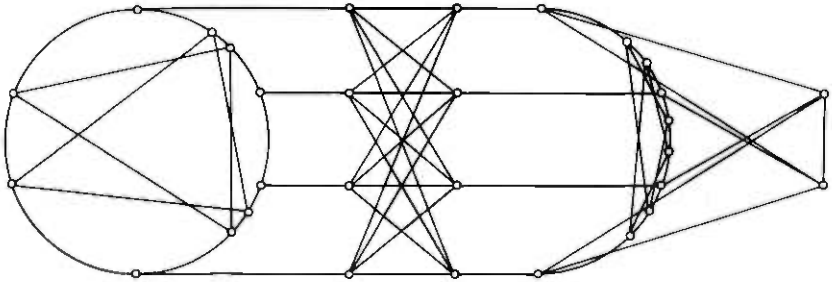


Fig. 6—Graph $M_4(14,16)$.

Manhattan k-graphs and are denoted by $M_k(n, m)$, where $n = |I|$, $m = |\Omega|$.

Because $M(k, n)$, $M(k, m)$ are rearrangeable, the k-rearrangeability of $M_k(n, m)$ follows immediately.

If $|\Omega \cap I| = 1$, then $M_k(I, \Omega)$ is the same as $M_k(|I|, |\Omega| - 1)$. If $|\Omega \cap I| = 2$ and $|\Omega \setminus I| \geq k$, then $M_k(I, \Omega)$ is $M_k(|I| - 1, |\Omega| - 1)$.

We notice that $M_n(n, m) = M(n, m)$.

Theorem 5: *Manhattan k-graphs are optimal rearrangeable graphs of capacity k for given I, Ω where $|\Omega| \geq |I| > k$, $|\Omega \cap I| \leq 2$, $|\Omega \setminus I| \geq k$.*

As we noted earlier, Manhattan graphs have considerably fewer edges than the corresponding complete bipartite graphs with the same vertex sets. This is also the case for Manhattan k-graphs as well. In particular, the number of edges saved is

$$[mn - \tfrac{1}{2}\{k(m+n) + \max[k, (n-k)(2k-n+1)] + \max[k, (m-k)(2k-m+1)]\}].$$

When $|\Omega \cap I|$ is large, alternate constructions of k-rearrangeable graphs for given I and Ω can be given by adding k additional vertices,

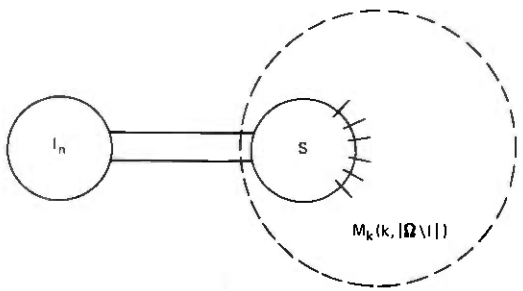


Fig. 7—k-rearrangeable graph with Steiner vertices.

called *Steiner vertices*. For example, we may consider the following graph with vertex set $I \cup \Omega \cup S$ and $|S| = k$ as shown in Fig. 7.

(*i*) The spanning subgraph of I and S is a complete bipartite graph $K_{n,k}$.

(*ii*) The spanning subgraph of S and $\Omega \setminus I$ is precisely the Manhattan graph $M_k(k, |\Omega \setminus I|)$.

This graph is clearly k-rearrangeable.

VI. GRAPH REPRESENTATIONS OF A SWITCHING NETWORK

Consider a graph G with vertex set $V(G)$. Let I and Ω be nonempty subsets of $V(G)$ and let $S = V(G) \setminus (I \cup \Omega)$, which we shall call the *Steiner set* of G.

The graph G corresponds to a switching network in the following way:

(*i*) $I \leftrightarrow$ inlet lines.

(*ii*) $\Omega \leftrightarrow$ outlet lines.

(*iii*) Edge $\{x, y\} \leftrightarrow$ a crosspoint between x and y.

(*iv*) $S \leftrightarrow$ additional lines.

For example, the rectangle network in Fig. 8 corresponds to the complete bipartite graph $K_{3,4}$.

The Manhattan graph $M(3, 5)$ of Fig. 1 corresponds to the rearrangeable network shown in Fig. 9.

An example of a network derived from the graph in Fig. 10a with a nontrivial Steiner set is shown in Fig. 10b.

In this way, a switching network can be represented by a graph. A rearrangeable graph then corresponds to a rearrangeable network. A k-rearrangeable graph corresponds to a rearrangeable network of capacity k.

Many problems in switching networks can in this way be viewed as graph-theoretic problems. Instead of minimizing the number of

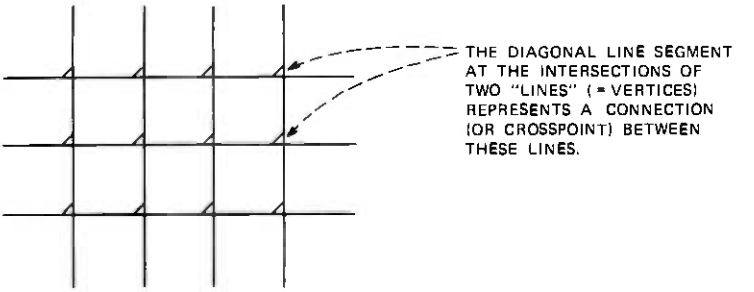


Fig. 8—Rectangle network of size 3×4.

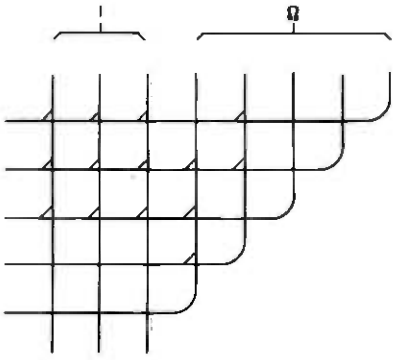


Fig. 9—Rearrangeable network corresponding to graph $M(3, 5)$.

crosspoints to reduce the cost of building a network, we consider the problem of finding a graph with the least possible number of edges. The size of S in the graph representation of a switching network determines how many lines we have to use in addition to the inlet and outlet lines. The Manhattan graph we have constructed then provides a model of a rearrangeable network with a minimum number of crosspoints for the case that the size of S is 0.

VII. CONCLUDING REMARKS

Almost all previous results on rearrangeable networks dealt with rearrangeable graphs having $|I| = |\Omega|$ and $|I \cap \Omega| = 0$. Beneš[3]

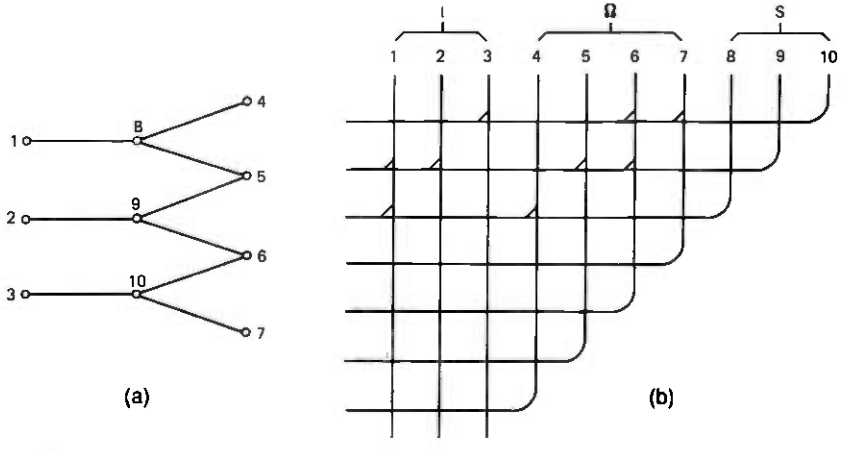


Fig. 10—(a) Graph with $I = \{1, 2, 3\}$, and $\Omega = \{4, 5, 6, 7\}$, and $S = \{8, 9, 10\}$. (b) Network corresponding to (a).

has shown that a rearrangeable network for $|I| = |\Omega| = n$ can be constructed with slightly more than $0(n \log n)$ crosspoints, which is just the information-theoretic lower bound. However, $|S|$ is required to be arbitrarily large to approach the $0(n \log n)$ bound. This result was later refined by Waksman[5] and Joel.[6]

When just one middle stage is allowed, Preparata[7] gave a lower bound on the number of crosspoints in a k-rearrangeable network and showed several optimal designs for arbitrary sizes of I and Ω.

With regard to nonblocking networks, we can define *nonblocking graphs* as those which satisfy the following property: The vertex set of a nonblocking graph G is $V(G) = I \cup \Omega \cup S$, where S is disjoint from I and Ω. For any assignment $A = \{(x_i, y_i) | i = 1, 2, \cdots, t\}$, we can find a path connecting x_i and y_i without disturbing the existing paths already connecting x_j and y_j, $1 \leq j < i$. In other words, there is always a path connecting x_i and y_i whose vertices and edges are disjoint from those of the previous paths.

If the vertex set of a nonblocking network is the union of I and Ω, one class of nonblocking graphs we can construct is formed from the union of a three-partite graph $K_{n,m,l}$ and a complete graph K_l, where $|I \cap \Omega| = l$, $|I| = n$, and $|\Omega| = m$, as shown in Fig. 5.

Bassalygo and Pinsker[8] have shown by a nonconstructive argument that there exist nonblocking networks with $0(n \log n)$ crosspoints, where $|I| = |\Omega| = n$ and the size of S approaches infinity. The best known construction, due to Cantor,[9] requires $0[n(\log n)^2]$ crosspoints.

VIII. ACKNOWLEDGMENT

The author wishes to thank F. K. Hwang for many helpful discussions.

REFERENCES

1. F. Harary, *Graph Theory*, Reading, Mass.: Addison–Wesley, 1972.
2. F. K. Hwang, personal communication.
3. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
4. P. Erdös and T. Gallai, "Graphs with Prescribed Degrees of Vertices," Nat. Lapok, *11*, 1960, pp. 264–274.
5. A. Waksman, "A Permutation Network," J.A.C.M. *15*, No. 1 (January 1968), pp. 159–163.
6. A. G. Joel, Jr. "On Permutation Switching Networks," B.S.T.J., *47*, No. 5 (May–June 1968), pp. 813–822.
7. F. P. Preparata, "On Multitransmission Networks," IEEE Trans. Circuit Theory, *CT-20*, No. 1 (January 1973), pp. 67–69.
8. L. A. Bassalygo and M. S. Pinsker, "On the Complexity of Optimal Switching Networks without Rearrangement," Problems Info. Transmission, *9*, 1973, pp. 84–87.
9. D. G. Cantor, "On Nonblocking Switching Networks," Networks, *1*, 1972, pp. 367–377.

Contributors to This Issue

Allen H. Cherin, B.E.E., 1961, City College of New York; M.S.E.E., 1965, University of Vermont; Ph.D. (E.E.), 1971, University of Pennsylvania; Bell Laboratories, 1965—. Mr. Cherin is engaged in studies associated with the characterization, splicing, and packaging of optical fibers. Member, IEEE, OSA.

Fan R. K. Chung, B.S., 1966, National Taiwan University; Ph.D., 1974, University of Pennsylvania; Bell Laboratories, 1974—. Mrs. Chung's current interests include combinatorics, graph theory, and the analysis of algorithms. She is presently investigating various problems in the theory of switching networks.

Daniel P. Heyman, B.Mgt.E., 1960, Rensselaer Polytechnic Institute; M.I.E., 1962, Syracuse University; Ph.D. (Operations Research), 1966, University of California at Berkeley; U.S. Air Force Logistics Command, 1960–63; Bell Laboratories, 1966—. Mr. Heyman has worked in various areas of operations research including queuing theory, economic modeling, and inventory theory. Member, ORSA, TIMS, Sigma Xi, Alpha Pi Mu.

Nuggehally S. Jayant, B.Sc., 1962, University of Mysore (India); B.E. (Distinction), 1965, and Ph.D., 1970, Indian Institute of Science, Bangalore; Research Associate, Stanford Electronics Laboratories, 1967–68; Visiting Scientist, Indian Institute of Science, January–March, 1972 and August–October, 1975; Bell Laboratories, 1968—. Mr. Jayant has worked on digital communication in the presence of burst-noise, on the detection of fading signals, on pattern discrimination problems, and on adaptive quantizers for waveform encoding.

Dietrich Marcuse, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–1957; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research and studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission

aspect of a light communications system. Mr. Marcuse is the author of three books. Fellow, IEEE; member, Optical Society of America.

Calvin M. Miller, B.S.E.E., 1963, North Carolina State University, M.S.E., 1966, Akron University; Bell Laboratories 1967—. Mr. Miller has developed equipment and methods for transmission line characterization. His present interests are in the area of fiber optics as a practical transmission medium. Member, Eta Kappa Nu, OSA.

Elizabeth J. Murphy, B.S. (Math.), 1964, Spring Hill College; M.S. (Math.), 1966, Auburn University; Bell Laboratories, 1967—. Ms. Murphy has worked on computation and numerical analysis associated with the characterization of transmission lines and optical fibers. Member, ACM.

Peter Noll, Dipl-Ing. (Electrical Engineering), 1964, Dr.-Ing. (Electrical Engineering), 1969, and venia legendi, 1974, Technical University of Berlin, Germany. Mr. Noll has been with the Heinrich-Hertz-Institut, Berlin, Germany, since 1964. He was involved in the development of electronic telephone exchanges and worked on problems of switching and routing of broadband signals. Since 1970, he has been engaged in research on adaptive speech encoding and on communication theory. He has also taught courses in digital signal processing and data reduction at the Technical University of Berlin. He was at Bell Laboratories during the summers of 1974 and 1975 on leave from the Heinrich-Hertz-Institut, Berlin. Member, Nachrichtechnische Gesellschaft, (NTG), Germany.