THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 56 March 1977 Number 3

Thick Dielectric Grating on Asymmetric Slab Waveguide

By D. MARCUSE

(Manuscript received September 13, 1976)

We discuss an approximate theory of scattering losses of a guided mode in an asymmetric slab waveguide with a thick grating on one side. The theory is an extension of an exact theory of thick dielectric gratings published previously. The results of the theory are presented in graphical form. The coupling coefficient between two guided waves traveling in opposite directions is calculated and compared to perturbation theory.

I. INTRODUCTION

Diffraction gratings deposited on top of a thin-film waveguide are useful as input and output couplers.[1,2] A guided wave traveling in the thin-film waveguide is scattered out into the two regions (air and substrate) adjacent to the film as it encounters the region of the diffraction grating. Ordinarily, the power that is scattered out of the thin-film guide splits up into several grating lobes; the number and strength of these lobes depends on the grating period D, the depth of the grating $2a$, and on the shape of its teeth, as shown in Fig. 1. The relationship between the propagation constant β of the guided wave, the index of refraction n_i of the medium into which the grating lobe escapes at angle θ_{mi}, and the grating period length D is expressed by the following equation,[2]

$$\cos \theta_{mi} = \frac{\beta - (2\pi m/D)}{n_i k}. \tag{1}$$

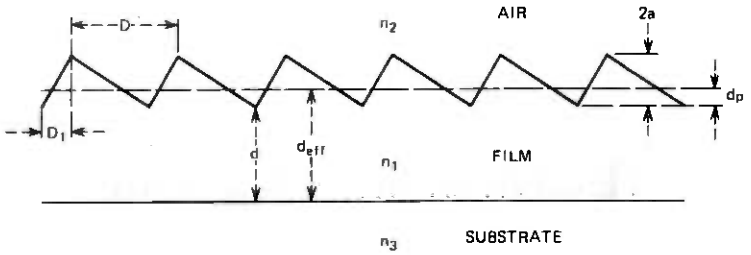


Fig. 1—Thick grating on a thin-film guide.

The integer m indicates the order of the grating lobe, $m = 1$ is the lobe of first order, $m = 2$, the lobe of second order, etc., and $k = 2\pi/\lambda$ is the free-space propagation constant. As the magnitude of the right-hand side of this equation exceeds the value unity, the scattering angle becomes imaginary and there is no scattered wave. (The angle θ_{mi} is measured with respect to the surface of the thin-film guide.) Thus, it is apparent that no scattered wave can escape from the film into medium i if $D < 2\pi/(\beta + n_i k)$. For values of D that are just larger than the right-hand side of the inequality (with n_i indicating the larger of the two refractive indices of the media adjacent to the film, the substrate say) a single grating lobe is radiated into the substrate. If D violates the inequality, with n_i being the refractive index of the air space (the region of lowest refractive index), a grating lobe is radiated into that region. If we let the value of D increase further, higher-order grating lobes begin to appear.

For purposes of coupling power from the outside into the film guide, a laser beam is directed at the grating and is aligned to coincide as closely as possible with one of the grating lobes.[2] If only one grating lobe exists, it is possible to capture most of the laser power and have it trapped in the thin-film guide. However, if other grating lobes exist, the laser power is split between the guided wave in the film and the other grating lobes so that the coupling efficiency for excitation of the trapped modes is reduced. It may be inconvenient to design a grating with only one lobe since this requires a grating with a very short period and also necessitates excitation of the thin-film guide through the substrate. For this reason, it is desirable to be able to control the amount of power radiated into undesirable grating lobes by shaping the form of the grating, that is, its teeth, in an appropriate way. Gratings with specially shaped teeth are known as blazed gratings.[3] An analysis of blazed gratings cannot be performed by using first-order perturbation theory because the grating, to be effective, must be thicker than is compatible with perturbation theory.

This paper proposes a new method of calculating grating responses

by an approximate method that, nevertheless, allows us to compute the response of thick gratings without having to search for the complex roots of a large determinant. Our approach is based on the exact grating theory described in Ref. 1. This exact theory is limited to TE waves (not an inherent limitation but one of convenience) and is applied to a grating defined as the boundary between two dielectric half spaces. A plane wave is incident from one side. The electromagnetic field outside of the grating is described as a superposition of infinitely many plane waves, most of which have propagation constants with one imaginary component. The field in the grating region is expressed as a double Fourier series. The unknown expansion coefficients are determined by matching of boundary conditions, not along the grating surface but along hypothetical planes adjacent to the grating.

This approach can easily be extended to the description of a grating on one side of a thin-film guide simply by adding the thin film to the structure and postulating plane waves in the film region. However, there is an important difference between the simple-grating and the waveguide-grating problems. The scattering problem of the grating between two half spaces is completely determined by the incident wave so that the amplitude coefficients of all the other waves can be obtained from an inhomogeneous equation system. The waveguide grating problem, on the other hand, leads to a homogeneous equation system. The distinction occurs because it is no longer possible to specify the direction of the incident wave, which is now the upward (or downward) traveling part of the standing wave pattern of the guided mode whose propagation constant is not known. In fact, the complex propagation constant would now be obtained as the solution of a determinantal eigenvalue equation.[2] However, the search for the complex solutions of a large determinantal equation is costly and time consuming and offsets the advantage of the original grating calculation.

To circumvent this problem, we propose a different approach. It is true that the exact eigenvalue of the determinantal equation is complex, but we know *a priori* that the real part of this complex solution, the propagation constant, is far larger than the imaginary part, the loss coefficient. This observation gives us confidence that it should be possible to determine the loss coefficient by computing the amount of radiated power once the problem has been approximately solved. The real part of the complex eigenvalue can be obtained by an approximation that is based on results obtained from the simpler grating theory described in Ref. 1. We have shown that the effective plane of reflection of the incident plane wave can be computed approximately by means of the WKB method. The comparison of the effective penetration depth computed from the WKB approximation with the exact result showed that the agreement was reasonable. We found that the penetration depth of the

wave was approximately given by the formula[1]

$$d_0 = \frac{4a\kappa_0^2}{3(n_1^2 - n_2^2)k^2} - \frac{1}{\kappa_0}\left[\frac{\pi}{4} - \arctan\frac{\gamma_0}{\kappa_0}\right]. \tag{2}$$

We use the definition

$$\kappa_0^2 = n_1^2 k^2 - \beta^2 \tag{3}$$

and

$$\gamma_0^2 = \beta^2 - n_2^2 k^2. \tag{4}$$

Figure 1 shows that $2a$ is the depth of the grating, n_1 the refractive index of the thin film, and n_2 represents the index of the medium above the film. The WKB solution that led to (2) fails (in the form used by us) as the grating becomes too thin. For this reason, we use as the penetration depth (see Figs. 12 and 13 of Ref. 1)

$$d_p = \begin{cases} d_0 \text{ if } d_0 < a \\ a \text{ if } d_0 > a \end{cases}. \tag{5}$$

The information gathered from Ref. 1 thus allows us to define an effective film thickness d_{eff} (see Fig. 1) as

$$d_{\text{eff}} = d + d_p \tag{6}$$

and thus enables us to calculate iteratively the propagation constant β from (3), (4), and[4]

$$\kappa_0 d_{\text{eff}} = \nu\pi + \arctan\frac{\gamma_0}{\kappa_0} + \arctan\frac{\delta_0}{\kappa_0}, \tag{7}$$

with

$$\delta_0^2 = \beta^2 - n_3^2 k^2 \tag{8}$$

(ν is the mode number of the guided wave; $\nu = 0$ for the lowest order TE mode.) The refractive index n_3 is the index of the medium on the other side of the film opposite the medium with index n_2.

Once the propagation constant of the guided mode is approximately known, we fix the value of that component of the standing wave inside of the thin film that approaches the grating and we solve the inhomogeneous equation system that results. It is clear that this equation system cannot provide an exact solution since we have already frozen the value of the propagation constant and have specified one of the two amplitudes associated with the guided wave to the right-hand side of the equation system, changing a homogeneous to an inhomogeneous equation system. However, we have checked that our approach gives precisely the same results as first-order perturbation theory for small values of the grating depth $2a$. Furthermore, the results obtained from this approximation

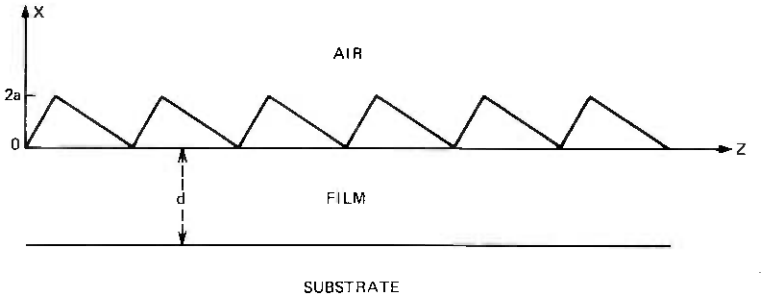


Fig. 2—The coordinate system in relation to the grating and the film.

agree well (in cases where agreement is to be expected) with the results of the exact grating theory. However, our present theory does not provide correct answers for the amplitudes of scattered waves inside the thin film that coincide with another guided mode. In this case, a "resonance" occurs and the results become meaningless. Coupling among guided modes thus cannot be handled by this theory and must be treated differently, as will be described later.

II. MATHEMATICAL FORMULATION OF THE PROBLEM

We use the following representation for the electric field[1] of the structure shown in Fig. 2:

$$E_y = \exp(-i\beta z) \sum_{m=-\infty}^{\infty} C_m \exp(-i\rho_m x) \exp\left(i\frac{2\pi}{D}mz\right) \quad \text{for } x \geqq 2a \quad (9)$$

$$E_y = \exp(-i\beta z) \sum_{n,m=-\infty}^{\infty} B_{nm} \exp[(i\pi/b)\, nx] \exp\left(i\frac{2\pi}{D}mz\right) \quad \text{for } 0 \leqq x \leqq 2a \quad (10)$$

$$E_y = \exp(-i\beta z) \sum_{m=-\infty}^{\infty} \{A_m^{(+)} \exp(-i\kappa_m x) + A_m^{(-)} \exp(i\kappa_m x)\} \exp\left(i\frac{2\pi}{D}mz\right) \quad \text{for } 0 \geqq x \geqq -d \quad (11)$$

$$E_y = \exp(-i\beta z) \sum_{m=-\infty}^{\infty} D_m \exp(i\sigma_m x) \exp\left(i\frac{2\pi}{D}mz\right) \quad \text{for } x \leqq -d. \quad (12)$$

We define

$$\beta_m = \beta - \frac{2\pi}{D}m \quad (13)$$

and express the parameters appearing in (9) through (12) as follows:

$$\rho_m^2 = n_2^2 k^2 - \beta_m^2 \tag{14}$$

$$\kappa_m^2 = n_1^2 k^2 - \beta_m^2 \tag{15}$$

$$\sigma_m^2 = n_3^2 k^2 - \beta_m^2. \tag{16}$$

The parameter b appearing in (10) is an arbitrary constant that should be larger than a. For our numerical evaluations we have used $b = a\sqrt{2}$. Equations (9), (11), and (12) are solutions of the wave equation but (10) is not. We solve our problem by substituting (10) into the wave equation, obtaining a set of equations for the determination of B_{nm}. However, these equations do not determine B_{nm} completely; in addition, we must satisfy boundary conditions by requiring that E_y and dE_y/dx remain continuous at $x = 2a$, $x = 0$, and $x = -d$. All these conditions lead to the following set of equation systems

$$\sum_{n',m'=-\infty}^{\infty} \left\{ N_{n'-n,m'-m} - \left[\left(\frac{\pi n'}{b} \right)^2 + \beta_{m'}^2 \right] M_{n,n'} \delta_{m,m'} \right\} B_{n',m'} = 0 \tag{17}$$

$$\sum_{n=-\infty}^{\infty} \left(\rho_m + \frac{\pi}{b} n \right) B_{nm} \exp\left(i \frac{2\pi}{b} na \right) = 0 \tag{18}$$

$$\sum_{n=-\infty}^{\infty} \left[\left(1 - \frac{\pi}{\kappa_m b} n \right) - \frac{\kappa_m - \sigma_m}{\kappa_m + \sigma_m} \exp\left(-2i\kappa_m d\right) \left(1 + \frac{\pi}{\kappa_m b} n \right) \right] B_{nm} = 0 \quad \text{for } m \neq 0. \tag{19}$$

If we remove the restriction $m \neq 0$ from (19), the combined equation system (17) through (19) would represent the exact formulation of our problem. However, since this would force us to solve the determinantal eigenvalue equation for complex β, we exclude the equation with $m = 0$ from (19) and add instead the following inhomogeneous equation to our set

$$\sum_{n=-\infty}^{\infty} \left[\left(1 - \frac{\pi}{\kappa_0 b} n \right) + \frac{\kappa_0 - \sigma_0}{\kappa_0 + \sigma_0} \exp\left(-2i\kappa_0 d\right) \left(1 + \frac{\pi}{\kappa_0 b} n \right) \right] B_{no} = 4A_0^{(+)}. \tag{20}$$

The equation system (17) stems from the substitution of (10) into the wave equation. The coefficients $M_{n,n'}$ and $N_{n'-n,m'-m}$ are defined in Ref. 1. Equations (18), (19), and (20) result from the boundary conditions. In fact, the left-most term in parenthesis in (20) as well as the term with the exponential function are each individually equal to $2A_0^{(+)}$. Equation

(20) is (twice) the arithmetic mean of these two equations and (19) (with $A_m^{(+)}$ instead of $A_0^{(+)}$) is their difference. We took the arithmetic mean of two equations, each expressing the relation between $A_0^{(+)}$ and B_{no}, to improve the accuracy of the approximation. The difference must, of course, be taken to eliminate $A_m^{(+)}$ from the exact equation system.

Equations (17) through (20) allow us to express B_{nm} in terms of $A_0^{(+)}$. For purposes of normalization, we express the amplitude coefficient $A_0^{(+)}$ in terms of the power P carried by the guided mode,

$$A_0^{(+)} = \left[\frac{\omega\mu_0 P}{\beta\left(d_{\text{eff}} + \frac{1}{\gamma_0} + \frac{1}{\delta_0}\right)}\right]^{1/2} \tag{21}$$

Finally, we need the amplitude coefficients of the scattered waves which may be expressed in terms of B_{nm} as follows,

$$C_m = \sum_{n=-\infty}^{\infty} B_{nm} \exp i\left(\rho_m + \frac{\pi}{b} n\right) 2a \tag{22}$$

and

$$D_m \exp(-i\sigma_m d) = \sum_{n=-\infty}^{\infty} \frac{\kappa_m(\kappa_m \cos \kappa_m d + i\sigma_m \sin \kappa_m d) - \frac{\pi}{b} n(\sigma_m \cos \kappa_m d + i\kappa_m \sin \kappa_m d) B_{nm}}{\kappa_m^2 - \sigma_m^2} \tag{23}$$

Knowing the amplitudes of all scattered waves, we can calculate the power that is carried away from the thin-film waveguide. We use the partial power attenuation coefficients

$$2\alpha_{2m} = \frac{\rho_m |C_m|^2}{2\omega\mu_0 P} \tag{24}$$

and

$$2\alpha_{3m} = \frac{\sigma_m |D_m|^2}{2\omega\mu_0 P} \tag{25}$$

and obtain the total power attenuation coefficient as the sum

$$2\alpha = \sum_m (2\alpha_{2m} + 2\alpha_{3m}), \tag{26}$$

where the summation extends over all real, propagating waves.

III. COUPLING COEFFICIENT BETWEEN GUIDED MODES

We are interested in finding the coupling coefficient for power transfer from the incident guided wave to its backward traveling counterpart. This is an important design parameter for distributed feedback lasers. The exact solution of our problem would give us this coefficient because we would know the amplitude coefficients of all the waves whether guided or not. Our approximate procedure fails if a principal grating lobe scatters power into the direction corresponding to another guided mode. For this reason, we use a different approach. If we want to couple the incident guided mode to the backward traveling mode via the first grating order, we need a grating period that is given by the formula,[5]

$$D = \frac{\pi}{\beta}. \tag{26}$$

A grating with such a short period does not scatter power out of the thin-film guide. We only need to know the amount of power scattered per unit length into the opposite direction. If the amplitude coefficient of this backward scattered wave is $A_1^{(-)}$, the coupling coefficient is defined as[6]

$$R = \frac{\kappa_0}{2\beta(d_{\text{eff}} + (1/\gamma_0) + (1/\delta_0))} \frac{A_1^{(-)}}{A_0^{(+)}}. \tag{27}$$

To first order of perturbation theory, we obtain from (27)

$$R = \frac{\kappa_0^2}{2\beta(d_{\text{eff}} + (1/\gamma_0) + (1/\delta_0))} a_1. \tag{28}$$

The factor a_1 is the Fourier coefficient of the spatial frequency component $2\pi/D$ of the grating function. For our triangular grating shape, we have

$$a_1 = \frac{2aD^2}{\pi^2(D - D_1)D_1} \sin \pi \frac{D_1}{D}. \tag{29}$$

We have stated the result of perturbation theory only for comparison purposes. We evaluate the coupling coefficient from (27) by calculating $A_1^{(-)}$ with the help of the exact grating theory developed in Ref. 1.

The simple, exact grating theory can be used to approximate waveguide losses by assuming that all waves that are scattered at the grating penetrate through the thin-film boundaries without any further reflection. We shall see that this assumption yields good results if the grating is on the side of the film with the greater index difference (the air side). For gratings on the substrate side, reflections from the opposite film boundary are important and the simple-minded approach yields unsatisfactory results. However, it is interesting to compare the results of the approximate theory presented here with loss calculations based

on the simple grating, since such a comparison can tell us when we can use the results of the simple grating theory directly and when we need the more sophisticated (if approximate) approach presented in this paper. A comparison of the two theories is also useful to give us confidence in the results of the approximate theory.

The partial waveguide losses can be computed from the simple grating theory by using (24) unchanged (except for the fact that the simple grating theory of Ref. 1 is now used to compute C_m) and by replacing D_m in (25) with A_m obtained from (16) of Ref. 1.

A discussion of the number of terms used in the series expansion of the field was given in Ref. 1.

IV. DISCUSSION OF RESULTS

Careful comparison of the results of our present theory with the perturbation theory[7] shows perfect agreement for small values of the grating depth $2a$. It is, of course, necessary to replace the amplitude of the sinusoidal grating (designated as σ in Ref. 7) with the Fourier amplitudes coefficient (29).

To show the difference of the scattering losses that result from using the present waveguide theory and to compare it to the simple grating theory, we have drawn in Fig. 3 the partial scattering loss of the first grating lobe for a grating with vanishingly small depth $2a$. The curves in this and subsequent figures are labeled accordingly. We normalize the loss coefficient by multiplying it with λ^3/a^2 to make it dimensionless and to reduce its dependence on a. To first order of perturbation theory the normalized attenuation coefficient should be independent of a.

The independent variable on the horizontal axis of all our figures is the scattering angle $\phi = 90 - \theta_{13}$ [see eq. (1)] of the first-order beam ($m = 1$), the wave corresponding to this angle escapes into the medium with the higher refractive index n_3. The angle ϕ is varied by varying the grating period D.

This practice of using the scattering angle of the substrate beam as the independent variable and defining it with respect to the direction normal to the film surface is taken from Ref. 7. Figure 3 and all subsequent figures use $n_1 = 1.59$, $n_2 = 1.0$, and $n_3 = 1.458$ (in some later figures, n_2 and n_3 will be interchanged). Furthermore, we use $d = d_{eff} = 0.571$; this choice was made to compare waveguides having the same effective width. Figure 3 applies to a symmetrical grating with $D_1/D = 0.5$. It is apparent how very similar the results of the two approaches are. The air beam disappears at an angle of 43.3°, because we have labeled all beams with the angle of the beam in the substrate and the angle of the air beam is related to the angle in the substrate by Snell's law.

A departure from the results obtained using the waveguide theory and the result calculated from the simple grating theory is discernible only

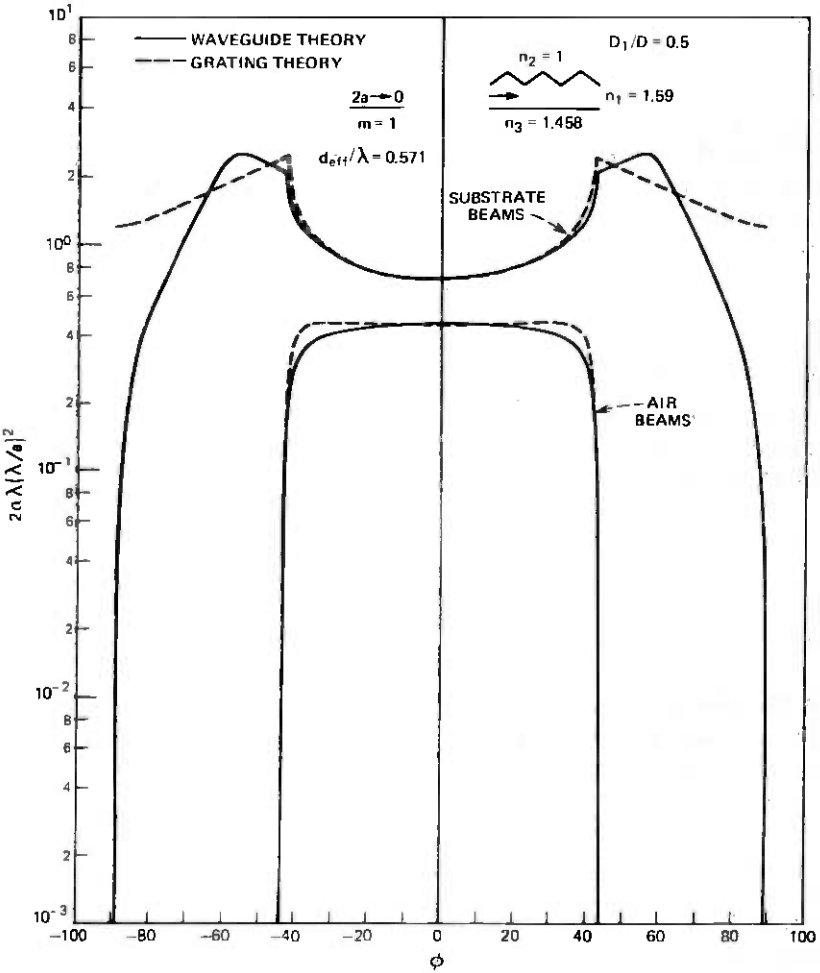


Fig. 3—Comparison of the waveguide grating theory with the simple grating theory for a symmetrical grating of vanishing depth ($2a/\lambda \to 0$). Shown is the normalized scattering loss coefficient of the first-order substrate beam. Film index $n_1 = 1.59$, air index $n_2 = 1$, substrate index $n_3 = 1.458$.

at beam angles that correspond to beams that nearly graze the surface. At these angles, reflection from the film-substrate interface becomes noticeable and indicates the difference in the solid and dotted curves. In particular, we see that the substrate beam, expressed by the solid line, vanishes at $\phi = \pm 90°$, whereas the dotted line remains at a finite value. This difference is caused by the fact that the substrate beam goes over into a guided mode in the waveguide case, but in the simple grating, where no guided modes exist, the scattering angle in the film can become still larger so that there is no discontinuity at the point where the actual

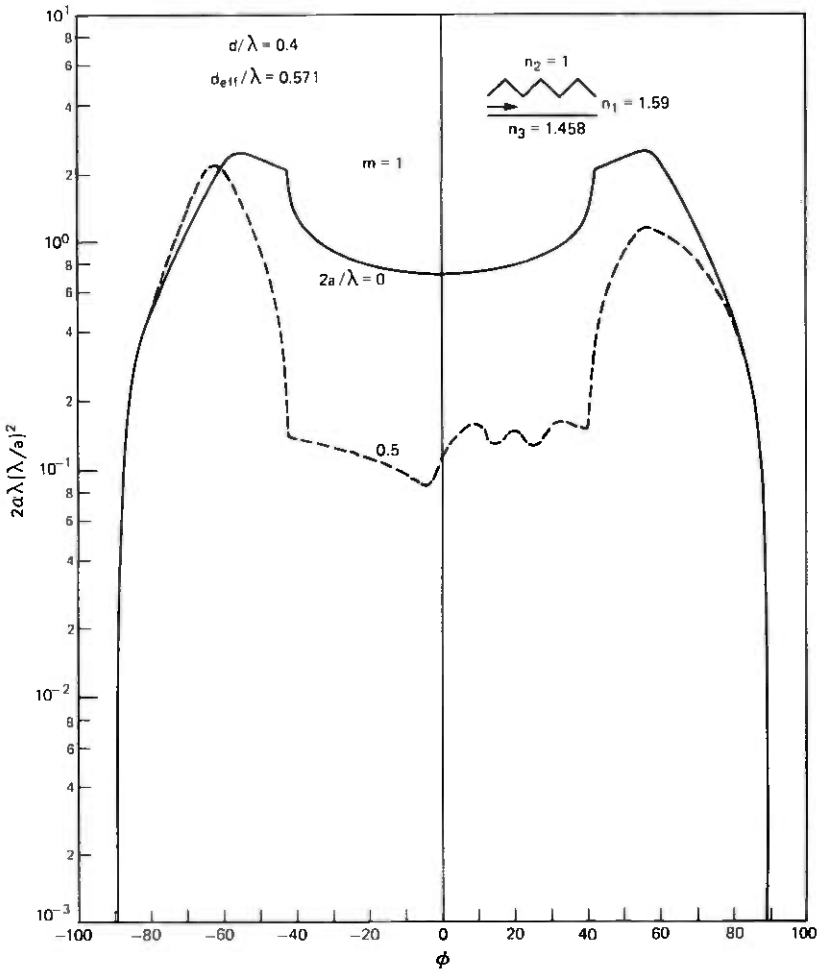


Fig. 4—Substrate beams: comparison of perturbation theory ($2a/\lambda \rightarrow 0$) and thick grating waveguide theory ($2a/\lambda = 0.5$) for a symmetrical grating on the air side of the film. The first-order substrate beam loss is shown.

substrate beam vanishes. However, the angle of the simple grating response has been adjusted by Snell's law to correspond not to the film but to the substrate angle, even though reflection at this interface does not exist in case of the simple grating.

Figure 4 provides a comparison between perturbation theory ($2a/\lambda \rightarrow 0$) and the first-order grating response for a grating on a thin-film waveguide with thickness $2a/\lambda = 0.5$. We have used a film thickness of $d/\lambda = 0.4$, but the thick grating increases the effective film thickness to $d_{\text{eff}}/\lambda = 0.571$. To have a meaningful comparison, we have used this film thickness also for the case $a \rightarrow 0$. Figure 4 shows clearly that per-

turbation theory overestimates the scattering losses of thick gratings. However, for ϕ near $\pm 90°$, the agreement between perturbation theory and the more precise theory is still remarkably close. This seems to be a general tendency which we shall encounter again. Figure 4 holds for the substrate beam while Fig. 5 shows a comparison between perturbation theory and the more precise theory for the air beam. Figure 6 applies to the same case, i.e., a symmetrical grating on the air side of the film, and shows the total scattering loss (power carried away by all grating orders in both media) as the solid line and compares it with the power carried away by the first-order grating response in the substrate indicated by the dotted line. The difference between the total amount of scattered power and the power in the first-order substrate beam is made

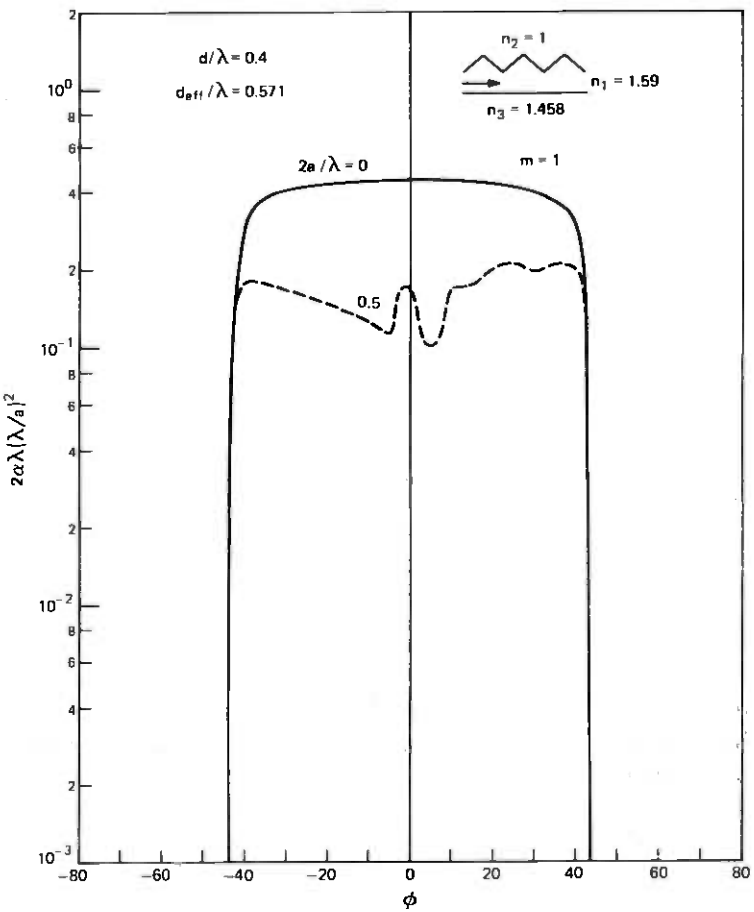


Fig. 5—Air beams: comparison of first-order perturbation theory ($2a/\lambda = 0$) with thick grating theory ($2a/\lambda = 0.5$) for the air beam with grating on air-film interface.

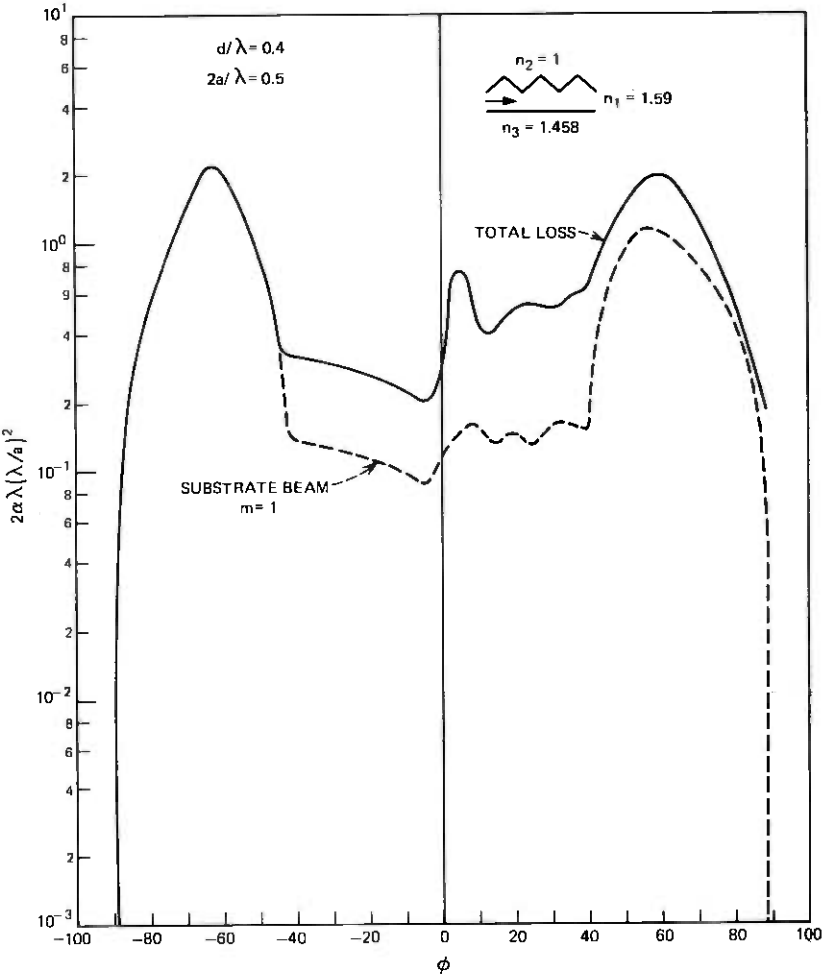


Fig. 6—The total loss is compared to the loss caused by the first-order substrate beam for $2a/\lambda = 0.5$ and a symmetrical grating on the air-film interface.

up partly by the power carried by the first-order air beam and partly by all the other grating orders. As the angle ϕ increases, more and more grating lobes appear. Rather than show each grating order separately we have added them all and have presented the total loss. The curve representing the total loss does not go to zero at $\phi = 90°$, because the grating responses of higher order do not vanish as the first-order substrate beam disappears inside of the thin film.

Fig. 7 shows the scattering losses of an asymmetric grating on the air side of the thin film with $D_1/D = 0$. We have included the total scattering loss as the topmost solid line, the first-order substrate beam as the dotted

line, and the first-order air beam as the lowest solid line. The most conspicuous aspect of this figure is the fact that so much more power is carried by the first-order substrate beam compared to the first-order air beam. The grating asymmetry is responsible for preferential scattering into the substrate. Comparison of Figs. 4 and 5 shows that the symmetrical grating scatters roughly equal amounts of power into air and substrate in the angular range where both beams exist simultaneously. Fig. 7 shows that a relatively small amount of power is scattered into higher-order grating modes, because the curve for the first-order substrate beam does not lie far below the total loss curve.

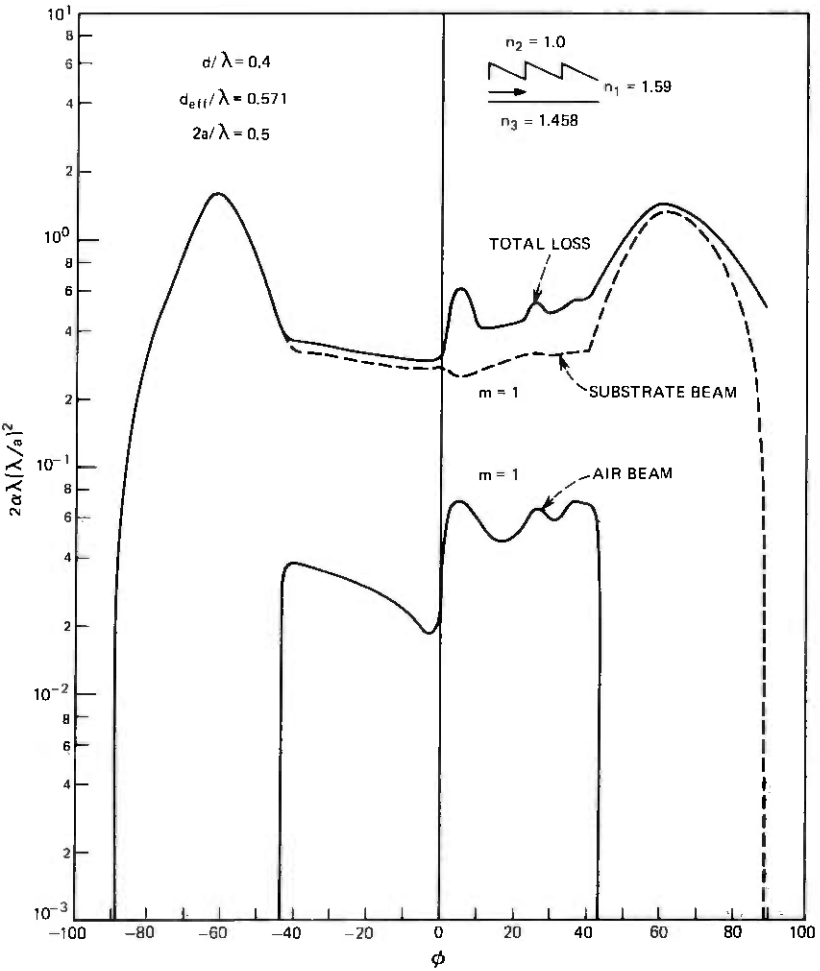


Fig. 7—Total loss, first-order substrate beam, and first-order air beam loss for an asymmetrical grating with $D_1/D = 0$ and $2a/\lambda = 0.5$.

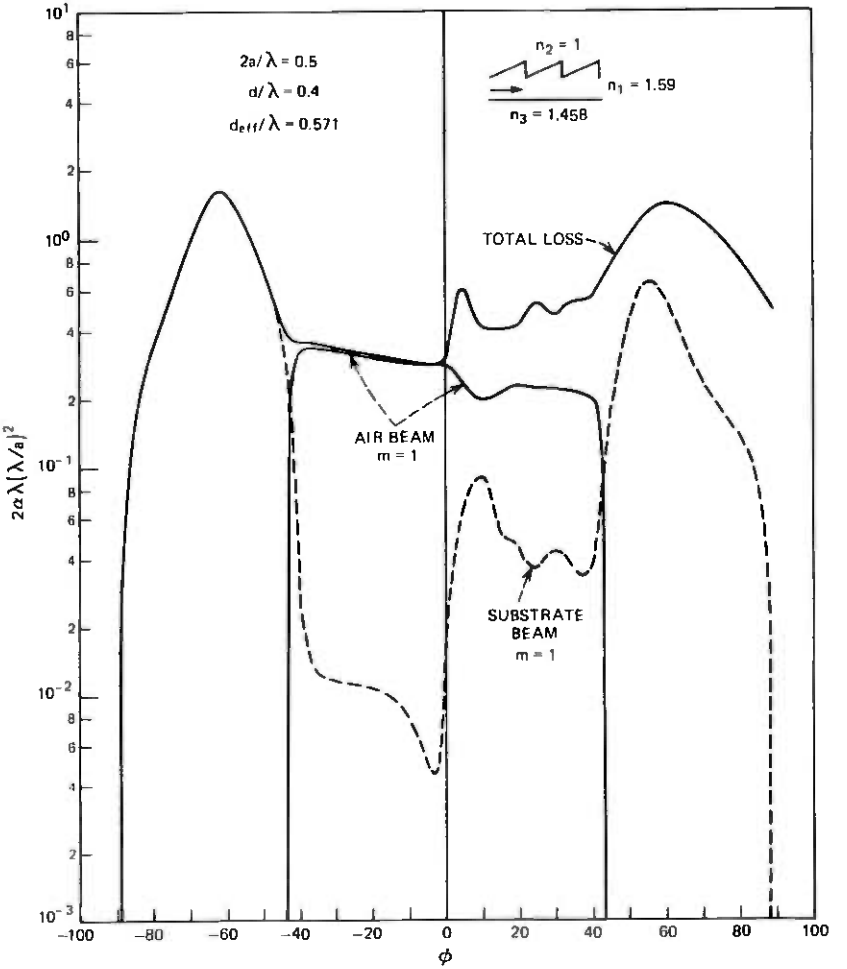


Fig. 8—Same comparison as in Fig. 7 for an asymmetrical grating with $D_1/D = 1$.

Figure 8 applies to a grating with the opposite asymmetry, $D_1/D = 1$. The total loss is the same as in Fig. 7 but the roles of substrate and air beams have been interchanged in the range $-43° < \phi < 43°$. For angles below this range, the substrate beam is identical to the corresponding beam for the grating with the opposite symmetry. For $\phi > 43°$, the substrate beam carries again significantly higher power than inside the range $-43° < \phi < 43°$ but higher-order modes now carry far more power at angles $\phi > 43°$ than in Fig. 7. An explanation of the influence of the grating shape in terms of geometrical optics is given in Ref. 1.

We have compared the results of the waveguide grating theory with the simple grating theory in Fig. 3 for the case of very thin gratings.

Figures 9 and 10 show such a comparison for a thick, asymmetric grating with $2a/\lambda = 0.5$ and $D_1/D = 1$. We see that we can obtain most of the scattering information from the simple grating theory. The two curves depart significantly only near the ends of the angular range of the substrate beam.

So far we have considered only gratings on the air side of the thin film. The next six figures apply to gratings on the substrate side of the film. We obtain these results from our theory simply by interchanging the roles of n_2 and n_3, with the values $n_1 = 1.59$, $n_2 = 1.458$, and $n_3 = 1.0$. For a deep grating with $2a/\lambda = 0.5$ and $d/\lambda = 0.4$, we now obtain a very

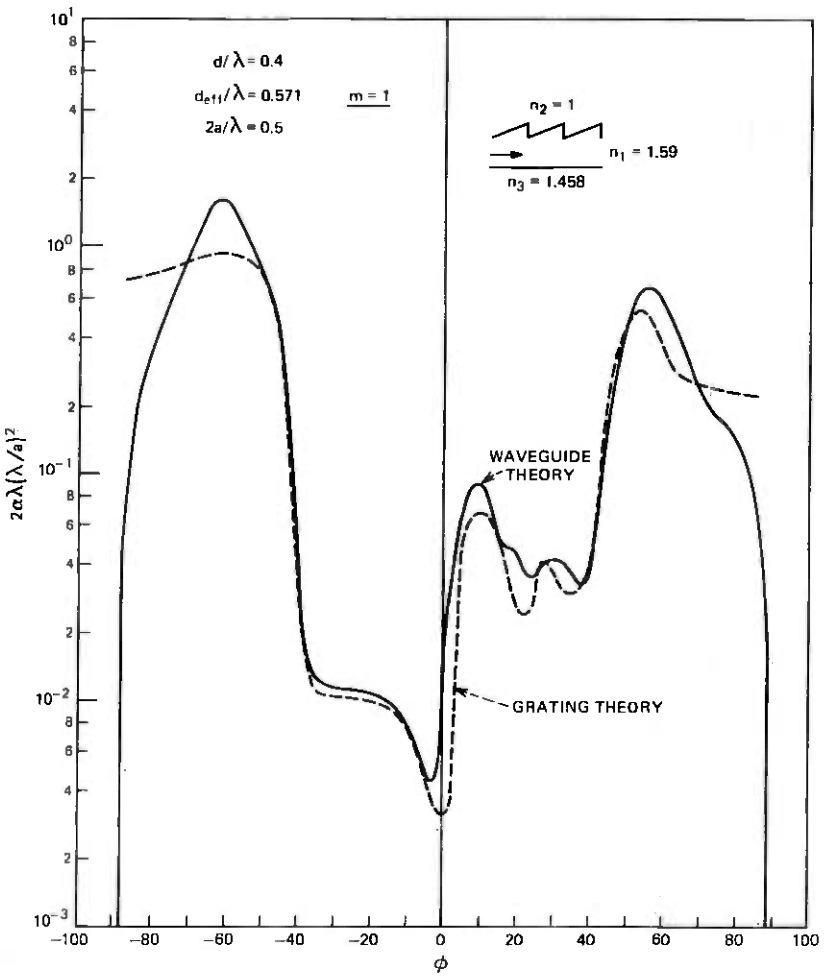


Fig. 9—Substrate beams: comparison of the grating guide theory with the simple grating theory for an asymmetrical grating with $D_1/D = 1$ for $2a/\lambda = 0.5$. The partial loss coefficient for the first-order substrate beam is shown.

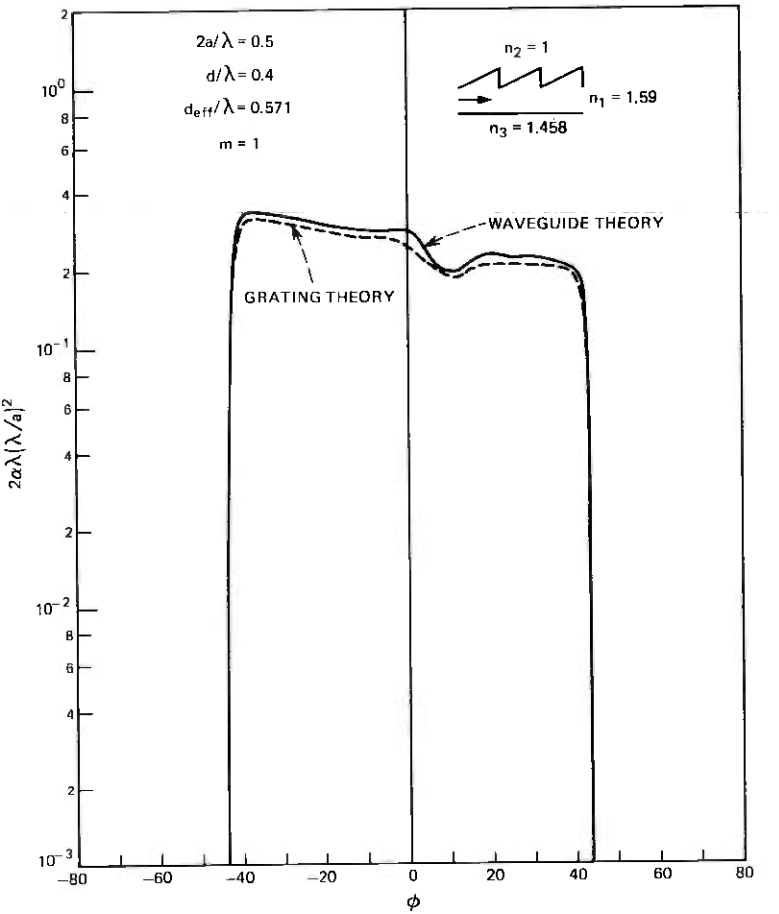


Fig. 10—Same as Fig. 9 for first-order air beams.

slightly different effective film thickness of $d_{\text{eff}}/\lambda = 0.569$. Figure 11 shows a comparison of perturbation theory ($2a/\lambda \to 0$) and thick grating theory for $2a/\lambda = 0.5$ for the first-order substrate beam for a symmetrical grating with $D_1/D = 0.5$. This figure should be compared with Fig. 4, because both cases are similar with the only difference being that the grating is now on the opposite side of the thin film. The thick grating theory is now in much closer agreement with perturbation theory, but both theories show a markedly different behaviour from the curves in Fig. 4, since there is obviously far more interference between the direct beam and the component that is reflected only once at the air-film interface. The deeper nulls discernible in the thick grating theory (dotted lines in Fig. 11) are caused by the fact that a slight shift has occurred that places the regions of destructive interference at angles where total in-

ternal reflection occurs at the air-film interface. Figure 12 shows a similar comparison for the first-order air beam for the same grating geometry. This figure should be compared with Fig. 5. Figure 12 is quite similar in shape to Fig. 5, but the curves are much lower, showing that air beam scattering is weaker if the grating is on the substrate side of the film. There are no pronounced interference effects, because the reflection from the film-substrate interface is much weaker. The dotted line in Fig. 12 labeled grating theory was computed with the help of the simple grating theory and shows remarkably close agreement with the grating guide theory.

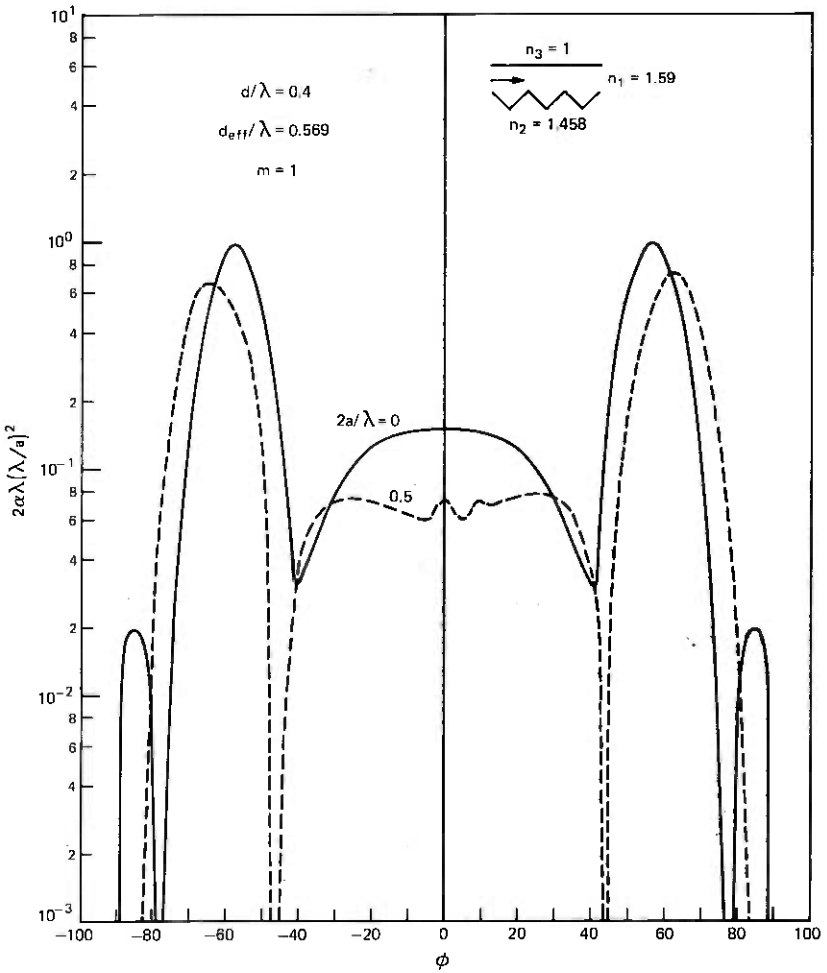


Fig. 11—Substrate beams: grating on film-substrate interface. Comparison between perturbation theory and thick waveguide grating theory ($2\alpha/\lambda = 0.5$) for a symmetrical grating, $D_1/D = 0.5$ for first-order substrate beams. Note the deep interference nulls.

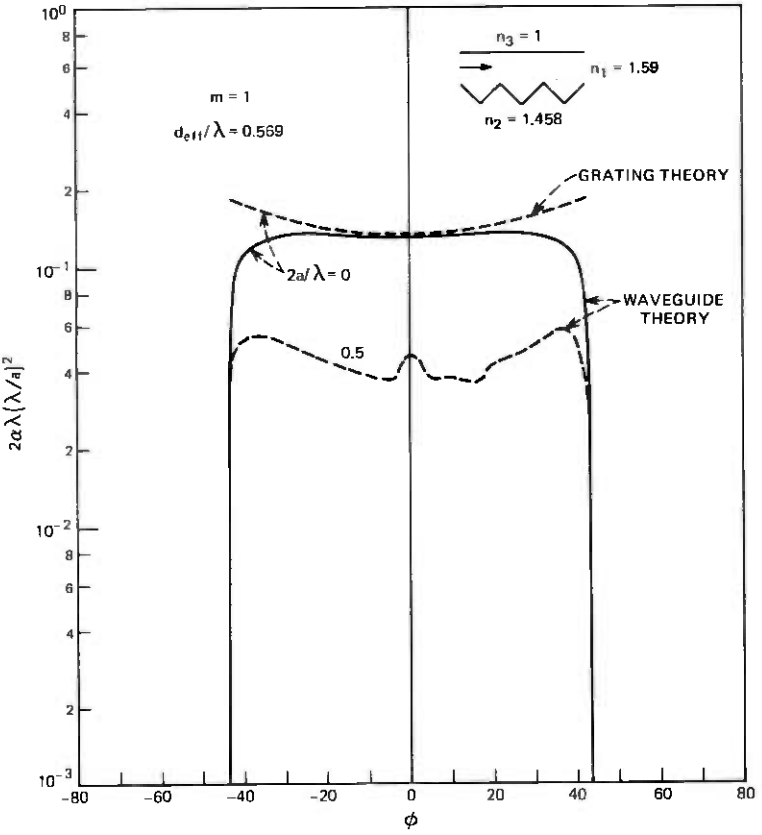


Fig. 12—Same as Fig. 11 showing the first-order air beams.

Figure 13 compares the total loss to the loss associated with power carried away by the first-order substrate beam for a thick grating with $2a/\lambda = 0.5$.

Figure 14 compares the theory of the simple grating with the waveguide grating theory for the first-order substrate beam for a thin grating ($2a/\lambda \rightarrow 0$) at the film-substrate interface. We see that the simple grating theory does not always suffice to predict the performance of a thin-film waveguide with a diffraction grating. The simple grating theory predicts the average loss correctly, but fails completely to account for interference effects. This figure should be compared with Fig. 3. The comparison shows that the simple theory is quite useful as long as interference effects between a direct and a reflected beam can be neglected, as in the case of the grating on the film-air interface (Fig. 3). For a grating on the film-substrate interface (Fig. 14), the simple grating theory is not applicable to the waveguide case. Figure 15 shows the comparison of the

two theories for a thick grating with $2a/\lambda = 0.5$ for the first-order substrate beam, whereas Fig. 16 compares the corresponding first-order beam in the air space. Just as in Fig. 12, the simple grating theory gives a good description of scattering for the air beam even if the grating is thick and is located on the film-substrate interface.

The last figure, Fig. 17, shows the normalized coupling coefficient R (multiplied by λ^2/a) as a function of the normalized grating thickness $2a/\lambda$ for gratings on the film-air interface (solid lines) and on the film-substrate interface (dotted lines) for symmetric ($D_1/D = 0.5$) and asymmetric gratings ($D_1/D = 0$ and 1). The two extreme cases of

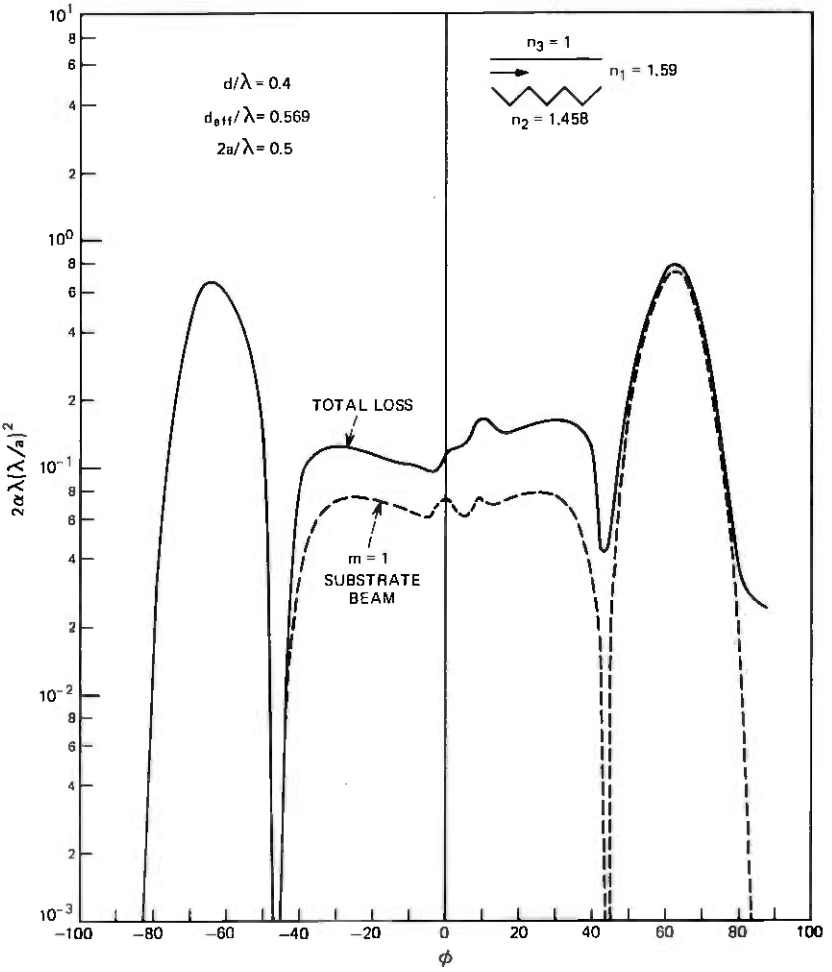


Fig. 13—Comparison of total loss and partial first-order substrate beam loss for a symmetrical grating on the film-substrate interface.

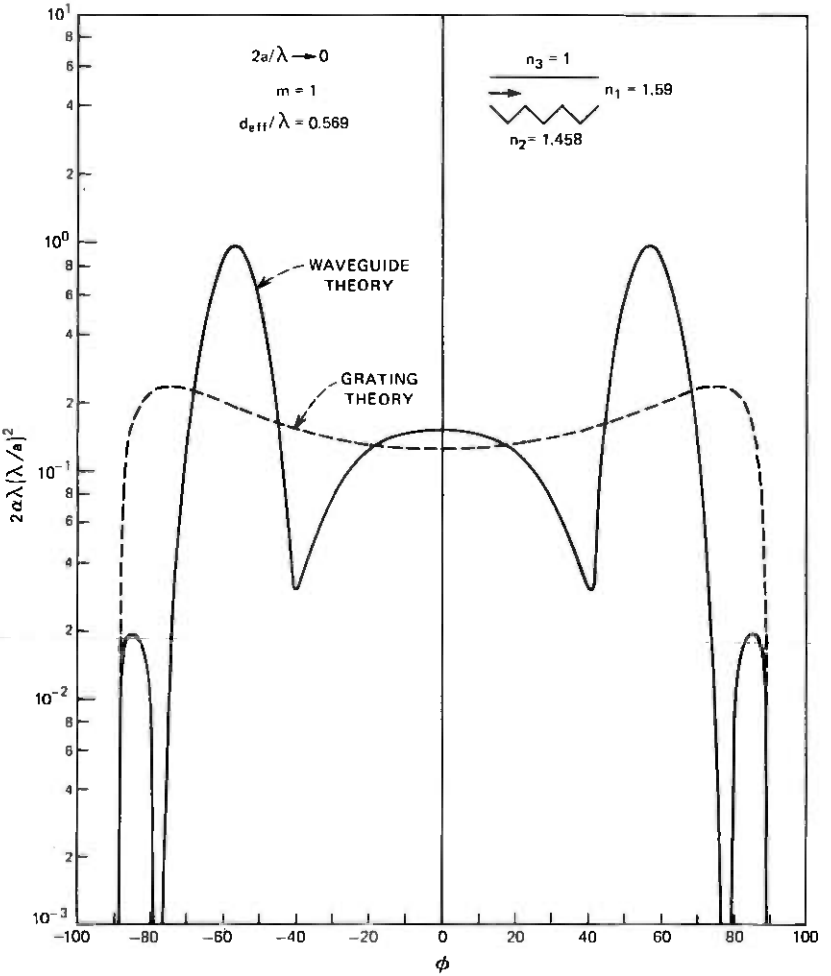


Fig. 14—Substrate beams: comparison of the waveguide grating theory and the simple grating theory for a thin grating ($2a/\lambda \to 0$) on the film-substrate interface.

asymmetry give exactly the same results. At $2a/\lambda = 0$ the curves agree, of course, with the perturbation theory (28). The most remarkable fact about the curves of Fig. 17 is their slight departure from the prediction of perturbation theory. Corresponding curves drawn from perturbation theory would be straight horizontal lines coinciding with our curves at $2a/\lambda = 0$. The exaggerated scale of the figure shows the downward slope for increasing grating thickness, but even for a grating whose thickness is equal to the vacuum wavelength of the wave, the results of the thick grating theory differ from perturbation theory by no more than 30%. This result is in agreement with the general trend that we observed in Fig. 4,

where we saw that the thick grating theory is in close agreement with perturbation theory near $\phi = -90°$. Coupling between a forward and backward traveling guided mode is an extreme case of backward substrate scattering, except that the beam does not escape into the substrate but is trapped in the film by total internal reflection. Figure 4 shows clearly how much perturbation theory and thick grating theory can differ at scattering angles that are more nearly normal to the film surface. Figure 17 thus shows that the perturbation formula (28) is remarkably accurate even for rather thick gratings whose thickness approaches the vacuum wavelength of the light wave. The curves in Fig. 17 were com-

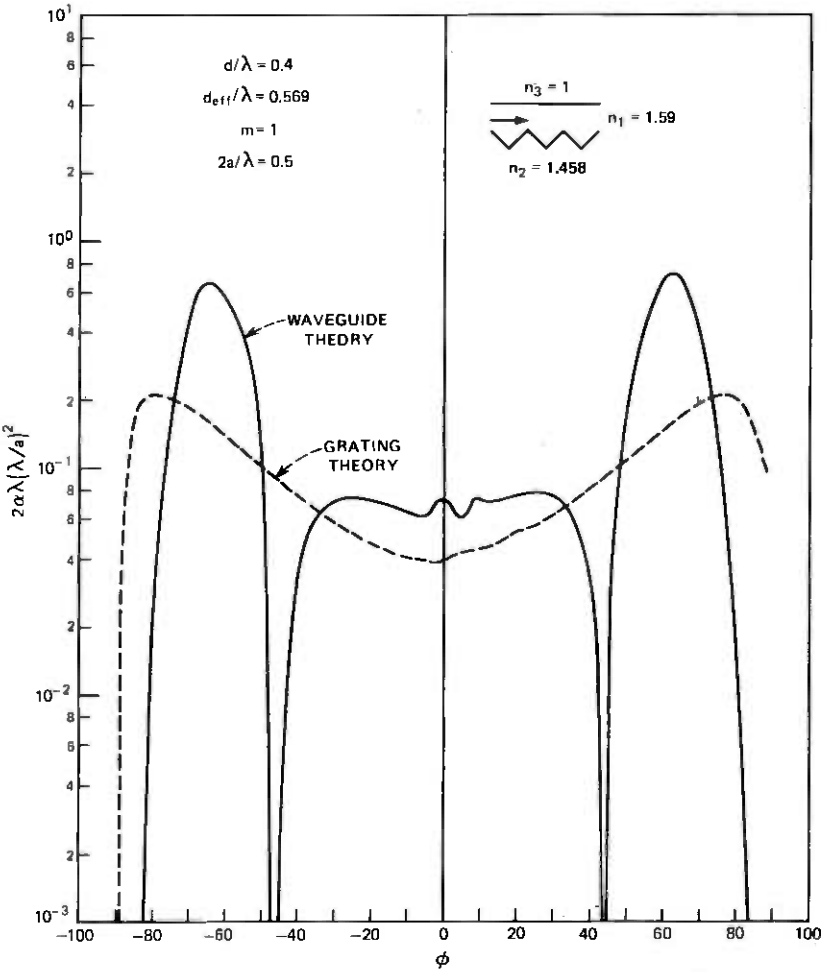


Fig. 15—Same as Fig. 14 for thick grating with $2a/\lambda = 0.5$.

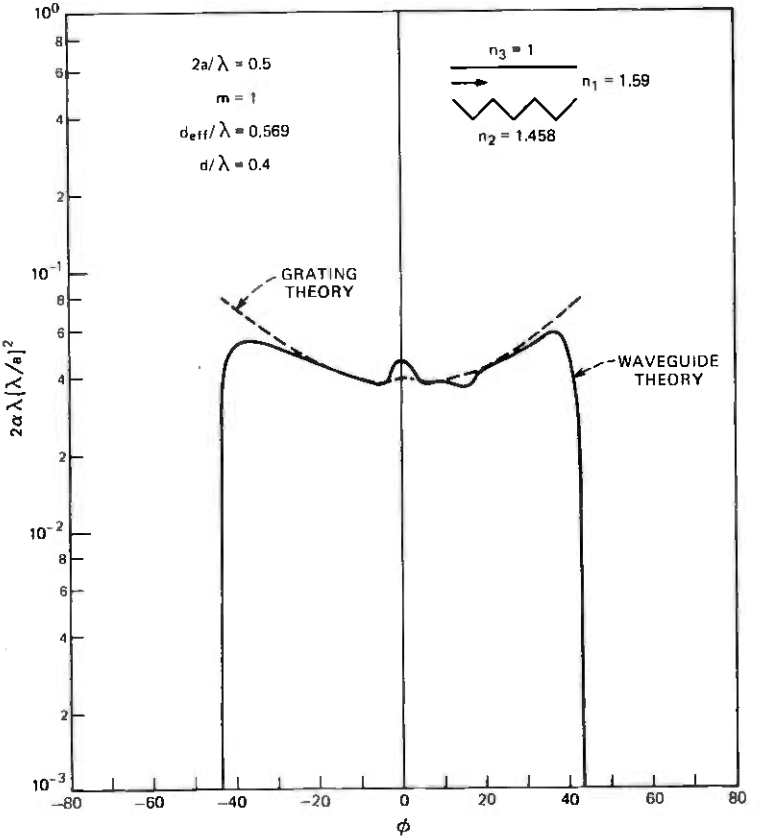


Fig. 16—Same as Fig. 15 showing the partial loss coefficients for first-order air beam scattering.

puted from (27), where the scattered wave amplitude $A_1^{(-)}$ is obtained from the simple grating theory.[1]

V. CONCLUSIONS

We have presented an approximate theory for scattering of power from a guided thin-film mode into the areas above and below the film by a thick diffraction grating deposited on one side. This theory has been compared with perturbation theory[7] and with the results of the exact, simple grating theory for a grating between two dielectric half spaces, and good agreement has been obtained in all cases where agreement can be expected. We are confident that our theory yields reasonable results for light scattering out of a thin film.

However, this theory does not give correct answers if applied to cou-

pling between two guided modes, even in the limit of very thin gratings where the correct answer is known from perturbation theory. This failure of the theory in the case of coupling among guided modes is understandable when we realize that a guided mode is at transverse resonance in the thin-film guide. The naive theory, that is based on the assumption that the mode amplitudes remain constant along the thin film, cannot account for a resonant situation where the power exchange may be complete and where mode coupling leads to new normal modes of the structure. On the other hand, it does not seem to hurt the calculation of the radiation loss coefficients if a minor grating lobe accidentally scatters power into the direction of a guided-film mode. Such "resonances" do occur, for example, over the angular range shown in Fig. 3

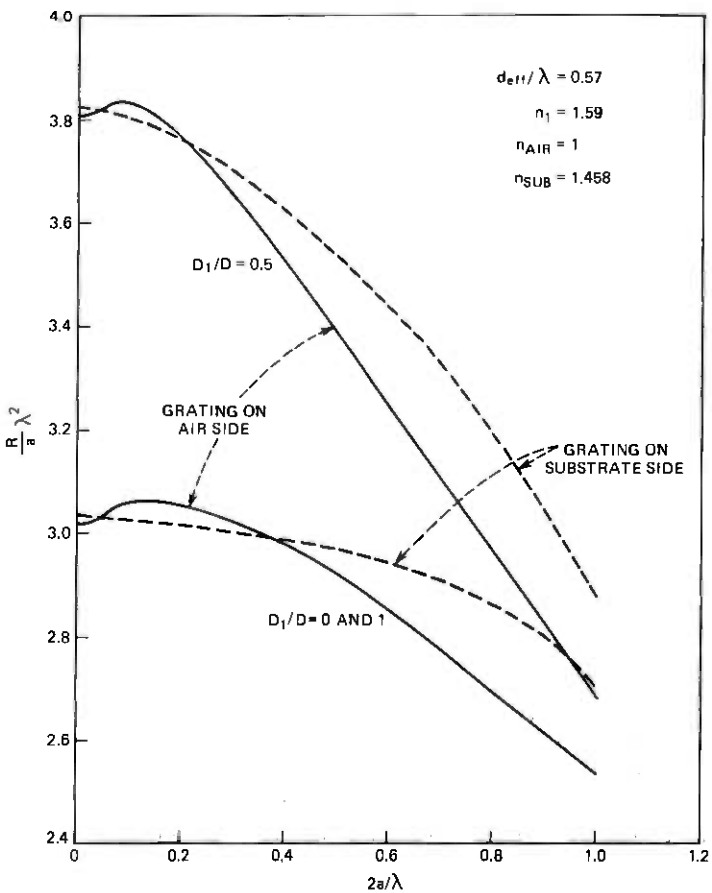


Fig. 17—Coupling coefficients between forward and backward guided mode. The solid lines hold for a grating on the film-air interface, the dotted lines describe a grating on the film-substrate interface. The curves for $D_1/D = 0$ and $D_1/D = 1$ are identical.

and the excellent agreement with perturbation theory and with the simple exact grating theory indicates that no difficulties have occurred.

Our theory has helped to clarify the areas where the simple grating theory[1] may be used to predict scattering losses even for film waveguides, but it has also shown that the simple grating theory does not help to predict waveguide losses if strong interference between a direct and a reflected beam may occur.[8]

Finally, we have used the simple grating theory to compute the coupling coefficient between two guided modes traveling in opposite directions. We found that perturbation theory holds approximately over a surprisingly wide range of grating thicknesses. Coupling between modes other than forward and backward modes could be treated very similarly.

Our approximate waveguide grating theory has the advantage of allowing direct calculations of power scattering without the need for a search routine for finding the complex roots of a large determinantal equation. It is, thus, a cheap and fast method for calculating the scattering properties of thick gratings on thin-film waveguides.

VI. ACKNOWLEDGMENT

The author acknowledges the contribution made to this paper by fruitful discussions with W. W. Rigrod.

REFERENCES

1. D. Marcuse, "Exact Theory of TE-Wave Scattering From Blazed Dielectric Gratings," B.S.T.J., *55*, No. 8 (October 1976) pp. 1295–1317.
2. T. Tamir, "Beam and Waveguide Couplers," in *Topics in Applied Physics*, Vol. 7 of *Integrated Optics*, New York: Springer Verlag, 1975, pp. 83–137.
3. M. Born and E. Wolf, *Principles of Optics*, Third ed., New York: Pergamon Press, 1965.
4. D. Marcuse, "Theory of Dielectric Optical Waveguides," New York: Academic Press, 1974, Eq. (1.2–12), p. 6.
5. H. Kogelnik and C. V. Shank, "Coupled Wave Theory of Stimulated Emission in Periodic Structures," J. Appl. Phys. *43*, No. 5 (May 1972), pp. 2327–2335.
6. Ref. 4, Eq. (4.3–33), p. 151.
7. W. W. Rigrod and D. Marcuse, "Radiation Loss Coefficients of Asymmetric Dielectric Waveguides with Shallow Sinusoidal Corrugations," IEEE J. Quant. Electron., *QE-12*, No. 11 (November 1976), pp. 673–685.
8. W. Streifer, R. D. Burnham, and D. R. Scifres, "Analysis of Grating-Coupled Radiation in GaAs: GaAlAs Lasers and Waveguides—II: Blazing Effects," IEEE J. Quant. Electron., *QE-12*, No. 8 (August 1976), pp. 494–499.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Photon Probe—An Optical-Fiber Time-Domain Reflectometer

By S. D. PERSONICK

(Manuscript received May 14, 1976)

This paper describes an optical time-domain reflectometer that incorporates a gated photomultiplier receiver. The instrument can detect extremely weak reflections from fiber breaks (more than 65 dB below the 4-percent reflection of a perfect break) with 0.5-m distance resolution. In addition, backward Rayleigh scattering, which occurs roughly uniformly along a fiber, can be used to estimate the attenuation vs position within a fiber. Therefore, regions of high attenuation can be located nondestructively from one end of the fiber.

I. INTRODUCTION

Time-domain reflectometers for locating breaks in optical fibers have been implemented by a number of researchers[1,2]. Typically, these instruments have incorporated GaAs injection lasers to produce pulses of light having a high peak power and a narrow width and have incorporated avalanche photodiode (APD) receivers. The present instrument incorporates a gated photomultiplier receiver, specifically designed for this application. The photomultiplier allows increased sensitivity compared to the APD receivers and its gating action allows the user to "look behind" large reflections that otherwise would cause undesirable saturation of the receiver.

The instrument is capable of detecting echos from breaks or imperfections that are 65 dB below the 4-percent echo from a perfect break. The distance resolution is 0.5 m.* By using a transmitted pulse that is wider or narrower (the present pulse is 5 ns in duration), a trade-off between sensitivity and resolution can be made.

In addition, the backward Rayleigh scattering, which occurs roughly uniformly along the fiber, can be used to estimate the loss as a function

* The delay per unit length of light in glass is 5 ns/m. The 5-ns *full-width* transmitted pulse gives a delay resolution of 0.5 m without special signal processing.

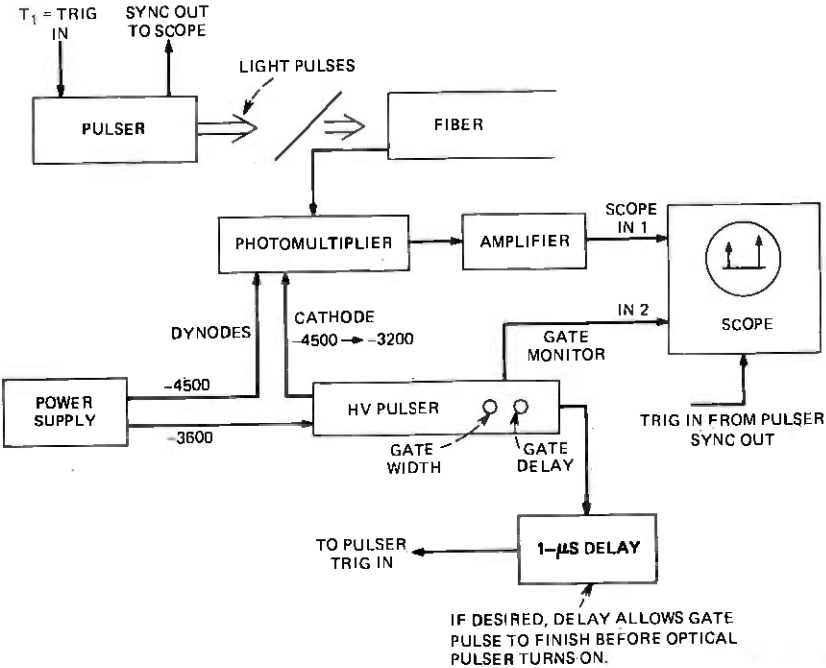


Fig. 1—Optical time-domain reflectometer.

of position along the fiber (see Section IV for a clarification of Rayleigh scattering). Thus, the loss uniformity can be estimated and regions of high loss can be identified nondestructively from one end of the fiber.

II. SYSTEM DESCRIPTION

The basic system is shown in Fig. 1. The light source is a large optical cavity, 8250-Å injection laser driven by an avalanche transistor. Drivers of this type have been described in the literature.[3] Pulse widths below 150 ps can be obtained, but the present system operates with a 5-ns transmitted pulse width. This was chosen as a compromise between resolution in the arrival time of the echo returns and pulse energy. The pulse-repetition rate is 5 KPPS. The laser output is collimated by a lens and passes through a beam splitter that serves as a directional coupler. The beam is then focused onto the fiber to be measured. (If only one end of the test fiber is available, the system is aligned with a short "pigtail" fiber of the same type as the test fiber; the test fiber is then spliced on to the pigtail.)

Echoes from the fiber (including echoes from the front face or from splices used to attach the test fiber to the pigtail) are directed by the beam splitter to the cathode of a gated electrostatic photomultiplier. The

photomultiplier cathode has a quantum efficiency of about 8 percent at this wavelength. It is mounted in a thermoelectrically cooled housing to minimize dark current and maximize cathode life. The multiplication factor of the tube is about 2×10^5. It is quantum noise limited when interfaced with commercial 50-ohm amplifiers. The tube has two features incorporated specifically for this application. The dynode chain draws current from a high-impedance divider network that limits the dynode current under high-light-level conditions. This minimizes the chances of damage due to light overload or abuse. In addition, the tube can be gated off by raising the cathode potential approximately 1000 V above its nominal −4500 V potential. This gating feature eliminates the saturation, caused by strong nearby echoes, that is present when using APD or PIN diode receivers.

The high-voltage gate pulses are obtained from a gate generator rated at 55-ns rise and fall times with a 30-pF, purely capacitive load. This is consistent with the performance measured with the present instrument. In this system, the photomultiplier turns on when the gate pulse turns off. The gate-pulse width is adjustable from 750 ns to 100 ms. Using the adjustable delays available on the generator, the gate pulse can be positioned to turn the tube on immediately after any undesired echo has arrived.

III. PERFORMANCE

Measurements of reflections (echo scans) using the optical TDR ("photon probe") are shown in Figs. 2 through 11.

Figure 2 shows the echo scan for a short fiber 70 m long. The fiber round-trip loss is negligible. The fiber numerical aperture (NA) is about 0.2. In the figure, we can see the saturated front-face reflection (no gating is applied) and the saturated back-face reflection, both of which appear as large exponentially decaying pulses. The saturation effects shown can be reduced somewhat by using a different amplifier following the photomultiplier. However, experiments reveal that without gating there is a long-term reduction (approximately 3 dB) of the photomultiplier gain which persists for tens of microseconds after overload of the tube with a large echo. When gating is used, this long-term reduction of the gain is eliminated. (See also the discussion of Figs. 6 through 8.) This long-term gain reduction is associated with the time constants of the photomultiplier burnout-prevention circuit. A double-round-trip reflection can also be seen at about 28 dB down (4 percent by 4 percent) from the back-face echo.

Figure 3 shows the echo scan for the 70-m fiber with gating of the front-face echo. The gating voltage was adjusted to leave a small remnant of the front-face echo, although complete gating is possible. The delay between the front-face echo remnant and the onset of Rayleigh-scat-

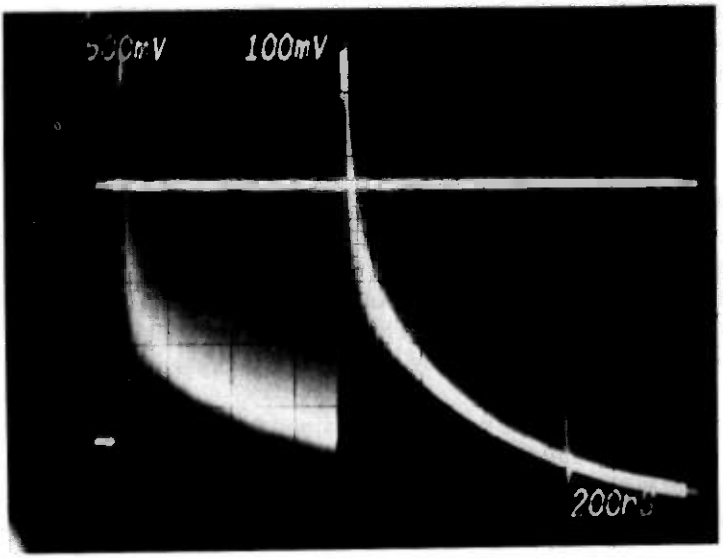


Fig. 2—Echo scan for a 70-m optical fiber.

tering reflections (the noise-like part of the trace) represents the resolution of the gating mechanism. The upper trace represents the position of the falling edge of the gating pulse. Fluctuations on the baseline of the Rayleigh scattering are due to pick up from the high-voltage gate gen-

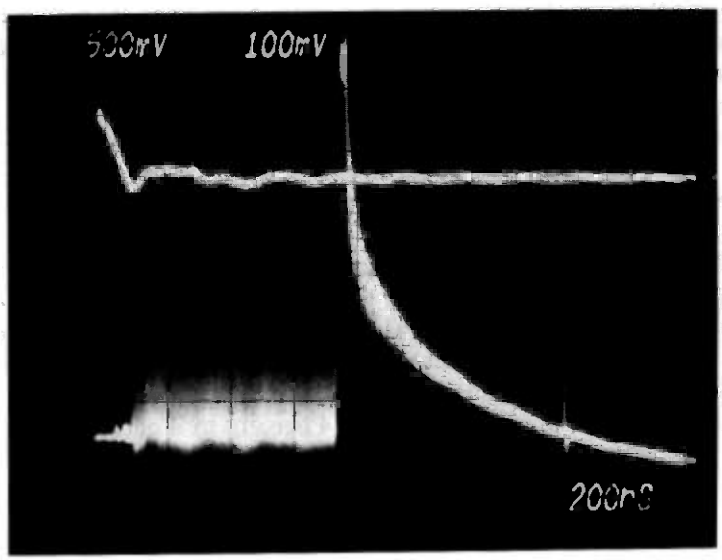


Fig. 3—Echo scan for a 70-m optical fiber with gating of front-face echo.

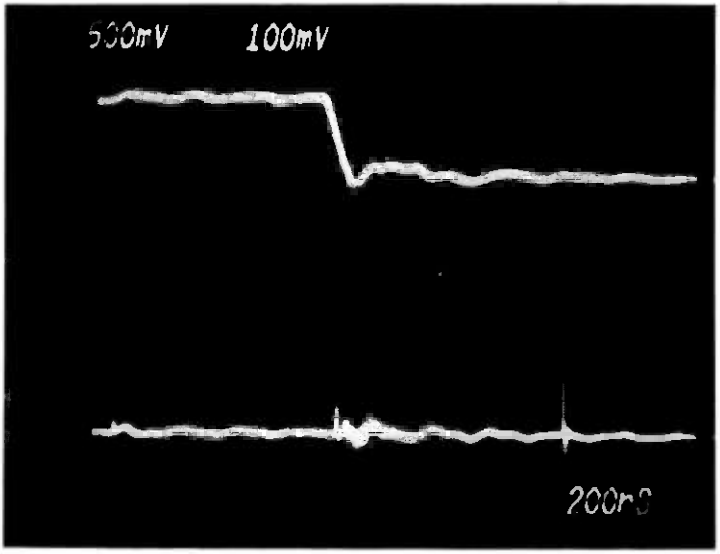


Fig. 4—Echo scan for a 70-m optical fiber with front- and back-face echoes gated out.

erator. Since the round-trip loss in this short fiber is much less than 1 dB, there is no decay in the Rayleigh scattering trace as a function of delay (distance along the fiber).

Figure 4 shows the echo scan for the 70-m fiber as above with front- and back-face echoes gated out. Note that there is no noise-like trace following the back-face echo. This is one confirmation that the noise-like trace is Rayleigh scattering and not dark current or a residue of the front-face echo.

Figure 5 shows the echo scan for the 70-m fiber with no gating and with 65 dB of optical pad between the beam splitter and the photomultiplier. The front- and back-face reflections are weak but visible.

Figure 6 shows the echo scan for a 600-m low-loss fiber with a 0.2 NA. No gating is applied. The amplifier following the photomultiplier in this measurement recovers quickly from overload, so the saturation effects due to the overload of the front-face echo are not as obvious. The decay in the Rayleigh scattering part of the trace is due to fiber attenuation vs length.

Figure 7 shows the same trace as Fig. 6 after boxcar averaging with a 0.5-s integration time and a 50-s total sweep time. No gating is applied.

Figure 8 shows the same trace as Fig. 7, but with gating of the front-face echo. The fine structure on this trace is system noise (laser pulse-amplitude fluctuations), and is not indicative of fiber details. However, the rate of decay of the roll off of the Rayleigh scattering and its shape

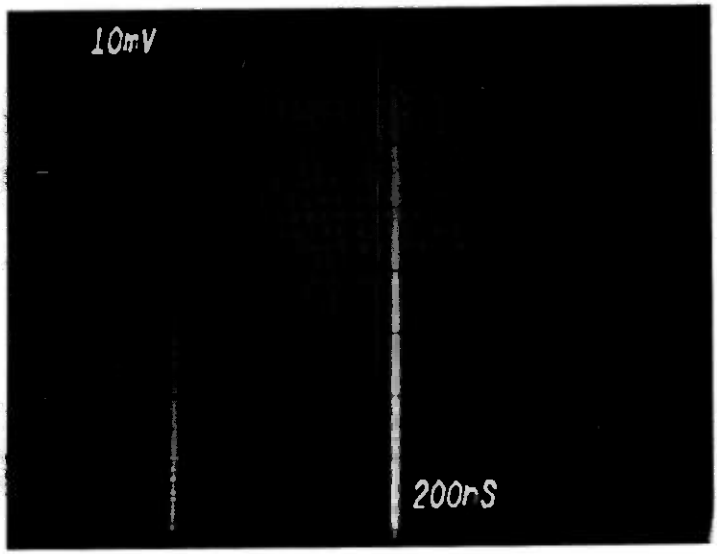


Fig. 5—Echo scan for a 70-m optical fiber with no gating and 65-dB optical pad.

are measures of the fiber loss and uniformity. With improvements in the laser pulser and the signal-processing technique, it is anticipated that fine structure in the loss vs length dependence will be obtainable.

Figure 9 shows the first attempt to use this instrument to analyze a

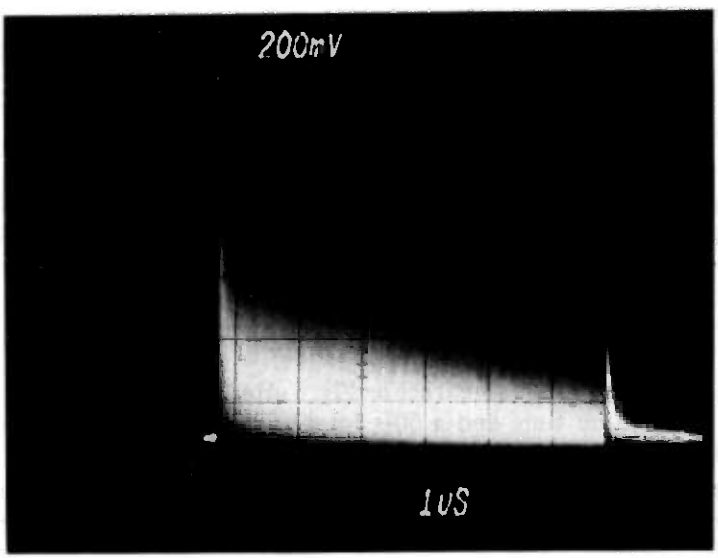


Fig. 6—Echo scan for a 600-m optical fiber before boxcar averaging.

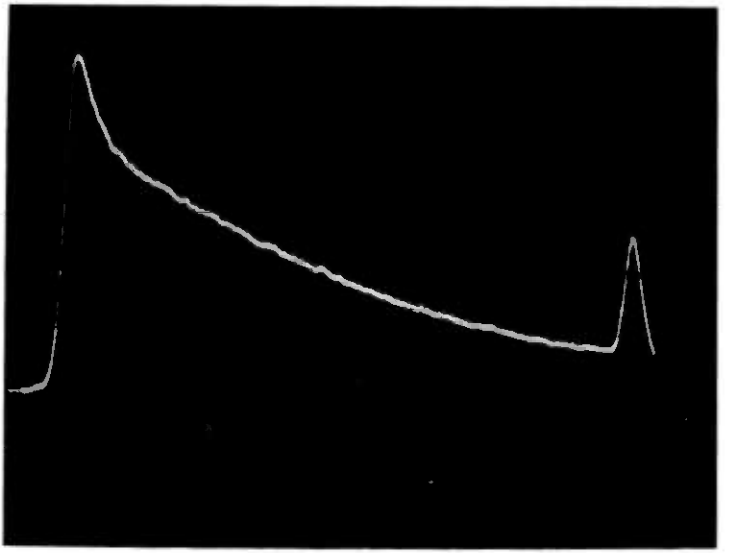


Fig. 7—Echo scan for a 600-m optical fiber after boxcar averaging.

fiber with unexplained high loss. An echo scan was obtainable from only one end and is shown before averaging. The total loss was stated as 20.8 dB, but no end-to-end transmission through the fiber was obtainable.

Figure 10 is the boxcar integrated version of Fig. 9 including front-face

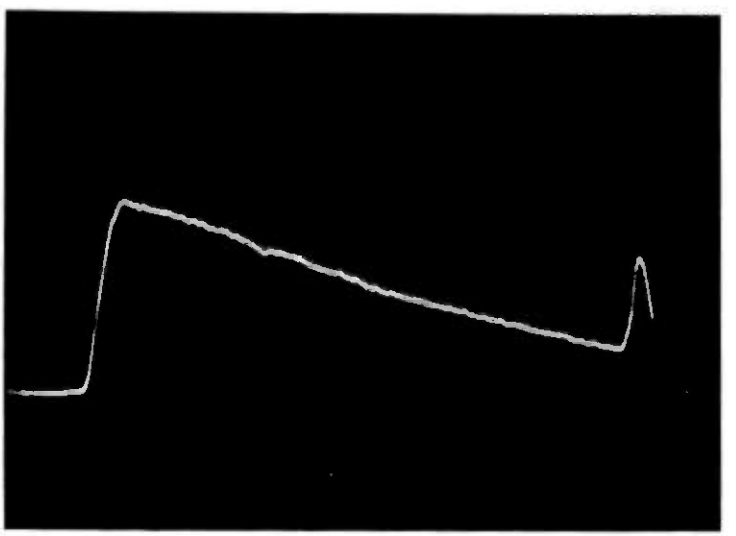


Fig. 8—Echo scan for a 600-m fiber after boxcar averaging and with gating of front-face echo.

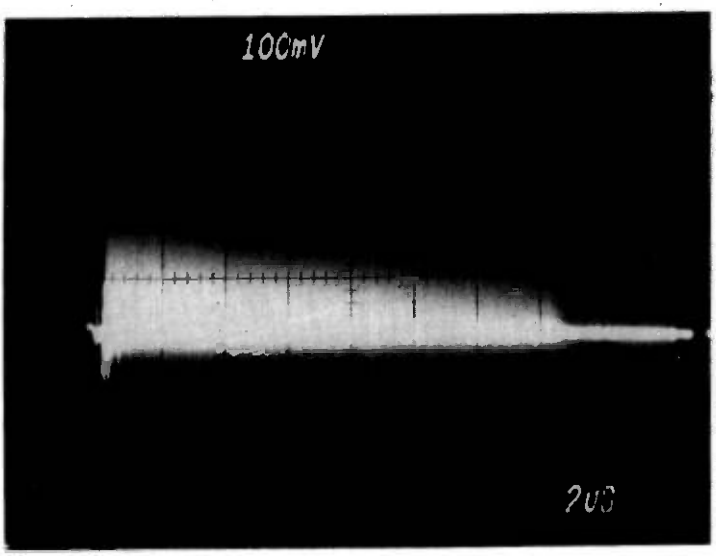


Fig. 9—First analysis of optical fiber with unexplained high loss.

echo gating. The integrating time is 0.5 s and the total sweep time is about 210 s. We see that the fiber has a rapid decay in Rayleigh scattering beginning at about 14.6-μs delay (obtained from the time scale on Fig. 9). Also, no back-face reflection is seen (compare Fig. 10 with Fig. 8). The

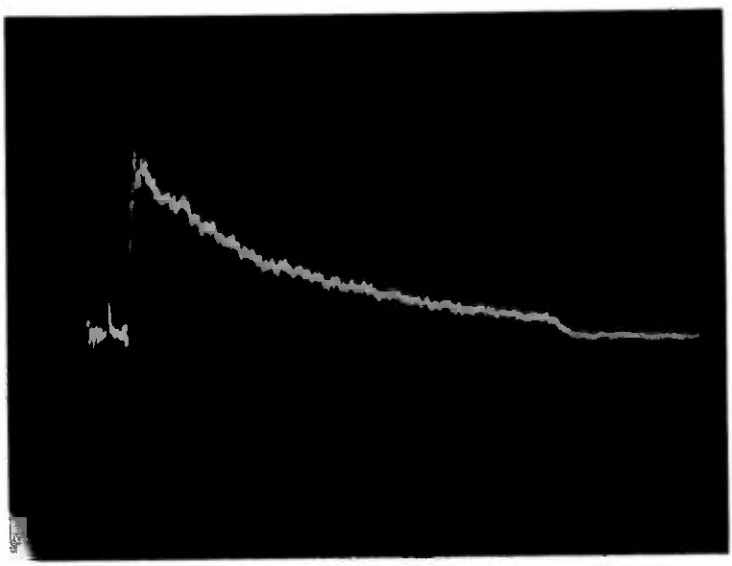


Fig. 10—Boxcar averaged version of echo scan of Fig. 9.

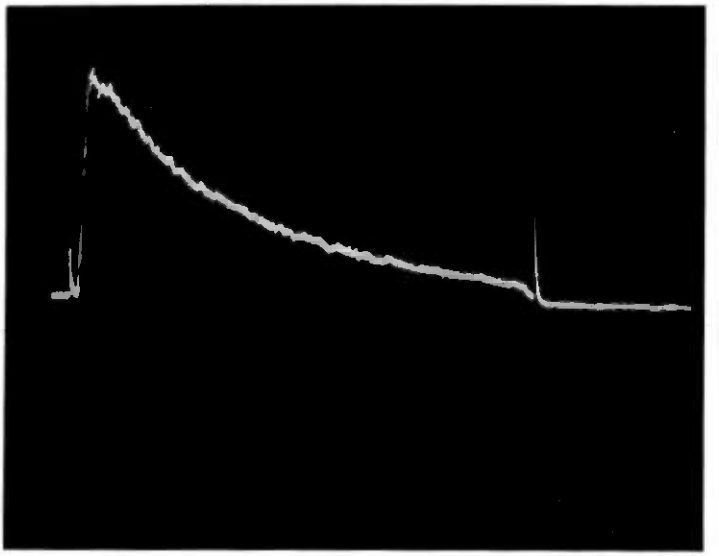


Fig. 11—Boxcar averaged scan of high-loss optical fiber after removal of lossy section.

fiber was stated as being 1500 m in length. The round-trip delay is about 10 ns/m. From this we deduce that most of the high fiber loss is concentrated in the last 50 m of one end of this fiber.

Figure 11 is the boxcar averaged trace of the high-loss fiber after removal of 120 m from the lossy end. A back-face echo is now visible. The fiber was apparently about 1570 m long. Note that not all of the high-loss region has been removed.

Figure 12 (upper trace) is a repeat of Fig. 11 using a logarithmic amplifier. The lower trace is a repeat of the upper trace with 3 dB of optical pad placed in front of the photomultiplier. This calibrates the loss-vs-length measurement (the spacing between the curves is 1.5 dB of *one-way* loss) and verified the accuracy of the logarithmic converter. The uniform loss-vs-length of the fiber can also be seen.

IV. THEORETICAL RESULTS

All of the fibers measured above were coated and wound loosely on a foam-plastic drum. No attempt was made to strip cladding modes, but this was done in the first few meters by the coating. A rough calculation of the expected level of Rayleigh backscattering was made as follows.*

* Rayleigh scattering is caused by variations in the density or composition of the fiber material on a scale that is small compared to the light wavelength. In state-of-the-art low-loss fibers, this is the dominant loss mechanism at 0.8-μm wavelength.

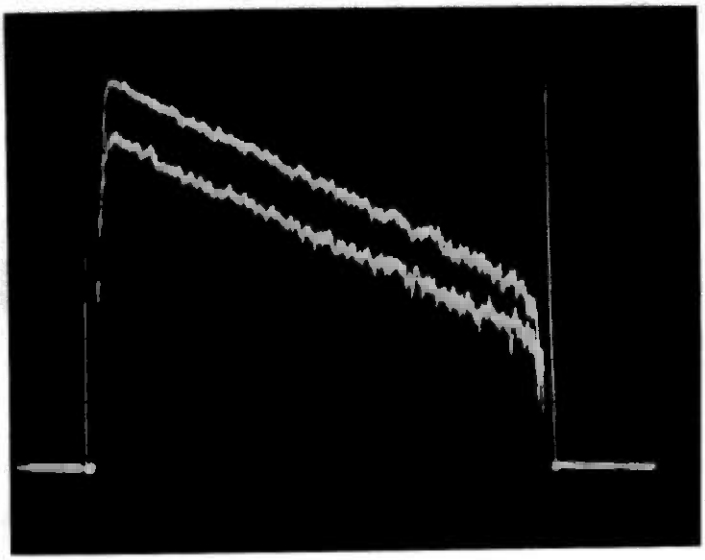


Fig. 12—Upper trace is same as Fig. 11 on logarithmic scale. Lower trace is same as upper trace with 3-dB added optical loss.

If an impulse of light of energy 1 joule is launched into the fiber, the pulse energy at a distance l (meters) along the fiber will be

$$E(l) = \exp[-\alpha(l)] \quad \text{(joules)}, \tag{1}$$

where $\alpha(l)$ is the cumulative attenuation up to distance l in nepers. In the distance interval $l + dl$, the Rayleigh-scattered energy from the fiber will be

$$dE_s(l) = \alpha_s \exp[-\alpha(l)]dl \quad \text{(joules)}, \tag{2}$$

where α_s is the Rayleigh-scattering loss in nepers per meter (assumed constant).† Of this scattered light, a fraction S will be recaptured traveling back down the fiber toward the transmitter.* At a time $t = 2l/c$, where c is the speed of light (m/s), this scattered return will arrive at the transmitter end with energy

$$\partial E_{s.echo} = S\alpha_s \exp[-2\alpha(l)]dl \quad \text{(joules)} \tag{3}$$

and will produce a response that has duration $2dl/c$. Thus, the amplitude (power) of the scattered echo at time t is

$$p(t) = \frac{cS\alpha_s}{2} \exp\left[-2\alpha\left(\frac{ct}{2}\right)\right] \quad \text{(watts)}. \tag{4}$$

* It is possible that the Rayleigh scattering from the fiber is not independent of position, and thus α_s should be replaced by $\alpha_s(l)$ in such cases. Also, there is a possibility of other (nonisotropic) types of scattering (e.g., Mie scattering from air bubbles) which do not have the same fractions, S, of recaptured light as Rayleigh scattering. Thus, in some cases S may be a function of l as well.

Thus, (4) represents the backscatter impulse response of the fiber. The fraction of scattered light recaptured is given by (assuming Rayleigh scattering is approximately isotropic)

$$S \cong \frac{\pi(\mathrm{NA})^2}{4\pi n^2} = \frac{(\mathrm{NA})^2}{4n^2}, \tag{5}$$

where NA = fiber numerical aperture (approximately 0.2 for the fibers used), n is the fiber index of refraction (approximately 1.5), and NA/n represents the half angle of the cone of captured rays. Thus, for the fibers used, S is approximately 0.005. For the fibers used in these experiments, α_s is about 4 dB/km or 0.0009 nepers/m at 0.82-μm wavelength. Thus, the backscatter impulse response is approximately

$$p(t) = 2.3 \times 10^{-6} c \exp\left[-2\alpha\left(\frac{ct}{2}\right)\right].$$

If the transmitted pulse in an actual measurement has peak power P_0 and width W, the backscattered power vs time is*

$$P_{\text{Total backscatter}} = P_0 S \alpha_s W c \exp\left[-2\alpha\left(\frac{ct}{2}\right)\right]$$

$$\cong P_0 \times 2.3 \times 10^{-6} \exp\left[-2\alpha\left(\frac{ct}{2}\right)\right]$$

$$\text{for} \quad c = 2 \times 10^8, \quad W = 5 \times 10^{-9} \text{ seconds.}$$

Thus, the backscatter power for a 5-ns transmitted pulse is about 43 dB below the 4-percent reflection of a perfect break.

The backscatter can be increased by using a wider transmitted pulse with a corresponding loss of resolution. The above result is consistent with Fig. 3 where the backscattering can be compared to the double-round-trip reflection, which is roughly 28-dB down from a perfect break reflection.

V. CONCLUSIONS

Using the above apparatus, we can detect extremely weak reflections from fiber breaks and imperfections. In addition, backward Rayleigh scattering can be used to estimate the loss as a function of position within the fiber nondestructively from one end.

The precise measurement of loss is a complicated process which requires great care to specify and control launching conditions and to obtain repeatability. The present apparatus is not intended as a sub-

* This equation assumes that $\exp[-2\alpha(ct/2)]$ is approximately constant for intervals of time t of duration W (pulse width). Otherwise one must use the average value of $\exp[-2\alpha(ct/2)]$ over the interval $(t, t + W)$.

stitute method of precise loss measurements, although refined versions of this apparatus could possibly serve that purpose. However, using this instrument, loss uniformity can be determined and concentrated loss sections in a fiber or cable can be located. In addition, fiber breaks can be located.

VI. ACKNOWLEDGMENT

The author wishes to thank R. Klein and R. Enck of Varian Corp. for their help in providing the gated photomultiplier and advice in its use. The author also wishes to thank J. S. Cook for encouragement in these experiments.

REFERENCES

1. C. Y. Boisrobert, "Some New Engineering Considerations For Fiber Optic Transmission Systems," Proceedings of the Topical Meeting on Optical Fiber Transmission, January 7–9, 1975, Williamsburg, Virginia. Sponsored by Laser and Electro-Optics Technical Group, Optical Society of America.
2. Y. Ueno and M. Shimizo, "An Optical Fiber Fault Locating Method," IEEE J. Quantum Electron., *QE-11*, No. 9, p. 77D.
3. J. R. Andrews, "Inexpensive Laser Diode Pulse Generator for Optical Waveguide Studies," Rev. Sci. Instrum., *45*, No. 1 (January 1974), pp. 22–24.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Computer Displays Optically Superimposed on Input Devices

By K. C. KNOWLTON

(Manuscript received August 3, 1976)

A set of pushbuttons on a console may appear to have computer-generated labels temporarily inscribed on them if the button set and computed display are optically combined, for example, by means of a semitransparent mirror. This combines the flexibility of light buttons with the tactile and kinesthetic feel of physical pushbuttons; it permits a user to interact more directly with a computer program, or a computer-mediated operation, in what subjectively becomes an intimately shared space.

A console of this design can serve alternately as a typewriter, computer terminal, text editor, telephone operator's console, or computer-assisted instruction terminal. Each usage may have several modes of operation: training, verbose, abbreviated, and/or special-privilege. Switching from one mode or use to another is done by changing the software rather than hardware; each program controls in its own way the momentary details of visibility, position, label, significance, and function of all buttons.

Several demonstrations are described, including a prototype of a proposed Traffic Service Position System (TSPS) console, and an interactive computer terminal resembling a Picturephone® set with a Touch-Tone® pad. Also suggested are combinations of computed displays with x-y tablets and other input devices.

In interactive use of computers, a large number of advantages result from virtually superimposing the computed display on an input device such as a two-dimensional array of pushbuttons.[1,2] A display so arranged can be used effectively to label buttons or relabel them with new meanings; indeed the buttons themselves may seem to appear and disappear according to their momentary significance or nonsignificance to the program. The same composite console—display plus input device—may have vastly different uses depending on the program that labels buttons and reacts to them. Thus combined are complete flexi-

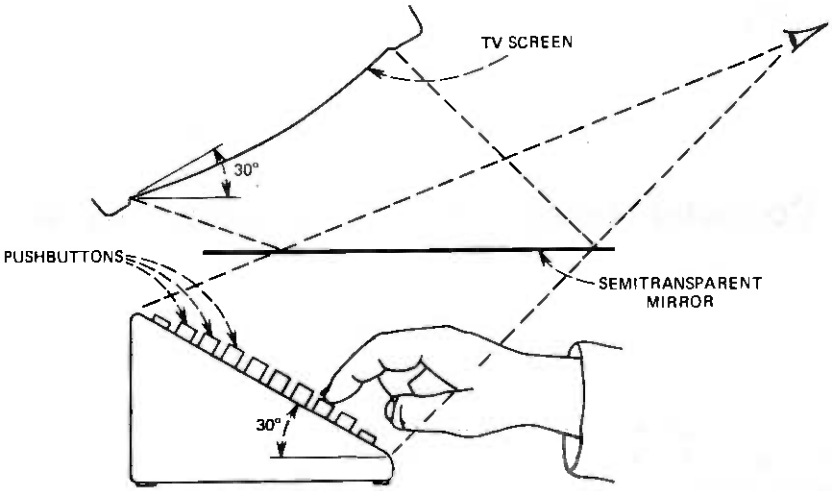


Fig. 1—Basic arrangement for superimposing a computed display on a two-dimensional array of buttons.

bility, normally associated with light buttons, and the tactile and kinesthetic feel of physical buttons that move, as on a typewriter. A button set may thus "be" a typewriter, calculator, telephone operator's console, computer-assisted instruction terminal, or music keyboard. An x-y tablet or other two-dimensional input device may likewise have a computed display superimposed on it. In all cases, the user enjoys a sense of close interaction with the computer in an intimately shared input-output space.

I. BASIC PRINCIPLES

A straightforward way of superimposing a display on a button set involves a semitransparent mirror, as illustrated schematically in Fig. 1. The user looks through the mirror and views his/her hand directly as it pushes buttons. The display—a television monitor in the illustration—and mirror are so arranged that the virtual image of displayed light buttons conforms in three-dimensional space exactly with the position of the physical keytops. Perceived spatial congruence is so precise that if the display is a bulging TV screen, then it is best for the button tops to conform to a convex envelope, as in Fig. 1, so that the central buttons do not seem too soft (i.e., so that the finger meets the physical button top at exactly the same depth as the image position). When displayed image, button set, and mirror are properly aligned, there is no parallax effect and bystanders perceive interactions exactly as the user does. In fact, the actual buttons need not be seen—it is best if they are painted a dull black so that they seem to disappear when the corresponding light

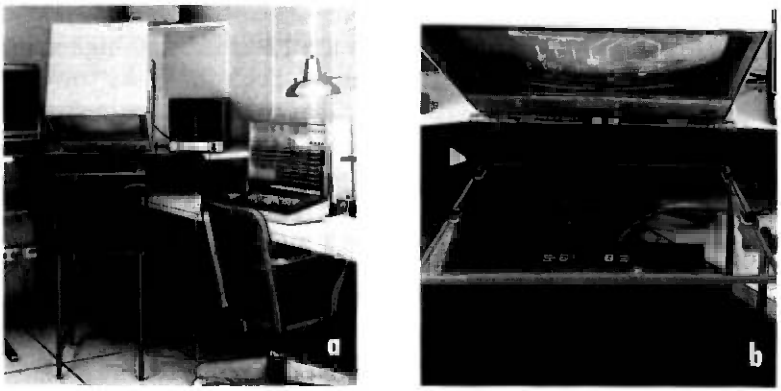


Fig. 2—(a) Laboratory setup for experimenting with displays superimposed on button tops. (b) Closeup of TV screen, mirror, and 12 by 10 set of buttons.

button is extinguished. For proper hand-eye coordination, the user does want to see his/her hand; therefore, lighting from the side is useful. If it is strong enough, then the mirror used need be only slightly transmitting (say 10 percent) and may thus be highly reflecting (say 75 percent) to maintain high visibility of the virtual display. Ambient light is no problem except that strong room illumination, direct or reflected, should be kept off the display screen.

The display, of course, needs to be generated upside down so that it appears right side up when viewed in the mirror. This is no fundamental problem except that one commonly imports or implements software in which the assumption of right-side-up generation (of alphabetic characters, for instance) may be embedded deep in the code.

One curiosity of these systems is that the hand seems transparent to light buttons, since it does not intervene in the path of reflected light. We can read through our fingertip the current label of the button pushed, as well as see other buttons beneath the hand. This is not in the least confusing to a user who has been at the machine for a few seconds; on the contrary, it is definitely helpful not to have to remove your hand to see what's beneath it.

Figure 2a is a photo of one generally useful laboratory prototype for experimenting with usages of virtual pushbutton consoles; Fig. 2b is a closeup of display, mirror, and button set as seen from farther away and lower than the user's normal head position. The computer used has 32K 24-bit words of core storage; programming is done in FORTRAN and an assembly language. The display is a normal 525-line TV monitor, with separate red, green, and blue (RGB) inputs, refreshed 30 times per second by specially built hardware from a separate core memory that holds 3 bits per picture cell.[3,4] (The displayed picture is only 496 lines of 528

pixels per line.) Each of the eight logical colors is program-definable to 128 levels per primary. The button set is a 12-wide by 10-high array of pushbuttons, ½-in. square on 1-in. centers, each with 3/16-in. travel. The computer reads only rows and columns in which buttons are momentarily depressed—all single hits are clearly decodable, as are multiple hits in the same row or column and some patterns produced by progressively adding buttons. The mirror is 16 in. square by ¼ in. thick; it is first-surface 75 percent reflecting and 10 percent transmitting.

II. PROTOTYPE FOR A TELEPHONE OPERATOR'S CONSOLE

The setup of Fig. 2 has been used to implement an experimental demonstration of a flexible telephone operator's console—in particular, a possible replacement of the present Traffic Service Position System (TSPS) station and/or future versions of it.[5] Figures 3 and 4 illustrate many of the features that such a console might have. In the demonstration, button tops are ½-in. green squares, containing green labels; lines connecting logical groups of them are blue—sometimes these alone appear where an entire set of button tops has vanished in order to preserve a sense of orientation and geography. Lights at the very top of the board, indicating what type of call is presently being processed, are red, as are occasional wide frames around buttons, pointing out mandatory operator actions.

The sequence of Figs. 3a through 3e illustrate the handling of a particular call, which is being charged to a third phone. Figure 3a shows the board before the call arrives, with only a few buttons present, indicating the limited number of things an operator can take initiative on with no call to process, such as to inquire as to the time of day (lower left). In Fig. 3b, a "zero +" call has arrived from a non-coin phone (the calling party has dialed the called number, but asked for operator assistance by the leading zero). The operator asks "May I help you?" and the calling party requests that the call be charged to a third phone, whereupon the operator prepares to push the special calling (SPL-CLG) button. With the class of charge thus declared (Fig. 3c), this button lights brightly and other class-of-charge buttons disappear; also, the key pulse special (KP-SPL) button lights with a red frame, in effect insisting that the operator enter a third phone or credit card number. (Notice that the KP-SPL with red frame can be read through the hand.) When the operator pushes this button, the keyset appears and is used for entering the third phone number (Fig. 3d). The number entered appears in the center of the panel and the SPL-NO display button appears, indicating that henceforth there is a special number, which may be redisplayed at a later time. The ST-TMG button with red frame means that no more information is needed; if it is pushed, the telephone machinery may start timing the call as soon as the called party answers the phone. When this button is pressed, it

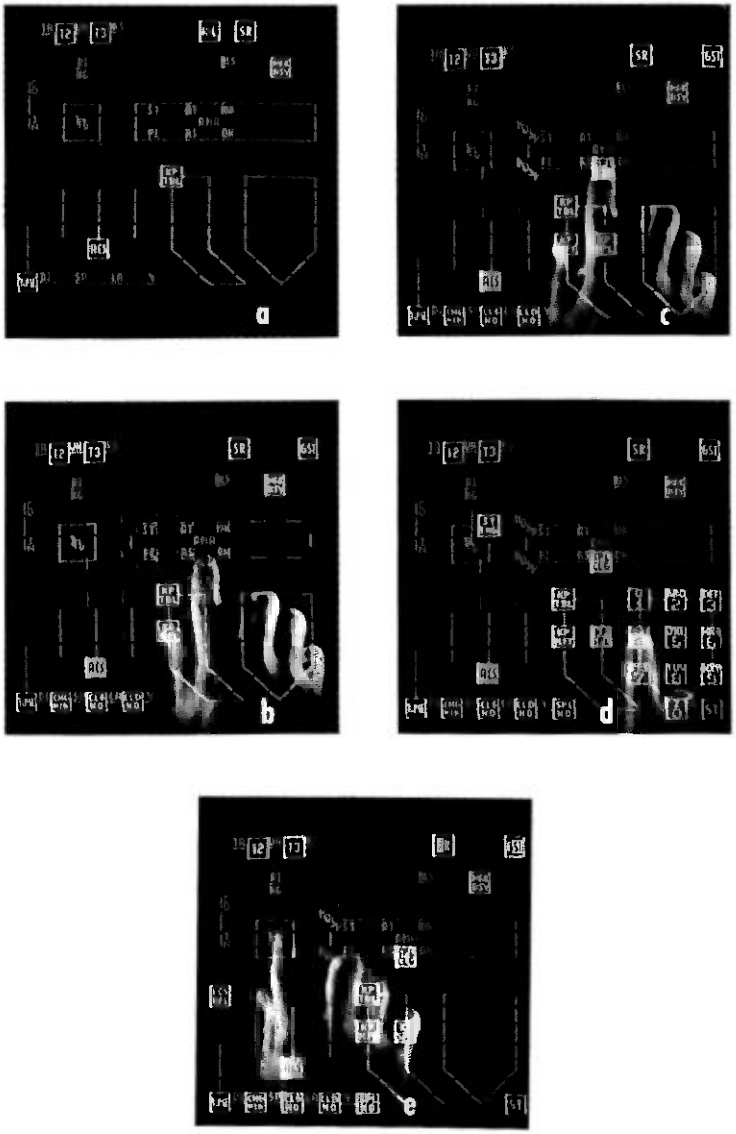


Fig. 3—Demonstration of a charge to third phone. (a) Quiescent board. (b,c) Operator pushes SPL-CLG button. (d) Third phone number is keyed in. (e) ST-TMG button pushed, permitting POS-REL.

disappears (Fig. 3e) and the POS-REL button appears, permitting the operator to release the call from this position, whereupon the board reverts to the quiescent state of Fig. 3a.

The sequence in Fig. 3 shows that this console is dynamic even during the processing of a phone call; all of the buttons that have meaning at

any moment, and only those buttons, are visible. The sequence in which they appear, in fact, tends to lead the operator through the required series of decisions and actions; this should be a significant help in the training of new operators (there could be a verbose mode, or a HELP button to spell out in words or phrases the meaning of buttons or situations encountered.)

Figure 4 illustrates more features of the TSPS demonstration. For the sake of comparison, Fig. 4a shows the complete set of buttons corresponding approximately with the currently used board. It is not immediately obvious here that the KP-TBL (meaning, "prepare to key in a trouble code") is one of the few meaningful actions when no call is present. (Figure 4d, on the other hand, is the recommended appearance of the quiescent board, with only valid buttons showing; here, the operator is inquiring as to the time, which is displayed while the TIME button is pushed.) Figures 4b and 4e compare the left- and right-handed boards; a left-handed operator presumably will want the keyset for entering phone numbers, etc., on the left. (Programming, as one should expect, is done in terms of logical buttons—*where* each button happens to be at any time is a matter of mapping. Individual operators might even be allowed to make their personal rearrangements of the board.) Figure 4c shows the entire board temporarily turned into a typewriter keyboard for possible future applications requiring alphanumeric input. Figure 4f repeats something that already appeared in Fig. 3: a call originates from a patient at a hospital, an institution which does not want to be the collecting agent for phone calls and therefore requests the phone company not to let the call be billed to the calling phone. Thus, in the spot where the operator might have expected a class-of-charge button PAID (by the calling phone) there is an explanatory note, HOSP, which neither appears nor functions like a button.

The key word is flexibility, including the option of introducing new buttons for new services or functions. All such alterations, including adding, modifying, rearranging, relabeling, or deleting buttons, are changes in *software;* they would be much easier to implement on any or all of the consoles than would be equivalent changes in hardware (once the changeover has been done).

One should finally note that, for the handling of phone calls, a great deal of electronic equipment is already needed, including circuitry and other means for detecting and decoding button pushes. The significant addition suggested by the present prototype is that the main call-handling mechanism could tell the console what buttons to light and extinguish, and where. In other words, the console would show what button presses are legal, something which is already implicit in the program. Operators would be less likely to do things out of order, simply because

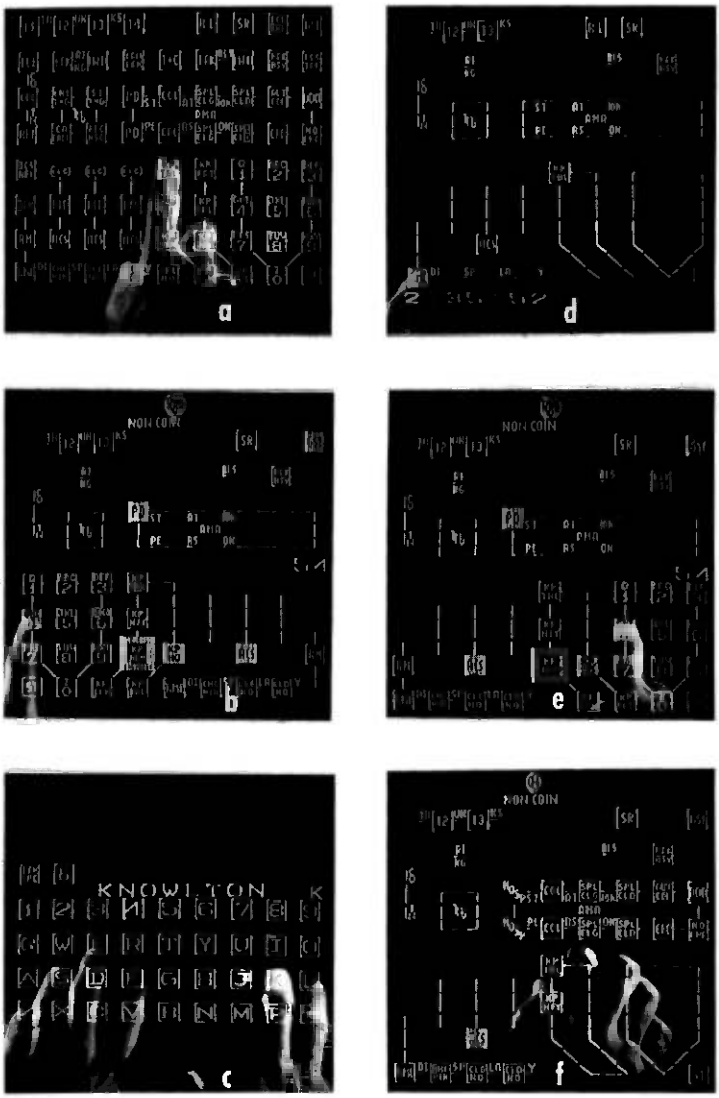


Fig. 4—Features of TSPS demonstration.

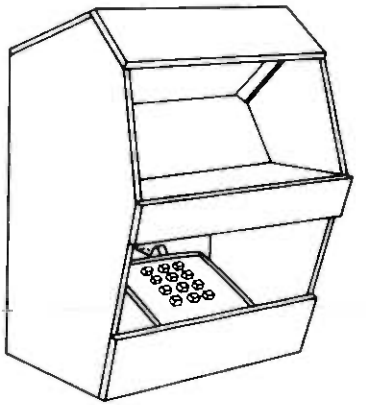
they would not expect anything to happen in response to pushing an unlighted button. A simple case in point is : if the start timing (ST-TMG) button is not present but the class-of-charge panel is entirely illuminated, it should be immediately obvious even to the beginner that the class of charge still needs to be declared (perhaps among other things) before it is legal or possible to start the timing of the call.

III. A RELABELABLE "TOUCH-TONE" PAD AS AN INTERACTIVE CONSOLE

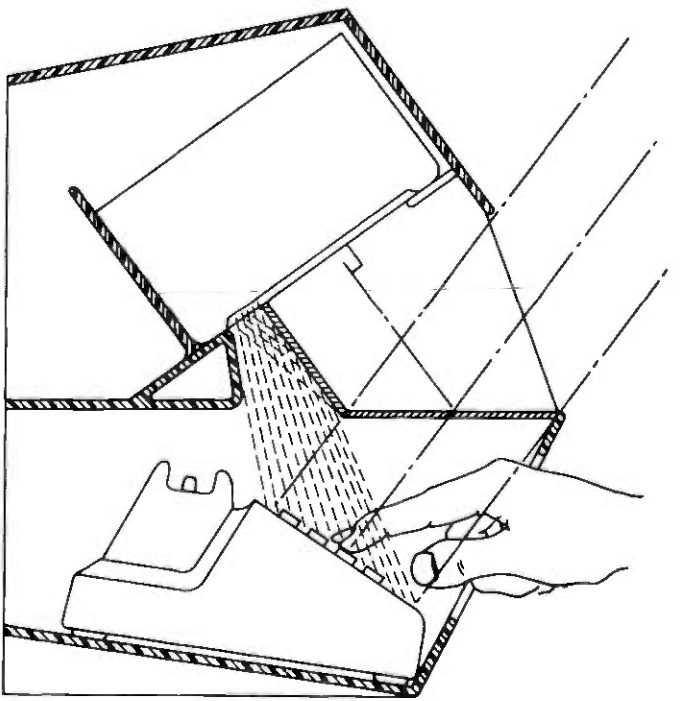
A set similar to a Picturephone set could be used as an interactive remote computer console, with the Touch-Tone pad as the input keyboard. A schematic for a mockup is shown in Fig. 5, where a semitransparent mirror effectively puts the computed image on the Touch-Tone buttons so that the 12 buttons can have several labelings and a correspondingly extended range of functions. A proposed new feature, as illustrated in Fig. 5b, is the use of the bottom quarter of the screen as a light source for illuminating the hand, but only when function buttons are displayed, not when some other graphic program result is being shown. In the latter instance, when the hand should not be seen, the screen "light" goes off. A partial cabinet hides the hand from room light.

Such a console has been built using a computer, a Touch-Tone pad as the keyboard, and a color television RGB monitor so modified that each of its three color signals is essentially a Picturephone signal.[6] The picture is again 3 bits per pixel with a total of 254 lines, 240 pixels per line. The front-surface mirror is 45 percent transmitting, 45 percent reflecting, with an antireflective magnesium fluoride coating on the second surface. Figure 6a shows a distant view of the button set and (inverted) display, whereas Fig. 6b shows the user's view for this same circumstance. These buttons are not black, yet the computer-generated image effectively obliterates the intrinsic labels on the buttons to the extent that one could turn the pad into a normal calculator, high numbers on top, with no confusion as to current numbering.

Two demonstration graphics programs have been written for this system. The first, whose three basic button labelings appear in Fig. 7, provides for drawing electronic circuit schematic diagrams, such as the one shown in Fig. 7d, by the juxtaposition and combination of basic patterns. A pattern is selected by pressing a key labeled by a small picture of the pattern (see Fig. 7a); the button marked NEXT EIGHT causes paging through several such sets of eight patterns. When pressed, the pattern is framed, as shown. The user may branch to that part of the program which places the new element on the faintly visible current picture by pushing PLACE. (Program branches involving relabeling buttons are indicated by symbols resembling miniature sets of 12 buttons with a label alongside.) Figure 7b shows the result: buttons for moving the new instance, for saying "OK, add it," and for seeing the result. The LET'S SEE button causes the button set and labels to go out, likewise the light illuminating the hand, whereas the faint circuit diagram in the background comes up to full brilliance, as in Fig. 7d. The speckling of unused buttons serves to obscure the intrinsic Touch-Tone labels. Another program branch provides for the redefinition of a pattern (see Fig.



(a)

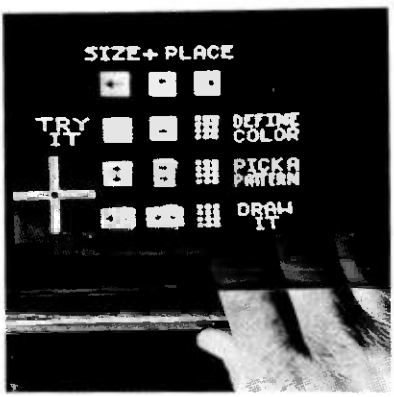


(b)

Fig. 5—(a) Console setup with a relabelable Touch-Tone pad as an iterative graphics console. (b) Side view of console showing extra mirror for reflecting bottom-of-screen light source.

7c). Here, an enlarged pattern appears on the right with a 3 by 3 window drawn on it. Contents of this window are displayed on the top nine buttons, where a button press flips the cell. The arrow buttons move the

(a)



(b)

Fig. 6—(a) Distant view of simulated Touch-Tone console setup showing inverted image on screen, mirror, and user's hand operating the buttons. (b) User's view (through mirror) of situation in (a) with Touch-Tone buttons effectively relabeled by the computed display.

window over the pattern; OK returns control to the rest of the program, with the pattern redefined.

A more elaborate program uses the capabilities of the color monitor to generate four-color designs like the one shown in Fig. 8 by selecting and applying variously stretched and positioned instances of basic patterns. Figure 9 shows its seven basic button labelings. The user may start with any of the four colors as background (Fig. 9a) and selects a pattern as before (Fig. 9b). In addition to placing it anywhere on the screen, the user may change its height and width independently (Fig. 9c). Before finally drawing the addition onto the picture, an optional

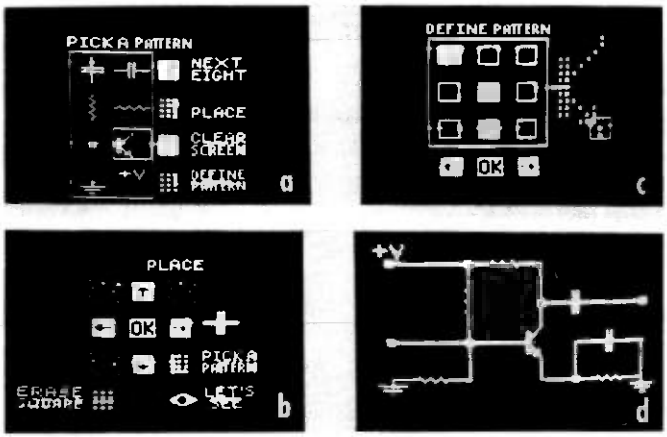


Fig. 7—(a) through (c). Button labelings of a program for composing electronic circuit diagrams. (d) Resulting diagram.

border, of width 0, 1, 2, or 3, and its color, are chosen (Fig. 9g). The colors themselves may be redefined (Fig. 9f) by increments or decrements of the three primary colors, the ± button serving to flip between modes ADD and SUBTRACT. Throughout the program, buttons which are temporarily meaningless disappear: if border width is zero, the color buttons vanish; if no more red can be added in defining the currently selected color, RED goes out; if position or size are extreme, the corresponding cursor arrow button disappears.

To summarize the Touch-Tone demonstrations, a complicated interactive graphics program can be run by means of a 12-button Touch-

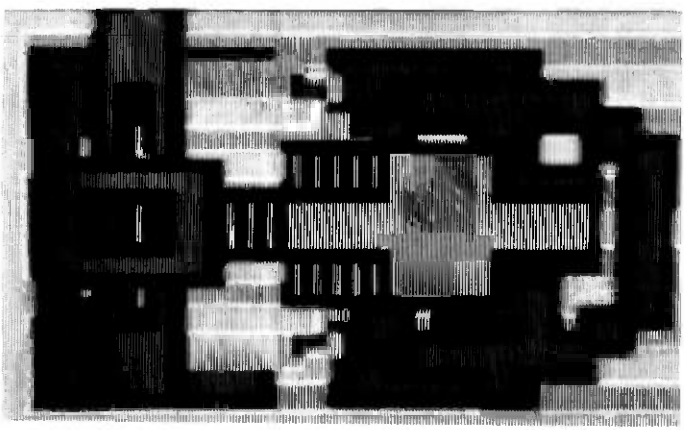


Fig. 8—Sample result of a more general program for production of four-color designs made of variously stretched and positioned geometric patterns.

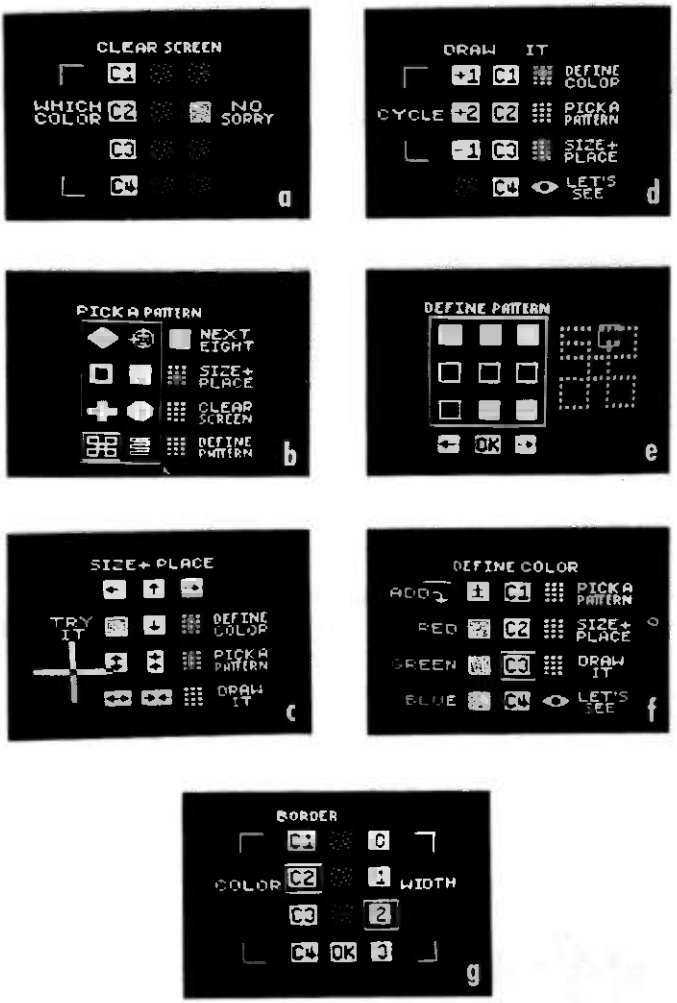


Fig. 9—The seven basic labelings for design-generating graphics system used to produce Fig. 8.

Tone pad if the buttons are easily relabeled to provide a rich variety of functions. Button forms and features found useful are:

Solid square with a 1- to 3-word label alongside.

Small pictures of symbols significant in the program.

Other iconic symbols:

An eye meaning "Let's see the picture."

Miniature 12-button set with label: a program goto involving relabeling.

Arrows for positioning a cursor and setting its size.

Picture cells in a basic pattern being defined.
Speckles to hide the intrinsic label of a nonfunctioning button.
A frame marking the current selection.
Alternation or cycling:
    Paging through sets of patterns.
    Add vs subtract for defining colors.
    Black vs white cells in pattern being defined.
Buttons that disappear when not meaningful:
    Cursor arrows when at extreme size or position.
    Color increments when at limit of range.
    Border color when width = 0.
Button set that disappears for viewing program result.
Frame around logical groups of buttons:
    Enlarged window of pattern cells.
    Pattern set.

IV. COMPARISONS WITH OTHER DEVICES

The existing system closest in form and function to this one is the touch panel[7,8] developed as a part of the Plato computer-assisted instruction project,[9] where the user's finger on the screen intercepts one vertical and one horizontal light beam, with position decoded to one spot out of 16 by 16. Virtual light buttons have the following advantages over the touch panel:

(*i*) The hand is transparent; it does not obscure buttons pressed or buttons below it.

(*ii*) The keyboard has tactile feedback through button motion, but it does *not* respond to fingers passively resting on unpressed buttons—both characteristics are very important in typing.

(*iii*) Keyboards can easily be made for the simultaneous detection of more than one button hit (as in defining a pattern, typing a capital letter, or inputting a musical chord).

(*iv*) Only one kind of electronic-detection circuitry is required, that for detecting contact closings. (Almost every commonly used system has a keyboard).

A light pen, like the touch panel, also is a single-position indicator, and it obscures part of the region pointed to, but it does permit drawing free-hand curves and precise positioning on the picture cell level. Virtual light buttons do not lend themselves well to these tasks; for such operations we might use instead an x-y tablet with virtually superimposed display. We can, however, use a button array for pointing with much finer resolution than button spacing, by any of a variety of protocols:

(*i*) A first button hit can position a cursor at the button center, and thereafter a small panel of four buttons in the left or right lower corner may serve to step the cursor in fine increments. In addition,

the cursor may slew in any of these four directions if a slew button is simultaneously depressed.

(*ii*) Alternatively, after a single button hit positions the cursor to the center of the button, a second nearby button depressed, before the first is released, can mean "so many subdivisions in this direction." This method is quickly learned and provides for easy positioning on a grid five or seven times finer in both directions than button spacing. Subsequent button hits, relative to either the first or second button, could mean picture cell displacements in the corresponding direction.

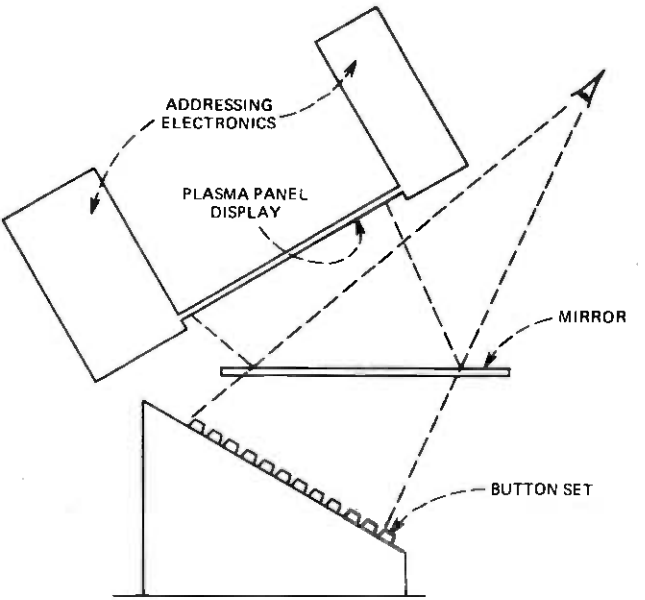
V. ONGOING WORK

Experimental and developmental work is continuing with hardware and with both general and specific software, as follows:

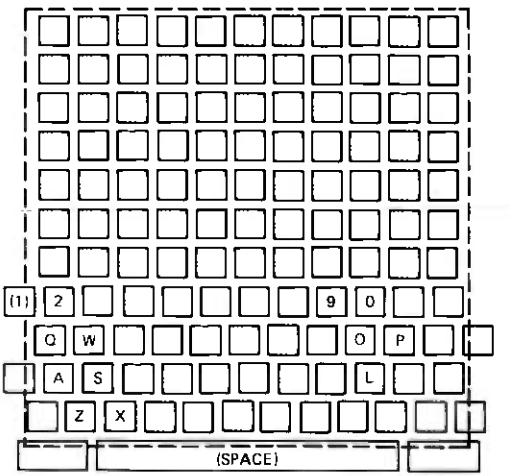
(*i*) The virtual light button setup is being considered as a possible form for a TSPS console. It would have the following advantages for operating companies: it would be expected to reduce training time; additions or changes in service or protocol would not require hardware changes to the consoles; one design would serve many purposes—handling phone calls, maintenance, traffic control, clerical work. Implications of the latter are that the operator's job could be restructured considerably, with periods of instruction or other jobs easily interleaved with normal call handling in off-peak hours.

(*ii*) A versatile and economical console is being designed, and a mockup built, using a 128-button panel, and a 512 by 512 60 pel/inch plasma panel[10] as the display, as shown schematically in Fig. 10. (The plasma panel needs no external refresh system—cells may be lit or extinguished individually, and each retains its state until changed. The panel is flat, permitting the button set to be flat, and since picture cell positions are defined by the structure of the device, the overall setup cannot become misaligned by electronic drift.) The board is arranged basically as an 11 by 11 array on ¾-in. centers (normal typewriter spacing) with slight adjustment of the bottom rows so they conform closely to a regular typewriter. Some buttons hang partially off the 8½ by 8½-in. display area; their meanings will normally be understood. The labels in Fig. 10 indicate how a typewriter keyboard is intended to be mapped onto the buttons set; the buttons can also be used as an 11 by 11 rectangular array but with some distortion in the lower lines, at least for the lowest 12-button row. Arrangements are being made to read all combinations of simultaneous button presses.

Commercially available plasma panels, sadly, are monochrome



(a)



(b)

Fig. 10—Console under construction using plasma panel display. (a) Side view. (b) Button layout: 128 buttons basically positioned on ¾-inch centers. Labels illustrate usage of lower board as a typewriter. Large square indicates virtual position of 8½-inch square plasma panel display area.

(neon orange). They are, however, much more economical than color TV monitors plus refresh buffers.

The ultimate flexible-but-economical console is expected to be

close to the above design of keyboard + plasma panel, with addressing and driving electronics of the plasma panel in the lower cabinet, not in a bulky frame around the display. It would contain a character-generator capable of generating (hierarchically) parts of buttons, button tops, and sets of button tops. The remote computer would then need to designate only which light button(s) to put up and/or extinguish, and where, whereas the console would report button hits, perhaps with local culling of hits on nonlighted buttons. Both are low-capacity channels; a twisted pair phone line would suffice. An 11 by 11-in. mirror would be big enough for a single user (larger mirrors are useful for demonstrations).

(*iii*) A general-purpose software package under development will ultimately facilitate the writing of specific usage programs. It will permit convenient design and labeling of light buttons, plus facilities for conveniently defining mappings between logical and physical buttons and describing changes in state: appearance and disappearance of buttons, changes in values of variables, and flow of control. It will also provide a testing ground for one of the author's basic attitudes about usage of such a system: that there should always be exact correspondence between buttons which appear and those responded to. This ground rule should aid the development of complex systems like TSPS where there is a huge number of combinatoric states of the board, and where it is a big and difficult job to define precisely and completely which button hits are or should be legal from instant to instant.

(*iv*) One application nearing completion is a text editor, where text being worked on appears in the top part of the screen, while the bottom part serves as a typewriter keyboard. The novel feature of the setup is that pointing (to lines or words or positions for deleting, changing, or inserting) is done by pointing into the text. Text is displayed with three lines of five characters on each button top, and one designates a character by a sequence of two button hits: first the button on which the character appears, followed by the same button if the character is centered on this button, or by a nearby button in the direction that the character is off-center (the second button is always one of the 3-high by 5-wide subarray centered on the first).

VI. ACKNOWLEDGMENTS

Dan Franklin designed and implemented the text editor. I thank Max Mathews and Peter Denes for facilitating the hardware developments and for suggesting the telephone operator console application, and Steve Bauman and Stu Silverberg for helping to formulate a meaningful TSPS demonstration. I thank Marie Hill, former Manager of Operator Services

at Morristown, N. J., for several helpful consultations and for arranging for me to observe and monitor operators processing actual calls.

REFERENCES

1. K. C. Knowlton, US Patent No. 3,879,722 Issued Apr. 22, 1975: "Interactive Input-Output Computer Terminal with Automatic Relabeling of Keyboard."
2. K. C. Knowlton, "Virtual Pushbuttons as a Means of Person-Machine Interaction," Proc. IEEE Conference on Computer Graphics, Pattern Recognition, and Data Structures, Beverly Hills, California, May 1975, pp. 350–351 (Abstract).
3. P. B. Denes, "A Scan-Type Graphics System for Interactive Computing," Proc. IEEE Conference on Computer Graphics, Pattern Recognition, and Data Structures, Beverly Hills, California, May 1975, pp. 21–24 (Abstract).
4. P. B. Denes, "Computer Graphics in Color," Bell Laboratories Record, *52*, No. 5 (May 1974), pp. 138–146.
5. R. J. Jaeger, Jr., and A. E. Joel, Jr., "TSPS No. 1—System Organization and Objectives," B.S.T.J., *49*, No. 10 (December 1970), pp. 2417–2443.
6. A. M. Noll, "Scanned Display Computer Graphics," Commun. ACM, *14*, No. 3 (March 1971) pp. 29–32.
7. F. A. Ebeling, R. S. Goldhor, and R. L. Johnson, "A Scanned Infra-red Light Beam Touch Entry System," Digest of Technical Papers, SID International Symposium, June 1972, pp. 134–135 (Abstract).
8. B. L. Richardson, "X-Y Coordinate Detection Using a Passive Stylus in an Infra-red Diode Matrix," Digest of Technical Papers, SID International Symposium, June 1972, pp. 132–133 (Abstract).
9. D. L. Bitzer and R. L. Johnson, "PLATO—A Computer Based System Used in Engineering Education," Proc. IEEE, 59, No. 6 (June 1971), pp. 960–968.
10. W. E. Johnson and L. J. Schmersal, "A Quarter Million Element AC Plasma Display With Memory," Proc. SID, *13*, No. 1 (First Quarter 1972).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

100-GHz Measurements on a Multiple-Beam Offset Antenna

By R. A SEMPLAK

(Manuscript received September 8, 1976)

Large-capacity satellite communication systems can be developed by equipping both satellites and earth stations with multiple-beam antennas. Such antennas can be produced from an offset Cassegrain with offset collectors as feeds.

I. INTRODUCTION

Large-capacity satellite communication systems can be achieved by equipping both the satellites and earth stations with multiple-beam antennas.[1] An offset Cassegrainian antenna (Fig. 1) fed by multiple, but separate, small corrugated horns has been proposed for such a system.[2] Measurements and theory have indicated that the offset geometry results

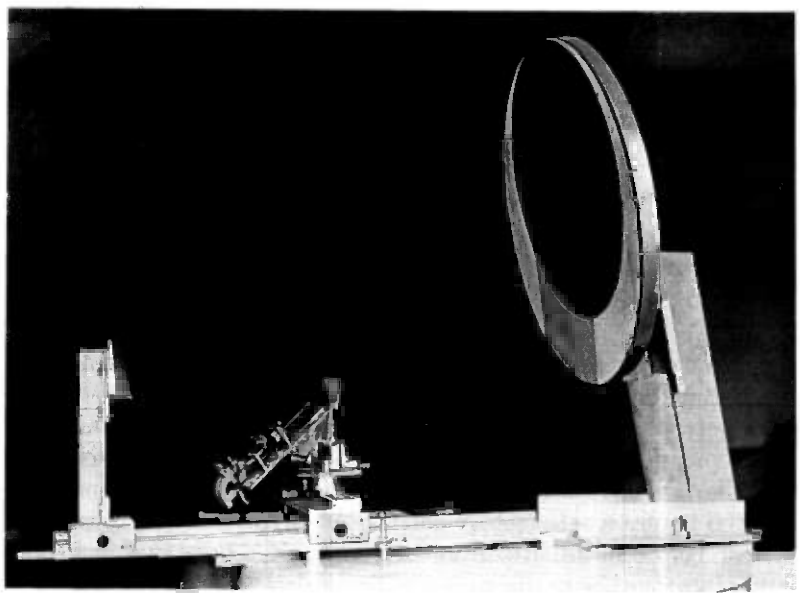


Fig. 1—100-GHz beam-scanning antenna.

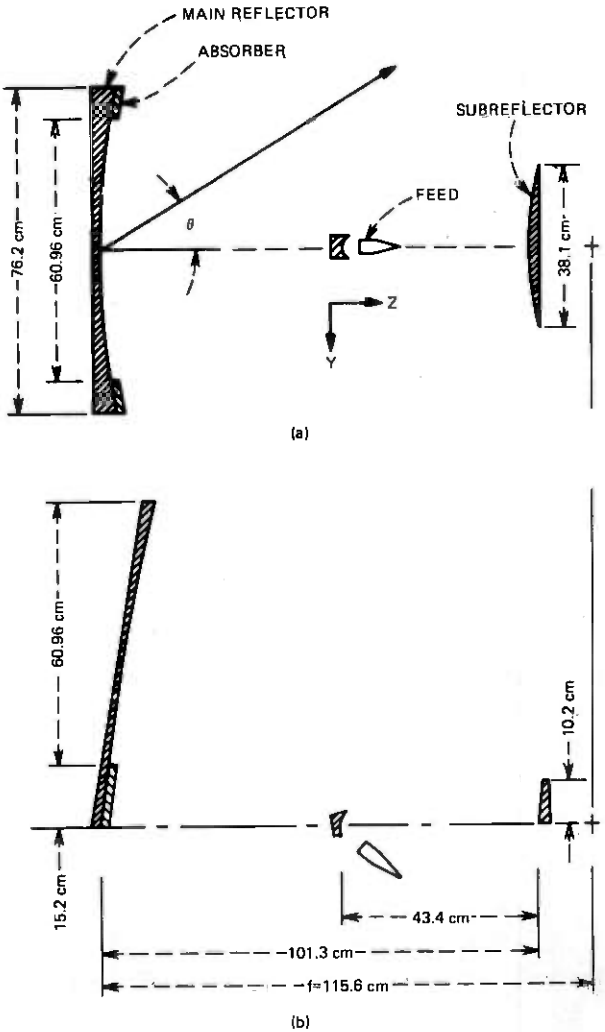


Fig. 2—Plan view of 100-GHz beam-scanning antenna. (a) Plan view. (b) Side view.

in an ideal configuration[3–5] for both earth-station and satellite antennas since the aperture has no blockage; this significantly reduces the side-lobe levels and, in turn, reduces adjacent station interference. The measurements to be discussed here were obtained with a dual-mode feed horn. The current state-of-the-art needs to be pushed to its maximum potential before a satisfactory corrugated (hybrid-mode) horn at 100 GHz can be produced; utilization of a corrugated horn as a feed should further improve the performance of the scaled model discussed here.

This 100-GHz exploratory study utilizes an antenna that scales to about 2 m at about 30 GHz on a satellite (shown in Fig. 1). The antenna consists of a 60.96-cm-diameter, numerically machined, parabolic section on the right; a confocal 38.1-cm-wide by 10.2-cm-high, numerically machined, hyperbolic subreflector on the left; and in the center, an offset collector feed that consists of a 5.08-cm, numerically machined, parabolic reflector illuminated by a dual-mode horn. The feed is mounted on a

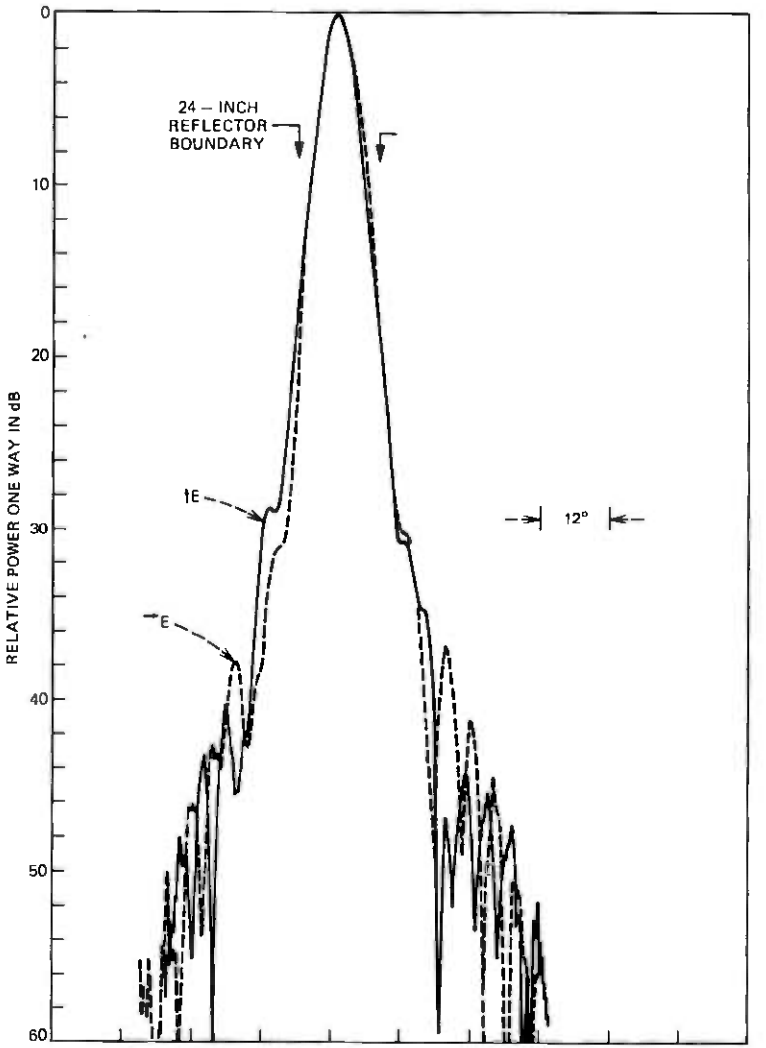


Fig. 3—Far-field radiation patterns of the offset collector feed for the principle linear polarization.

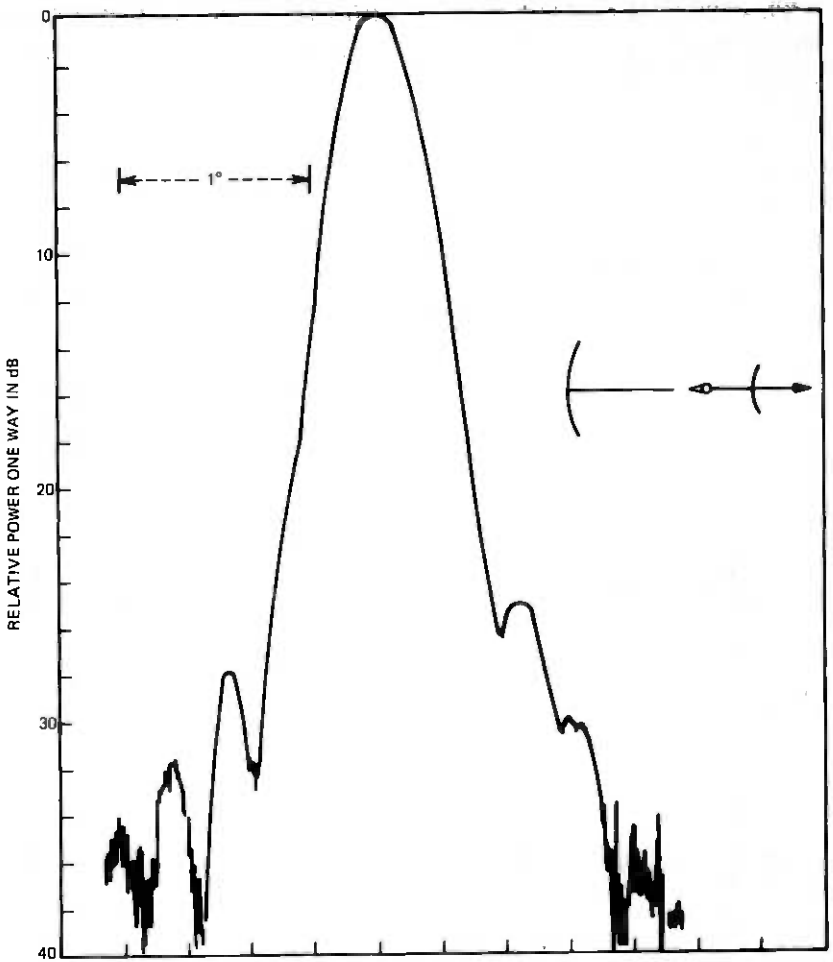


Fig. 4—Azimuth scan of the far-field radiation pattern of the antenna for zero feed displacement (y = 0, z = 0); this is the reference position $\theta = 0°$.

device that provides coarse and fine adjustments in both the y and z axis (into the plane of the photograph and along the axis of the main paraboloid, respectively). Fine adjustments are provided in the direction of the x axis and rotation about this axis is also provided.

The 60.96-cm-diameter main reflector was obtained by reducing a 76.2-cm-diameter reflector with absorbing material whose one-way transmission loss at 100 GHz is of the order 26 dB. The absorber is positioned as shown in Fig. 1 to avoid blockage. Hence, contributions to the far-field radiation pattern from the masked area are small.

Figures 2a and 2b are the plan view and side view, respectively, of the scale model showing various dimensions. From these dimensions, we observe that the aperture is large in wavelengths and the f/D (i.e., the ratio of prime focal length to diameter) is large, 1.9; hence, we should be able to scan tens of beamwidths by lateral displacement of the feed.[6]

The radiation characteristics of the feed, shown in Fig. 3, are for the principle linear polarizations in the azimuthal plane. We observe from this figure that the amplitude is very similar for both polarizations. The tick marks on this figure, which denote the edges of the apodized main reflector, indicate an illumination taper of the order 18 dB.

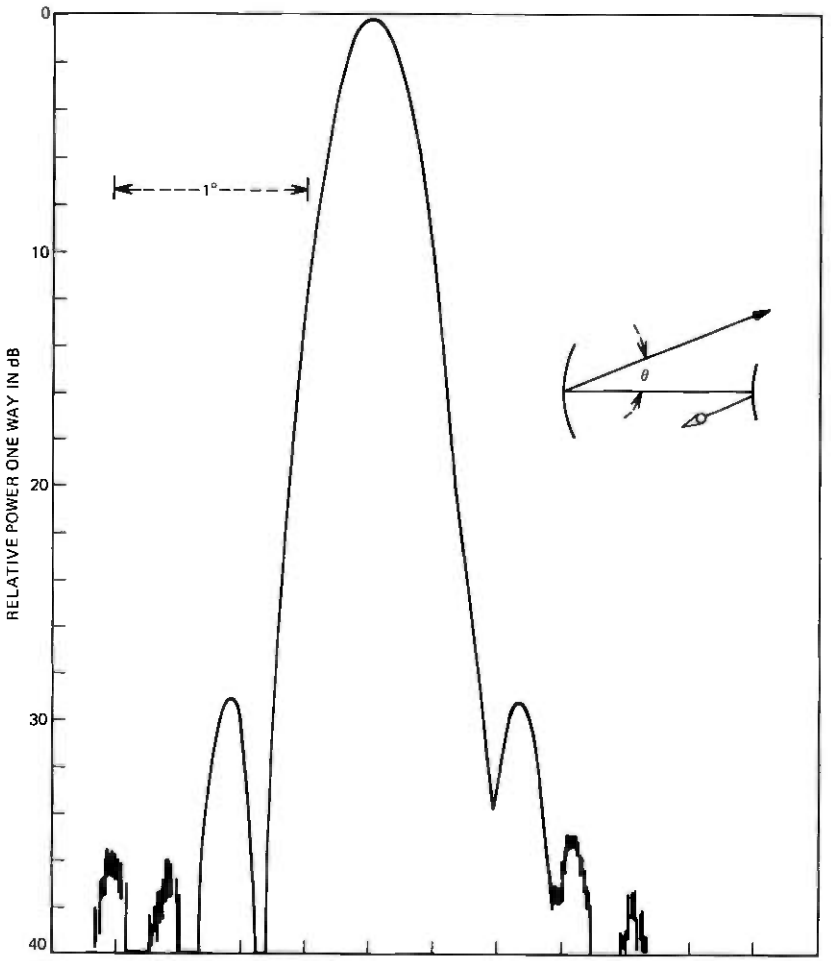


Fig. 5—Azimuth scan with beam displaced $\theta = 3.4°$.

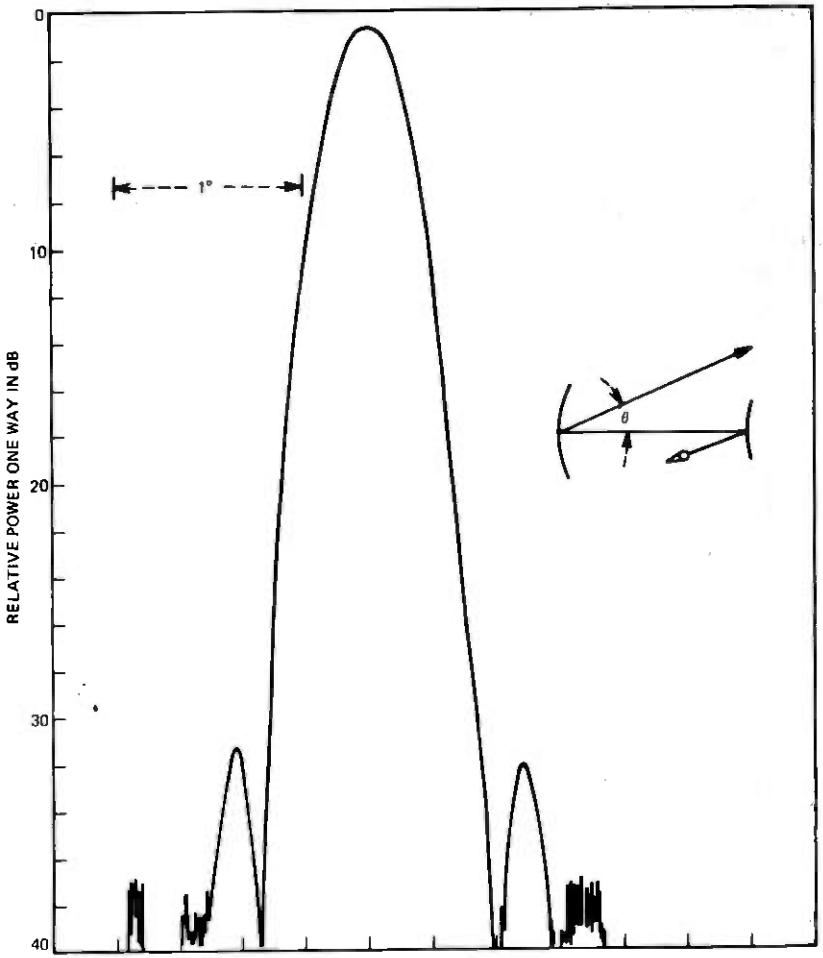


Fig. 6—Azimuth scan with beam displaced $\theta = 4.3°$.

II. FAR-FIELD RADIATION MEASUREMENTS

As determined by using Bell Laboratories radio range, measurement of the incident field presented to the aperture of the beam-scanning antenna indicates an amplitude variation of the order 0.5 dB. Also, variations in the refractive index of the air along the 480-meter range produce scintillations. These scintillations are time-of-day dependent and their effects were minimized by measuring in the early morning and late afternoon. Overcast days were ideal. The siting of the antenna is such that radiation-pattern measurements in the elevation plane are perturbed by ground reflections and therefore are not discussed.

The measurements shown in Figs. 4 through 11 are made in the azimuthal plane for vertical polarization. For Fig. 4, the feed is on axis and all subsequent beam-scanning measurements are referred to the on-axis measurement. Here in Fig. 4, we see a well-behaved far-field pattern with side lobes about 27 dB down located two and one-half beamwidths from axis. The insert of this figure is a sketch showing the relative location of the feed with respect to the subreflector and main reflector. The arrow indicates the direction of the transmitting source.

Figures 5 through 10 show measurements obtained by displacing the feed laterally ($y > 0$). The feed was adjusted in all its degrees of move-

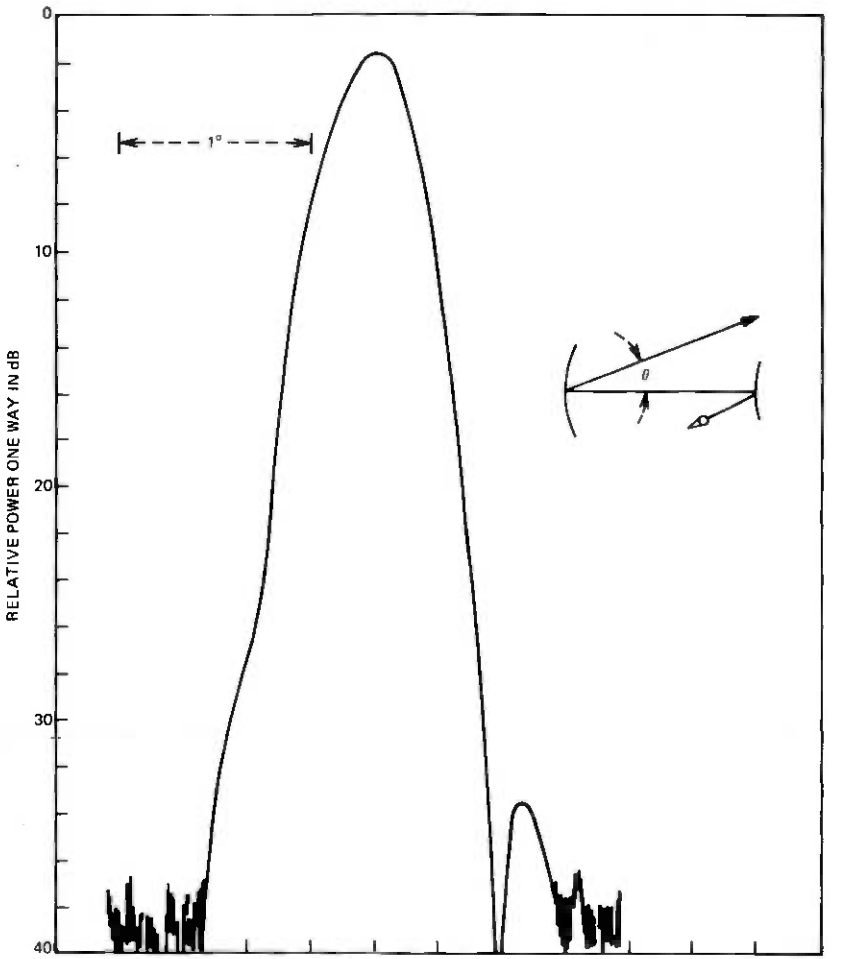


Fig. 7—Azimuth scan with beam displaced $\theta = 5.0°$.

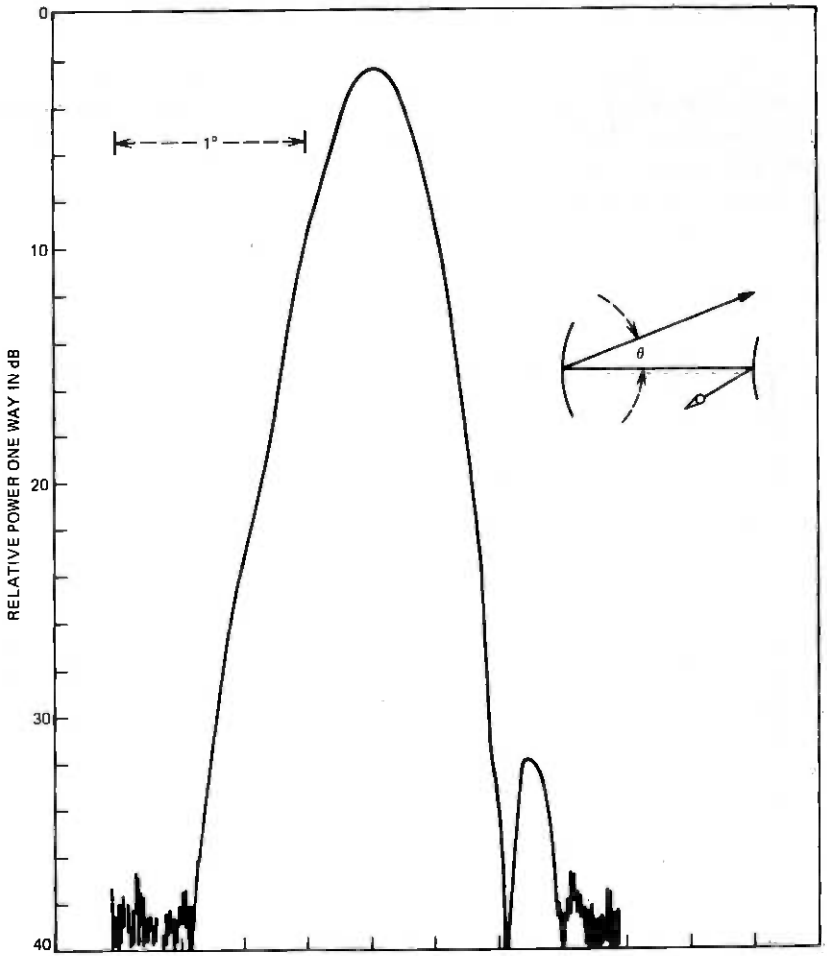


Fig. 8—Azimuth scan with beam displaced $\theta = 6.25°$.

ment to optimize the gain for each displacement of the feed, including x and z. The angular beam displacement θ is indicated on each figure. We observe the decrease in gain (at zero angle) and also that the side lobes continuously decrease with increase in scan angle.

Measurements were also made by displacing the feed in the opposite direction ($y < 0$) and Fig. 11 is typical; this pattern is almost mirror symmetric to that of Fig. 7. One notes the similarity and therefore concludes that the feed is on the electrical axis of the antenna for the $\theta = 0$ pattern. Further examination and comparison of Figs. 4 through 11 disclose the lack of deep minima for the first nulls of the on-axis position

(Fig. 4); the minima are much more pronounced in Figs. 5 and 6. Improvement in performance for the "on-axis" position was explored by moving the feed in small increments, but the absence of sharp nulls for this reference position (Fig. 4) can not be explained at this time.

III. BEAM-SCANNING MEASUREMENTS

Figure 12 is a plot of gain (relative to the on-axis gain) versus scan angle. The footed vertical bars indicate the maximum and minimum gain measured for the indicated scan angle. The solid dot is the average of

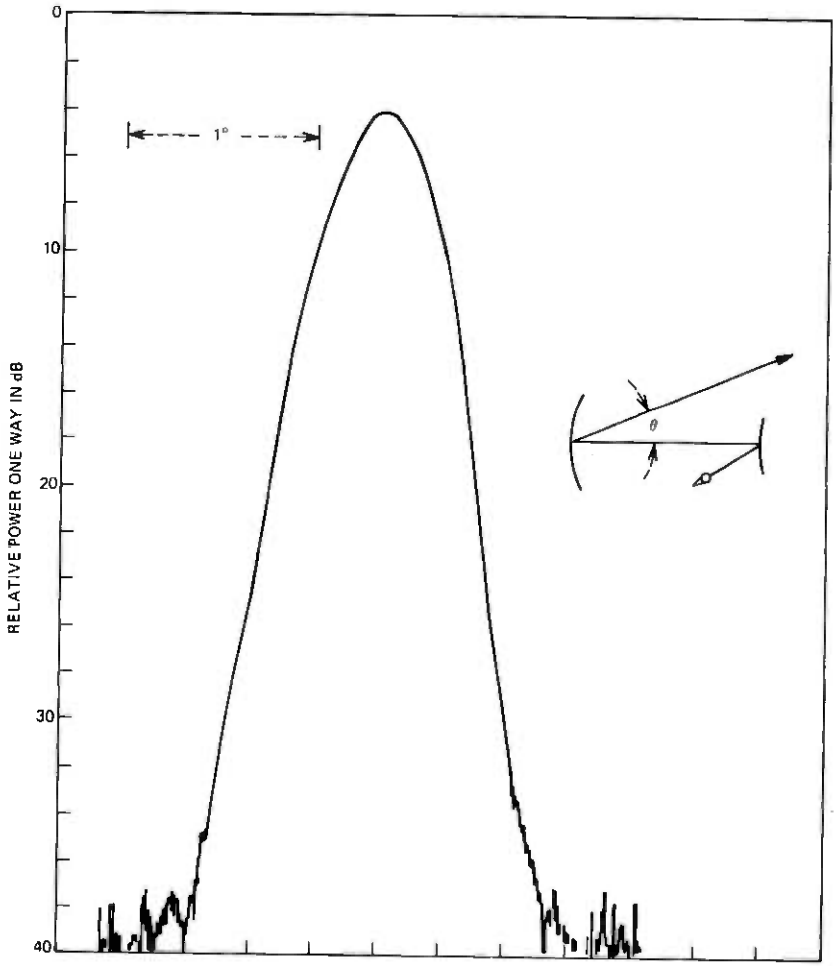


Fig. 9—Azimuth scan with beam displaced $\theta - 7.24°$.

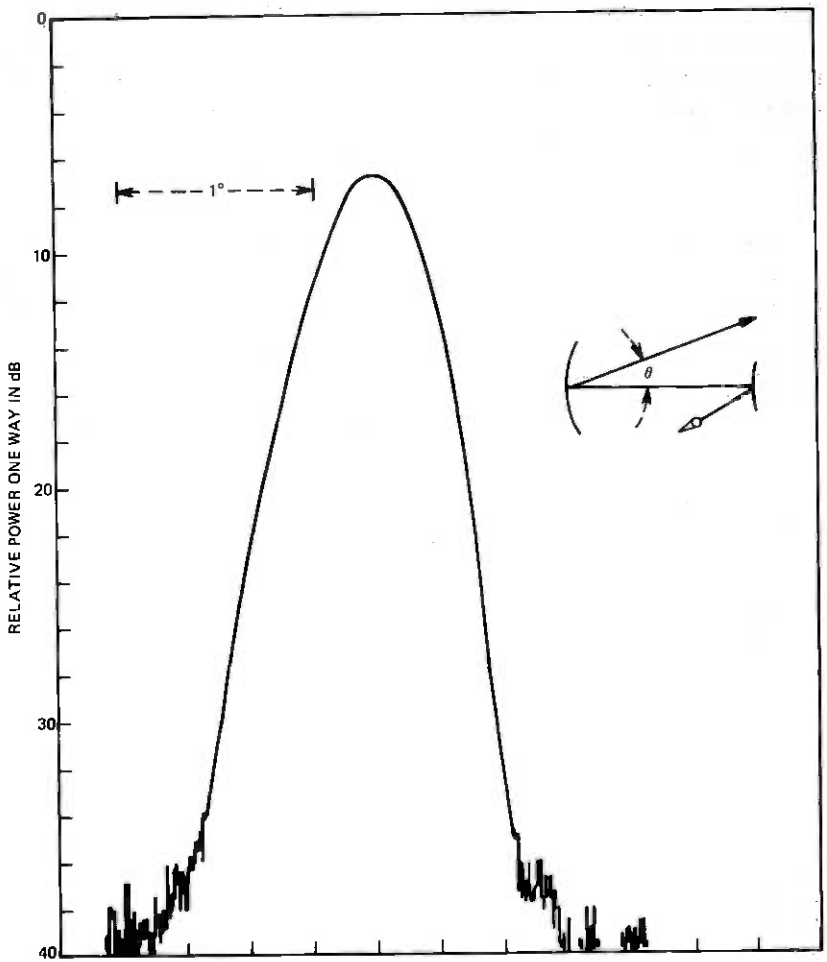


Fig. 10—Azimuth scan with beam displaced $\theta - 8.01°$.

measurements within a 0.2 degree interval. The open circles, measurements made in the opposite scan direction ($y < 0$), are similar to those for $y > 0$. Scan-angle measurements of less than 3 degrees were not made since the change in relative gain is small. The curve is well behaved out to a scan angle of 4.5 degrees (of the order of 12 beamwidths). Many measurements were made in the region of 4.5 to 6.0 degrees to confirm the presence of the shoulder in the curve. Clearly, the antenna can be scanned ±3.5 degrees (18 beamwidths) with degradation in gain of only 0.25 dB. A scan of ±4.5 degrees (24 beamwidths) degrades the gain only 1 dB. Thus, the contiguous United States could be served by a syn-

chronous satellite with a multiple-beam antenna of this type and suffer less than 0.5-dB degradation of gain.

Figure 13 is a plot of feed displacement in the y and z directions required to optimize the gain for a given scan angle of the beam. In a sense, this is the plot of the locus of focus—it is not a straight line. It was found that no displacement in x was necessary.

Figure 14 shows plots of the 3, 10, and 20-dB beamwidths as a function of scan angle; the beamwidths increase with increasing scan angle. This increase in beamwidth limits the number of feeds that can be used on this type of antenna.

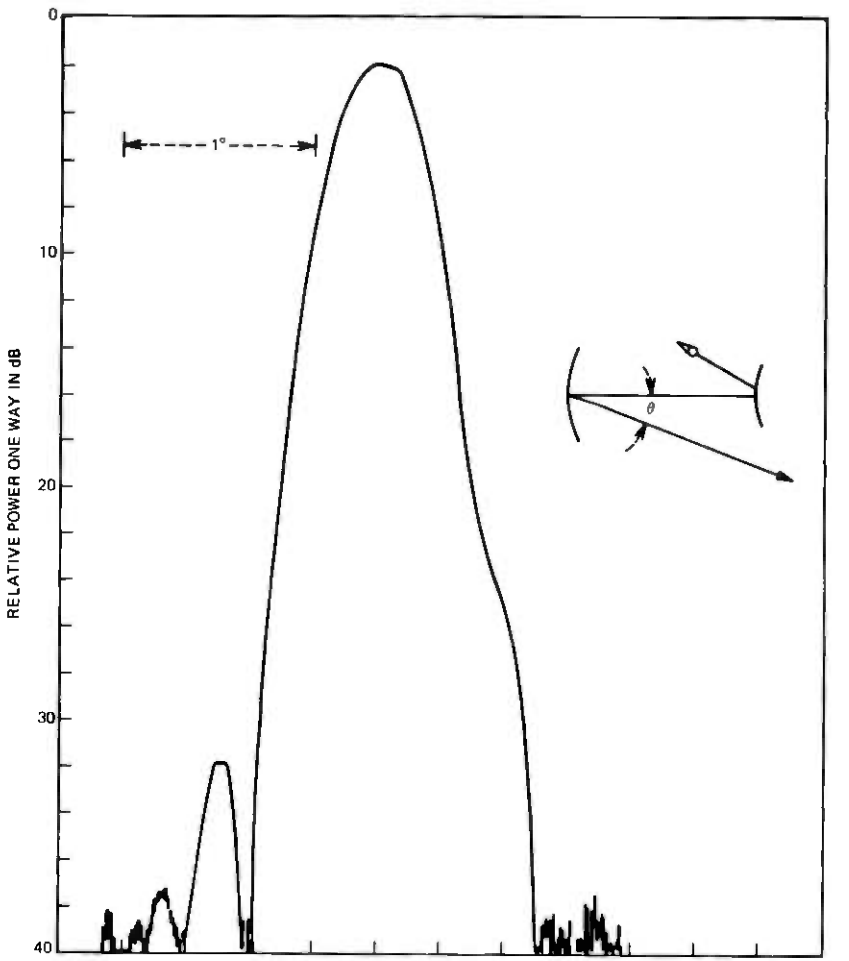


Fig. 11—Azimuth scan with beam displaced in opposite direction $\theta = -5.32°$.

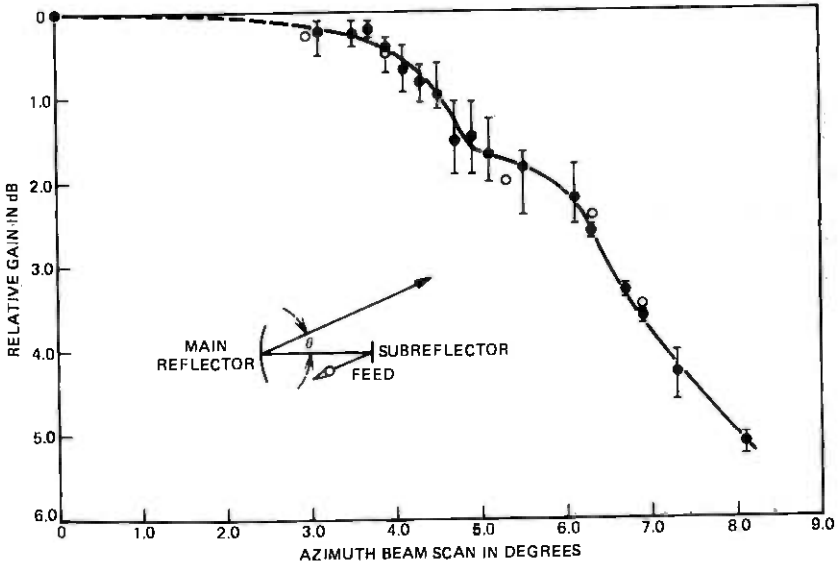


Fig. 12—Relative gain measured as a function of beam scanning θ.

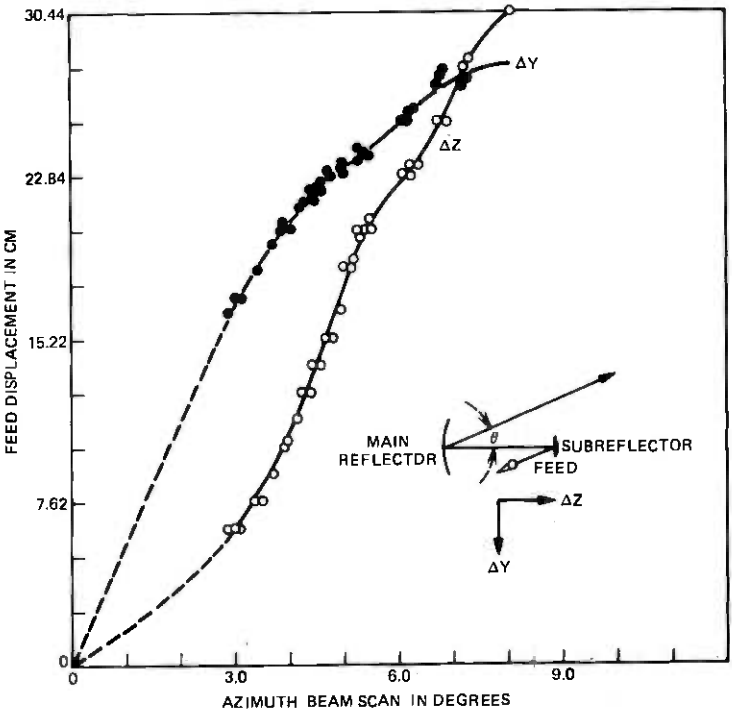


Fig. 13—Feed displacement as a function of beam scanning θ.

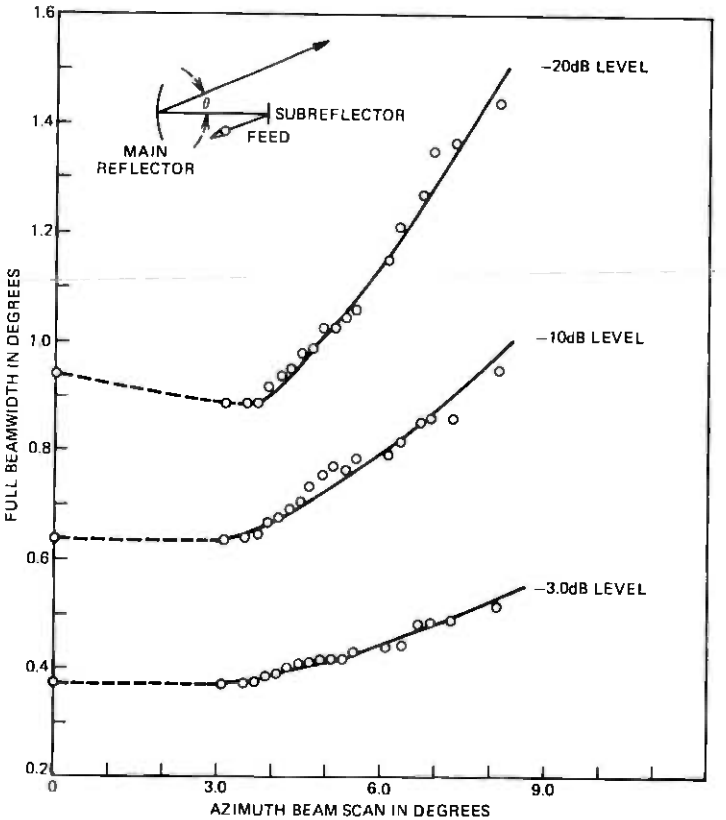


Fig. 14—Plot of 3, 10, 20 dB level beamwidths as a function of beam scanning θ.

IV. CONCLUSIONS

From the measurements, we conclude that multiple-beam antennas can be produced by using an offset Cassegrain with offset collectors as feeds. Further, they are suitable for both earth stations and satellites. We can scan a total of 18 beamwidths with only 0.25-dB degradation in gain; a scan of this order produces little change in radiation characteristics.

V. ACKNOWLEDGMENTS

The assistance of W. E. Legg in measuring the feed system is greatly appreciated. Thanks to R. H. Turrin and E. A. Ohm for providing the subreflector, to D. A. Gray for the help with the 100-GHz equipment, and to D. C. Hogg for his encouragement.

REFERENCES

1. L. C. Tillotson, "A Model of a Domestic Satellite Communication System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.
2. E. A. Ohm, "A Proposed Multiple-Beam Microwave Antenna for Earth Stations and Satellites," B.S.T.J., *53*, No. 8 (October 1974), pp. 1657–1665.
3. C. Dragone and D. C. Hogg, "The Radiation Pattern and Impedance of Offset and Symmetrical Near-Field Cassegrainian and Gregorian Antennas," IEEE Trans. Ant. Propag., *AP-22*, No. 3 (May 1974), pp. 472–475.
4. M. J. Gans and R. A. Semplak, "Some Far-Field Studies of an Offset Launcher," B.S.T.J., *54*, No. 7 (September 1975), pp. 1319–1340.
5. Ta-Shing Chu and R. H. Turrin, "Depolarization Properties of Offset Reflector Antennas," IEEE Trans. Ant. Propag. *AP-21*, No. 3 (May 1973), pp. 339–345.
6. John Ruze, "Lateral Feed Displacement in a Paraboloid," IEEE Trans. Ant. Propag. *13*, No. 5 (September 1965), pp. 660–665.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Detection and Selective Smoothing of Transmission Errors in Linear PCM

By R. STEELE and D. J. GOODMAN

(Manuscript received May 5, 1976)

We consider detection of transmission errors in PCM by means of statistical hypothesis testing of the received quantized sequence. When errors are detected, a median filter is used to smooth waveform discontinuities. We describe two error detectors, one (CDC), based on correlation measurements, and the other (DDC), based on sample-to-sample difference measurements. While both offer s/n advantages over conventional PCM in the presence of errors, DDC is more promising both in terms of performance and simplicity of implementation.

I. INTRODUCTION

An acceptable decoded signal-to-noise ratio (s/n) can be maintained in a pulse code modulation (PCM) system in the presence of transmission errors if error detecting and correcting codes are added to the transmitted PCM signal. This approach[1,2] when combined with a Huffman coder in juxtaposition to the PCM and channel encoders offers the best theoretical solution, given that the properties of the transmission channel can be specified. By best solution we mean that for a specified channel bandwidth and error rate, the highest decoded s/n can be achieved. However, this approach is not usually justified economically, and partial solutions may be appropriate.

One solution is to retain a conventional transmitter and to modify the receiver of a PCM system to make provision for the possibility of transmission errors. Jayant[3] has observed that when the channel error rate is high, a linear or non-linear filter prior to the desampling filter reduces the noise due to transmission errors, but only at the expense of a degradation of speech quality when the channel error rate is low. This approach is analogous to reducing the bandwidth of a high-frequency receiver or introducing a noise filter in a high-fidelity system.

In this paper, we discuss a system which, like one of those described by Jayant, uses a median filter[4,5] to squelch channel error noise. How-

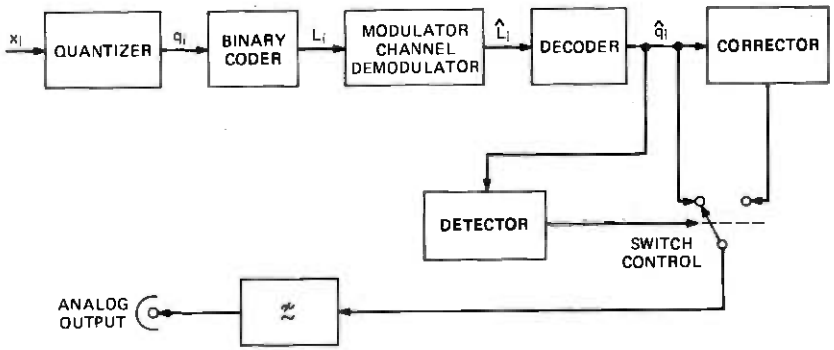


Fig. 1—PCM system with error protection.

ever, in our system this filter is introduced selectively under the control of an error detector, which measures certain statistics of the received PCM signal and makes inferences about whether individual samples or short blocks of samples have been contaminated by channel errors. Only when errors are detected is the median filter introduced. We use the results of computer simulations to show that this approach can lead to significant improvements in system signal-to-noise ratio. Another detection and correction system has been proposed for differential PCM systems operating substantially above the Nyquist rate[6].

II. ERROR DETECTION AND CORRECTION

We have investigated the system shown in Fig. 1, using a third-order median filter as a corrector. When the detector infers the presence of an error in either an individual sample or a block of samples, it causes the switch to be in the right-hand position, thus introducing the median filter, which, at time m, replaces the sample $\hat{q}_m$ with the median value of the samples $\hat{q}_{m+1}$, $\hat{q}_m$, $\hat{q}_{m-1}$. With no error detected, the switch is in the left-hand position and $\hat{q}_m$ goes directly to the low-pass filter, which transforms the sequence of quantized samples to a continuous waveform. Higher-order median filters,[5] though less satisfactory in the absence of an error detector,[3] may well improve the performance of a selective correction scheme. Linear smoothing is also likely to be effective.

The detector is essentially looking for an unexpected event in $\{\hat{q}_i\}$, and the more unexpected the event, the greater is the likelihood of detection. Correspondingly, large errors are more likely to be observed than small ones. Very small errors are unlikely to be detected due to their statistical similarity to the sequence $\{q_i\}$ at the transmitter. This characteristic is a limitation, although not too serious as it is the large errors that cause the greatest degradation of s/n.

We have studied two detectors, both of which process blocks of M samples and compute a statistic characteristic of each block. They also

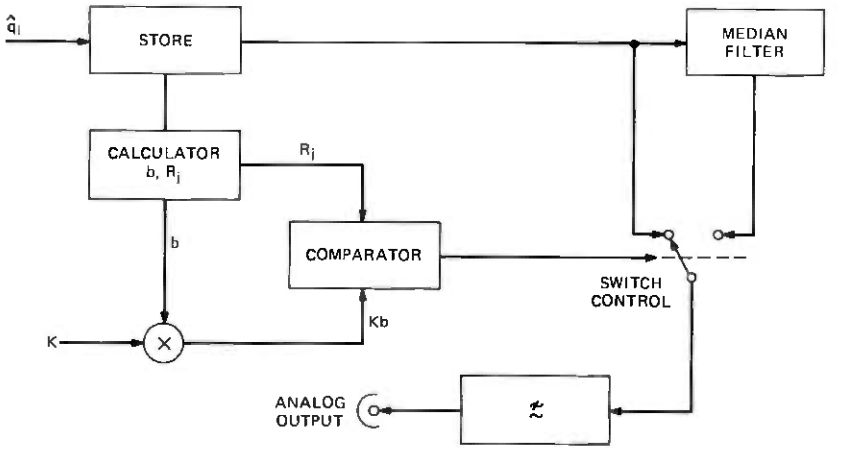


Fig. 2—Correlation detection and correction.

process shorter blocks (of length SB) imbedded in each long block, compute corresponding block statistics, and infer errors in a short block whenever its statistic is substantially different from that of the long block. In the correlation detection and correction system (CDC), the statistic is the first-order autocorrelation coefficient. In the difference detection and correction system (DDC), the statistic is the rms value of the sample-to-sample difference. In both cases, the long block length is $M = 64$ for speech sampled at 8 kHz. In CDC, the short block has $SB = 16$ samples while in DDC, $SB = 2$ and the block statistic is simply the magnitude of a single sample-to-sample difference.

III. TWO ERROR DETECTORS—DEFINITIONS

3.1 Correlation detection and correction

Figure 2 is a schematic representation of CDC. Correlation coefficients b and R_j are computed over blocks of two different sizes. The coefficient b, computed over a long block of samples of length M, provides an estimate of the local correlation of the transmitted sequence. It is compared with $\{R_j\}$, a sequence of correlation coefficients computed over short blocks of length SB, which lie within the long block. The presence of one or more errors in a short block can result in correlations substantially lower than b because channel errors are independent of the signal source. Thus, if R_j is substantially lower than b, the system infers the presence of an error in the block of length SB and replaces the samples in that block:

$$\hat{q}_j, \hat{q}_{j+1} \cdots, \hat{q}_{j+(SB-1)}$$

with the corresponding outputs of the median filter.

Specifically, for the first M received samples, we have

$$b = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1} \hat{q}_i \hat{q}_{i+1}}{\frac{1}{M}\sum_{i=1}^{M} (\hat{q}_i)^2}. \tag{1}$$

Within this block, the detector computes R_1, the correlation coefficient of samples $\hat{q}_1$ to $\hat{q}_{SB}$; R_2 is based on $\hat{q}_2$ to $\hat{q}_{SB+1}$; etc. In general,

$$R_j = \frac{\hat{q}_j\hat{q}_{j+1} + \hat{q}_{j+1}\hat{q}_{j+2} + \ldots + \hat{q}_{j+SB-2}\hat{q}_{j+SB-1}}{(SB-1)\,S_j{}^2}, \tag{2}$$

where

$$S_j = \frac{1}{SB}\sum_{i=1}^{SB} (\hat{q}_{j+i-1})^2; \qquad j = 1,2,\ldots,M - SB + 1. \tag{3}$$

The second long block begins at $k = M - SB + 2$ and provides a value of b to be compared with $R_{M-SB+2}, R_{M-SB+3}, \ldots, R_{2(M-SB+1)}$. The third long block begins at $k = 2M - 2SB + 3$ etc. The time windows defining short blocks move one sample at a time; these defining long blocks slide over $M - SB + 1$ samples.

A particular block j of SB samples is deemed to contain errors if R_j is sufficiently smaller than b; i.e., if

$$R_j < Kb, \tag{4}$$

where $K < 1$ is a design parameter of the CDC system.

3.2 Difference detection and correction

This scheme, shown in Figure 3, is based on the notion that the differences between successive samples of a correlated input source tend to be relatively small. The detector infers that an unusually large sample difference is the result of a transmission error. Over the nth block of M samples the detector computes σ_b, the rms difference between successive samples, where

$$\sigma_b{}^2 = \frac{1}{M}\sum_{nM+1}^{(n+1)M} (\hat{q}_i - \hat{q}_{i-1})^2. \tag{5}$$

It then examines the magnitudes, $|\Delta_k|$, of individual sample-to-sample differences and if

$$|\Delta_k| \geqq L\sigma_b, \tag{6}$$

the system replaces $\hat{q}_k$ with the corresponding output of the median filter. We use for Δ_k

$$\Delta_k = \hat{q}_k - Q_k, \tag{7}$$

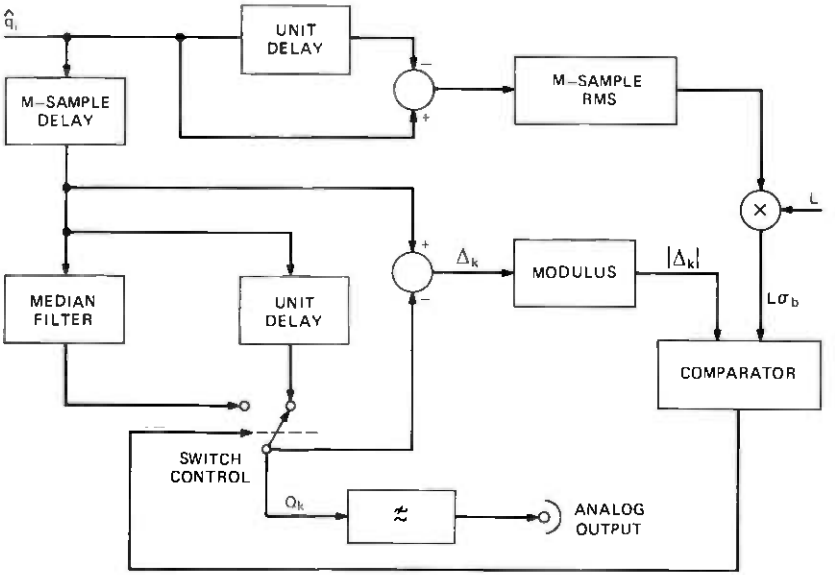


Fig. 3—Difference detection and correction.

where Q_k is the actual system output. It is $\hat{q}_{k-1}$, if $|\Delta_k| < L\sigma_b$; otherwise, it is a median filter output.

Notice that when an error is detected, the DDC system replaces an *individual* sample with a median filter output, while the CDC system uses the median filter to modify an *entire block* of samples. This greater selectivity of the DDC system results in the modification of fewer correctly received samples than with CDC. This property accounts for the fact that DDC provides better measured performance than CDC.

IV. PERFORMANCE EVALUATION

To gain insight into the detection and correction mechanisms, we implemented both detectors on a general-purpose computer and studied their operation on PCM samples derived from an artificial, statistically stationary source. The initial simulations were efficient computationally and demonstrated the effects of design parameters and signal characteristics on s/n. They also indicated that DDC performs better than CDC in the presence of isolated channel errors. These simulations were followed by investigations (using both software and special-purpose hardware) of DDC operating on PCM-coded speech transmitted over binary symmetric channels. With speech transmission, the error suppression provided by DDC is clearly audible and the s/n characteristics are similar to those observed with the artificial source.

In the next two sections, we present the results of the first simulations.

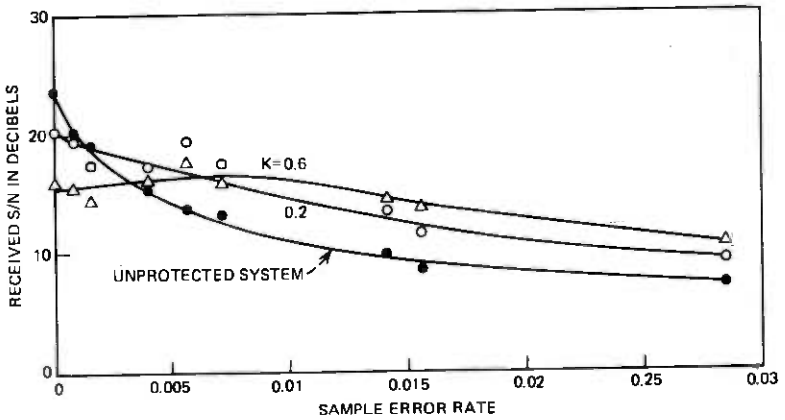


Fig. 4—CDC performance, 32-level quantizer.

The source was a Gauss-Markov sample sequence with a correlation of 0.85, chosen to be representative of speech sampled at 8 kHz. Channel errors were introduced by replacing source samples with quantized samples of a white Gaussian process independent of the source.

4.1 CDC, Gauss-Markov source

With applications to speech communication in mind, we adopted, for the length of the long blocks $M = 64$, an interval (6–10 ms for typical sampling rates) over which correlation properties are expected to change slowly. The choice of the length, SB, of the short blocks reflects a compromise between the aims of obtaining: (*i*) reliable correlation measures (achieved with SB large), and (*ii*) accurate error localization (achieved with SB small). Unreliable correlation measures result in false alarms—spurious error detections—while imprecise error localization leads to the modification of a large number of correctly received samples. The other detector design parameter is K in eq. (4), which sets the threshold of error detection. A low value of K provides a stringent criterion, leading to fewer false alarms, but also fewer correctly detected errors than a high value of K.

After studying the influence of SB and K on the false alarm and correct detection probabilities, we arrived at $SB = 16$ as an appropriate compromise. Effective values of K range from 0.2 to 0.6, depending on the sample error rate (η).

The dependence of s/n on error rate is shown in Fig. 4 for a 32-level quantizer and $K = 0.2$ and 0.6. Observe that for $\eta > 0.002$, the CDC system having $K = 0.2$ is preferable to an unprotected PCM system. For $\eta > 0.006$, the improvement in the received s/n is approximately 3 dB. The CDC system offers an improvement of a further 1 to 2 dB when $\eta > 0.01$, K

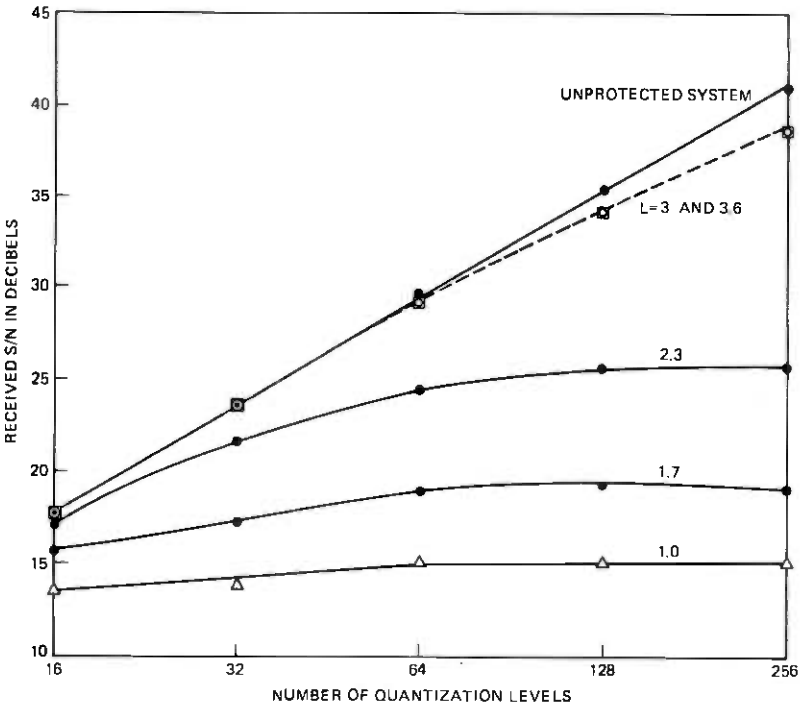


Fig. 5—DDC performance, error-free channel.

= 0.6. On the other hand, when the channel quality is reasonably high ($\eta < 0.002$), the CDC degrades s/n performance relative to an unprotected system. Even with $K = 0.2$, the deterioration is more than 3 dB in an error-free channel, and it is more substantial when the number of quantizer levels exceeds 32. The extra noise arises from the replacement of a block of 16 correctly received samples by median filter outputs whenever a false alarm occurs. This high cost of false alarms is a principal disadvantage of CDC. It does not exist in the DDC system, where an isolated false alarm introduces only one median filter output.

4.2 DDC, Gauss-Markov source

In this system, the detector design parameters are the block size M over which the rms difference signal, σ_b, is calculated and L in eq. (6), which determines the criterion of error detection. As in the CDC system, we used $M = 64$ to provide a syllabic measure of the rms sample-to-sample difference signal, and investigated the effects on s/n of several values of L under various transmitter and channel conditions.

There is an important improvement in the error-free condition compared to CDC. Figure 5 shows that for the criteria $L = 3$ and 3.6, the

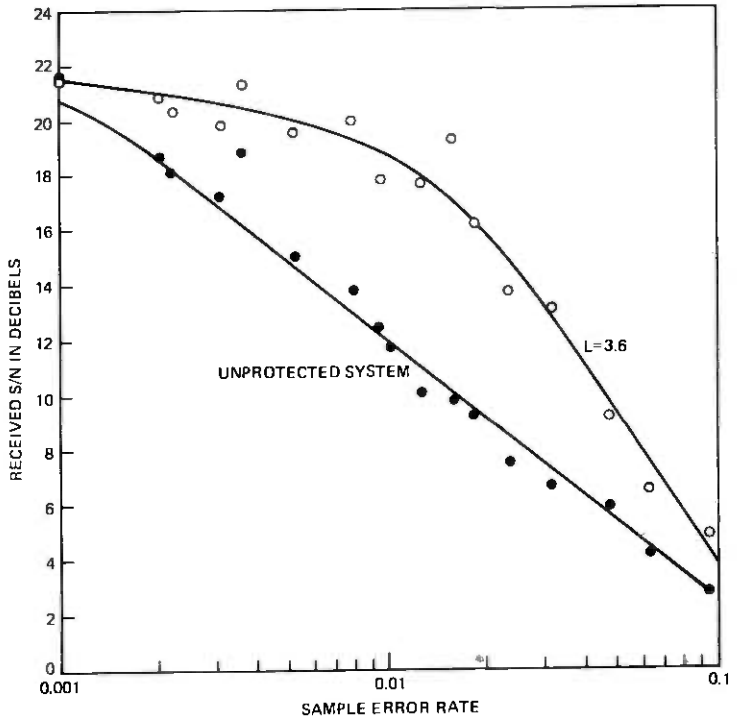


Fig. 6—DDC performance, 32-level quantizer.

degradation in received s/n due to false alarms is negligible for quantizers with 16 to 64 levels, and is only 2 dB for a quantizer with 256 levels.

The variation of s/n with η is displayed in Fig. 6 for a 32-level quantizer and $L = 3.6$. The DDC system is superior to the unprotected linear PCM system. With DDC, the s/n decreases by approximately 3 dB compared to the 9-dB reduction of the unprotected system when η increases from 0.001 to 0.01. At a 1 percent error-rate DDC provides a 7-dB improvement in s/n. With larger quantizers, the dependence of s/n on η follows the curves in Fig. 6 for $\eta > 0.003$. At these error rates the major part of the received noise is due to transmission errors rather than quantization.

Figure 7 shows s/n as a function of input power for a 32-level quantizer and $\eta = 0.016$. At low levels of input power, the quantized samples at the transmitter are generally quite small while transmission errors can cause very large samples to appear at the receiver. The resulting large sample-to-sample differences are reliably detected making DDC particularly effective at low input levels, which accounts for the fact that with DDC, s/n depends only to a small extent on input power. This property of DDC is in strong contrast to unprotected PCM in which the noise due to channel errors is essentially independent of signal level and s/n decreases

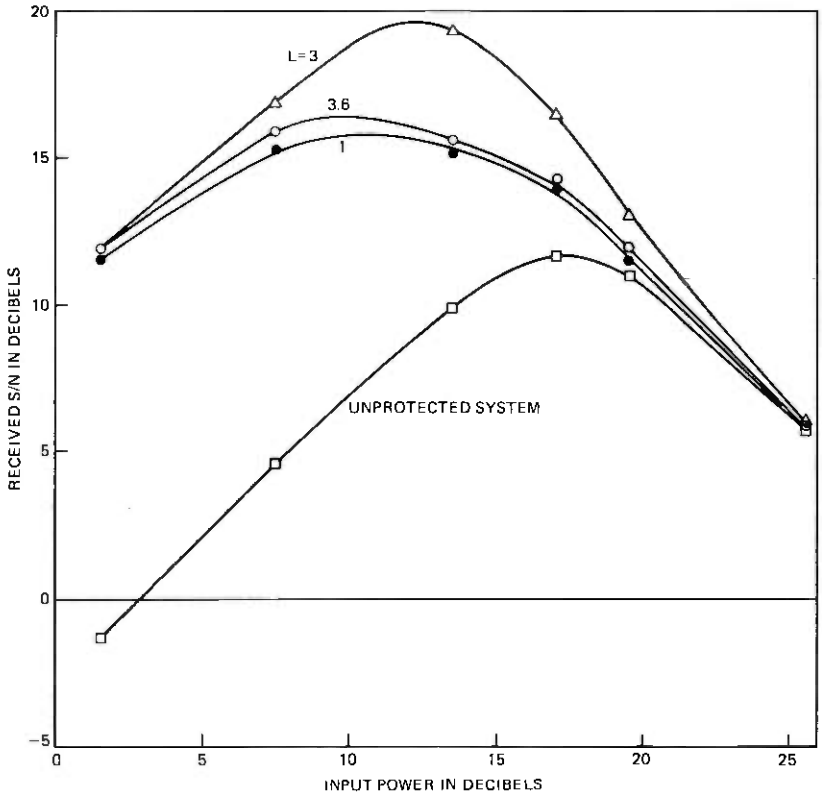


Fig. 7—DDC performance, 32-level quantizer, $\eta = 0.016$.

1 dB for each dB decrease in input power, as shown in Fig. 7. Because the s/n improvement at low signal levels does not depend on significant sample-to-sample correlations, the DDC system can be expected to perform well with speech signals which consist of essentially two waveform types: (*i*) high-level, highly correlated waveforms of voiced sounds, and (*ii*) low-level, uncorrelated waveforms of unvoiced sounds. Errors in both waveform types are detectable with DDC.

4.3 DDC, speech inputs

Encouraged by performance with Gauss-Markov inputs, we studied, by means of a computer simulation and a laboratory model, DDC systems operating on received PCM samples derived from a speech source. In the simulation, quantized samples were coded in a 5-bit sign-magnitude format and transmitted over a binary symmetric channel. The s/n performance, measured over an entire 2-second sentence, is shown in Fig.

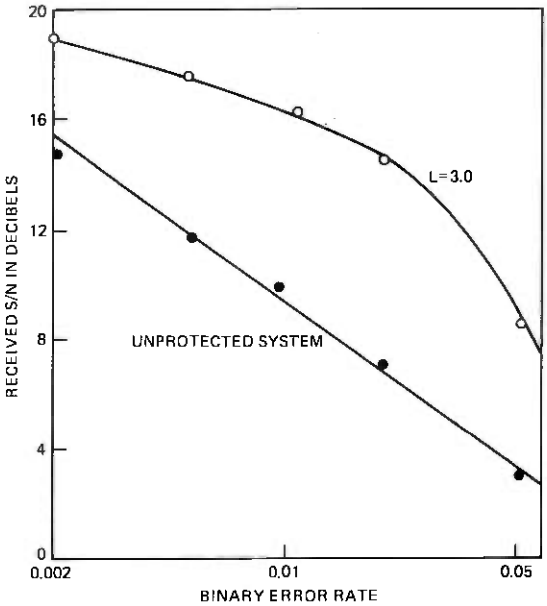


Fig. 8—DDC performance, speech transmission, 32-level quantizer.

8. The improvement introduced by DDC is immediately audible in the decoded outputs of DDC and the unprotected PCM receiver.

V. DISCUSSION

Two detection systems in connection with a smoothing filter corrector have been described and shown to give improvement in the s/n of a linear PCM codec in the presence of transmission errors. This improvement has been achieved without recourse to error detecting and correcting codes. The key element of both systems is a detector that allows the smoothing filter to be used selectively on the basis of inferences about errors in the received sequence. To date we have experimented with only one corrector, the third-order median filter. Although it increases s/n relative to an unprotected system, it is possible that other smoothing schemes offer even greater improvements.

Of the two detectors, DDC has better s/n performance and is easier to implement. It does not involve the calculation of correlation coefficients and the rms value of the quantized difference signal is easy to measure. It also is well-matched to the characteristics of speech waveforms.

There are existing and anticipated digital transmission systems in which performance is limited by channel quality. We hope that the approach taken here will be of value in upgrading performance at the expense of a tolerably small increase in receiver cost. Our method is also

applicable to signal enhancement in systems other than PCM in which distortions are characterized by short, severe signal discontinuities. In such systems (FM signals exhibiting clicks is an example), the disturbed signal can be digitized, selectively smoothed, and desampled in the manner described here to produce an improved output.

REFERENCES

1. R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.
2. A. D. Wyner, "Another Look at the Coding Theorem of Information Theory—a Tutorial," Proc. IEEE, *58*, No. 6 (June 1970), pp. 894–913.
3. N. S. Jayant, "Average- and Median-Based Smoothing Techniques for Improving Digital Speech Quality in the Presence of Transmission Errors," IEEE Trans. Commun., *COM-24*, No. 9 (Sept. 1976), pp. 1043–1045.
4. J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," 1974 EASCON Record, p. 673.
5. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. on Acoust. Speech and Signal Processing, *ASSP-23*, No. 6 (December 1975) pp. 552–557.
6. R. Steele and M. A. Yeoman, "Detection and Partial Correction of Isolated Errors From the Received Data in a First-Order DPCM Coder," Electron. Lett., *11*, No. 11, (May 29, 1975).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

On the Angle Between Two Fourier Subspaces

By J. E. MAZO

(Manuscript received September 3, 1976)

Examining an approximation inspired by equalization theory, we consider the minimum angle Ω_N between the subspaces of Hilbert space generated by the sequences $\{e^{ik\omega}\}_{-N}^{N}$ and $\{e^{ik\omega}\}_{|k|>N}$. Here $\omega \in [-\pi,\pi]$ and the inner product for the Hilbert space involve a positive, bounded weight function $r(\omega)$. The finite Toeplitz matrices R and Γ generated by $r(\omega)$ and $1/r(\omega)$, respectively, play a crucial role, and, in fact, $\sin^2\Omega_N$ is the reciprocal of the largest eigenvalue of $R\Gamma$. In general, $\sin^2\Omega_N$ is shown to be bounded away from unity as N becomes large. The geometry of the problem enables us to give some results concerning the product matrix $R\Gamma$, which, out of the present context, may seem surprising.

I. INTRODUCTION AND SUMMARY

Let H be the Hilbert space of square-integrable functions on $[-\pi,\pi]$ with an inner product† given by

$$(f,g)_r = \frac{1}{2\pi}\int_{-\pi}^{\pi} f^*(\omega)g(\omega)r(\omega)d\omega, \tag{1}$$

where the weight function $r(\omega)$ is bounded and strictly positive; i.e.,

$$0 < r \leqq r(\omega) \leqq R. \tag{2}$$

We call a Fourier subspace of H any subspace generated by a finite or infinite collection of functions of the form $e^{in\omega}$, n an integer. In particular, we shall be interested in the Fourier subspaces [relative to the metric $r(\omega)$]:

$$\begin{aligned} F_N &= \left\{\sum_{-N}^{N} f_n e^{in\omega}\right\}_r \\ G_N &= \left\{\sum_{|n|>N} g_n e^{in\omega}\right\}_r \end{aligned} \tag{3}$$

† The subscript r, as in (1), will be used when we wish to emphasize that the weight function $r(\omega)$ is being used. No subscript will refer to an arbitrary inner product, while the case $r(\omega) = 2\pi$ will be called the "usual" metric. The usual inner product will be written without a subscript as well.

for $N \geqq 0$.† If $r(\omega)$ is constant, then the subspaces F_N and G_N are orthogonal. We will be concerned with the minimum angle between F_N and G_N for a general weight function satisfying (2), and with the limiting behavior of this angle as $N \to \infty$. (The concept of the angle between subspaces is not new to the engineering literature. See for example Ref. 1.)

The main results of our investigation are stated in terms of two finite Toeplitz matrices, R and Γ, which are generated by the weight functions $r(\omega)$ and $g(\omega) = 1/r(\omega)$, respectively [see eqs. (30) and (33) for precise definitions]. We also need the Fourier coefficients r_n and g_n of $r(\omega)$ and $g(\omega)$. Then, we show

(*i*) $$\sin^2\Omega_N = \frac{1}{\text{largest eigenvalue of } R\Gamma}.$$

(*ii*) $$\lim_{N\to\infty} \sin^2\Omega_N \leqq 2\left[1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\omega)d\omega \times \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{d\omega}{r(\omega)}\right]^{-1} < 1.$$

(*iii*) All eigenvalues of $R\Gamma$ are $\geqq 1$.

(*iv*) $$\sum_{-\infty}^{\infty} |n| r_n g_n^* < 0.$$

This is not a trivial inequality in the sense that $\sum_{-\infty}^{\infty} r_n g_n^* = 1$ is.

The case $r(\omega) = 1 + a\cos\omega$ is solved exactly in Section V, showing that the obvious bound $\sin^2\Omega_N \geqq r_{\min}/r_{\max}$ is often loose. A better bound, still involving only this ratio, is given in (68).

In somewhat general terms, this problem arose in the mean-square equalization theory of data transmission, where the question is one of bounding the effect of replacing tap weight values by certain Fourier coefficients. To be specific, let us ignore the effects of noise and note that the job of the equalizer is to invert the Nyquist equivalent channel. That is, if we had an infinite number of taps at our disposal, we would take the transfer function of the equalizer $C_\infty(\omega)$ to be

$$C_\infty(\omega) = \frac{1}{X(\omega)} = \sum_{-\infty}^{\infty} \epsilon_n e^{in\omega}.$$

The equalizer transfer function $C_N(\ )$ when only the usual $(2N + 1)$ taps are available can always be written as

$$C_N(\omega) = \frac{1}{X(\omega)} - \sum_{-N}^{N} \delta_n e^{in\omega} - \sum_{|n|>N} \epsilon_n e^{in\omega}.$$

In the above expression δ_n, $|n| < N$ are "corrections" to the Fourier coefficients ϵ_n, $|n| < N$. The mean-square error resulting from the

† Any function of the form (3) is in H if and only if the associated coefficient sequence is square-summable.

equalizer $C_N(\omega)$ can then be shown to be given by

$$\frac{1}{2\pi}\sum_{-\pi}^{\pi}|X(\omega)C_N(\omega)-1|^2d\omega$$

$$=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{-N}^{N}\delta_k e^{ik\omega}+\sum_{|k|>N}\epsilon_k e^{ik\omega}\right|^2|X(\omega)|^2d\omega.$$

The minimum mean-square error $E^2_{\min}$ is the minimum of the above expression over the δ_k. Now we can imagine taking the (fixed) vector

$$\sum_{|k|>N}\epsilon_k e^{ik\omega}$$

and decomposing into a vector in the space F_N and one perpendicular to it. The part in F_N can be "subtracted off" by the choice of δ's, leaving the remainder. The fraction subtracted off can never be greater than $\cos^2\Omega_N$, where Ω_N is the angle between the two subspaces F_N and G_N when $X_{\text{eq}}(\omega)^2$ is used as the weight function $r(\omega)$ for inner products. Thus,

$$\sin^2\Omega_N\times\|\epsilon_N\|_r^2\leqq E^2_{\min}\leqq\|\epsilon_N\|_r^2, \tag{4}$$

where

$$\|\epsilon_N\|_r^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{|k|>N}\epsilon_k e^{ik\omega}\right|^2|X_{eq}(\omega)|2d\omega.$$

The point of replacing the exact tap values of the finite equalizer by Fourier coefficients is not to replace one calculation by another. Rather, it is to supplement calculation by insight, since much is known about properties of Fourier coefficients. This will be done for a specific equalization problem involving timing recovery for finite equalizers in a later work.

II. WARMUP EXERCISES

The angle θ between two fixed vectors, f and g, is defined by

$$\theta=\cos^{-1}\frac{Re(f,g)}{\|f\|\,\|g\|},\quad\theta\in[0,\pi]. \tag{5}$$

If f and g are restricted to be in subspaces F and G, respectively, the infimum of (5) (call it Ω) over all f and g (so restricted) is called the angle between the two subspaces. We easily see that

$$\|f-g\|^2\geqq\|f\|^2+\|g\|^2-2\|f\|\times\|g\|\cos\Omega, \tag{6}$$

and thus by minimizing the right member of (6) with respect to the norm of f, we have

$$\inf_f\|f-g\|^2\geqq\sin^2\Omega\|g\|^2. \tag{7}$$

In fact, we can also calculate $\sin^2\Omega$ *via* the formula

$$\sin^2\Omega = \inf_g \inf_f \frac{\|f-g\|^2}{\|g\|^2}. \tag{8}$$

When (2) holds, the infimum angle between our subspaces F_N and G_N [given by (3)] is actually attained and its value is strictly positive. In fact, it follows from an application of a theorem by Paley and Wiener[2] that the two sequences (usual metric)

$$\begin{aligned} \{\phi_n\} &= \left\{\sqrt{\frac{r(\omega)}{2\pi}}\, e^{in\omega}\right\} \\ \{\psi_n\} &= \left\{\frac{1}{\sqrt{2\pi r(\omega)}}\, e^{in\omega}\right\} \end{aligned} \tag{9}$$

form a complete biorthogonal pair, i.e.,

$$(\phi_n,\psi_m) = \delta_{nm} \tag{10}$$

and

$$h = \sum_{-\infty}^{\infty} (\phi_n,h)\psi_n = \sum_{-\infty}^{\infty} (\psi_n,h)\phi_n \tag{11}$$

for any h in L_2. Thus, either sequence in (9) forms a basis for L_2, or, equivalently, $\{e^{in\omega}\}_r$ forms a basis for H [with weight function $r(\omega)$]. Now if $f \in F_N$, $g \in G_N$,

$$\inf_g \|f-g\|_r$$

is attained when g is the orthogonal projection of f on G_N, and is a continuous function of the finite dimensional f. Therefore,

$$\inf_f \left[\inf_g \frac{\|f-g\|_r^2}{\|f\|_r^2}\right]$$

is attained, since we may restrict $\|f\|_r = 1$ and thus are minimizing over a compact set. The basis property of $\{e^{in\omega}\}_r$ in H assures that the minimum is not zero.

There are several ways to get at the minimum angle Ω_N between F_N and G_N. We shall begin by using (8) and the calculus of variations.

However, before we begin to work on this, let us review an old problem of linear prediction (really, linear interpolation) theory. We are required to find the minimum value of

$$E^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|1 - \sum_{m\neq 0} a_m e^{im\omega}\right|^2 r(\omega)\,d\omega \tag{12}$$

over all l_2 sequences $\{a_m\}$, under assumption (2) for $r(\omega)$. We let

$$a(\omega) = \sum_{m\neq 0} a_m e^{im\omega}$$

be any element of $L_2(-\pi,\pi)$ which has zero for its zeroth Fourier coefficient. We then have

$$E^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |1 - a(\omega)|^2 r(\omega)d\omega. \tag{13}$$

The calculus of variations yields

$$\int \delta a^*[1 - a(\omega)]r(\omega)d\omega = 0 \tag{14}$$

or

$$\int e^{in\omega}[1 - a(\omega)]r(\omega)d\omega = 0, \quad n \neq 0. \tag{15}$$

Thus,

$$[1 - a(\omega)]r(\omega) = \text{const} = k. \tag{16}$$

From (13) and (16), on the one hand, we have

$$E^2_{\min} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{k^2}{r^2(\omega)} r(\omega)d\omega = \frac{k^2}{2\pi}\int_{-\pi}^{\pi} \frac{d\omega}{r(\omega)} \tag{17}$$

and, on the other,

$$E^2_{\min} = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega(1 - a^*)[(1 - a)r] = \frac{k}{2\pi}\int_{-\pi}^{\pi} (1 - a^*)d\omega = k, \tag{18}$$

since a^* has no $m = 0$ term. Equating the results of (17) and (18) enables us to solve for k [note that $k = 0$ must be excluded under (2)], yielding

$$E^2_{\min} = \frac{1}{\dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \frac{d\omega}{r(\omega)}} = k. \tag{19}$$

If $r(\omega)$ has a zero somewhere in such a manner that $\int 1/r = \infty$, then (19) says $E^2_{\min} = 0$. This turns out to be the correct conclusion for the infimum, but this infimum is never attained. As is well known, the calculus of variations can only be applied if the infimum is attained. In fact, in the present problem, if we set $k = 0$ in (16), we would conclude that there is an l_2 sequence $\{a_m\}$ such that

$$1 - \sum_{m \neq 0} a_m e^{im\omega} = 0 \quad \text{a.e.}, \tag{20}$$

which obviously cannot be.

Another way to do this problem is to use the biorthogonal sequences (9). We note that all vectors of the type

$$\sum_{m \neq 0} a_m \phi_m \tag{21}$$

form the subspace orthogonal to the vector ψ_0. Hence, $E^2_{\min}$must simply be the squared norm of the projection of ϕ_0 onto ψ_0. This is (in the usual norm)

$$\|\phi_0\|^2\cos^2(\phi_0,\psi_0) = \|\phi_0\|^2 \frac{|(\phi_0,\psi_0)|^2}{\|\phi_0\|^2\|\psi_0\|^2}$$

$$= \frac{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \sqrt{r(\omega)} \times \frac{1}{\sqrt{r(\omega)}}\, d\omega\right)^2}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{d\omega}{r(\omega)}} = \frac{1}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{d\omega}{r(\omega)}}. \quad (22)$$

III. FINDING THE MINIMUM ANGLE

We proceed with the calculus-of-variations approach to finding $\sin^2\Omega_N$ via (8). Let

$$f(\omega) = \sum_{-N}^{N} f_k e^{ik\omega} \in F_N$$

$$g(\omega) \in G_N. \quad (23)$$

Then, if we vary g^* in $\|f - g\|_r^2$, we obtain

$$\int \delta g^*(w)[f(\omega) - g(\omega)]r(\omega)d\omega = 0 \quad (24)$$

for all allowed variations. Thus, (24) means $[f - g]r \in F_N$, or, in other words,

$$[f(\omega) - g(\omega)]r(\omega) = \sum_{-N}^{N} b_k e^{ik\omega} \equiv b(\omega) \quad (25)$$

for some numbers b_k. As in Section II, (25) permits us to write two expressions for $\min\|f - g\|_r^2$. They are

$$\min_g \|f - g\|_r^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|b(\omega)|^2}{r(\omega)}\, d\omega \quad (26)$$

and

$$\min_g \|f - g\|_r^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega f^*(\omega)b(\omega) = \mathbf{f}^* \cdot \mathbf{b}. \quad (27)$$

The vector notation in (27) refers to a row of $(2N + 1)$ numbers. Letting $\mathbf{b} = k(\mathbf{f} + \mathbf{b}_\perp)$, $\mathbf{f}^* \cdot \mathbf{b}_\perp = 0$, we may equate (26) and (27), solve for k, and obtain

$$\min_g \|f - g\|_r^2 = \frac{\left(\sum_{-N}^{N} |f_k|^2\right)^2}{\min\limits_{\substack{\gamma \\ \mathbf{f}\cdot\gamma = 0}} \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|\sum_{-N}^{N} f_k e^{ik\omega} + \sum_{-N}^{N} \gamma_k e^{ik\omega}\right|^2}{r(\omega)}\, d\omega}. \quad (28)$$

Thus, from (8) and (28)

$$\sin^2\Omega_N = \min_{\substack{\mathbf{u} \\ \sum_{-N}^{N} |u_i|^2 = 1}} \frac{1}{\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{-N}^{N} u_k e^{ik\omega}\right|^2 r(\omega)d\omega \times \min_{\substack{\mathbf{w} \\ \mathbf{w}\cdot\mathbf{u}=0}} \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega \frac{\left|\sum_{-N}^{N}(u_k + w_k)e^{ik\omega}\right|^2}{r(\omega)}\right]} \tag{29}$$

If we introduce the Toeplitz matrix

$$\Gamma_{nm} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(n-m)\omega}\frac{1}{r(\omega)}\,d\omega, \quad |n|, \quad |m| \leqq N, \tag{30}$$

the second term in the denominator of (29) has the form

$$u^+\Gamma u + w^+\Gamma w + 2u^+\Gamma w. \tag{31}$$

Expression (31) may be minimized over the appropriate $\mathbf{w}$ using a Lagrange multiplier, yielding

$$\frac{1}{u^+\Gamma^{-1}u}$$

Thus,

$$\sin^2\Omega_N = \min_{\mathbf{u}} \frac{u^+\Gamma^{-1}u}{u^+Ru}, \tag{32}$$

where R is the Toeplitz matrix corresponding to $r(\omega)$, i.e.,

$$R_{nm} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(n-m)\omega} r(\omega)d\omega, \quad |n|, \quad |m| \leqq N. \tag{33}$$

Finally, the minimization of (32) yields

Theorem I. Let the matrices R and Γ be defined as in (30) and (33). Then the minimum angle Ω_N between the Fourier subspaces F_N and G_N, defined in (3), is

$$\sin^2\Omega_N = \frac{1}{\text{largest eigenvalue of } R\Gamma} \tag{34}$$

This theorem implies that Ω_N is invariant under the replacement $r(\omega) \to 1/r(\omega)$.

Since similarity transformations preserve eigenvalues, we note that the eigenvalues of $R\Gamma$ are the same as those of $\sqrt{\Gamma}\,R\sqrt{\Gamma}$, which is, by (2), (30), and (33), a strictly positive definite Hermitian matrix.

Before exploring consequences of (34), we shall rederive it from a geometric point of view. Let $\{\phi_i\}_1^\infty$ and $\{\psi_i\}_1^\infty$ be complete biorthogonal sequences of vectors for a Hilbert space. Let

$$\begin{aligned} V &= \{\phi_i\}_1^N, \quad W = \{\phi_i\}_{N+1}^\infty \\ T &= \{\psi_i\}_1^N \end{aligned} \tag{35}$$

be subspaces generated by the indicated vectors. Note that the orthogonal complement of W, $W_\perp$, is given by $W_\perp = T$. Also note that our problem is equivalent to that of finding the minimum angle between V and W. If $v \in V$, and α is the angle between v and W (i.e., the angle between v and its projection on W), and β is the angle between v and $W_\perp$, we have†

$$\alpha + \beta = \frac{\pi}{2}. \tag{36}$$

Thus, the minimum angle between V and W (call it Ω) is the complement of the maximum angle between V and T, called θ_M. Thus, we have

$$\sin^2\Omega = \cos^2\theta_M. \tag{37}$$

The spaces V and T both have dimension N here.

Let P represent the orthogonal projection operator onto T, and Q the orthogonal projection operator onto V. It can be shown that if $V \in V$ is a vector in V which attains the minimum or maximum angle between V and T, we must have ‡

$$QPv = \lambda v, \tag{38}$$

where $\lambda \geqq 0$ is the square of the cosine of the indicated angle. We shall see shortly that (34) is a form of (38) when we represent P and Q by matrices that are representations of the restrictions of P to V and Q to T.

We begin by deriving these matrix representations. For general biorthogonal sequences $\{\phi_i\}$, $\{\psi_i\}$, let

$$\begin{aligned} R_{nk} &= (\phi_n,\phi_k) \\ \Gamma_{nk} &= (\psi_n,\psi_k) \end{aligned} \qquad n,k = 1,2,\cdots,N. \tag{39}$$

† If we call the projection of v on W by v_1, then $v - v_1$ is the projection of v on $W_\perp$. Since v, v_1, and $v - v_1$ all lie in a plane, (36) follows immediately from a simple diagram depicting these three vectors.

‡ Let α be any vector in the space such that $Q\alpha \neq 0$. Then if θ is the angle between $Q\alpha \in V$ and W, we have $\cos^2\theta = \|PQ\alpha\|^2/\|Q\alpha\|^2 = (\alpha,QPQ\alpha)/(\alpha,Q\alpha)$. Vectors α which yield stationary values of this ratio of quadratic forms can be obtained by differentiating $(\alpha,QPQ\alpha)$ holding $(\alpha,Q\alpha)$ constant (via a Lagrange multiplier λ). This procedure yields (38) upon setting $Q\alpha = v$.

Any vector x can be written uniquely as a vector in V plus a vector in the orthogonal complement of V, i.e.,

$$x = \sum_{1}^{N} a_i \phi_i + \sum_{N+1}^{\infty} b_i \psi_i. \tag{40}$$

If we form the inner product of (40) with $\phi_j, j = 1,2,\cdots,N$, we can calculate

$$a_k = \sum_{l=1}^{N} (R^{-1})_{kl}(\phi_l, x). \tag{41}$$

Thus, given any x, its projection onto V is simply

$$\sum_{1}^{N} a_i \phi_i \tag{42}$$

with a_i given by (41). Similarly, the projection of x onto T is

$$\sum_{1}^{N} b_i \psi_i \tag{43}$$

with

$$b_k = \sum_{l=1}^{N} (\Gamma^{-1})_{kl}(\psi_l, x) \tag{44}$$

Hence, if we start with any vector $v \in V$,

$$v = \sum_{1}^{N} v_i \phi_i, \tag{45}$$

the result of projecting it onto T and then projecting this vector back to V is another vector $v'' \in V$ with components v_m'' given by

$$v_m'' = \sum_{i=1}^{N} (R^{-1}\Gamma^{-1})_{mi} v_i. \tag{46}$$

Hence, the operator equation (38) becomes the $N \times N$ matrix equation

$$(\Gamma R)^{-1} v = \lambda v. \tag{47}$$

Equation (34) is thus rederived, after an appropriate relabeling of indices.

Since the reciprocals of the largest and smallest eigenvalues of $R\Gamma$ have interpretations as squares of cosines of angles, we have

Theorem II. Let matrices R and Γ be defined as in (30), (33). Then all eigenvalues of $R\Gamma$ are ≥ 1.

IV. IMPLICATIONS CONCERNING $\sin^2\Omega_N$

From (30) and (33), we see that the matrix elements of $r\Gamma$ are given in terms of the Fourier coefficients r_j and g_j of $r(\omega)$ and $g(\omega) = 1/r(\omega)$.

Thus, (33) reads $R_{nm} = r_{m-n}$. We see that, in this notation

$$(R\Gamma)_{nk} = \sum_{m=-N}^{N} g_{k-m} r_{m-n}, \quad |n|, |k| \leqq N. \tag{48}$$

This can also be written

$$(R\Gamma)_{nk} = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{i(n\omega - k\omega')} \frac{r(\omega)}{r(\omega')} \frac{\sin \frac{2N+1}{2}(\omega - \omega')}{\sin \frac{1}{2}(\omega - \omega')} d\omega d\omega'. \tag{49}$$

Equation (49) follows easily from (36), (33), and the identity

$$\sum_{m=-N}^{N} e^{im\psi} = \frac{\sin \frac{2N+1}{2}\psi}{\sin \frac{\psi}{2}}.$$

If the largest eigenvalue of $R\Gamma$ approaches 1 as N becomes large, then $\sin^2\Omega_N \to 1$, and the subspaces F_N and G_N described in (3) eventually become orthogonal. Equivalently, we have seen that the question becomes the following: Does the largest angle between the subspaces (with the usual metric)

$$F_N = \left\{ e^{in\omega} \sqrt{\frac{r(\omega)}{2\pi}} \right\}_{-N}^{N}$$

and

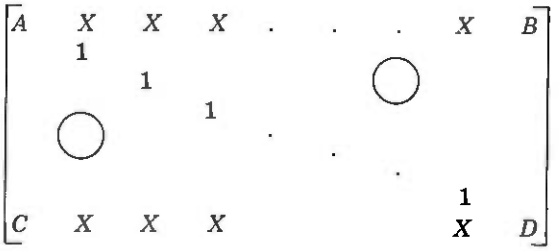
$$G_N = \left\{ e^{in\omega} \frac{1}{\sqrt{2\pi r(\omega)}} \right\}_{-N}^{N}$$

approach zero? If we *set* $N = \infty$, the two generated spaces are identical† (all of L_2), so from this point of view it comes as a surprise that the limiting angle between F_N and G_N is bounded away from zero.

Let us assume that $r(\omega)$ has only a finite number of Fourier coefficients; that is, assume $r_j = 0$ if $|j| > k$. For this case, the reader may verify, using either (48) or (49), that the $(2N+1) \times (2N+1)$ matrix $R\Gamma$ has the form (once $N \geqq k$)‡

† F_N is never G_N for finite N unless $r(\omega) = \text{const}$. This follows from using (11) to show that you cannot expand each ϕ_n, $n \leqq |N|$ in terms of the ψ_k, $|k| < N$.

‡ To see this, let us evaluate $(R\Gamma)_{ab}$ from (48). If $|a| < (N-k)$, the summation in (48) may be extended from $N = -\infty$ to $N = +\infty$, since $r_j = 0$ if $j > k$. Using the duality between l_2 and L_2, the resulting sum is then $1/2\pi \int_{-\pi}^{\pi} r(\omega) exp(ia\omega)][g(\omega) exp(ib\omega)]^* \, d\omega = \delta_{ab}$, since $g(\omega) = 1/r(\omega)$.



(50)

That is, the first k rows and the last k rows are nonvanishing. The remaining $2(N-k)+1$ diagonal elements are exactly unity, while all other matrix elements vanish. The four $k \times k$ matrices in the corners are singled out for special attention and are labeled A, B, C, D. The capital X's in the first and last rows are inserted only to indicate that elements in the first and last k rows are not vanishing, in general.

As an example, we write the elements of A explicitly, labeling the elements of **A** by a_{rs}, $r, s = 0,1, \cdots, k-1$; i.e., $a_{00} = (R\Gamma)_{-N,-N}$, etc. Then

$$a_{rs} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sum_{l=-r}^{k} f_l e^{il\omega} e^{-is\omega}}{f(\omega)} d\omega. \tag{51}$$

The elements of D can be determined from A using the property

$$(R\Gamma)^*_{m,l} = (R\Gamma)_{-m,-l}. \tag{52}$$

It is important to note that (for $N \geqq k$) the elements of A and D do not depend on N. However those of B and C do. For example, the upper right corner of B is the element $(R\Gamma)_{-N,N}$, given from (48) as

$$(R\Gamma)_{-N,N} = \sum_{l=0}^{k} r_l g_{2N-l}. \tag{53}$$

Not only does (53) depend on N, but, by the Riemann-Lebesgue lemma, $g_{2N-l} \to 0$ as N increases for l bounded. As N increases, all elements of B and C similarly vanish.

We now look further at the problem of calculating the eigenvalues of (50). If one is not an eigenvalue then the matrix $R\Gamma - \lambda I$ has $2(n-k)+1$ diagonal elements $(1-\lambda)$. By multiplying a row in which such an element occurs by the appropriate constant and adding the result to one of the first or last rows, all the elements indicated by "X" in (50) can be made to vanish. Clearly then, the eigenvalues that are not unity are given by the nonunity eigenvalues of the $2k \times 2k$ matrix

$$K \equiv \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{54}$$

where B and C depend on N. All other eigenvalues of (49) are one. Since B and C vanish in the limit of large N, we have†

$$\lim_{N\to\infty} \lambda_{\max}(N) = \text{largest eigenvalue of } A. \tag{55}$$

The simple fact that F_N is never G_N for finite N implies that the largest eigenvalue of $R\Gamma$, and hence K, is strictly greater than one. Also the fact that all eigenvalues of $R\Gamma$ are greater than or equal to 1 implies $tr\, K = tr\, A + tr\, D > 2k$. But A and D do not depend on N and have the same trace. Hence, $tr\, A > k$, and A has an eigenvalue strictly greater than unity. Thus, from (34) and (55), $\lim \sin^2\Omega_N < 1$.

The above discussion implies the following:‡

Theorem III. If $r(\omega)$ has only a finite number of nonvanishing Fourier components and is not constant, then $\lim_{N\to\infty} \sin^2\Omega_N < 1$.

Theorem IV. Let $1 \geq r(\omega) > 0$ on $[-\pi,\pi]$ have only a finite number of nonvanishing Fourier coefficients, r_n. Set $g(\omega) = 1/r(\omega)$ and call its Fourier coefficients g_n. Then

$$\sum_{n=-\infty}^{\infty} |n| r_n g_n^* < 0,$$

unless $r(\omega)$ is constant.

Proof. Using (49) for the product $R\Gamma$ and the identity immediately following it, we calculate

$$tr\, R\Gamma - (2N+1) = (2N+1)\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{r(\omega)}{r(\omega')} K_N(\omega,\omega')\, d\omega d\omega' - 1\right], \tag{56}$$

where $K_N(\omega,\omega')$ is the well-known Fejer kernel[3]

$$K_N(\omega,\omega') = \frac{1}{2\pi(2N+1)} \frac{\sin^2 \frac{2N+1}{2}(\omega-\omega')}{\sin^2 \frac{1}{2}(\omega-\omega')}. \tag{57}$$

It has the following property. Let $X(\omega) \in L_2(-\pi,\pi)$, and let

$$\tilde{X}(\omega) = \int_{-\pi}^{\pi} K_N(\omega,\omega') X(\omega')\, d\omega'. \tag{58}$$

† The $k \times k$ matrices A and D have the same eigenvalues.

‡ The restrictions in Theorem III and Theorem IV to only a finite number of nonvanishing components of $r(\omega)$ is removed in Section V.

Call the Fourier coefficients of $\tilde{X}(\omega)$ and $X(\omega)$, $\tilde{X}_n$ and X_n, respectively. Then

$$\tilde{X}_n = X_n \left[1 - \frac{|n|}{2N+1}\right], \quad |n| \leqq 2N+1 \tag{59}$$
$$= 0 \text{ otherwise.}$$

Thus, if $r(\omega) = \sum_{-k}^{k} r_n e^{in\omega}$,

$$\int_{-\pi}^{\pi} K_N(\omega,\omega')r(\omega)d\omega = r(\omega') - \frac{1}{2N+1}\sum_{-k}^{k} r_n|n|e^{in\omega'} \tag{60}$$

if only $(2N+1) \geqq k$. Substituting (60) into the right member of (56), we obtain

$$\sum_{-k}^{k} r_n|n|\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{in\omega'}}{r(\omega')}d\omega' = \sum_{-k}^{k} r_n|n|g_n^*. \tag{61}$$

Noting that we have already established that the left-hand side of (56) is strictly positive [$r(\omega) \neq$ const], the theorem follows.

V. EXAMPLE AND FURTHER COMMENTS

A particular example is provided by choosing

$$r(\omega) = 1 + a\cos\omega, \quad |a| < 1, \tag{62}$$

and, thus, $k = 1$. We calculate

$$A = D = \frac{1}{2}\left[1 + \frac{1}{\sqrt{1-a^2}}\right]$$
$$B(N) = C(N) = \frac{\rho^{2N-1}}{\sqrt{1-a^2}}\left[\frac{a}{2} + \rho\right] > 0, \quad N > 0, \tag{63}$$
$$\rho = \frac{-1+\sqrt{1-a^2}}{a}.$$

When $N = 0$, we have, from (49),

$$\lambda_{\max}(N=0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\omega)d\omega \times \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{r(\omega)} = \frac{1}{\sqrt{1-a^2}},$$

and otherwise

$$\lambda_{\max}(N) = A + B(N). \tag{64}$$

From (29) it follows that

$$\sin^2\Omega_N \geqq \frac{r_{\min}}{r_{\max}}, \tag{65}$$

where $r_{\min}$ and $r_{\max}$ denote the minimum and maximum of $r(\omega)$, and it is interesting to compare numerically (64) with (65). Set $a = 0.6$, so that $r_{\min}/r_{\max} = 0.25$. Then $\Omega_N \geqq 30$ degrees from (65). On the other hand $\lambda_{\max}(N) = 1.125 + 0.125/9^N$, $N \geqq 0$, which means Ω_N starts at about 63 degrees and increases to 70 degrees. Equivalently, while (65) allows a factor of four between the upper and lower bounds of (4) for this example, the more exact evaluation has them differing by only 12 to 25 percent depending on N.†

Exact solutions, as we have just found, may be useful for estimating Ω_N for some particular $r(\omega)$. If we already know $\tilde{\Omega}_N$ for some other $\tilde{r}(\omega)$ and if it is true that there are constants $\mu,\mu' \geqq 0$ so that

$$\frac{\tilde{r}(\omega)}{1+\mu} \leqq r(\omega) \leqq (1+\mu')\tilde{r}(\omega), \tag{66}$$

then, in a similar way to which (65) was derived, we can show that

$$\frac{\sin^2\tilde{\Omega}_N}{(1+\mu)(1+\mu')} \leqq \sin^2\Omega_N \leqq (1+\mu)\,(1+\mu')\sin^2\tilde{\Omega}_N. \tag{67}$$

Equation (67) could be useful when $r(\omega)$ has a large or infinite number of Fourier coefficients.

Proceeding further along the direction of bounds, we note that, using only $r_{\min}$ and $r_{\max}$, (65) can be considerably sharpened. One can show, in fact, letting $E = r_{\min}/r_{\max}$, that

$$\sin^2\Omega_N \geqq \left[\frac{1}{2} + \frac{1}{4}\left(E + \frac{1}{E}\right)\right]^{-1}. \tag{68}$$

The two basic ingredients are that [see (32) through (34)]

$$\lambda_{\max}(N) = \max_{\psi} \frac{\psi^+ R\psi}{\psi^+ \Gamma^{-1}\psi} \tag{69}$$

and

$$\frac{\|\psi\|^4}{\psi^+ A\psi} \leqq \psi^+ A^{-1}\psi \tag{70}$$

for any positive definite Hermitian matrix.[4] Combining (69) and (70) yields

$$\lambda_{\max}(N) \leqq \sup_{\psi} \frac{(\psi^+ R\psi)(\psi^+ \Gamma\psi)}{\|\psi\|^4}. \tag{71}$$

The right member of (71) is further upper bounded by

$$\max_{u(\omega)} \frac{1}{2\pi} \int_{-\pi}^{\pi} |u(\omega)|^2 r(\omega) d\omega \times \frac{1}{2\pi} \int_{-\pi}^{\pi} |u(\omega)|^2 \frac{1}{r(\omega)} d\omega, \tag{72}$$

† Another bound involving only the ratio $r_{\min}/r_{\max}$ is given in (68). For the present example it yields $\Omega_N \geqq 53$ degrees, a considerable improvement over (65).

where

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|u(\omega)|^2 d\omega = 1, \tag{73}$$

$u(\omega)$ being any L_2 function, not just one having Fourier components u_n restricted to $|n| \leqq N$.

To maximize (72), consider maximizing

$$Q = (\sum p_i a_i)\left(\sum p_i \frac{1}{a_i}\right) \tag{74}$$

with $\sum p_i = 1$, $p_i \geqq 0$, $a_i > 0$ and distinct. Introducing a Lagrange multiplier λ and differentiating, we obtain

$$a_l \sum \frac{p_j}{a_j} + \frac{1}{a_l}\sum p_j a_j = \lambda \tag{75}$$

for all nonvanishing p_l. Whatever values the optimum p_i's take, we may regard the sums in (75) as fixed numbers, independent of the index l. The resulting quadratic equation in a_l can be satisfied by at most two values of a_l and, hence, only two p_l are nonvanishing, and are easily seen to correspond to the maximum and minimum a_l if we are to maximize Q. Also the two p_l have equal values. Thus, (for a maximum Q)

$$Q_{\max} = \frac{1}{4}(a_{\max} + a_{\min})\left(\frac{1}{a_{\max}} + \frac{1}{a_{\min}}\right) \tag{76}$$

and (68) follows.

Our next theorem says something about the limiting behavior of $\lambda_{\max}(N)$, and in fact bounds the latter away from unity in the general case.

Theorem V. If $r(\omega) \neq$ *const.*

$$\lim_{N\to\infty} \lambda_{\max}(N) \geqq \frac{1}{2}\left[1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\omega)d\omega \times \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{r(\omega)}\right] > 1. \tag{77}$$

This immediately removes the restriction in Theorem III. Also, the left side of (56) is now, in general, bounded away from zero, and, by simple limiting procedures, Theorem IV also follows without restricting $r(\omega)$ as was previously required. Of course, "$-\infty$" is included in the statement "<0."

Proof. We begin with a modified form of (69) (let $\psi = \Gamma\phi$) which states

$$\lambda_{\max}(N) = \max_{\phi} \frac{\phi^+ \Gamma R \Gamma \phi}{\phi^+ \Gamma \phi}. \tag{78}$$

Thus, any particular choice for ϕ provides a lower bound. We choose for

the components ϕ_k of ϕ, $|k| \leqq N$, $\phi_k = \delta_{-N,k}$. Inserting this choice into (78) yields

$$\lambda_{\max}(N) \geqq \frac{(\Gamma R \Gamma)_{-N,-N}}{\Gamma_{-N,-N}} = \frac{1}{g_0} \sum_{n,m=-N}^{N} g_{n+N} g_{-N-m} r_{m-n}$$

$$= \frac{1}{g_0} \sum_{s,t=0}^{2N} g_s^* g_t r_{s-t} = \frac{1}{2g_0} \left[\sum_{s,t=-2N}^{2N} g_s^* g_t r_{s-t} + g_0^2 r_0 \right]. \tag{79}$$

Since† $\lim \sum_{-N}^{N} g_s^* g_t r_{s-t} = g_0$, the theorem follows.

REFERENCES

1. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—II," B.S.T.J., *40*, No. 1 (January 1961), pp. 65–84.
2. F. Riesz and B. Sz.-Nagy, *Functional Analysis*, New York: Ungar, 1955, pp. 208–210.
3. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, New York: Interscience, 1953, pp. 102–103.
4. E. F. Beckenbach and R. Bellman, *Inequalities*, New York: Springer-Verlag, 1965, p. 69.

†Namely, $\sum g_e r_{s-t}$ becomes, in the limit, the sth component of $1/r(\omega) \times r(\omega)$ which is, of course, δ_{s0}.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Rate vs Fidelity for the Binary Source

By S. P. LLOYD

(Manuscript received May 5, 1976)

Errors are deliberately introduced in the output of a binary message source to reduce the entropy rate. The errors depend on the source sequence in a deterministic shift-invariant manner. The tradeoff between error rate permitted and reduction of entropy rate is of interest. It is shown that the ideal bound cannot be attained. If the errors are required to be produced causally, then a bound stronger than the ideal bound takes over. The quantities of interest are found explicitly for the example: change all 0's in 0-runs of length 1 to 1's.

If a transmission channel has capacity C bits/second and a message source has entropy rate H bits/second satisfying $H \leqq C$, then the source can be encoded, fed to the channel, decoded at the channel output, and recovered essentially without error after such handling. The rate-distortion theory is concerned with the case where $H > C$; we try to minimize some measure of the errors that are necessarily present.[1]

We treat here a special class of systems in which the errors are deliberately introduced before submission to the channel to reduce the entropy rate to that of the channel; the mutilated source is then handled without further error by the channel. The usual treatment involves use of block codes, but we will examine the more interesting sliding (or shift-invariant) codes.

The source in Fig. 1 emits letters x_n, $-\infty < n < \infty$, at rate 1 per unit time. The letters are drawn from alphabet $A = \{0,1\}$ with probability distribution $P\{x_n = 0\} = P\{x_n = 1\} = 1/2$, the same for all n, and the draws are statistically independent. We denote by $x = (x_n: -\infty < n < \infty)$ a sample sequence of the source process X. The entropy rate of the source is $H(X) = 1$ bit per unit time.

The error generator operates on a source sequence x to produce a sequence $e = (e_n: -\infty < n < \infty)$ of A valued random variables $e_n = e_n(x)$. The error at time n is a deterministic function $e_n = \eta(\cdots, x_{n-1}; x_n, x_{n+1}, \cdots)$ of the whole sample sequence x. The measurable function η is the same for all n, so that the dependence of e on x is shift

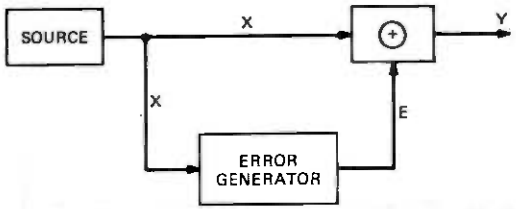


Fig. 1—Reducing the entropy rate by introducing errors.

invariant: if sequence x is shifted m places, the sequence $e = e(x)$ shifts m places with it.

The output of the adder box "$\oplus$" in Fig. 1 is simply $y_n = x_n \oplus e_n$, with $\oplus$ the usual addition mod 2. We regard the output process Y as X corrupted by the errors E. Now, depending on how E is generated, process Y can have entropy rate $H(Y) < H(X)$, and so can be handled by a channel of correspondingly smaller capacity at the price of the errors introduced. We are concerned with the tradeoff between the error rate and the decrease in entropy rate. Explicitly, suppose error rate ϵ is specified, $0 \leqq \epsilon \leqq 1/2$, and that $\eta(\cdots;;\cdots)$ is a stationary error-generating function with the property $P\{e_n = 1\} = \epsilon$. The resulting Y process will have a certain entropy rate $H(Y) \leqq H(X)$ determined by η. What is the least value that $H(Y)$ can have for all such η?

I. THE IDEAL BOUND

Let us consider the joint process $Z = (Y,E)$, where the Z alphabet is $\{(0,0),(0,1),(1,0),(1,1)\}$ and each z_n in a sequence $z = (z_n: -\infty < n < \infty)$ is the pair $z_n = (y_n,e_n)$. The mapping Θ: $X \to Z$, which sends a sample sequence x to sequence $z = \Theta x$, is obviously shift invariant. The map Θ is also measure preserving by definition; the probability measure on the space of sequences z is that induced by Θ and the X distribution. In the other direction, $x_n = y_n \oplus e_n$, $-\infty < n < \infty$ is the inverse map Φ: $Z \to X$, which recovers the source sequence x if the compressed version y and the errors e are known. This map is also shift invariant and measure preserving. Since processes X and Z are isomorphic in the above sense, their entropy rates are the same: $H(Z) = H(X) = 1$.

From the general theory (Section 6 in Ref. 3), the entropy rate $H(Z)$ is the average conditional entropy

$$\begin{aligned} H(Z) &= H(z_1|\cdots,z_0) \\ &= H[(y_1,e_1)|\cdots,(y_0,e_0)] \end{aligned}$$

of letter z_1, given the preceding letters $\cdots,z_0$. Using the addition law for conditional entropy, we find

$$H(Z) = H[e_1|\cdots,(y_0,e_0)] + H[y_1|\cdots,(y_0,e_0) \text{ and } e_1]$$
$$\leqq H(e_1) + H(y_1|\cdots,y_0)$$
$$= h(\epsilon) + H(Y),$$

since $H(e_1) = h(\epsilon) \triangleq \epsilon \log_2(1/\epsilon) + (1-\epsilon)\log_2[1/(1-\epsilon)]$ when $P\{e_1 = 1\} = \epsilon$, $P\{e_1 = 0\} = 1 - \epsilon$. Using $H(Z) = 1$, we have the lower bound $H(Y) \geqq 1 - h(\epsilon)$, $0 \leqq \epsilon \leqq 1/2$, for any such compression scheme.

Our first result is:

Theorem 1: *For error rate* $0 < \epsilon < 1/2$ *it is not possible to attain the bound* $H(Y) = 1 - h(\epsilon)$.

Proof: For each fixed $N \geqq 1$, there holds $NH(Z) = H(z_1,\cdots,z_N|\cdots,z_0)$, by induction from

$$H(z_1,\cdots,z_N|\cdots,z_0) = H(z_1|\cdots,z_0) + H(z_2,\cdots,z_N|\cdots,z_1).$$

Arguing as before, we find

$$N = H[(y_1,e_1),\cdots,(y_N,e_N)|\cdots,(y_0,e_0)]$$
$$= H[y_1,\cdots,y_N|\cdots,(y_0,e_0)]$$
$$+ H[e_1,\cdots,e_N|\cdots,(y_0,e_0) \text{ and } y_1,\cdots,y_N];$$

moreover,

$$H[y_1,\cdots,y_N|\cdots,(y_0,e_0)] \leqq H(y_1,\cdots,y_N|\cdots,y_0)$$
$$= NH(Y)$$

and

$$H(e_1,\cdots,e_N|\cdots,e_0 \text{ and } \cdots,y_N) \leqq NH(e_1) = Nh(\epsilon).$$

Now, equality in this last step holds *iff* $e_1,\cdots,e_N, f(\cdots,e_0 \text{ and } \cdots,y_N)$ are mutually independent, f is any measurable function of the variables indicated. (Equality in the first step requires that $y_1,\cdots,y_N$ be conditionally independent of $\cdots,e_0$ given $\cdots,y_0$, but we will not need this.)

For real valued variable u, let us define

$$u^{(\alpha)} = \begin{cases} u & \text{if } \alpha = 0, \\ 1-u & \text{if } \alpha = 1; \end{cases}$$

we put also $u^{(\alpha)(\beta)} = [u^{(\alpha)}]^{(\beta)} = [u^{(\beta)}]^{(\alpha)}$ for all $\alpha,\beta \epsilon A$; note that $u^{(0)(0)} = u^{(1)(1)} = u$, $u^{(0)(1)} = u^{(1)(0)} = 1 - u$. From

$$x_j = y_j \oplus e_j$$
$$= y_j e_j + (1-y_j)(1-e_j)$$
$$= y_j^{(0)} e_j^{(0)(0)} + y_j^{(1)} e_j^{(1)(0)}$$

and

$$\begin{aligned} 1 - x_j &= y_j \oplus (e_j \oplus 1) \\ &= y_j(1 - e_j) + (1 - y_j)e_j \\ &= y_j^{(0)} e_j^{(0)(1)} + y_j^{(1)} e_j^{(1)(1)}, \end{aligned}$$

it is apparent that

$$x_j^{(\alpha)} = \sum_{\beta} y_j^{(\beta)} e_j^{(\beta)(\alpha)},$$

where α,β are variables in the set A. Multiplying these equations together for $1 \leqq j \leqq N$ gives

$$x_1^{(\alpha_1)} \cdots x_N^{(\alpha_N)} = \sum_{\beta_1} \cdots \sum_{\beta_N} y_1^{(\beta_1)} \cdots y_N^{(\beta_N)} e_1^{(\alpha_1)(\beta_1)} \cdots e_N^{(\alpha_N)(\beta_N)},$$

for each of the 2^N choices for $\alpha_1, \cdots, \alpha_N$.

If $H(Y) = 1 - h(\epsilon)$, then $e_1, \cdots, e_N, [y_1^{(\beta_1)} \cdots y_N^{(\beta_N)}]$ are mutually independent for each choice of the β's. Since $E\{e_j^{(\gamma)}\} = \epsilon^{(\gamma)}$, $1 \leqq j \leqq N$, we find

$$\begin{aligned} \frac{1}{2^N} &= E\{x_1^{(\alpha_1)} \cdots x_N^{(\alpha_N)}\} \\ &= \sum_{\beta_1} \cdots \sum_{\beta_N} \epsilon^{(\alpha_1)(\beta_1)} \cdots \epsilon^{(\alpha_N)(\beta_N)} E\{y_1^{(\beta_1)} \cdots y_N^{(\beta_N)}\}, \text{ all } \alpha\text{'s}. \end{aligned}$$

Using now the assumption $\epsilon \neq 1/2$, let c be the number $c = -\epsilon/(1 - 2\epsilon)$, so that $c^{(1)} = (1 - \epsilon)/(1 - 2\epsilon)$. From

$$\sum_{\alpha} c^{(\alpha)(\gamma)} = 1, \quad \sum_{\alpha} c^{(\alpha)(\gamma)} \epsilon^{(\alpha)(\beta)} = \delta_{\gamma,\beta},$$

we obtain

$$\begin{aligned} \frac{1}{2^N} &= \sum_{\alpha_1} \cdots \sum_{\alpha_N} c^{(\alpha_1)(\gamma_1)} \cdots c^{(\alpha_N)(\gamma_N)} \times \frac{1}{2^N} \\ &= \sum_{\alpha_1} \cdots \sum_{\alpha_N} \sum_{\beta_1} \cdots \sum_{\beta_N} c^{(\alpha_1)(\gamma_1)} \epsilon^{(\alpha_1)(\beta_1)} \cdots c^{(\alpha_N)(\gamma_N)} \epsilon^{(\alpha_N)(\beta_N)} \\ &\qquad \times E\{y_1^{(\beta_1)} \cdots y_N^{(\beta_N)}\} \\ &= E\{y_1^{(\gamma_1)} \cdots y_N^{(\gamma_N)}\}, \quad \text{all } \gamma\text{'s}. \end{aligned}$$

If this holds for all $N \geqq 1$, then the $\{y_n: -\infty < n < \infty\}$ are independent identically distributed random variables with distribution $P\{y_n = 0\} = P\{y_n = 1\} = 1/2$. The entropy rate of this process is $H(Y) = 1 \neq 1 - h(\epsilon)$. □

II. THE CAUSAL BOUND

We now consider the case where each e_n depends only on the present and past values of the x's. That is, $e_n = \eta(\cdots,x_{n-1};x_n)$, $-\infty < n < \infty$, for η a measurable function of the variables indicated. The relation between Z and X is thus bicausal: z_n depends only on $\cdots,x_n$ and x_n depends only on $\cdots,z_n$. It follows that conditionals given $\cdots,z_n$ agree w.p.1 with conditionals given $\cdots,x_n$.

Theorem 2: If the dependence of the error process E on X is causal, then $H(Y) \geqq 1 - 2\epsilon$, $0 \leqq \epsilon \leqq 1/2$.

Proof: Setting A variant form of the basic inequality is $H(Y) \geqq 1 - H(e_0|\cdots,x_{-1})$, obtained from

$$\begin{aligned} 1 = H(Z) &= H[(y_0,e_0)|\cdots,(y_{-1},e_{-1})] \\ &= H[e_0|\cdots(y_{-1},e_{-1})] + H[y_0|\cdots,(y_{-1},e_{-1}) \text{ and } e_0] \\ &\leqq H(e_0|\cdots,x_{-1}) + H(y_0|\cdots,y_{-1}); \end{aligned}$$

we have used only that $y_n \oplus e_n$ is less informative than (y_n,e_n), $-\infty < n \leqq -1$. The assumption that η is causal is not involved.

Let us partition the space of sample sequences x into the four disjoint subsets:

$$\begin{aligned} A_1 &= \{x\colon \eta(\cdots,x_{-1};0;x_1,\cdots) = 0 \text{ and } \eta(\cdots,x_{-1};1;x_1,\cdots) = 0\} \\ A_2 &= \{x\colon \eta(\cdots,x_{-1};0;x_1,\cdots) = 1 \text{ and } \eta(\cdots,x_{-1};1;x_1,\cdots) = 1\} \\ A_3 &= \{x\colon \eta(\cdots,x_{-1};0;x_1,\cdots) = 0 \text{ and } \eta(\cdots,x_{-1};1;x_1,\cdots) = 1\} \\ A_4 &= \{x\colon \eta(\cdots,x_{-1};0;x_1,\cdots) = 1 \text{ and } \eta(\cdots,x_{-1};1;x_1,\cdots) = 0\}. \end{aligned}$$

The random variable $\kappa(x)$ is defined as the part namer for this partition; i.e., $\kappa(x) = j$ *iff* $x \in A_j$, $1 \leqq j \leqq 4$. Since $\kappa(x)$ depends only on coordinates $\cdots,x_{-1}$ and $x_1,\cdots$ of x, the conditional distribution of x_0 given κ is $P\{x_0 = 0|\kappa\} = P\{x_0 = 1|\kappa\} = 1/2$ w.p.1. The resulting random conditional entropies of e_0,y_0 are seen to be

$$\begin{aligned} h(e_0|\cdots,x_{-1} \text{ and } x_1,\cdots) &= h(e_0|\kappa) \\ &= \begin{cases} 0 \text{ for } \kappa = 1,2 \\ 1 \text{ for } \kappa = 3,4 \end{cases} \\ h(y_0|\cdots,x_{-1} \text{ and } x_1,\cdots) &= h(y_0|\kappa) \\ &= \begin{cases} 1 \text{ for } \kappa = 1,2 \\ 0 \text{ for } \kappa = 3,4. \end{cases} \end{aligned}$$

Putting $a_i = P\{A_i\}$, $1 \leqq i \leqq 4$, the average conditional entropies are then

$$H(e_0|\cdots,x_{-1} \text{ and } x_1,\cdots) = H(e_0|\kappa) = a_3 + a_4$$

$$H(y_0|\cdots,x_{-1} \text{ and } x_1,\cdots) = H(y_0|\kappa) = a_1 + a_2.$$

The error rate is

$$\epsilon = P\{e_0 = 1\} = \frac{1}{2}(a_3 + a_4) + a_2,$$

so we have

$$H(e_0|\cdots,x_{-1} \text{ and } x_1,\cdots) = 2\epsilon - 2a_2 \leqq 2\epsilon.$$

Assume now that E depends causally on X; then, e_0 is conditionally independent of $x_1,\cdots$ given $\cdots,x_{-1}$, implying $H(e_0|\cdots,x_{-1} \text{ and } x_1,\cdots) = H(e_0|\cdots,x_{-1})$. Combining the inequalities, we obtain $H(Y) \geqq 1 - 2\epsilon$. □ This bound is strictly above the ideal bound when $0 < \epsilon < 1/2$, since $h(\epsilon) > 2\epsilon$ on this interval.

III. EXAMPLE

The following example is mentioned in Ref. 4, but a solution is not given. Let the errors be $e_n = \eta(x_{n-1};x_n;x_{n+1})$, $-\infty < n < \infty$, with η the function

$$\eta(1;0;1) = 1$$

$$\eta(x_{-1};x_0;x_1) = 0 \text{ if } x_{-1}x_0x_1 \neq 101.$$

The error rate is $P\{e_n = 1\} = 1/8$. We will compute $H(Y)$ and $H(E)$ explicitly and compare $H(Y)$ with the bounds of Sections I and II.

A graphical representation of η is given in Fig. 2. The vertices of the directed graph are the state pairs $x_{-1}x_0$, the arrows represent the transitions from $x_{-1}x_0$ to x_0x_1, and the value $\eta(x_{-1};x_0;x_1)$ is shown on the arrow from $x_{-1}x_0$ to x_0x_1. The corresponding graph of $y_0 = x_0 \oplus \eta(x_{-1};x_0;x_1)$ appears in Fig. 3.

We now compute $H(Y)$. Examination of Fig. 3 reveals that process Y is a renewal process, with renewal at the beginning of each run, either

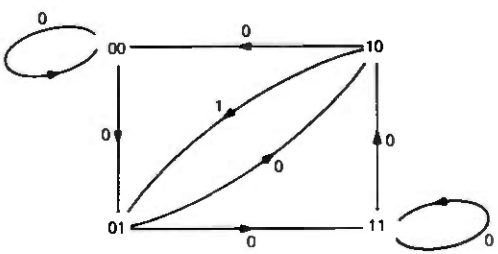


Fig. 2—$e_0 = \eta(x_{-1};x_0;x_1)$. Values of e_0 as function of the transition.

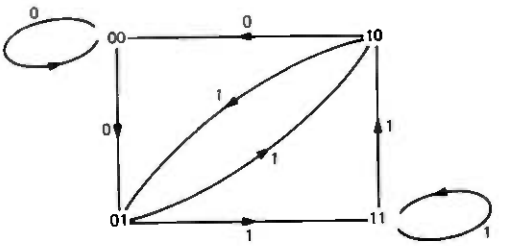


Fig. 3—$y_0 = x_0 \oplus \eta(x_{-1},x_0,x_1)$. Values of y_0 as function of the transition.

of 0's or of 1's. Moreover, the length $R^{(0)}$ of a 0-run has the geometric distribution

$$P\{R^{(0)} = j\} = \frac{1}{2^{j-1}}, \quad j = 2,3,\cdots.$$

The mean and entropy of this distribution are easily found to be $E\{R^{(0)}\} = 3, H(R^{(0)}) = 2$.

The 1-runs of Y involve the subgraph shown in Fig. 4, relabelled for convenience. A 1-run results from a path (driven by x) which starts at A and follows lettered arrows until exit occurs at B along the dotted arrow. If the length $R^{(1)}$ of the run has value $R^{(1)} = j$, the driving x's have probability $2^{-(j+1)}$ per path, so

$$P\{R^{(1)} = j\} = \frac{\nu_j}{2^{j+1}}, \quad j = 1,2,\cdots,$$

where ν_j is the number of paths of length j from A to B along lettered arrows.

For $j \geq 4$, we classify the paths of length j from A to B according to the earliest appearance of arrow a:

(i) One path $c(d)_{j-2}e$ which does not contain a.

(ii) Paths which start $ba\cdots$.

(iii) For each $0 \leq k \leq j-4$, paths which start $c(d)_k ea\cdots$.

In (ii) the continuations "$\cdots$" are just the paths from A to B of length

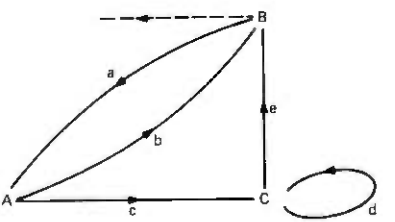


Fig. 4—Subgraph for 1-runs of process Y.

$j-2$, one each, and in (*iii*) the lengths of the continuations are $j-(k+3)$ for each $0 \leqq k \leqq j-4$. In consequence,

$$\nu_j = 1 + \nu_{j-2} + \sum_{k=0}^{j-4} \nu_{j-(k+3)}, \quad j \geqq 4.$$

The initial terms for the recursion are $\nu_1 = 1$, $\nu_2 = 1$, $\nu_3 = 2$, clearly, and it is convenient to define $\nu_0 = 0$. Then from

$$\begin{aligned} \nu_j &= 1 + \nu_{j-2} + \nu_{j-3} + \cdots + \nu_0, \quad j \geqq 2, \\ \nu_{j-1} &= 1 + \nu_{j-3} + \cdots + \nu_0, \qquad\quad j \geqq 3, \end{aligned}$$

it is apparent that

$$\nu_j = \nu_{j-1} + \nu_{j-2}, \quad j \geqq 3.$$

That is, $\nu_1, \nu_2, \cdots$ is the Fibonacci sequence $1,1,2,3,5,8, \cdots$.

The generating function for the Fibonacci sequence is $\Sigma_1^\infty \nu_j x^j = x/(1-x-x^2)$, as is well known, so the distribution of $R^{(1)}$ has generating function

$$\sum_{j=1}^{\infty} x^j P\{R^{(1)} = j\} = \frac{x}{4-2x-x^2}, \quad |x| \leqq 1 < \sqrt{5}-1.$$

Taking $(d/dx)_{x=1}$, we obtain $E\{R^{(1)}\} = 5$ for the mean length of 1-runs in the Y process. For numerical evaluation of the entropy $H(R^{(1)})$ of the $R^{(1)}$ distribution, we obtain the $r_j = P\{R^{(1)} = j\}$, $j \geqq 1$, from the recursion

$$r_j = \frac{1}{2} r_{j-1} + \frac{1}{4} r_{j-2}, \quad j \geqq 3;$$

$$r_1 = \frac{1}{4}, \quad r_2 = \frac{1}{8}.$$

The numerical result is

$$\begin{aligned} H(R^{(1)}) &= \sum_{j=1}^{\infty} r_j \log_2 \frac{1}{r_j} \\ &= 3.593946 \text{ bits per run.} \end{aligned}$$

Starting at the beginning of a run, suppose y is truncated after M runs of both kinds have occurred. The total number of coordinates y_n is the sum $\Sigma_1^M [R_m^{(0)} + R_m^{(1)}]$ of M independent samples each of $R^{(0)}, R^{(1)}$. The total random entropy is the corresponding sum $\Sigma_1^M [h(R_m^{(0)}) + h(R_m^{(1)})]$ for the samples. Omitting the detailed arguments, we obtain from the strong law of large numbers

$$H(Y) = \lim_{M\to\infty} \frac{\sum_1^M [h(R_m^{(0)}) + h(R_m^{(1)})]}{\sum_1^M [R_m^{(0)} + R_m^{(1)}]} \quad \text{w.p.1}$$

$$= \frac{H(R^{(0)}) + H(R^{(1)})}{E\{R^{(0)}\} + E\{R^{(1)}\}}$$

$$= 0.699243 \text{ bit per letter.}$$

As a check, note

$$P\{y_0 = 1\} = \frac{E\{R^{(1)}\}}{E\{R^{(0)}\} + E\{R^{(1)}\}} = \frac{5}{8},$$

which is clear from Fig. 3. The entropy of y_0 is $H(y_0) = h(3/8) = 0.954434$, and the difference

$$h(3/8) - H(Y) = H(y_0) - H(y_0|\cdots,y_{-1})$$

$$= I(y_0,\{\cdots,y_{-1}\})$$

$$= 0.255191 \text{ bit per letter}$$

is the amount by which Y fails to be a Bernoulli process. The ideal bound is

$$H(Y) \geqq 1 - h(1/8)$$

$$= 0.456436 \text{ bit per letter.}$$

The bound of Section II is easily worked out to be

$$H(Y) \geqq 1 - H(e_0|\cdots,x_{-1})$$

$$= 1 - (1/2)h(1/4)$$

$$= 0.594361 \text{ bit per letter.}$$

The entropy rate $H(E)$ of the errors can also be obtained from run-length considerations. Indeed, $\{e_n = 1\}$ is just the event $\{x_n = 0$ is a 0-run of length 1$\}$ in process X. The 0-run lengths $S^{(0)}$ and the 1-run lengths $S^{(1)}$ in process X each have the geometric distribution

$$P\{S^{(0)} = j\} = P\{S^{(1)} = j\} = \frac{1}{2^j}, \quad j = 1,2,\cdots,$$

as is well known. Let the run lengths after an occurrence of $\{S^{(0)} = 1\}$ be $S_1^{(1)}, S_1^{(0)}, S_2^{(1)}, S_2^{(0)}, \cdots$, and let random variable J be the smallest $\nu \geqq 1$ for which $S_\nu^{(0)} = 1$. Since $P\{S^{(0)} = 1\} = P\{S^{(0)} > 1\} = 1/2$, we again have (1/2, 1/2) Bernoulli trials, i.e.,

$$P\{J = j\} = \frac{1}{2^j}, \quad j = 1,2,\cdots.$$

The number of intervening x_n's is the 0-run length $V^{(0)} = S_1^{(1)} + S_1^{(0)} + \cdots + S_{j-1}^{(0)} + S_j^{(1)}$ in the E process. The generating function for each $S^{(1)}$ is

$$\sum_1^\infty x^j P\{S^{(1)} = j\} = \frac{x}{2-x},$$

and the generating function for $S^{(0)}$ conditional on $S^{(0)} > 1$ is

$$\sum_2^\infty x^j P\{S^{(0)} = j \mid S^{(0)} > 1\} = \frac{x^2}{2-x},$$

so we have

$$\sum_1^\infty x^k P\{V^{(0)} = k\} = \sum_{j=1}^\infty \frac{1}{2^j}\left(\frac{x}{2-x}\right)^j\left(\frac{x^2}{2-x}\right)^{j-1}$$

$$= \frac{x(2-x)}{8-8x+2x^2-x^3}, \quad |x| \leqq 1 < 1.13968.$$

Taking $(d/dx)_{x=1}$ gives $E\{V^{(0)}\} = 7$, and since the 1-runs in E have length $V^{(1)} = 1$ w.p.1, we have the check

$$P\{e_n = 1\} = \frac{E\{V^{(1)}\}}{E\{V^{(0)}\} + E\{V^{(1)}\}} = \frac{1}{8}.$$

For numerical evaluation of $H(V^{(0)})$, we use the recurrence

$$v_k = v_{k-1} - \frac{1}{4}v_{k-2} + \frac{1}{8}v_{k-3}, \quad k \geqq 4;$$

$$v_1 = \frac{1}{4}, \quad v_2 = \frac{1}{8}, \quad v_3 = \frac{1}{16}$$

satisfied by $v_k = P\{V^{(0)} = k\}$, $k \geqq 1$. The numerical result is

$$H(V^{(0)}) = \sum_1^\infty v_k \log_2 \frac{1}{v_k}$$

$$= 4.061168 \text{ bits per run},$$

giving

$$H(E) = \frac{H(V^{(0)}) + H(V^{(1)})}{E\{V^{(0)}\} + E\{V^{(1)}\}} = \frac{1}{8}H(V^{(0)})$$

$$= 0.507646 \text{ bit per letter}$$

as the entropy rate of the errors. The entropy of e_0 being $H(e_0) = h(1/8) = 0.543564$ bit per letter, the difference

$$\begin{aligned} h(\epsilon) - H(E) &= H(e_0) - H(e_0|\cdots,e_{-1}) \\ &= I(e_0,\{\cdots,e_{-1}\}) \\ &= 0.035918 \text{ bit per letter} \end{aligned}$$

is the amount by which E fails to be Bernoulli.

IV. ACKNOWLEDGMENTS

The author wishes to thank Aaron D. Wyner for bringing the problem to his attention. A paper by Berger and Lau[2] came to the author's attention after the present paper was written. Some of the results overlap; the methods are different.

REFERENCES

1. T. Berger, *Rate Distortion Theory*, Englewood Cliffs N.J.: Prentice-Hall, 1971.
2. T. Berger and J. K.-Y. Lau, "On Binary Sliding Block Codes," IEEE Trans. Inform. Theory, *IT-23*, No. 3 (May 1977).
3. P. Billingsley, *Ergodic Theory and Information*, New York: John Wiley, 1965.
4. R. M. Gray, D. L. Neuhoff, and D. S. Ornstein, *Nonblock Source Coding With a Fidelity Criterion*, Ann. Probability, *3* (1975) pp. 478–491.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Pitch-Adaptive DPCM Coding of Speech With Two-Bit Quantization and Fixed Spectrum Prediction

By N. S. JAYANT

(Manuscript received June 9, 1976)

This paper is concerned with the utilization of speech waveform periodicities in differential pulse code modulation (DPCM) coding with 2-bit adaptive quantization and time-invariant spectrum prediction. Our work is based on computer simulations of DPCM codes. We have studied pitch detectors based on autocorrelation and an average magnitude difference function (AMDF), and we have measured the benefits of predicting from a previous pitch period as functions of pitch-period-updating frequency and periodicity-indicating thresholds (for autocorrelation and the AMDF). We have compared several alternative methods of utilizing past quantized samples (in the present and previous pitch periods) for providing speech sample predictions. We find the following combination to be attractive for waveform coding at bit rates in the neighborhood of 16 kb/s: 2-bit adaptive quantization with a one-word (2-bit DPCM word) memory, pitch detection performed on unquantized speech (preferably with an AMDF criterion) and a prediction scheme that uses fixed three-tap (short-term) prediction for nonperiodic waveform segments, but switches to an appropriate one-tap (long-term) predictor upon the detection of strong periodicity. With four sample utterances, the latter procedure results in an average SNR (signal-to-noise ratio) gain of 3.75 dB over a non-pitch-adaptive encoder.

I. INTRODUCTION

An important subclass of speech waveform encoders is characterized by the use of adaptive quantization and predictive (DPCM) encoding.[1] Time-invariant spectrum predictors are simple to implement and robust in the context of coarse quantization. The benefits of adaptive prediction are, however, well recognized and documented,[2,3] and the greatest

achievements in bit-rate reduction have in fact depended on the use of adaptive short-term (spectrum) prediction as well as adaptive long-term (pitch) prediction, as seen in the paper by Atal and Schroeder.[4]

This paper is concerned with the relatively less documented combination of *adaptive pitch prediction and nonadaptive spectrum prediction*. The study of this kind of prediction is motivated by the observation that speech waveforms abound in highly periodic segments and by the conjecture that the use of this periodicity may provide a prediction potential that is substantial enough to obviate the need for adaptive short-term (spectrum) prediction. The attraction in this approach will evidently depend on the complexity of pitch detection itself. The pitch detectors used in this paper are based on autocorrelation and AMDF (average magnitude difference function) and are quite simple to implement; they are indeed much simpler than the mean-squared-error-minimizing pitch detector described in Ref. 4. Moreover, as discussed in Section IV, the success of pitch-adaptive DPCM does not depend critically on accurate pitch detection in the sense in which the term is used in formal speech research.[5]

A thesis by Trottier[6] considers the possibility of simplifying the Atal-Schroeder encoder.[4] Among other things, this thesis discusses simple pitch-detection algorithms, the criticality of a well-designed adaptive quantizer, and the inefficiency of approaches seeking to simplify adaptive spectrum prediction through the use of very few predictor taps, say two. An unpublished work of Grizmala[7] provides one of the first proposals for a simple pitch-adaptive DPCM that entirely avoids adaptive spectrum prediction. Grizmala discusses AMDF-based pitch detection and fixed three-tap spectrum prediction for nonperiodic waveform segments. More recently, Xydeas and Steele report an instance of a 6-dB SNR gain for a fixed-spectrum DPCM encoder arising from the utilization of waveform periodicities.[8] Finally the detection of periodicity based on autocorrelation and AMDF is documented in speech papers[5,9,10] as well as in coding literature.[11]

One of the contributions of the present paper is the demonstration that fixed-spectrum pitch-adaptive DPCM is useful in the context of a specific type of adaptive quantizer that has received considerable attention in recent coding work.[12,13] This paper also shows that AMDF-based pitch detection is slightly more effective than an autocorrelation-based procedure. The paper also demonstrates that, during periodic waveform segments, a simple one-tap predictor across the pitch period is more efficient than several multitap predictors involving many past samples in the present and previous pitch periods. Finally, the paper includes formal measurements of pitch prediction gain as a function of (*i*) pitch-period-update frequency, and of (*ii*) thresholds that the AMDF and correlation functions should exceed for a waveform segment to be

judged as periodic. Our results are all based on computer simulations of DPCM encoders.

The results of this paper are expected to be relevant to speech waveform coding at bit rates in the order of 16 kb/s. At this bit rate, the use of fixed spectrum prediction and adaptive quantization results typically in a quantization noise level that is quite easily perceived, while the sophistication of adaptive spectrum prediction is often unwarranted, because undesirable quantizer-predictor interactions begin showing up at around 16 kb/s in practical waveform coder designs.[14,15] Adaptive pitch prediction, on the other hand, appears to be a useful and robust sophistication at 16 kb/s. With this bit rate in mind, this paper will deal exclusively with two-bit quantizers for the DPCM coding of Nyquist-sampled (8-kHz) telephone-quality (200–3200 Hz) speech. Our numerical results refer to two female utterances, "The chairman cast three votes" and "The boy was mute about his task," and two male utterances "A lathe is a big tool," and "The boy was mute about his task." These utterances will henceforth be labeled F1, F2, M1, and M2.

The organization of the paper is as follows. Section II recommends a slowly adaptive quantizer with a one-word memory, and Section III proposes a three-tap spectrum predictor. Section IV discusses pitch detection by means of AMDF- and autocorrelation-type procedures, and points out how pitch analysis can be performed either on quantized speech or on the original unquantized speech. Section V compares different prediction algorithms for periodic segments, including the important example of an appropriate one-tap predictor. Section VI measures the gains of pitch-adaptive DPCM as a function of (*i*) the pitch-detection procedure, (*ii*) AMDF and autocorrelation thresholds used in hypothesizing periodicity, (*iii*) pitch-period-updating time, and (*iv*) prediction algorithms used for periodic waveform segments. Section VII summarizes performance figures for the four sample sentences and discusses results in the context of 16-kb/s waveform-coding.

II. TWO-BIT ADAPTIVE QUANTIZER

Figure 1 shows a uniform four-level quantizer used for pitch-adaptive DPCM coding. The step-size Δ is adaptive. The adaptations are based on a one-word memory.[12,13] Specifically, the step-size is modified at every sampling instant by a multiplier that depends only on whether the magnitude of the previous quantizer output was $0.5\Delta_r$ or $1.5\Delta_r$. Respective step-size multipliers make $\Delta_{r+1} = E_1 \cdot \Delta_r$ or $E_2 \cdot \Delta_r$. In the context of quantizing prediction errors across a pitch period, we have found that the most useful adaptations were 'slow' adaptations of the form:[12]

$$E_1 = 0.95; E_2 = 1.10. \tag{1}$$

As discussed at length in Ref. 12, values of optimal step-size multipliers

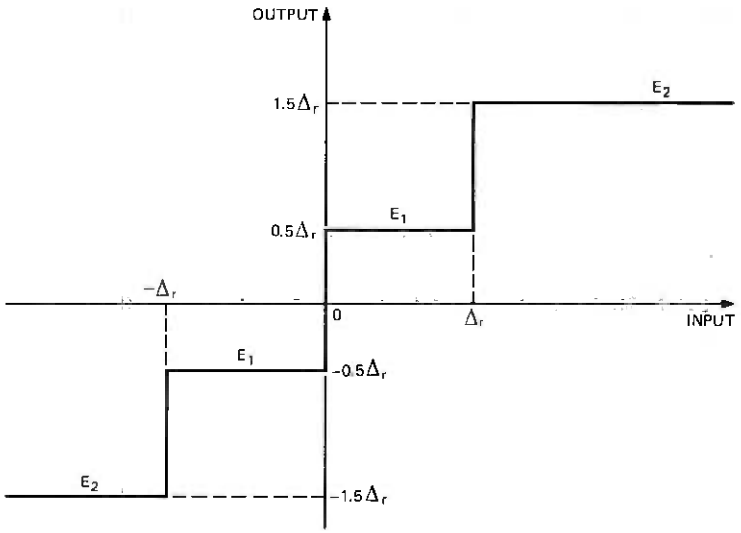


Fig. 1—A 2-bit adaptive quantizer.

reflect the nature of the input signal spectrum, and the stationarity of the input variance. The step-size adaptations were subject to maximum and minimum values that were appropriate for the given peak speech amplitude of ±1024:

$$\Delta_{MAX} = 192, \quad \Delta_{MIN} = 1.5. \tag{2}$$

Finally, nonuniform quantizers were not found to be very effective in pitch-adaptive DPCM using adaptive quantization. This had to do with the effect of DPCM predictions on the probability density function (PDF) at the quantizer input. The observation that nonuniform quantization is not very beneficial reflects the fact that predictions in DPCM cause a quantizer-input PDF that is more gaussian than the PDF of the original speech amplitudes. The latter, for example, can be modelled by a gamma-PDF for which nonuniform quantization is very useful.[2,3]

III. TIME-INVARIANT SPECTRUM PREDICTION

A T-tap spectrum predictor is represented by

$$X_r = \sum_{s=1}^{T} a_s \cdot XQ_{r-s}, \tag{3}$$

where X and XQ refer to input and quantized speech samples.

In time-invariant (fixed) prediction, the coefficients a are matched to the long-term spectrum of speech via the corresponding autocorrelation function, as described in Ref. 1.

Using a typical long-term spectrum characterization,[7] the following designs have been used for fixed one-tap and three-tap spectrum predictors:

$$a_1 = 0.85 \qquad \text{for } T = 1 \tag{4}$$

and

$$a_1 = 1.10; \quad a_2 = -0.28; \quad a_3 = -0.08 \qquad \text{for } T = 3. \tag{5}$$

These predictor coefficients are rounded values resulting from a spectrum model where the speech autocorrelations are 0.825, 0.562, and 0.308 for delays of one, two, and three 8-kHz samples, respectively. These autocorrelations are reported in Ref. 16 as the result of a study on a very large speech-sample base, and constitute slight revisions of very similar autocorrelations reported in Ref. 17.

In coding our speech waveforms, the three-tap predictor provided a typical SNR gain of nearly 1 dB over the one-tap predictor. Spectrum predictions in this paper will henceforth refer to a time-invariant three-tap design, as in eq. (5).

IV. MEASUREMENT OF PITCH PERIOD

This section defines the AMDF- and autocorrelation-based pitch measurements used in our work, discusses the use of unquantized speech samples X or quantized samples XQ for the pitch analysis, and provides illustrations of pitch measurements. In general, pitch analysis will be based on a window $\mathcal{W}$ containing W contiguous speech samples Z ($Z = X$ or XQ). The sampling instant when a pitch period is measured is denoted by r, so that a current speech sample will be Z_r (X_r or XQ_r, as appropriate). The pitch period is denoted by P, and P is assumed to have minimum and maximum values P_{MIN} and P_{MAX}, respectively. G_1 and G_2 are thresholds that can be used to hypothesize waveform periodicity with varying degrees of confidence. V is the pitch period updating time (see Section VI).

4.1 AMDF-based pitch measurement

Consider the average magnitude difference function

$$\text{AMDF}(p) = \text{AVERAGE}|Z_u - Z_{u-p}|;$$
$$p = P_{\text{MIN}}, P_{\text{MIN}} + 1, \cdots, P_{\text{MAX}}, \tag{6}$$

where the averaging is over all pairs $(u, u-p)$ such that both Z_u and Z_{u-p} are in $\mathcal{W}$.

The AMDF pitch detector estimates the pitch period P to be

$$P = p_{EST}$$

$$\text{if } \text{AMDF}(p_{EST}) < \text{AMDF}(p) \tag{7}$$

for all p in the range (P_{MIN}, P_{MAX}) with the exception of p_{EST}, and if

$$\text{AMDF}(p_{EST}) < G_1 \cdot \text{AVERAGE}(|Z_u|), \text{ for} \tag{8}$$

all u in $\mathcal{W}$.

The value of G_1 is discussed in detail in Section VI. Typically, $G_1 = 0.5$. With Nyquist-sampled (8-kHz) speech and for a single pitch-analysis procedure that should cover the expected range of p in both male and female speech, the following numbers seem appropriate:[5]

$$P_{MIN} = 16, \quad P_{MAX} = 160, \quad W = 256. \tag{9}$$

Notice that P_{MIN} excludes the obvious minimum AMDF (0) at $p = 0$, and that the window length W is well in excess of the maximum anticipated pitch period P_{MAX}. It turns out that this requirement ($W > P_{MAX}$) is quite important for efficient pitch prediction and waveform coding. The range of the pitch-period search ($16 < p < 160$) is wide enough to cause frequent problems with multiple peaks in the AMDF function, and multiples of the fundamental pitch period are often picked up as P. Fortunately, however, this kind of error in pitch tracking appears to be quite harmless as far as pitch-adaptive waveform codes are concerned: the need is for a sequence of waveform samples $\{XQ\}$ that provide good predictions of a current sequence $\{X\}$ in periodic segments, and it seems to be immaterial whether $\{X\}$ and $\{XQ\}$ are one pitch period apart or n (>1) pitch periods apart.

4.2 Autocorrelation-based pitch measurement

Consider the autocorrelation function

$$C(p) = \text{AVERAGE}(\text{sgn } Z_u \cdot \text{sgn } Z_{u-p});$$
$$p = P_{MIN}, P_{MIN} + 1, \cdots, P_{MAX}, \tag{10}$$

where the averaging is over all pairs $(u,u-p)$ such that both Z_u and Z_{u-p} are in $\mathcal{W}$ and, furthermore, both $|Z_u|$ and $|Z_{u-p}|$ exceed an appropriate speech-clipping level

$$Z_{CLIP} = 0.64 \text{ MAX}(|Z|^1_{MAX}, |Z|^3_{MAX}), \tag{11}$$

where $|Z|^1_{MAX}$ is the maximum speech magnitude in the first one-third part of $\mathcal{W}$ and $|Z|^3_{MAX}$ is the maximum speech magnitude in the third one-third part of $\mathcal{W}$.

The autocorrelation-pitch detector estimates the pitch period P to be

$$P = p_{EST} \tag{12}$$

if

$$C(p_{EST}) > C(p)$$

for all p in the range of (P_{MIN}, P_{MAX}) with the exception of p_{EST}, and if

$$C(p_{EST}) > G_2. \tag{13}$$

The role of G_2 is discussed at length in Section VI. Typically $G_2 = 0.2$. Appropriate values of P_{MIN}, P_{MAX}, and W follow (9). The nonzero value of P_{MIN} excludes the obvious maximum $C(0)$ at $p = 0$.

The center-clipping operation described by (11) is quite effective in mitigating spurious peaks in the $C(p)$ function, such as peaks representing a low first-formant frequency. Typically, autocorrelation pitch detectors work with speech that is low-pass filtered to, say, 900 Hz,[5] but such filtering was not used in our waveform coding program.

The pitch-measurement techniques based on (6) and (10)—especially the autocorrelation method (10)—are easier to implement than the mean-squared-error-minimizing pitch detector described in Ref. 4, which is based on computing the autocorrelation of Z [this involves computing products of real numbers, instead of taking differences as in (6) or using one-bit numbers as in (10)]. The efficacies of AMDF- and autocorrelation-based pitch detectors have recently been calibrated in terms of the performance of several other pitch-tracking procedures.[5]

4.3 Pitch analyses based on X and XQ

Figure 2a demonstrates pitch analysis based on original, unquantized speech samples X. We see how the analysis window can be aligned so as to extend equally on either side of the current sample X_r to be encoded

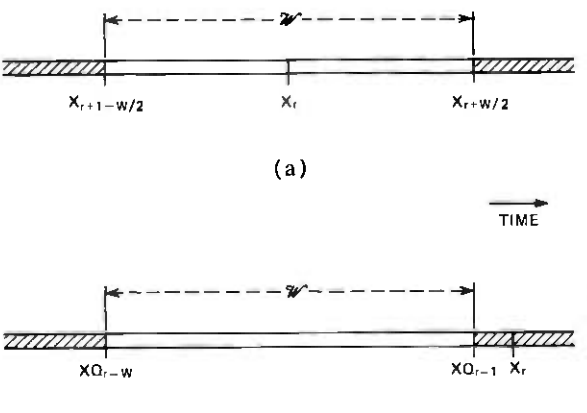


Fig. 2—Pitch analysis based on (a) unquantized speech X and (b) quantized speech XQ.

Table I — Local and global minima/maxima in pitch-period search ($G_1 = 0.84$, $G_2 = 0.2$; speech sample: M2, analysis based on unquantized speech)

p^*	Minimization of Normalized AMDF	Maximization of Autocorrelation
29	—	0.33
30	0.66	0.34
31	—	0.35
34	0.62	—
37	—	0.38
38	—	0.39
95	—	0.45
96	0.40	0.46

* Pitch-period estimate = 96 samples

(quantized); such alignment turns out to be quite critical for realizing the maximum potential of pitch-adaptive waveform codes.

Figure 2b shows the analysis of pitch based purely on past quantized samples XQ_{r-s} ($s > 0$). Figures 2a and 2b apply equally to AMDF or autocorrelation analysis.

4.4 Illustrative measurements of pitch

Table I demonstrates examples of AMDF- and autocorrelation-based searches for the pitch period P. Entries in the table represent those local minima/maxima in the AMDF/C functions, which were below/above all previous local minima/maxima in the search for P ($16 < p < 160$). Also, only those minima/maxima that cross the G_1/G_2 thresholds, eqs. (8) and (13), are listed. For both the AMDF and C functions, a global peak appears at the pitch period $P = 96$.

Table II provides a typical time plot of P (number of 8-kHz samples) for four different pitch-tracking techniques. The analysis refers to a sample segment from the utterance F1. Notice the remarkable closeness of X-based contours in columns 1 and 3. Notice also that with XQ-based analyses, the AMDF function tends to preserve pitch information much better than the autocorrelation measurement.

V. PREDICTION ALGORITHMS FOR PERIODIC WAVEFORMS

Figure 3 sketches a periodic waveform segment. P is the 'pitch period', X_r is a current waveform sample to be encoded, and XQ denotes an already quantized sample in the present 'pitch period' or in an earlier 'very similar segment' of the periodic waveform.

Our prediction algorithms for periodic waveforms are linear, and they are of the general form

$$\hat{X}_r = \sum_{u=1}^{3} a_u \cdot XQ_{r-u} + \sum_{v=0}^{3} a_{P+v} \cdot XQ_{r-P-v}. \tag{14}$$

Table II — Pitch-period contours from four pitch-tracking techniques (speech sample: F1). Entries along columns are successive values of P (number of 8-kHz samples)

AMDF of X	AMDF of XQ	Autocorrelation of X	Autocorrelation of XQ
2	2	2	19
39	39	2	19
78	39	78	19
39	39	39	19
39	39	39	39
39	39	39	38
39	39	39	39
39	2	39	41
43	2	43	44
40	2	40	35
41	42	41	25
132	132	132	2
134	134	134	2
135	134	135	2
57	135	57	2
78	80	78	48
157	157	157	50
35	35	2	19
2	2	2	19
2	2	2	19
2	2	2	2
2	2	2	18
2	2	2	2
2	2	2	18
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	31
35	2	35	34
35	2	35	34
35	35	35	35
36	35	36	36
36	35	36	36
36	36	36	36
37	36	37	37
37	37	37	37
37	37	37	37
37	37	37	37
37	37	37	37
37	37	37	37
37	37	75	37
75	75	37	37
37	75	37	37

We have considered many special cases of the general algorithm (14); Table III summarizes three interesting examples.

The seven-tap predictor attempts a clever combination of spectrum prediction [see (5) in Section III] and pitch prediction. This approach was proposed by Grizmala,[7] who in turn was simplifying a formal procedure of Atal and Schroeder.[4] The three-tap predictor is the simplest

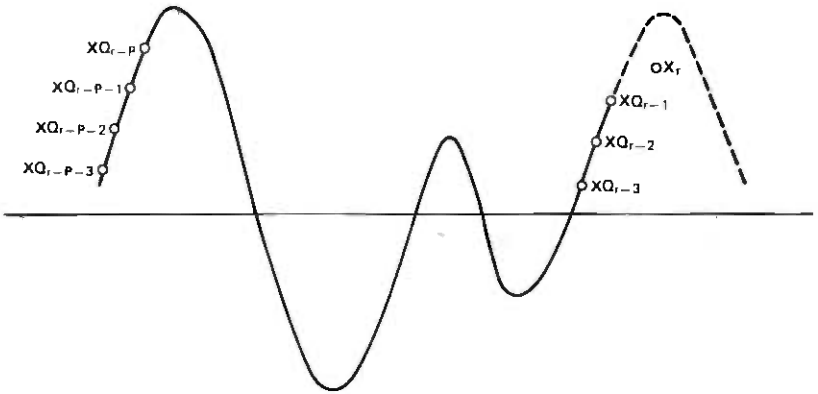


Fig. 3—Prediction algorithms for periodic waveforms.

nontrivial combination of the two types of prediction. It is suggested by a simple geometrical procedure of completing an idealized parallelogram with vertices at the topmost four dots in Fig. 3. Finally, the one-tap predictor is the simplest approach to pitch-adaptive coding and is suggested by the very strong correlations that are observed between X_r and X_{r-P} in highly periodic waveform segments.

VI. DESIGN AND PERFORMANCE OF PITCH-ADAPTIVE DPCM CODER

Figure 4 provides a block diagram of the pitch-adaptive DPCM coder. It is different from conventional DPCM[1] in the inclusion of a special predictor for encoding the periodic segments of the input waveform. The spectrum predictor is formally defined by (5) and the pitch predictor by (14). The switching between the two predictors is controlled by the crossings of appropriate thresholds G_1 and G_2 (Section IV) by the AMDF or autocorrelation functions, respectively. The test for periodicity is done once every V samples. If the waveform is decided to be "periodic" as a result of the test, the pitch period P (coming out of the AMDF or autocorrelation measurement) is used in the predictive encoding of a current block of V samples. (Both the binary "periodic/nonperiodic" decision and the pitch period, if any, are updated for the next block of V samples.)

6.1 SNR, SNRV, and SNRSEG

The design and utility of pitch-adaptive coders will be discussed using the following signal-to-noise ratio as a performance criterion

$$\text{SNR(dB)} = 10 \log_{10}\left[\sum_{r=1}^{N} X_r^2 \Big/ \sum_{r=1}^{N} (X_r - XQ_r)^2\right], \qquad (15)$$

Table III — Three prediction algorithms for periodic waveforms

Name of Predictor	a_1	a_2	a_3	a_P	a_{P+1}	a_{P+2}	a_{P+3}
AVERAGER	0.5	0	0	0.5	0	0	0
"7-Tap"	1.1	−0.28	−0.08	1	−1.1	0.28	0.08
"3-Tap"	1	0	0	1	−1	0	0
"1-Tap"	0	0	0	1	0	0	0

where N is the total number of input samples.

In deference to the fact that the pitch-adaptive coding is performed in blocks of V samples, we consider an additional measure of performance for the Sth block

$$\text{SNRV}(S)(\text{dB}) = 10 \log_{10} \left[\sum_{r=V(S-1)+1}^{V \cdot S} X_r^2 \Big/ \sum_{r=V(S-1)+1}^{V \cdot S} (X_r - XQ_r)^2 \right]. \quad (16)$$

The average value of SNRV over the total input signal duration (over N/V input blocks) will be called the 'segment-signal-to-noise ratio' SNRSEG (Ref. 18)

$$\text{SNRSEG} = \frac{1}{N/V} \sum_{S=1}^{N/V} \text{SNRV}(S). \quad (17)$$

SNRV is an obvious indicator of local encoding quality; its average value SNRSEG reflects aspects of quantizer performance that do not

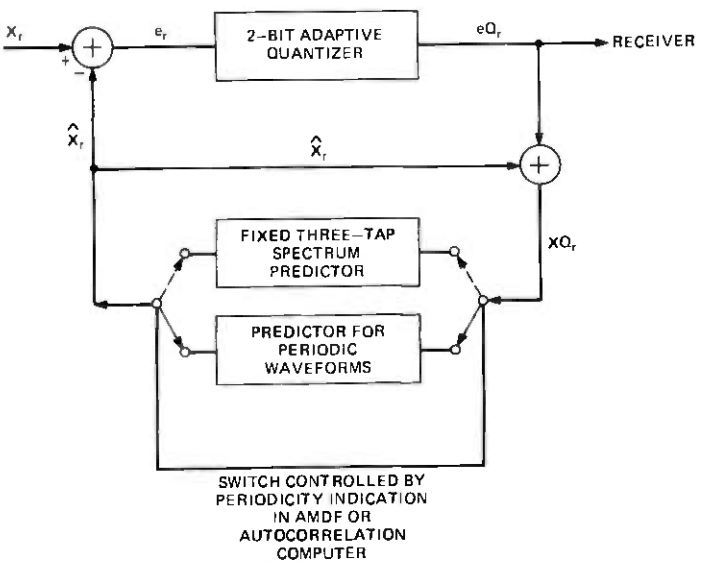


Fig. 4—Block diagram of pitch-adaptive coder.

Table IV — Comparison of prediction algorithms (utterance: F1; number of blocks: 134; block length V: 64; pitch-detector: based on unquantized speech and AMDF; $G_1 = 0.71$)

Predictor	Averager	7-Tap	3-Tap	1-Tap
SNR(dB)	10.3	13.1	13.3	14.4
SNRSEG(dB)	15.0	16.5	16.8	16.8

always come out from the conventional SNR measure.[18] For example, the time variation of SNRV would provide an appropriate indication of the differential treatment of voiced and unvoiced waveform segments (this is seen in Fig. 5); also, occasional large samples of SNRV (associated with pitch-adaptive coding of highly periodic segments) would have a better chance of showing up in the final result if the performance measure is SNRSEG, rather than the conventional SNR.

6.2 Comparison of the prediction algorithms of Table III

Table IV compares the performances of the four predictors in Table III for the DPCM encoding of a typical position of utterance F1. It is very interesting that the simplest of these predictors, the one-tap predictor, provides the best encoding. In fact, the rest of this paper will uniformly assume an appropriate one-tap predictor for periodic segments.

6.3 Choice of decision thresholds G_1 and G_2

Table V illustrates AMDF-based coding as a function of the periodicity-decision threshold G_1 [see (8)]. A choice of $G_1 = 0.84$ appears to provide the best combination of SNR and SNRSEG. This value of G_1 corresponds to a 1.5-dB prediction gain [ratio of average magnitude of input X to average magnitude of prediction error e (see Fig. 4)]. The value of $G_1 = 0.71$ (corresponding to a 3-dB prediction gain) provides a performance that is very close to the maximum. In fact, Grizmala[7] recommends the latter value of $G_1 = 0.71$.

Table VI shows corresponding results for autocorrelation-based DPCM with G_2 as parameter. One notes a broad optimum, with $G_2 = 0.2$ rep-

Table V — Effect of G_1 on AMDF-based pitch-adaptive DPCM (all parameters are the same as for Table I except that G_1 is now a variable)

G_1	0	0.50	0.71	0.84	1.0
SNR(dB)	9.3	14.2	14.2	14.4	14.5
SNRSEG(dB)	12.5	15.2	16.6	16.8	14.9

Table VI — Effect of G_2 on autocorrelation-based pitch-adaptive DPCM (all parameters are the same as for Table I except that the pitch detection is now correlation-based)

G_2	0.1	0.2	0.3	0.4	0.6
SNR(dB)	13.6	13.8	13.6	13.3	10.3
SNRSEG(dB)	14.6	14.5	14.3	15.8	14.3

resenting a reasonable autocorrelation threshold for hypothesizing periodicity; it is interesting that an SNRSEG criterion would dictate $G_2 = 0.4$.

6.4 Comparison of pitch detectors: AMDF vs autocorrelation; X-analysis vs XQ-analysis

Table VII compares, for optimal settings of G_1 and G_2, the encoding performances of AMDF- and autocorrelation-based pitch measurements. Notice the slight superiority of the AMDF approach, especially from an SNRSEG point of view. Notice also that pitch analyses based on X (Fig. 2a) are distinctly superior to those based on quantized speech XQ (Fig. 2b). Finally, it is very significant that, in the case of XQ-based analyses, the value of SNRSEG is 3- to 5-dB higher than that of SNR. This indicates that even with XQ-based designs, many periodic segments get encoded very well in a short-term sense (leading frequently to very good SNRV values that tend to boost the average SNRV-value SNRSEG). The above observation has been confirmed in informal listening tests. These tests have also shown that the quantization noise in XQ-based AMDF-coding tends to be "whiter" than the noise obtaining with the other three pitch-detection schemes of Table VII.

6.5 Pitch-period update-time V

Table VIII shows coder performance as a function of how frequently the periodicity test is made, and a possible pitch period recomputed.

Table VII — Comparison of four pitch detectors (all parameters are the same as for Table I, except that four pitch detectors are involved, and G_1 and G_2 are optimized for each case)

Type of Pitch Analysis	AMDF		Autocorrelation	
Basis of the analysis	X	XQ	X	XQ
SNR-optimizing G-values (G_1 for AMDF, G_2 for correlation)	0.84	0.84	0.20	0.30
SNR(dB)	14.4	10.0	13.8	10.1
SNRSEG(dB)	16.8	15.0	14.5	13.2

Table VIII — Dependence of performance on update time V; entries are SNR values in dB (female utterance: F1; number of blocks: 134; male utterance: M1; number of blocks: 134; pitch detector: based on unquantized speech and AMDF; $G_1 = 0.71$)

V	32	64	128	192
Male	—	12.1	11.4	9.8
Female	15.1	14.4	12.8	—

Recall that the update time assumed in Tables IV through VII was $V = 64$ samples (8 ms). Previous researchers[4–7] have usually recommended V-values like 40 or 50.

VII. SUMMARY AND CONCLUSIONS

Table IX compares, for the complete utterances F1, F2, M1, and M2, the performance of pitch-adaptive DPCM coding with that of DPCM with a fixed three-tap spectrum predictor. Note that both of these coders use adaptive quantization. The conventional encoder uses a fixed spectrum predictor while the pitch-adaptive encoder includes a second adaptive one-tap predictor, which is switched in whenever an AMDF analysis on X suggests sufficient periodicity ($G_1 = 0.84$).

We note that there exists across the four sample sentences an average 3.8-dB SNR gain with pitch-adaptive coding. The better performance with female speech is not surprising, since for a given duration of a voiced speech utterance, the high-pitched female utterances have a greater number of pitch periods.

Figure 5 provides a typical time-plot of pitch period P and local signal-to-noise-ratio SNRV in pitch-adaptive coding. The example refers to a segment from F2. A pitch-period of zero in Fig. 5 indicates absence of periodicity. Notice the low values of SNRV for these nonperiodic blocks. Also, notice the cluster of three values of $P \simeq 133$. These three estimates are obviously three times a true pitch period $\simeq 44$.

As mentioned earlier, the work in this paper was motivated by the desire to improve waveform encoder performance at bit rates in the order

Table IX — Summary of DPCM encoder performance

Sample Utterance	Median Pitch (Number of 8-kHz Samples)	Number of Speech Blocks ($V = 64$)	DPCM With no Pitch Tracking SNR(dB)	Pitch-Adaptive DPCM SNR(dB)
F1	36	240	10.0	15.0
F2	40	288	14.0	18.0
M1	90	192	11.0	13.5
M2	92	245	11.0	14.5

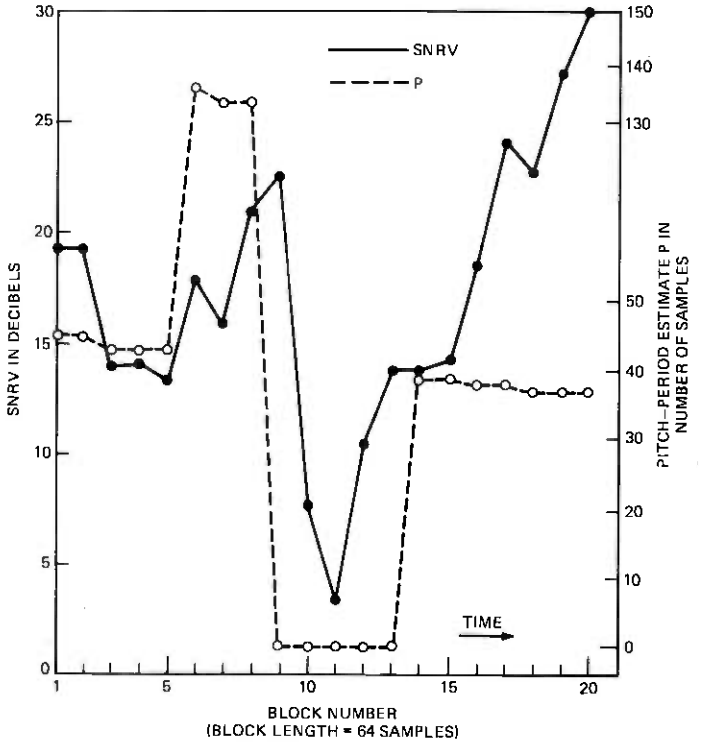


Fig. 5—Typical time variations of pitch period and local signal-to-noise ratio SNRV. (Data refers to a segment from utterance F2)

of 16 kb/s. The 2-bit pitch-adaptive coders discussed need 16 kb/s to transmit prediction-error information; and if pitch-analysis is to be performed on uncoded speech, the transmission of this information to a receiver will entail an additional channel capacity of about 1 kb/s. This assumes that pitch-period samples are coded with 7-bit accuracy and updated (and transmitted once, say, every 56 samples (8 kHz × 7 bits/56 = 1 kb/s). Alternatively, the coder can be used on a 16-kb/s channel if the sampling rate can be restricted to 15 kb/s/2 bits = 7.5 kHz.

VIII ACKNOWLEDGMENT

The author wishes to thank Mrs. I. Sondhi and Dr. Ing. P. Noll for their help with computer programs for pitch-adaptive coders.

REFERENCES

1. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers," Proc. IEEE, *62*, (May 1974), pp. 611–632.
2. P. Noll, "Non-adaptive and adaptive DPCM of speech signals," Polytech. Tijdschr. Ed. Elektrotech./Electron. (The Netherlands), No. 19, 1972.

3. P. Noll, "A Comparative Study of Various Quantization Schemes for Speech Encoding," B.S.T.J., *54*, No. 9 (November 1975), pp. 1597–1614.
4. 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.
5. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Audio and Speech Signal Processing, *ASSP-24* (October 1976), pp. 399–418.
6. L. I. Trottier, "An Investigation of Digital Vocoders," Masters Thesis, Department of Electrical Engineering, McGill University, Montreal, Quebec, January 1973.
7. F. Grizmala, "Application of Linear Predictive Coding to Long Haul Facilities—Results of a Simulation Study," unpublished work, January 1972.
8. C. S. Xydeas and R. Steele, "Pitch Synchronous 1st-Order Linear D.P.C.M. System," Electron. Lett., *12* (19 February 1976), pp. 93–95.
9. M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. Audio Electroacoust., *16* (June 1968), pp. 262–266.
10. H. Fujisaki, "Pitch Measurement Using Autocorrelation Techniques," J. Accoust. Soc. Amer., *28* (1956), p. 1518(A).
11. M. J. Ross et al., "Absolute Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust. Speech Signal Process. *22* (October 1974), pp. 353–362.
12. N. S. Jayant, "Adaptive Quantization With a One-Word Memory," B.S.T.J., *52*, No. 7 (September 1973), pp. 1119–1144.
13. 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.
14. S. U. H. Qureshi and G. D. Forney, "A 9.6/16 KBPS Speech Digitizer," Proc. IEEE Inter. Conf. Commun., San Francisco, June 1975, pp. 30-31 to 30-36.
15. D. L. Cohn and J. L. Melsa, "The Residual Encoder—An Improved ADPCM System for Speech Digitization," pp. 30-26 to 30-31 of Proceedings in Ref. 14.
16. R. P. Crane and E. T. Hedin, Jr., Bell Laboratories, unpublished work.
17. R. A. McDonald, "Signal-to-Noise and Idle Channel Performance of DPCM systems—Particular Application to Voice Signals," B.S.T.J., *45*, No. 7 (September 1966), pp. 1123–1151.
18. P. Noll, "Adaptive Quantizing in Speech Coding Systems," Int. Zurich Seminar on Digital Communication (IEEE), March 1974, pp. B3.1–B3.6.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech

By L. R. RABINER, C. E. SCHMIDT, and B. S. ATAL

(Manuscript received September 24, 1976)

Recently, a statistical-decision approach to the problem of voiced-unvoiced-silence detection of speech was proposed by Atal and Rabiner. This method was found to perform well on high-quality speech. However, the five speech parameters used in the analysis were not found to be as good for telephone-quality speech. Thus, an investigation was undertaken to determine a suitable set of parameters that would provide a reliable voiced-unvoiced-silence decision across a variety of standard telephone connections. A large number of parameters (70) were included in the investigation, including 12 LPC coefficients, 12 correlation coefficients, 12 parcor coefficients, 12 LPC partial error terms, etc. Many of the parameters were immediately eliminated because they provided almost no separability between the three decision classes. The remaining parameters were used in a knockout optimization to determine the five best parameters to use for a voiced-unvoiced-silence analysis. Various error weights were investigated to see what types of errors occurred and how they could be minimized. Finally, the use of the Itakura two-pole spectral normalization was investigated to see its effect on the error scores.

I. INTRODUCTION

In a recent paper, Atal and Rabiner described a fairly sophisticated method for reliably classifying segments of a waveform as voiced speech, unvoiced speech, or silence.[1] The analysis method used a statistical pattern-recognition approach to make this three-class decision. In another investigation, Rabiner et al. showed that the accuracy of the classification algorithm was quite high when the input signal was wideband; however, for telephone speech inputs, the accuracy of the classification degraded quite significantly.[2] The reason for this result

was not that the method inherently broke down for telephone inputs, but instead that the particular parameter set effective for wideband inputs was not equally effective for band-limited inputs. Thus, the motivation for the work to be presented in this paper is to investigate the suitability of a large number of parameters as features for reliable voiced-unvoiced-silence classification for telephone-quality speech.

Figure 1 shows a block diagram of the basic voiced-unvoiced-silence analysis algorithm. As shown in this figure, there are three steps in the method. First the speech is preprocessed. Generally, this preprocessing is a simple filtering operation; e.g., in the earlier work, a 200-Hz highpass filter was used to remove dc, hum, or low-frequency noise components present in the input signal. For telephone line inputs, we have considered somewhat more sophisticated preprocessing; namely, we have studied the use of a second-order inverse filter (as originally proposed by Itakura[3]) to normalize out the effects of varying telephone lines.

The second step in the algorithm is the feature measurement stage. For wideband inputs, only five parameters were considered, namely:

(*i*) Energy of the signal
(*ii*) Zero-crossing rate of the signal
(*iii*) Autocorrelation coefficient at unit sample delay
(*iv*) First predictor coefficient
(*v*) Energy of the prediction error.

These measurements were shown to provide a high degree of separability between the three classes of signal for wideband inputs.[1] However, for telephone-quality inputs, the band-limiting of the telephone line considerably reduces the effectiveness of all of the parameters in separating the classes of voiced speech, unvoiced speech, and silence. For example, the absence of signal energy above about 3 kHz significantly reduces the number of zero crossings for unvoiced speech.

To find an effective set of parameters that would be capable of reliably distinguishing between the three signal classes for telephone line inputs, a large number of parameters (70 in total) were studied. Using a set of training data, the probability-density functions for each of the parameters were estimated. Those parameters that provided little or no separation between voiced speech, unvoiced speech, and silence were eliminated from consideration. The remaining 36 parameters were studied as to their effectiveness in classifying telephone line inputs. A knockout type optimization was used to obtain the five most effective parameters for classifying signals according to an error-weighting scheme. Several combinations of different test sets of data and error weights were investigated.

The final step in the analysis method of Fig. 1 is a distance computa-

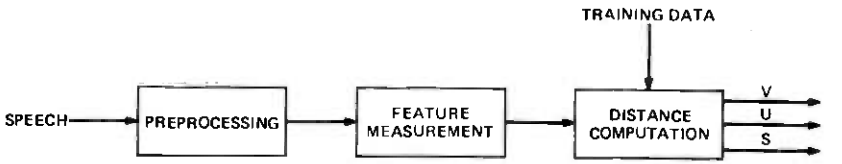


Fig. 1—Block diagram of silence-unvoiced-voiced classification system.

tion to determine whether a test signal is voiced, unvoiced, or silence. For this step, the non-Euclidean distance metric of Ref. 1 was retained because of its invariance properties to linear transformations of the data.[4]

Before presenting the results of the investigation, it is worthwhile reviewing the major distortions of telephone line signals as compared to wideband signals recorded with a high-quality microphone. These distortions include:

(*i*) Band limitation—The frequency response of a telephone line is approximately band limited between 300 Hz and 3000 Hz.

(*ii*) Phase distortion—For the frequency band between 300 and 3000 Hz, the magnitude of the incoming signal remains relatively flat; however, the phase is altered significantly in this band.

(*iii*) Nonlinear effects—Various nonlinearities occur in telephone transmission, including amplitude distortion (signal fading), peak and center clipping, impulse and/or gaussian noise addition, crosstalk, etc.

The effects of the first type of distortion are the most significant as far as this analysis method is concerned.* However, the other types of distortion can, and often do, play a role in determining an effective set of parameters for classifying telephone line signals.

The organization of this paper is as follows. In Section II, we present a description of the techniques used to determine the most effective sets of five parameters for classifying the incoming signals. In Section III, we present the results of the knockout optimization tests for each of the test sets of data and for each set of error weights. Finally, in Section IV, we compare the results on telephone inputs to those obtained with wideband inputs. A typical example showing how the method ultimately performed is presented to illustrate the types of problems that occur with telephone inputs.

* In this work we are considering only those distortions that occur within a local PBX; thus, one would expect a minimum of phase distortion and other nonlinear effects to occur. The place in which such distortions can become significant is in long-distance transmissions.

II. TELEPHONE SIGNAL ANALYSIS SYSTEM

For the preprocessing step of the analysis, a single highpass filter was always used to eliminate hum, dc offset, and low-frequency noise. This filter is described in Ref. 1. A second type of preprocessing was also investigated: the spectrum normalization technique as originally proposed by Itakura.[3] In this technique, the gross long-time spectrum of the signal is estimated using a two-pole LPC model, and then the signal is inverse filtered to remove the gross spectral tilt. Using the two-pole spectral normalization to reduce the spectral variability should, theoretically, also make the feature estimates more reliable. The rationale for considering this form of preprocessing is that for telephone speech the individual telephone transducer and line responses vary greatly across different handsets and telephone lines. Thus, any features estimated over such varying conditions may be adversely affected by the inherent variability of the transmission medium.

The way in which the two-pole spectral normalization was implemented is shown in Fig. 2. For each frame (10 ms of data), three correlation coefficients, $R(m)$, $m = 0,1,2$, are computed using the relation

$$R_j(m) = \sum_{n=0}^{N-m} s_j(n)s_j(n+m) \quad m = 0,1,2, \tag{1}$$

where N is 100, the sampling frequency is 10 kHz, and j is a frame counter that goes from 1 to NF, the number of frames in the utterance. The weighted normalized averages of the first two correlation coefficients (the $m = 1$, $m = 2$ terms) are computed as

$$\overline{R(m)} = \frac{1}{NC} \sum_{j=1}^{NF} \frac{R_j(m)}{R_j(0)} W_j(m), \qquad m = 1,2, \tag{2}$$

where $W_j(m)$ is a weight on the correlation function of the form

$$W_j(m) = \begin{cases} 1 & \text{if } R_j(0) > T \\ 0 & \text{otherwise} \end{cases}; \tag{3}$$

i.e., only frames whose energy $[R_j(0)]$ exceeds a fixed threshold T are

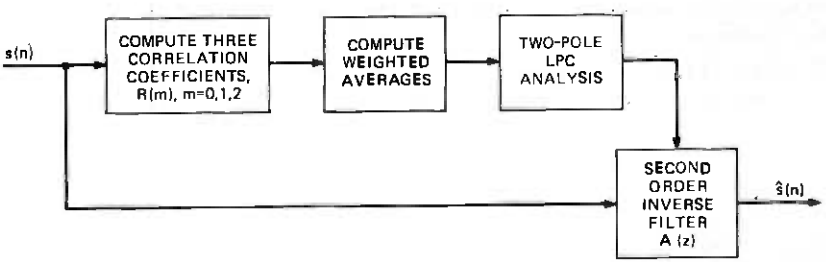


Fig. 2—Block diagram of two-pole spectral-normalization system.

included in the computation of the average correlations. The factor NC in eq (2) is the number of frames that exceeds the threshold of eq. (3). The weighting is used in computing the average correlations to eliminate unvoiced sounds and silence for which the correlation values are significantly different from those for voiced frames.

The third step in the normalization procedure is to compute the predictor coefficients of a two-pole linear predictive coding (LPC) match to the long-time average gross spectrum. If we denote the two LPC coefficients as a_1 and a_2, then the inverse filter needed to normalize the speech spectrum has a transfer function

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2}. \tag{4}$$

On a frame-by-frame basis the inverse filter can be applied directly to the autocorrelation coefficients of a high-order LPC analysis of the signal by convolving them with the autocorrelation coefficients of the second-order inverse filter.[3]

2.1 Features used in the analysis

The parameters (features) studied in the course of this work included the following:

Parameter	Description
1–12	The LPC coefficients of a 12th-order analysis using the Burg lattice method with a 10-ms frame[5,6]: a(1) to a(12).
13–24	The first 12 autocorrelation coefficients of the signal using a 10-ms frame: $\phi(0,1)$ to $\phi(0,12)$.
25–36	The first 12 parcor (partial correlation) coefficients of the signal: $k(1)$ to $k(12)$.
37–48	The first 12 partial normalized error coefficients of the LPC analysis: $E(1)$ to $E(12)$.
49–60	The first 12 cepstral coefficients of the signal as obtained by transforming the LPC coefficients: c(1) to c(12).
61	The log energy of the signal: LE.
62	The number of zero crossings per 10-ms frame: NZ.
63	The log normalized error of the 12-pole LPC analysis: LNE.
64	The maximum value minus the minimum value of the signal during the frame: ML.
65	The absolute energy in the first difference of the signal: ED.
66	The number of zero crossings per 10-ms frame for the first difference signal: NZD.
67	The maximum value minus the minimum value for the first difference signal: MLD.
68	The absolute energy of a smoothed version of the signal: ES.
69	The number of zero crossings per 10-ms frame for the smoothed signal: NZS.
70	The maximum value minus the minimum value for the smoothed signal: MLS.

Figure 3 shows the basic measurement scheme. For each 10-ms frame of the signal, an LPC analysis was performed using the Burg lattice method[5,6] giving a set of 12 LPC coefficients, 12 parcor coefficients, and 12 partial normalized errors defined as

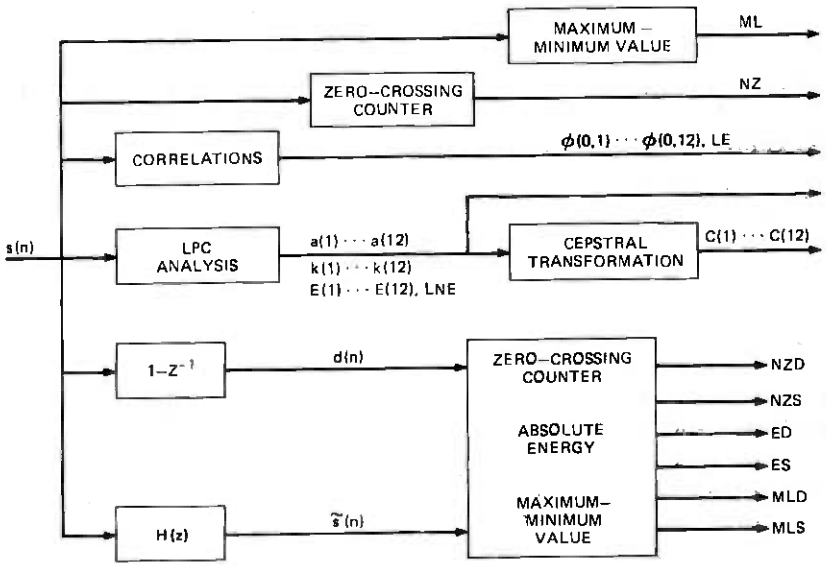


Fig. 3—Block diagram of feature-measurement system.

$$E(l) = \prod_{i=1}^{l} [1-k^2(i)], \qquad l = 1,2,\cdots,12, \tag{5}$$

i.e., $E(l)$ is the normalized error of an l-pole LPC analysis. Since the lattice method does not require the set of correlations directly, they are computed on the signal from the equation

$$\phi(o,i) = \sum_{n=0}^{N-1} s(n)s(n-i), \qquad i = 1,2,\cdots,12, \tag{6}$$

i.e., a nonstationary correlation function is computed. The cepstral coefficients are computed directly from the LPC coefficients using the recursion relation

$$c(i) = a(i) - \sum_{m=1}^{i-1} \frac{m}{i} c(m)a(i-m), \qquad 1 \leqq i \leqq 12. \tag{7}$$

Two other measurements are made directly on the signal $s(n)$. These are the zero-crossing count defined as the number of zero crossings per 10-ms interval, and a computation of the difference between the maximum and minimum signal amplitudes in the frame.

In addition to the above parameters, six additional measurements are made on the first difference of the signal, $d(n)$, defined as

$$d(n) = s(n) - s(n-1) \tag{8}$$

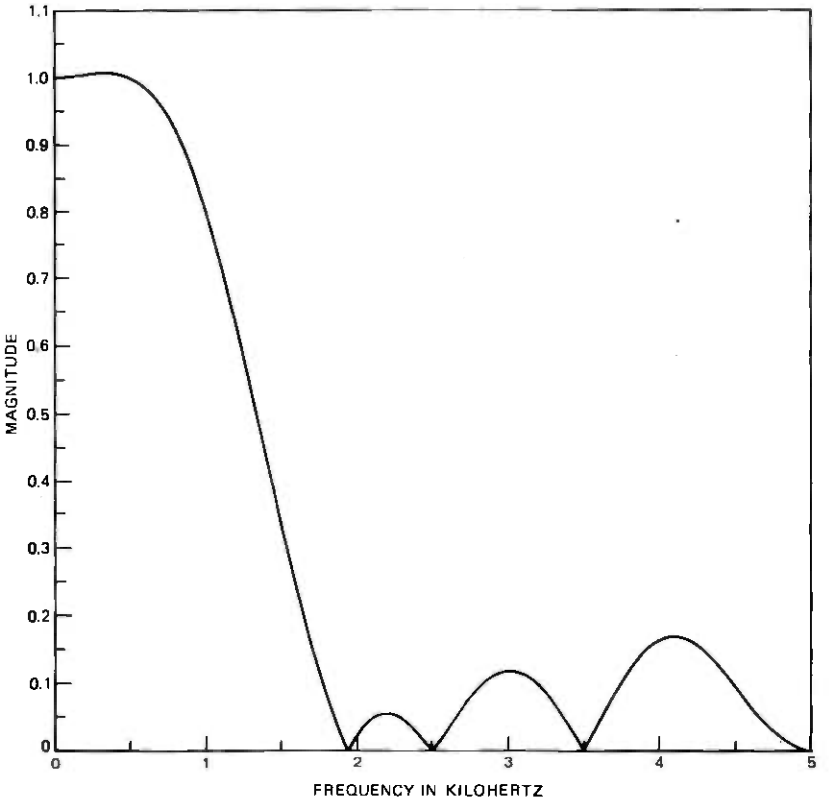


Fig. 4—Frequency response of lowpass smoothing filter.

and a smoothed version of the signal obtained via the filtering relation*

$$\tilde{s}(n) = -s(n) + s(n-2) + 2s(n-3) + 4s(n-4) + 4s(n-5) + 4s(n-6) + 2s(n-7) + s(n-8) - s(n-10). \quad (9)$$

It can be seen that the filtering of eq (9) can be accomplished without the need for a multiplier and, thus, can be implemented quite efficiently. Figure 4 shows the frequency response of this filter. It can be seen that the filter provides a small amount of high-frequency attenuation and therefore can be considered as a lowpass smoothing filter. The measurements made on $d(n)$ and $\tilde{s}(n)$ are zero-crossing count, absolute energy, and difference between maximum and minimum signal levels in the frame.

* This filter as well as parameters 65–70 were suggested by D. R. Reddy for inclusion in this work.

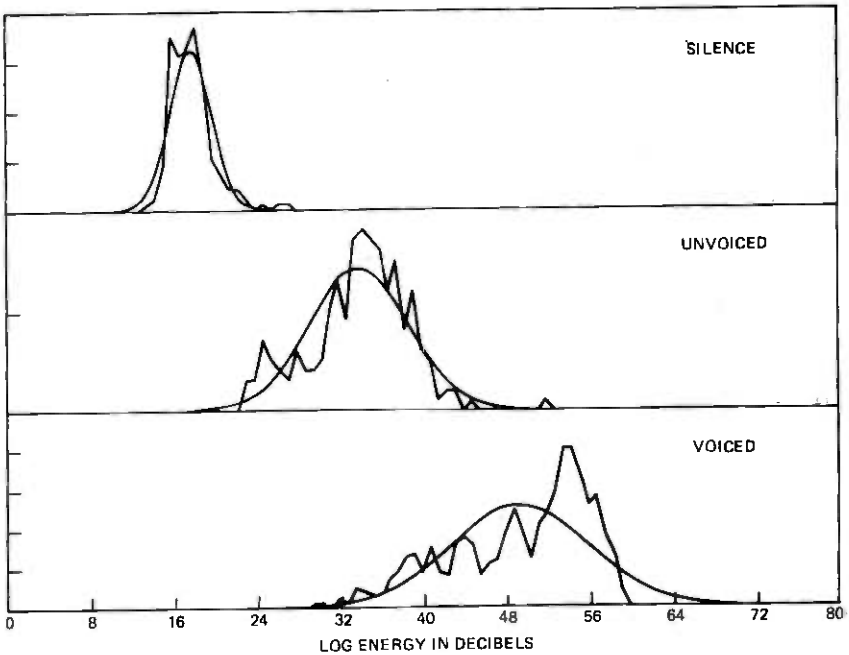


Fig. 5—Probability distributions for log energy of the signal for silence, unvoiced, and voiced classes. Both estimated and gaussian fits to the distributions are shown.

Once the initial set of 70 parameters was chosen, a training set of data was used to estimate one-dimensional probability functions for each of the parameters and for each signal classification. A one-dimensional gaussian curve having the same mean and standard deviation as the measured distributions was also computed for each parameter.* Figures 5 through 7 show three typical distributions for the parameters log energy (feature 61), first LPC coefficient (feature 1), and twelfth LPC coefficient (feature 12), respectively. For the log-energy parameter (Fig. 5), the distributions for silence, unvoiced, and voiced speech were fairly well separated with means of 18, 34, and 49 dB, respectively. Similarly the distributions for the first LPC coefficient (Fig. 6) were also well separated with means of −0.19, −0.66, and −1.9 for silence, unvoiced, and voiced speech, respectively. However, as shown in Fig. 7, the distributions for all parameters were not well separated across the different classes. In this case, the distributions for all three signal classes overlapped considerably. It seems reasonable that features in which such behavior is observed will not be effective in the classification procedure. Therefore,

* For the distance metric used in this work, it is not critical that the one-dimensional distributions of the parameters be well approximated by a simple gaussian curve. It is important, however, that the distributions be unimodal.

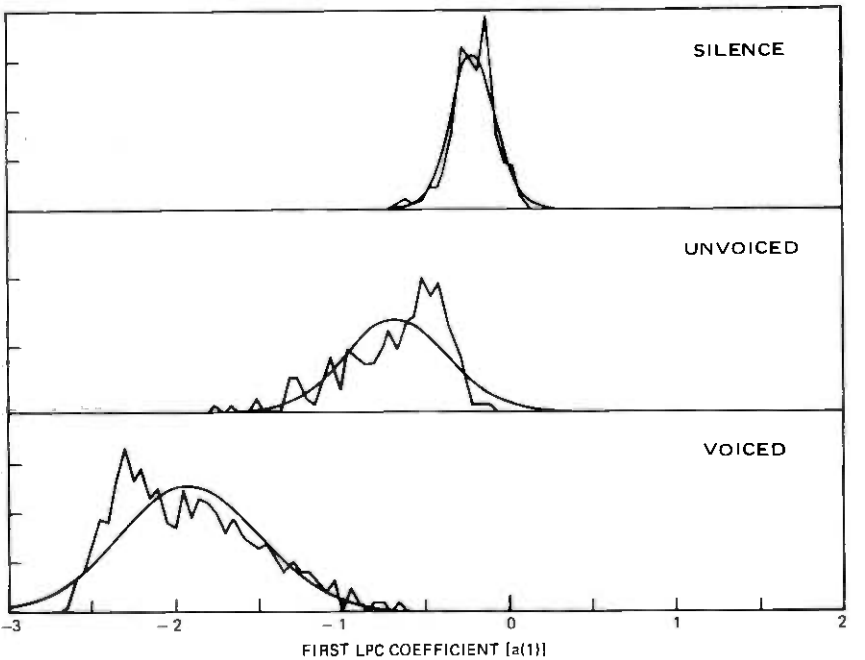


Fig. 6—Probability distributions for first LPC coefficient for silence, unvoiced, and voiced classes.

such parameters were not considered in the testing to be described in this paper.

A total of 34 of the 70 parameters were eliminated in this manner. The parameters eliminated were the higher LPC coefficients [a(5) to a(12)], the higher autocorrelation coefficients [$\phi(0,5)$ to $\phi(0,12)$], the higher parcor coefficients [$k(5)$ to $k(12)$], the higher cepstral coefficients [$c(5)$ to $c(12)$], and the last two partial normalized LPC error coefficients [$E(11)$ and $E(12)$]. The remaining 36 parameters were used in all the optimization tests described in the next section.

2.2 Knockout optimization procedure

To choose the set of five parameters out of the remaining 36 features that best (most accurately) classified signal intervals as silence, unvoiced, or voiced speech, a knockout optimization procedure was used.[7] Figure 8 shows a flow diagram of the procedure. Using a testing set of data (see Section 2.3) and an objective error measure, the knockout optimization proceeded first to find the single best parameter for separating the three classes. The best parameter is knocked out and used in combination with each of the remaining 35 features to find the best pair of parameters for the signal classification. This process of knocking out the best parameter

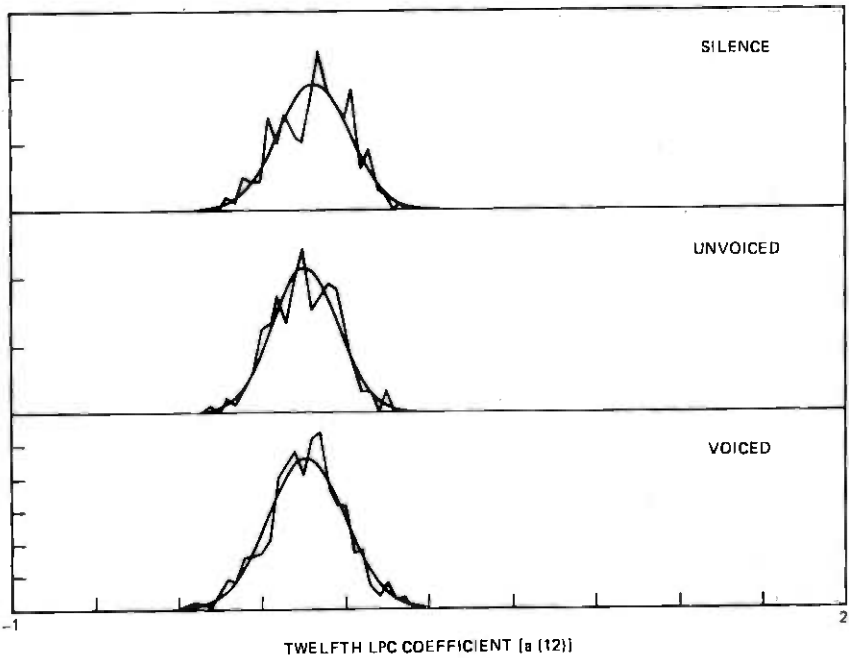


Fig. 7—Probability distributions for twelfth LPC coefficient for silence, unvoiced, and voiced classes.

and combining all knocked-out features with the ones remaining in the parameter set was iterated until a total of five parameters were obtained.

Several comments should be made about this procedure. First, it is noted that this method does *not* necessarily yield the optimum set of five parameters for making the silence, unvoiced, voiced decision. In general, the resulting parameter set is suboptimal since only a very small subset of the total number of combinations of 36 parameters taken five at a time are considered in this method. In defense of the method, however, one can argue that, within the constraints of the procedure, an optimal set of the 36 parameters is chosen. Furthermore, at least theoretically, the addition of each new knocked-out feature reduces the error score. Finally, it is argued that the resulting feature sets provide significantly better accuracy for signal classification than almost any randomly chosen set of five of the 36 parameters.

2.3 Distance computation

The distance computation used throughout this investigation was the non-Euclidean distance metric defined in Ref. 1. For the feature vector $\mathbf{x} = [\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(5)]$ with mean vector $\mathbf{m}_i = [\mathbf{m}_i(1),\mathbf{m}_i(2),\ldots,\mathbf{m}_i(5)]$

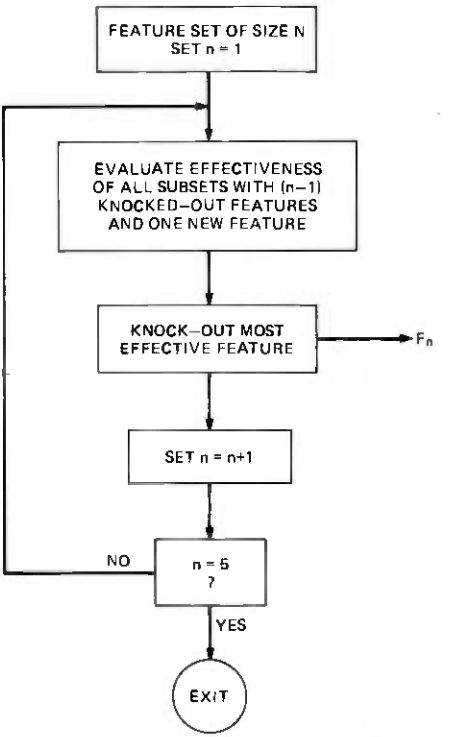


Fig. 8—Flow chart of knockout optimization algorithm.

and covariance matrix Λ_i, the distance computation was of the form

$$d_i = (\mathbf{x} - \mathbf{m}_i)\Lambda_i^{-1}(\mathbf{x} - \mathbf{m}_i)^t, \tag{10}$$

where $i = 1$ (silence), 2 (unvoiced), or 3 (voiced), and t denotes the transpose of a vector. For each signal class, d_i is computed and the decision rule is to select class i such that $d_i < d_j$ for all $j \neq i$; i.e., choose the class with the minimum distance to vector **x**.

To implement the distance computation in eq. (10) during the knockout optimization required the computation of a new covariance matrix Λ_i for each subset of parameters being considered. Thus, on the order of 420 covariance matrices had to be estimated in a typical optimization run. This represented a substantial amount of computation.

2.4 Experimental procedure

The formal evaluation of the feature sets was made by choosing a fairly large data base of different utterances and different speakers, using part of the data base for training the system, and using the remainder of the data base for testing the system.

A total of five speakers (two male, three female) were used in the telephone-line evaluation. Each speaker recited three utterances,* each one over a new dialed-up connection and thereby ensuring that a different telephone-transmission path was obtained for each utterance. Two different telephone transmitters (carbon microphones) were also used in the test. One utterance from each speaker was used in the training set; the remaining two utterances were used in the testing set.

For each recorded utterance, a manual analysis was performed on each 10-ms interval to classify it as voiced, unvoiced, or silence based on both the acoustic waveform and a phonetic transcription of the utterance. Each signal classification was further modified with a label as to the certainty of the manual classification. The labels used were:

(*i*) Absolutely certain—clear characteristics of the class to which it was assigned.

(*ii*) Moderately certain—generally a boundary interval between classes in which two types of signal were present.

(*iii*) Uncertain—classified primarily on linguistic information about the utterance. Included in this class were voiced fricatives, voiced stops, and certain transients (including some telephone-line transients).

Figure 9 shows an example in which uncertain intervals occurred. This section of speech is from the beginning of the word *cowboys*. The initial intervals should linguistically be labelled as either silence or unvoiced speech corresponding to the stopgap and burst of the voiceless stop /k/. However, acoustically the initial seven intervals (as marked in Fig. 9) show properties more similar to voiced speech than to silence or unvoiced sounds. These intervals were treated as uncertain intervals and were marked as unvoiced speech for testing purposes.

For the training set, only those intervals for which the classification was absolutely certain were used. For the testing set, three sets of data were used. One set contained only those intervals for which the classification was absolutely certain (TS1). The second set contained both the moderately certain as well as the absolutely certain intervals (TS2). The third set contained all the intervals, regardless of the certainty of manual classification (TS3). In the next section, we present results for each of these testing sets of data.

III. RESULTS

The knockout optimization procedure described in Section II was run on the three sets of test data using 10 different error-weighting matrices.

* Each utterance was a carefully chosen sentence containing a mixture of voiced, unvoiced, and silence intervals.

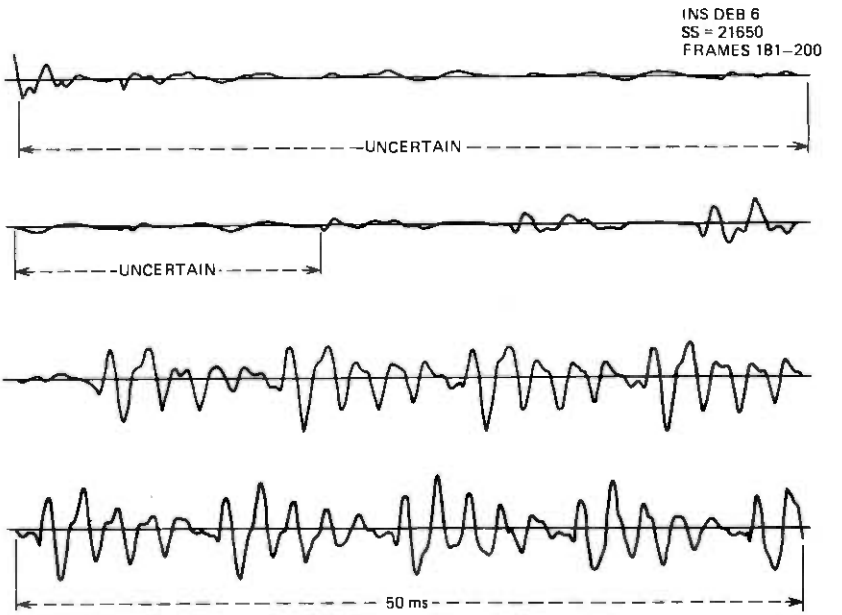


Fig. 9—The acoustic waveform for a section of speech in which an interval was uncertain.

In addition, the entire experiment was rerun on data preprocessed using the two-pole spectral normalization method described in Section II. Table I provides a summary of the three test sets of data, the 10 error-weight matrices, and the two processing conditions.

The error-weight matrices were used to study the effects of weights for each type of classification error on the overall error rate and the choice of the optimal features. The definition of a general error-weight matrix is as follows. If we let E denote the overall error score in classifying the data of a test set, then

$$E = N_{ss}W_{ss} + N_{su}W_{su} + N_{sv}W_{sv} + N_{us}W_{us} + N_{uu}W_{uu} + N_{uv}W_{uv} + N_{vs}W_{vs} + N_{vu}W_{vu} + N_{vv}W_{vv}, \quad (11)$$

where N_{ab} is the number of frames of a class a which were classified as belonging to class b, and W_{ab} is the weight attached to this pair of classifications. It should be clear from eq. (11) that

$$\begin{aligned} N_s &= N_{ss} + N_{su} + N_{sv} \\ N_u &= N_{us} + N_{uu} + N_{uv} \\ N_v &= N_{vs} + N_{vu} + N_{vv}, \end{aligned} \quad (12)$$

where N_a is the number of frames in the test set in class a. Table II shows

Table I — Summary of factors considered in the investigation

Data Test Sets	TS1	Absolutely certain intervals
	TS2	Moderately certain intervals added to TS1
	TS3	Uncertain intervals added to TS2
Error-Weight Matrices	WM1	Uniform matrix
	WM2	Silence weighting matrix
	WM3	Unvoiced weighting matrix
	WM4	Voiced weighting matrix
	WM5	Silence-to-unvoiced weighting matrix
	WM6	Unvoiced-to-silence weighting matrix
	WM7	Silence-to-voiced weighting matrix
	WM8	Voiced-to-silence weighting matrix
	WM9	Voiced-to-unvoiced weighting matrix
	WM10	Unvoiced-to-voiced weighting matrix
Preprocessing	P1	Direct transmission
	P2	Two-pole spectral normalization

the 10 weight matrices described in Table I. Each matrix is expressed in the form

$$W = \begin{bmatrix} W_{ss} & W_{su} & W_{sv} \\ W_{us} & W_{uu} & W_{uv} \\ W_{vs} & W_{vu} & W_{vv} \end{bmatrix}, \tag{13}$$

where W_{ab} is not generally the same as W_{ba}.

As seen in Table II, error weight-matrix 1 (WM1) attaches equal weight to all six types of misclassifications and, therefore, is the canonic error matrix for the three-class problem. Error matrices 2–4 (WM2–WM4) each choose a subset in which one of the three classes is essentially merged with another class. For example, error matrix 4 (WM4) gives 0 weight to errors between the classes of silence and unvoiced speech; however, the other four types of error have unity weight. Thus, this matrix serves to distinguish most effectively between voiced speech and nonvoiced (either silence or unvoiced) speech. As another example, error matrix 2 (WM2) gives 0 weight to errors between the classes of voiced and unvoiced speech. Thus, this matrix serves to distinguish between speech (voiced or unvoiced) and silence. As such, it would be useful for speech-detection applications. Error matrices 5 through 10 each focus on only one of the six sets of misclassifications. The results for these cases give a lower bound on the error rate for special cases in which only a single type of misclassification is considered.

For each of the sets of data of Table I, the knockout optimization procedure was used giving the five best features and the resulting overall misclassification rate, defined as

$$E_N = \frac{E}{(N_s + N_u + N_v)}. \tag{14}$$

Table II — Error-weight matrices used in the investigation

$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ WM1 (a)	$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ WM2 (b)	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ WM3 (c)
$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ WM4 (d)	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM5 (e)	$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM6 (f)
$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM7 (g)	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ WM8 (h)	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ WM9 (i)
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ WM10 (j)		

For error weights 5 through 10 (where only a single misclassification was counted) the overall misclassification rated was defined as

$$E_N = \frac{E}{N_a}, \tag{15}$$

where

$$N_a = \begin{cases} N_s & \text{for WM5 and WM7} \\ N_u & \text{for WM6 and WM10} \\ N_v & \text{for WM8 and WM9} \end{cases}. \tag{16}$$

The results of these experiments are presented in Tables III through VI. Tables IV through VI present the misclassification rate results, and Table III gives both the parameter numbers and the mnemonics of the five parameters chosen by the optimization procedure. The results in these tables are presented sequentially; i.e., the results obtained using only l of the five parameters (l = 1,2,3,4) are indicated in the appropriate rows of the tables.

Two comments should be made about the data. In many cases, it was found that the overall misclassification rate did not monotonically decrease as more features were knocked out of the parameter set. For these

Table III — Optimal features chosen by the knockout optimization for telephone inputs

Test Set	Parameter	Weight Matrix									
		WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS1	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	70-MLS*	67-MLD	68-ES*	63-LNE	46-E(10)
	2	14-ϕ(0,2)	26-k(2)	68-ES	44-E(8)	27-k(3)				2-a(2)	42-E(6)
	3	63-LNE	16-ϕ(0,4)	69-NZS	68-ES	67-MLD*				68-ES	43-E(7)*
	4	67-MLD			4-a(4)					65-ED	
	5	64-ML								27-k(3)*	
TS2	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	64-ML	67-MLD	68-ES*	63-LNE	46-E(10)
	2	50-c(2)	26-k(2)	68-ES	66-NZD	27-k(3)	70-MLS*			2-a(2)	42-E(6)
	3	16-ϕ(0,4)		69-NZS	45-E(9)	67-MLD*				68-ES	62-NZ
	4				70-MLS					65-ED	25-k(1)
	5				67-MLD					27-k(3)*	45-E(9)*
TS3	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	50-c(2)	67-MLD	68-ES*	14-ϕ(0,2)	46-E(10)
	2	50-c(2)	26-k(2)	68-ES	66-NZD	27-k(3)	52-c(4)		43-E(7)*	70-MLS	42-E(6)
	3	16-ϕ(0,4)		69-NZS	45-E(9)	67-MLD*	68-ES*			63-LNE	40-E(4)
	4	3-a(3)			70-MLS					68-ES	43-E(7)
	5	64-ML			61-LE						66-NZD

* Other features provided the same overall misclassification rate.

Table IV — Error rates for telephone inputs

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
	TS1 Without Two-Pole Spectral Normalization									
1	15.1	4.7	9.8	2.8	2.4	0.7	2.4	0	0.9	1.0
2	11.0	4.7	5.7	2.3	0.6	0			0.5	0.7
3	7.5	4.4	5.3	1.9	0				0.4	0.3
4	6.9			1.9					0.3	
5	6.4									

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
	TS1 With Two-Pole Spectral Normalization									
1	13.8	5.1	11.0	5.1	2.4	4.6	2.4	0	3.2	6.8
2	9.7		7.1	4.5		0.7			3.2	2.7
3	8.6		6.0	4.4					2.1	2.3
4	7.6		6.0	3.9					1.9	
5	7.6								1.8	

Size of Training and Testing Sets for TS1

	Number of Frames	
	Training	Testing
S	207	328
U	210	306
V	539	1180
	956	1834

Table V—Error Rates for Telephone Line Inputs

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS2 without two-pole Spectral Normalization										
1	16.2	5.8	12.2	4.5	2.4	0.5	2.6	0	1.8	2.7
2	11.7	5.4	7.4	4.3	0.8	0			1.0	2.1
3	10.8		7.1	3.6	0				0.7	1.3
4				3.5					0.5	1.1
5				3.5						

Parameter	Weight Matrix WM1	WM2	NM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS2 with two-pole Spectral Normalization										
1	18.1	7.2	14.4	8.6	2.9	3.9	2.7	0.3	4.5	9.5
2	13.4		10.4	8.2		1.1		0	5.2	6.1
3	12.4		9.4	7.8					3.9	5.0
4	11.6			7.4					3.8	4.2
5	11.5			7.2					3.6	3.7

Size of Training and Testing Sets for TS2

	Number of Frames Training	Testing
S	207	375
U	210	378
V	539	1196
	956	1949

cases data are presented up to the number of parameters at which the error rate kept decreasing. The second comment concerns the specific features knocked out in the optimization (as given in Table III). In many cases, a large number of features (other than the ones presented) provided essentially the same overall misclassification rate as the feature that was knocked out. These cases are indicated by an asterisk after the feature number in these tables. For such cases, features other than the ones indicated in the table may be equally appropriate.

IV. ANALYSIS OF THE RESULTS

Several important observations can be made by examining carefully the results of Tables III though VI. First, it can be seen by comparing error rates for matrix WM1 to those for matrices WM2 through WM4 that most of the overall error rate for the canonic error matrix was due to misclassifications between the classes of silence and unvoiced speech* (compare results for WM1 and WM3). This result is certainly not unan-

* Further evidence of this result is given in Table VII, which shows a breakdown of the error components. This table is discussed later in this section.

Table VI — Error rates for telephone line inputs

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
				TS3 Without Two-Pole Spectral Normalization						
1	18.5	6.3	13.4	6.3	2.4	0.9	2.7	0	3.3	5.9
2	13.2	5.6	9.0	6.1	0.8	0.2			2.1	5.4
3	12.2		8.7	6.1	0	0			1.3	5.0
4	12.0		8.7	5.2					1.1	4.1
5	11.7			5.0						3.4

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
				TS3 With Two-Pole Spectral Normalization						
1	20.3	8.2	16.2	10.3	2.9	4.5	2.9	0.2	4.9	12.5
2	15.6		12.2	9.5		1.3		0	5.9	14.3
3	14.3		11.2	9.1					4.8	11.4
4	13.7			8.7					4.5	9.8
5	13.4			8.5					4.4	8.7

Size of Training and Testing Sets for TS3

	Number of Frames	
	Training	Testing
S	207	379
U	210	443
V	539	1264
	956	2086

ticipated since the band limiting of telephone speech has the most severe effect on unvoiced sounds whose spectral components often fall above the high-frequency cutoff of the telephone transmission.

Based on the above result it would seem reasonable to compare error scores using error matrices 1 and 4. It can be seen from the tables that if one does not consider distinctions between silence and unvoiced speech, then an improvement of somewhat more than 2 to 1 in error score is obtained. For the case of absolutely certain classifications, an error rate of 1.9 percent is obtained for error matrix 4. For test sets TS2 and TS3, the error rate for error matrix 4 increases to 3.5 and 5.0 percent, respectively.

The results using error matrix 2 (the speech detection matrix) show that, in the case of absolutely certain classifications (Table IV) an error rate of 4.3 percent is obtained. For test sets TS2 and TS3, the error rate for matrix 2 increases slightly to 5.4 and 5.6 percent, respectively.

The results of using error matrices WM5–WM10 show that the most frequent misclassification occurs between silence and voiced speech in which error rates on the order of 2 to 3 percent were obtained for all three test cases. The problem here occurs during low-level sounds, such as voiced stops where the silence regions are often classified as voiced due to the presence of low-frequency components of the signal. Unfortunately, such signals do not fall neatly into either category and the decision algorithm consistently classified them as voiced sounds whereas the manual classification was silence.

Comparisons of the results of Tables IV, V, and VI showed that the error scores increased with the complexity of the test set as anticipated. However, it is difficult to attach too much meaning to the absolute error rates for TS2 and TS3, since the frames which were added constituted boundary frames and frames which were subject to classification error in the manual classification. The results are presented to provide information as to the sizes of the increases in error rate that are to be expected with such input test sets.

The data of Table III (the optimal feature list) are also quite interesting. The influence of the weight matrix is evident by scanning across the rows of the table. Each weight matrix had its own set of optimal features, which were different from those of any other weight matrix. By scanning down the columns of this table, however, it is seen that the influence of the data test set was fairly weak in that the optimal-feature set remained substantially the same for all three test sets across the three sets of data.

An interesting result shown in Tables IV through VI is that the two-pole normalization scheme did not provide essentially any improvement in the classification accuracy across any of the test conditions studied. This result is a little surprising in light of the work of Itakura who found

that it compensated different telephone transmission conditions quite adequately.[3] One possible reason for this result is that the non-Euclidean distance metric to some extent compensates automatically for the variable telephone transmission conditions by appropriate linear transformation of the feature space. Thus, for this classification method, the use of a two-pole spectral normalization is of little value.

An additional breakdown of the error analysis for the most important error-weight matrices (WM1, WM2, and WM4) is given in Table VII in which the percentage of each type of misclassification is presented. It can be seen in this table that certain types of errors dominated the scores. For example, no cases occurred throughout the test in which a voiced interval was classified as silence. It can also be seen that, as mentioned previously, the error rate for silence-to-unvoiced speech dominated the overall error rates for error matrices WM1 and WM2, whereas no single component of the error dominated the overall error rate for matrix WM4.

4.1 Comparison with wideband results

Although some numerical scores were presented in Ref. 1 for misclassification rates using the analysis method on wideband (high-quality) data, a set of companion results were obtained in this study to compare and contrast the error results for wideband and telephone signals. Using the identical procedures discussed in Sections II and III, a set of optimal features and error rates were obtained for wideband test sets of signals. The results of these runs are presented in Tables VIII and IX. Comparisons of the error rate tables (VII and VIII) show the following:

(*i*) For error weight matrix WM1, the scores for wideband data were from two to three times lower than for telephone data. This is due to the vastly improved scores on the category of silence-to-unvoiced errors. The error rates for many of the other possible misclassifications were quite comparable.

(*ii*) For error weight WM4, the scores for wideband data were only slightly better than for telephone data, indicating that a voiced-not voiced decision can be as reliably made over a telephone line as for high-quality inputs. However, the speech–not speech decision is much more difficult for telephone data than for wideband signals.

(*iii*) For error weight matrix WM2, the scores for wideband data were from two to eight times lower than for telephone data. This result is again due to the improved performance in discriminating between silence and unvoiced speech for wideband data.

Table VII — Breakdown of error percentages for telephone line inputs

Test Set	Error-Weight Matrix WM1									
	SU	SV	US	UV	VS	VU	S	U	V	Overall
TS1	16.8	6.1	4.6	5.2	0	0.9	22.9	9.9	0.9	6.4
TS2	30.1	5.9	2.9	5.0	0	4.1	36.0	8.0	4.1	11.0
TS3	28.8	5.8	2.9	15.4	0	2.5	34.6	18.3	2.5	11.7
Test Set	Error-Weight Matrix WM4									
	SU*	SV	US*	UV	VS	VU	S**	U**	V	Overall
TS1	93.0	4.6	0	3.0	0	0.9	97.6	3.0	0.9	1.9
TS2	88.9	6.7	0.3	6.1	0	1.6	95.7	6.4	1.6	3.5
TS3	43.1	6.1	3.2	8.6	0	3.4	49.2	11.8	3.4	5.0
Test Set	Error-Weight Matrix WM2									
	SU	SV	US	UV*	VS	VU*	S	U**	V**	Overall
TS1	11.6	5.2	7.6	17.5	0.1	4.7	16.8	25.1	4.8	4.4
TS2	14.8	5.4	7.5	27.7	0.2	7.0	20.2	35.2	7.2	5.4
TS3	15.4	5.6	7.9	27.6	0.2	9.4	21.0	35.5	9.5	5.6

* These results had 0 weight in the overall error score and, therefore, did not affect the choice of features.
** Only a single component of this error score is included in the overall score.

Table VIII — Breakdown of error percentages for wideband inputs

	Error-Weight Matrix WM1									
Test Set	SU	SV	US	UV	VS	VU	S	U	V	Overall
TS1	2.3	0	3.9	2.6	0	1.5	2.3	6.6	1.5	2.2
TS2	8.7	5.8	5.7	6.7	0.1	2.3	14.5	12.4	2.4	5.3
TS3	8.9	5.9	7.2	6.8	0.1	2.3	14.8	14.0	2.5	5.7
	Error-Weight Matrix WM4									
Test Set	SU*	SV	US*	UV	VS	VU	S**	U**	V	Overall
TS1	29.5	0	5.3	2.5	0	0.9	29.5	7.9	0.9	1.1
TS2	39.9	5.8	4.8	9.1	0	1.4	45.7	13.9	1.4	3.1
TS3	33.3	5.9	5.0	5.9	0	2.0	39.3	10.9	2.0	3.1
	Error-Weight Matrix WM2									
Test Set	SU	SV	US	UV*	VS	VU*	S	U**	V**	Overall
TS1	3.4	0	2.0	2.0	0	2.0	3.4	4.0	2.0	0.5
TS2	7.3	7.3	4.3	11.0	0	2.7	14.5	15.3	2.7	2.3
TS3	7.4	7.4	5.9	11.3	0	3.0	14.8	17.2	3.0	2.6

* These results had 0 weight in the overall error score and therefore did not affect the choice of features.
** Only a single component of this error score is included in the overall score.

Table IX — Optimal features for wideband test sets and error-weight matrices WM1, WM4, and WM2

Test Set	WM1	WM4	WM2
TS1	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	25 k(1)	70 MLS
	14 $\phi(0,2)$	69* NZS	61 LE
	61 LE	67* MLD	14 $\phi(0,2)$
	68* ES		68 ES
TS2	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	25 k(1)	70 MLS
	14 $\phi(0,2)$	16 $\phi(0,4)$	61 LE
	61 LE	64 ML	68 ES
	68 ES		67 MLD
TS3	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	15 $\phi(0,3)$	70 MLS
	14 $\phi(0,2)$	26 k(2)	61 LE
	61 LE	13 $\phi(0,1)$	68 ES
	68 ES	52 c(4)	67 MLD

4.2 Typical test example

Figures 10 and 11 show the results of applying the classification method to the utterance, "Few thieves are never sent to the jug," spoken by a male speaker. The contour shown in (a) of each figure is a manual classification of each frame. Part (b) shows the results of analysis using parameters obtained from WM1 (Fig. 10) and WM4 (Fig. 11). Part (c) shows the results of nonlinearly smoothing the analysis contours using a median smoother.[8] Parts (d), (e), and (f) show plots of the probability of correct classification based on the distance calculation for each class; i.e., if we denote the distance calculated for silence as D_s, the distance calculated for unvoiced as D_u, and the distance calculated for voiced as D_v, then

$$P(S) = \frac{D_u D_v}{D_s D_u + D_s D_v + D_u D_v} \tag{17}$$

$$P(U) = \frac{D_s D_v}{D_s D_u + D_s D_v + D_u D_v} \tag{18}$$

$$P(V) = \frac{D_s D_u}{D_s D_u + D_s D_v + D_u D_v}. \tag{19}$$

It can be seen that $P(S)$, $P(U)$, and $P(V)$ define a probability measure, since

$$P(s) + P(u) + P(v) = 1 \tag{20}$$

and

$$0 \leqq P(s), P(u), P(v) \leqq 1 \tag{21}$$

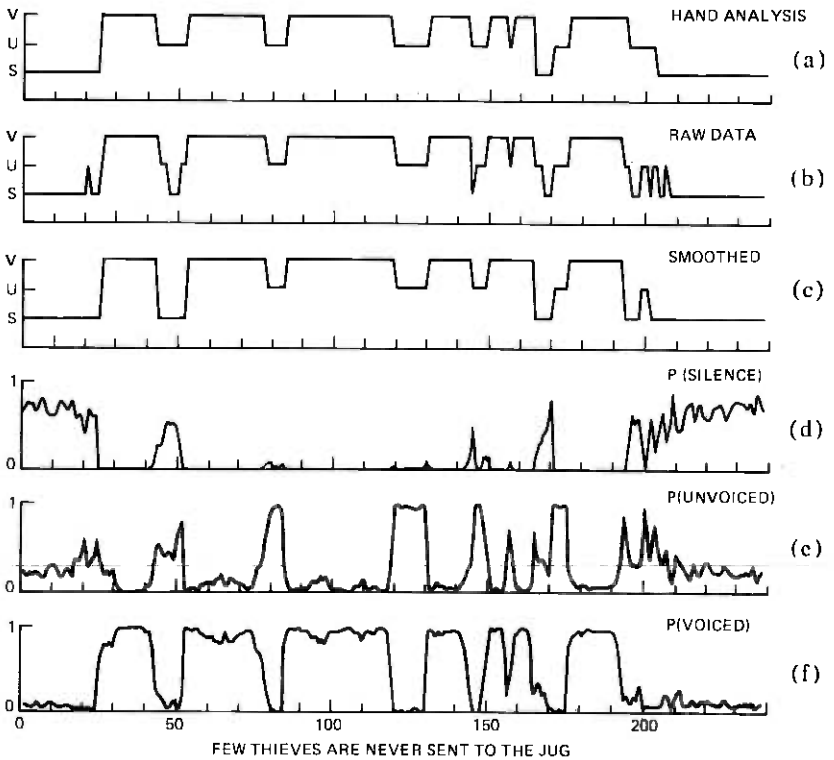


Fig. 10—The analysis results for the utterance, "Few Thieves are Never Sent to the Jug," using optimal features from TS1 with weight matrix WM1.

for all values of D_s, D_u, and D_v. Furthermore, the probabilities satisfy the relation

$$\lim_{Da \to 0} P(a) \to 1 \tag{22}$$

and

$$\lim_{D_a \to \infty} P(a) \to 0. \tag{23}$$

Thus, as the distance increases, the probability measures decrease.

Contrasting the silence-unvoiced-voiced contours of Figs. 10 and 11, the following observations can be made:

(*i*) The results obtained using features derived from matrix WM4 essentially never classified frames as silence. Instead all silence frames

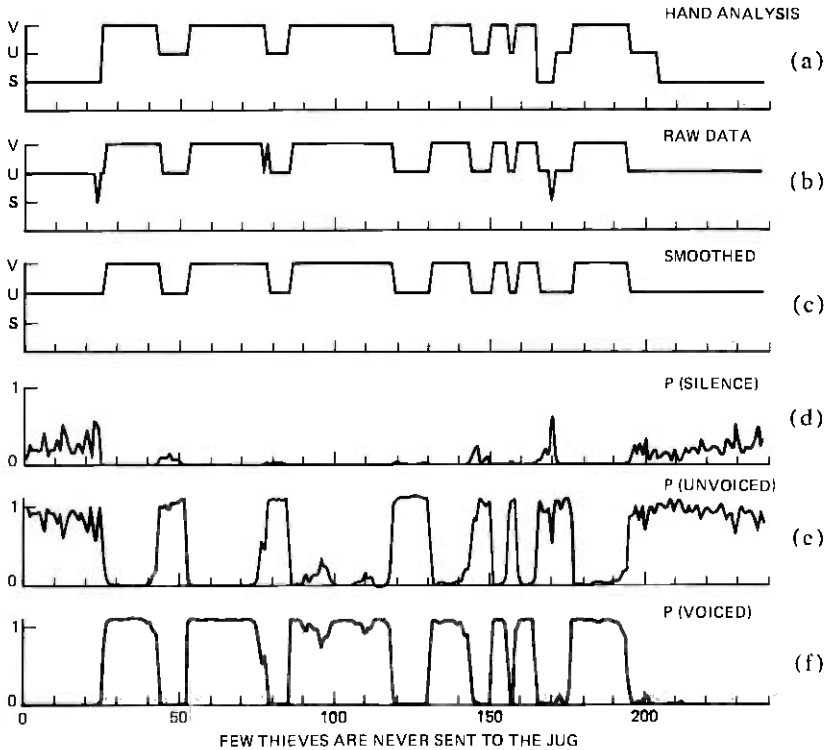


Fig. 11—The analysis results for the same utterance as Fig. 10 using optimal features from TS1 with weight matrix WM4.

were classified as unvoiced, consistent with the zero weight given to this type of error.

(*ii*) Both sets of results contain only a small number of misclassifications of voiced intervals. All but one of these voiced misclassifications occurred at boundaries between voiced and nonvoiced speech.

(*iii*) The probability measures for voiced speech using features derived from WM4 were somewhat higher throughout the voiced regions than corresponding results derived from WM1 features. This indicates that a somewhat better feature set for voiced sounds is obtained at the tradeoff of the high error rate for silence-to-unvoiced errors (and vice versa).

Results similar to those discussed above have been obtained for a wide variety of utterances tested on the system using these sets of features. It is concluded that if one is willing to forego any attempt at distin-

guishing between unvoiced sounds and silence, then reliable voiced-nonvoiced decisions can be obtained over telephone lines.

V. SUMMARY

Through a series of fairly extensive tests, we have investigated quite thoroughly the potential of a fairly sophisticated silence-unvoiced-voiced classification system. We have shown that, depending on the weight attached to various types of misclassifications, a set of optimal features can be found that minimizes the weighted misclassification error rate. For telephone line inputs, the results showed that reliable discrimination between silence and unvoiced sounds is quite difficult; however, reliable discrimination between voiced and nonvoiced sounds (silence or unvoiced speech) can be achieved at error rates fairly close to those obtained with wideband input signals.

Extensive testing of the optimal feature sets obtained from the analysis showed the method to be reliable enough for use in several typical applications in the area of man-machine communication by voice.[9,10]

One aspect of the analysis system which was not varied was the distance metric used in the final classification. Although the non-Euclidean distance metric is a very powerful one for the features studied, other distance metrics have been proposed based on fixed parameter sets, such as the LPC parameters, etc.[3,11] Investigations into the applicability of such distance metrics to the silence-unvoiced-voiced classification problem are currently in progress.

REFERENCES

1. B. S. Atal and L. R. Rabiner, "A Pattern-Recognition Approach to Voiced-Unvoiced-Silence Classification With Applications to Speech Recognition," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 3 (June 1976), pp. 201–212.
2. L. R. Rabiner, M. J. Chang, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 5 (October 1976), pp. 399–418.
3. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-23*, No. 1 (February 1975), pp 67–72.
4. B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., *55* (June 1974), pp 1304–1312.
5. J. Burg, "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands, 1968.
6. J. Makhoul, "New Lattice Methods for Linear Prediction," Proceedings 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, April 1976, pp 462–465.
7. M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-23*, No. 2 (April 1975), pp 176–182.
8. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans Acoust., Speech, and Signal Process., *ASSP-23*, No. 6 (December 1975), pp 552–557.

9. L. R. Rabiner and M. R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 2 (April 1976), pp 170–182.
10. J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE, *64*, No. 4 (April 1976), pp 405–415.
11. A. H. Gray, Jr., and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech, and Signal Process. *ASSP-24*, No. 5 (October 1976), pp 380–391.

Contributors to This Issue

Bishnu S. Atal, B.S. (Honors) (Physics), 1952, University of Lucknow; Diploma, 1955, (Electrical Communication Engineering) Indian Institute of Science, Bangalore; Ph.D. (Electrical Engineering), 1968, Polytechnic Institute of Brooklyn; Bell Laboratories, 1961—. From 1957 to 1960, Mr. Atal was a Lecturer in Acoustics at the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore. At Bell Laboratories, he has worked on a wide range of topics in acoustics, such as computer simulation of sound transmission in rooms, new measurement techniques for concert halls, fading in mobile radio, automatic speaker recognition, and speech coding. More recently, his research interests have centered on new methods for analysis and synthesis of speech signals. He is the author of a number of technical papers in architectural acoustics and speech communication. Fellow, Acoustical Society of America.

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D. (E.E.), 1967, Imperial College, London; Bell Laboratories, 1967—. Mr. Goodman has studied various aspects of digital communications, including analog-to-digital conversion, digital signal processing, assessment of the quality of digitally coded speech, and error mechanisms in digital transmission lines. In 1974 and 1975, he was a Senior Research Fellow at Imperial College, London, England. Member, IEEE.

Nuggehally S. Jayant, B.Sc. (Physics and Mathematics), 1962, Mysore University; B.E., 1965, and Ph.D., 1970 (Electrical Communication Engineering), Indian Institute of Science, Bangalore; Research Associate at Stanford University, 1967–1968; Bell Laboratories, 1968—. Mr. Jayant was a Visiting Scientist at the Indian Institute of Science January–March 1972 and August–October 1975. He has worked on coding for burst error channels, detection of fading signals, statistical pattern discrimination, spectral analysis, and problems in adaptive quantization and prediction, with special reference to speech signals.

K. C. Knowlton, B.E.P., 1953, and M.S. (Engineering Physics), 1955, Cornell University; Ph.D. (Communication Sciences), 1962, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Knowlton is a member of the Speech and Communication Research Department at Bell Laboratories. From 1971–1972, he was Visiting Lecturer of computer graphics and computer art at the University of California (Santa Cruz). He has specialized in the area of computer graphics and has also been interested in the relationship of computer technology to cinematography. He is the author of the following programming languages: L^6, BEFLIX, TARPS, and a coauthor of ATOMS. Member, EXPLOR, ASFA, ACM, AAAS.

Stuart P. Lloyd, B.S., 1943, University of Chicago; Ph.D., 1951, University of Illinois; Institute for Advanced Study, 1952; Bell Laboratories, 1952—. Mr. Lloyd has worked on problems in information theory, ergodic theory, and other branches of probability theory. He is presently concerned with shift invariant coding of information sources. Member, American Mathematical Society, Mathematical Association of America, Institute of Mathematical Statistics, AAAS.

Dietrich Marcuse, 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 working in transmission research and the study of coaxial cables and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of three books. Adjunct professor at the University of Utah; topical editor for integrated and fiber optics for the Optical Society. Fellow, IEEE; member, Optical Society of America.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics, 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964–. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematical Research Center. Member, American Physical Society, IEEE.

S. D. Personick, B.E.E., 1967, City College of New York; S.M., 1968, and Sc.D. 1969, Massachusetts Institute of Technology; Bell Laboratories, 1967—. Mr. Personick is engaged in the research and development of optical fiber transmission systems. He is presently supervisor of the fiberguide repeater group.

Lawrence R. Rabiner, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. Coauthor, *Theory and Application of Digital Signal Processing.* Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; former President, IEEE G-ASSP Ad Com; member, G-ASSP Technical Committee on Digital Signal Processing, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Technical Committee on Speech Communication of the Acoustical Society; former Associate Editor of the G-ASSP Transactions.

Carolyn E. Schmidt, B.S. (Mathematics), 1974, Lafayette College; Bell Laboratories, 1974—. Miss Schmidt is a member of the Acoustic Research Department and is currently involved in work on speech communications. Member, Phi Beta Kappa.

Ralph A. Semplak, B.S. (Physics), 1961, Monmouth College, N.J.; Bell Laboratories, 1955—. Mr. Semplak, a member of the Antennas and Propagation Research Department, has done research on microwave antennas and propagation. He participated in the Project Echo and *Telstar®* communications satellite experiments. He currently is concerned with the attenuation effects of rain on propagation at 18.5 and 30.9 GHz. Member, Sigma Xi and Commission F of International Scientific Radio Union (URSI).

Raymond Steele, B.Sc. (E.E.), 1959, University of Durham, England; Ph.D., 1975 Loughborough University of Technology, England. Mr. Steele was a lecturer at Royal Naval College, Greenwich, London from 1965 to 1968 when he became senior lecturer at Loughborough University. Mr. Steele has been engaged in source encoding of speech and picture signals and is the author of a book on delta modulation systems. He served as a consultant at Bell Laboratories from July through October, 1975.