THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 55 November 1976 Number 9

An Optical Apparatus for Very-Small-Angle Light Scattering—Design, Analysis and Performance

By J. B. LASTOVKA

(Manuscript received April 4, 1976)

We describe an optical apparatus designed and built to extend conventional light-scattering measurements to the very-small-angle regime. The present instrument covers the angular range $0.003° \leqq \theta \leqq 0.15°$ with an instrumental resolution (HWHM) of 0.00045° (1.6 arc seconds), and exhibits an exceptionally low stray-light background. The theoretical and practical considerations important in achieving this performance are analyzed in detail. Besides its primary purpose of studying long-wavelength (0.01 cm to 1 cm) thermally driven fluctuations, the present type of apparatus should also prove quite useful in other areas where long-wavelength perturbations must be probed, such as, (i) holographic and optical memory imaging, (ii) surface roughness testing, and (iii) index of refraction profiling.

I. INTRODUCTION

Laser light scattering has, over the past decade, been developed[1–3] into an extremely powerful tool for probing the long-wavelength ($\Lambda_f \approx 2 \times 10^{-5}$ cm to 2×10^{-3} cm) elementary excitations of liquids, gases, and solids. Combined with diffraction grating, Fabry-Perot, or optical mixing spectrometers the technique is capable of spanning an impressive range of more than 13 decades in energy or frequency measurement. Yet, in contrast to what has happened in the field of inelastic X-ray scattering,[4,5] very little has been done to utilize very-small-angle (VSA) light scattering to probe longer-wavelength ($\Lambda_f \approx 10^{-3}$ cm to 1 cm) excitations. With a few notable exceptions,[6–11]

most light-scattering experiments have been limited to the scattering-angle range $\theta > 1°$.

There have been a number of reasons for this apparent lack of progress in the very-small-angle scattering regime. On the one hand, experimentalists in the field, encountering a seemingly divergent stray-light level at small angles, have assumed that attempts to work in the VSA region would present insurmountable problems. On the other hand, there did not appear to be any physical phenomena where the important elementary excitations were confined to the corresponding longer-wavelength regime. Or, in cases where they were, it seemed that the use of more conventional macroscopic experimental techniques represented a satisfactory experimental approach.

Recently, however, there has been a resurgent interest in problems involving general hydrodynamic instabilities[12] both in normal liquids and liquid crystals.[13–26] The "critical wavelengths" involved in the onset of these instabilities are, in general, controlled by some macroscopic dimension of the sample chamber and tend to fall in the range $100\ \mu m < \Lambda_c < 1$ cm. Light scattering is the only technique offering the possibility of probing these wavelengths without physically disturbing the sample and with a sensitivity sufficient to detect the thermally driven critical fluctuations. However, probing the excitation wavelength region $100\ \mu m < \Lambda < 1$ cm requires the capability of resolving and detecting the scattered light at very small angles, $0.3° \geqq \theta \geqq 0.003°$.

This paper describes the experimental progress which has been made in extending the light-scattering technique to this very-small-angle, long-wavelength regime.

In Section II, we describe the physical configuration of a light-scattering apparatus that has been constructed for use in the VSA region. This section also summarizes the measured performance characteristics of the instrument in terms of angular resolution and stray light. Section III is a detailed presentation of the basic diffraction and aberration considerations that influence the design of a VSA light-scattering apparatus. Section IV outlines various empirical observations made during the course of construction of the present instrument, relating to the stray-light behavior of optical components at small angles.

II. AN APPARATUS FOR VERY-SMALL-ANGLE LIGHT SCATTERING

2.1 Introduction

In this section, we present a general description of the physical layout and performance of a light-scattering apparatus that has been constructed for the VSA regime. The theoretical background and practical considerations necessary to analyze the detailed characteristics of the instrument are deferred to Sections III and IV. Although

designed specifically for the study of the Bénard convective instability, this apparatus embodies solutions to most of the problems to be encountered in the general small-angle light-scattering experiment.

2.2 Performance goals

The following performance goals were established for the present instrument and evaluated at the various stages of construction and modification:

(*i*) The ability to make quantitative measurements of both the scattered intensity and the temporal intensity autocorrelation function for scattering angles ranging from a few mrad down to at least 50 μrad. (We will, in general, specify angular deflections in μrad; Table I lists conversion factors to other common units of angular measure.)

(*ii*) A stray-light level per coherence solid angle in the scattered field ($d\mathcal{P}_{sl}/d\Omega_{COH}$) that was less than 10^{-6} of the incident beam power.

(*iii*) An angular instrumental resolution of less than 15 μrad.

(*iv*) The capability of continuously scanning the instrument over a reasonable range in scattering angle without the need for realignment.

(*v*) The attainment of near-diffraction-limited performance using customary spherical optics of reasonable cost.

Taken individually, each of the above goals can be met or bettered by existing optical instruments. To cite just two examples, the 200-inch Mount Palomar telescope has a diffraction-limit angular resolution of about 0.1 μrad; and, in a typical $\theta = 90$ degrees light-scattering experiment, the desired stray-light level would be considered a straightforward achievement. Insofar as the angular range is concerned, we can easily show that the scattered light observed at these angles is contributed by plane-wave components of the refractive-index per-

Table I — Conversion factors between various common units of angular measure

	Deg	Rad	mrad	μrad	Arc Min.	Arc Sec
1 Deg	1	0.0174	17.45	17,453	60	3600
1 Rad	5.73	1	10^3	10^6	3438	2.06×10^5
1 mrad	0.0573	10^{-3}	1	10^3	3.438	206.3
1 μrad	5.73×10^{-5}	10^{-6}	10^{-3}	1	0.0034	0.2063
1 arc min.	1/60	2.91×10^{-4}	0.291	291	1	60
1 arc sec.	1/3600	4.85×10^{-6}	4.85×10^{-3}	4.848	1/60	1

turbations in the sample whose wavelengths, Λ, are given by the small-angle Bragg condition

$$\Lambda = \lambda_0/\theta, \tag{1}$$

where λ_0 is the incident-beam wavelength. Therefore, probing the scattering-angle range from 50 μrad to 3 mrad gives information about Fourier components of the refractive index having wavelengths between 1.0 cm and 0.016 cm, respectively. Here we can point out that this spatial-frequency region is routinely examined by common interferometric checking methods and holographic techniques.

The instrument described in this paper is unique in that it meets *all* of the performance criteria simultaneously. In being able to probe perturbations with wavelengths as long as 1.0 cm, it represents a 100-fold improvement on previous low-stray-light-level scattering instrumentation, while its small stray-light background gives it a 1000-fold sensitivity advantage over conventional interferometric and holographic equipment. On a per-unit-aperture-size basis, its ability to resolve closely spaced faint (10^{-6}) and strong (1) features is about 50 times better than the Mount Palomar telescope.

2.3 Optical components and physical configuration of the instrument

Figure 1 sketches the optical configuration of the most recent version of the apparatus designed to meet the performance criteria set out in the preceding paragraphs. For brevity, we refer to this particular optical system as the MK VI instrument.

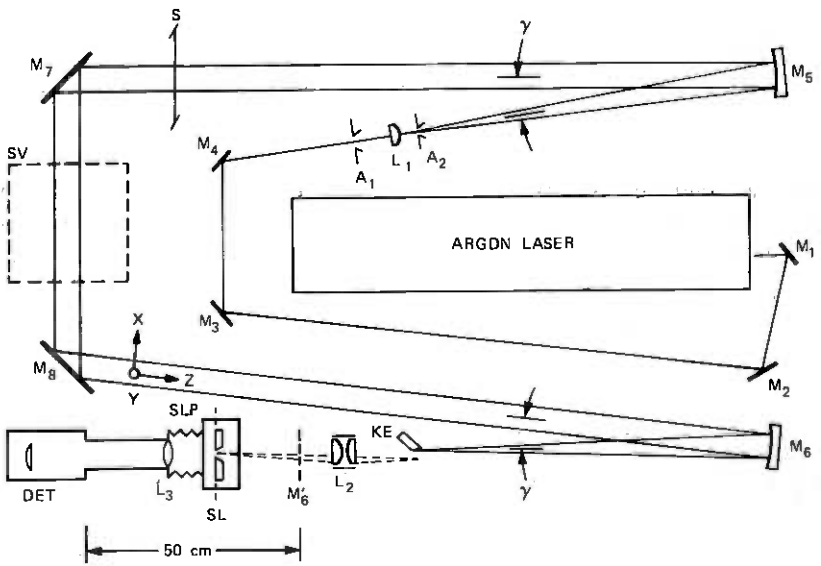


Fig. 1—Optical component layout of the MK VI small-angle-scattering instrument. Component sizes and spacings are shown approximately to scale.

Before discussing the specific function of the various elements of the spectrometer, we present below, for reference purposes, a brief description of each of these elements and their mounting following the identification scheme used in Fig. 1. Whenever spatial or angular displacements are specified, they are to be interpreted according to the conventions illustrated in Fig. 2. The $\hat{x}$ (or θ) and $\hat{y}$ (or φ) axes are taken to be mutually orthogonal cartesian (angular) coordinates perpendicular to the axial ray at the point in question. The $\hat{x}(\theta)$ direction will always lie in the plane of Fig. 1, the instrument's tangential plane, while $\hat{y}(\varphi)$ will denote the vertical or sagittal plane. The direction of beam travel defines the local $\hat{z}$ axis. The basic hardware components of the MK VI instrument are the following:

A_1—An adjustable circular diaphragm stop with an aperture diameter $d_{A1} \approx 5$ mm.

A_2—A fixed, precision-pinhole aperture, $d_{A2} = 100\ \mu$m. A_2 is mounted with $\hat{x}$ and $\hat{y}$ vernier adjustments relative to L_1.

Argon ion laser—The laser is normally adjusted to provide between 50 mW and 200 mW of output at either $\lambda_0 = 5145$ Å or $\lambda_0 = 5017$ Å. The laser used has a flat-long radius spherical resonator, placed at about $\frac{1}{4}$ hemispherical spacing, and oscillates in TEM_{00} modes only. The output is a well-collimated beam with a slight spheroidal distortion. The beam has a gaussian intensity profile with a diameter of 1.4 mm as measured to the $1/e^2$ points.

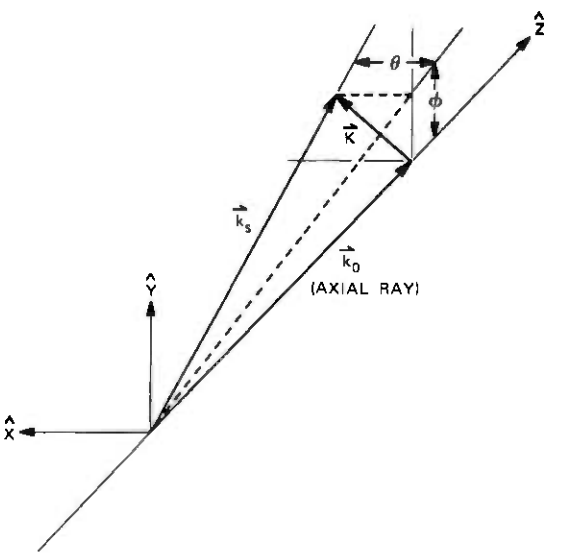


Fig. 2—Cartesian coordinate system for (x, y, z) showing the cartesian angular deflection θ and φ.

DET—A silicon-diode photodetector. The diode used is a photovoltaic device operated without bias as a current source. The detector has a 1-cm-diameter active area.

KE—A precision knife-edge custom-fabricated from neutral-density "black glass" plate. The 40-mm-long edge is straight to within 1 μm and nick free.

L_1—Plano convex achromat with a focal length $f_{L1} = 132$ mm. L_1 and A_2 share a common mount with vernier $\hat{y}$ and $\hat{z}$ degrees of freedom.

L_2—An anastigmatically mounted pair of plano convex achromats with an effective focal length $f_{L2} = 94.77$ mm. L_2 has $(\hat{x}, \hat{y}, \hat{z}, \hat{\theta}, \hat{\varphi})$ vernier adjustability.

L_3—An achromatic lens having $f_{L3} = 150$ mm and a mounted free aperture diameter of 35 mm.

M_1, M_2, M_3, M_4—Flat mirrors $1\frac{1}{2}$ inches in diameter with $\lambda/10$ surface figure.

M_5, M_6—Dielectrically coated, concave, spherical mirrors fabricated of fused quartz. They have a radius of curvature of 2 m and a surface conformity of $\lambda/10$. The mounted free aperture is 6.5 cm in diameter.

M_7, M_8—Aluminized, first-surface, fused-quartz, flat mirrors. They have a mounted free aperture of 13 cm and a surface figure of $\lambda/20$.

S—A bilateral slit with straight jaws that can be used to reduce the $\hat{y}$ dimension of the probe beam.

SL—A commercial, precision, bilateral slit. The jaws have a 50 mm usable height and an accurately adjustable opening range from 3 μm to 3 mm. The slit assembly is mounted on a precision x-z translational stage positioned by large-barrel micrometer heads with a maximum conforming error of about 1 μm. The $x(\theta)$-axis micrometer can be manually positioned or can be driven by a digitally controlled stepping motor.

SV—The location of the scattering sample.

These optical components are mounted on a 3-inch-thick aluminum slab that forms a stable base for the instrument. Because random laboratory air currents and temperature gradients can cause angular beam deflections comparable to the instrumental resolution, the entire apparatus is covered by an essentially air-tight Plexiglas* enclosure.

2.4 Functional description of the apparatus

We can most easily describe the basic optical characteristics of the instrument by following the beam path through the system starting at the laser source.

* Registered trademark of Rohn & Haas Company.

Mirrors M_1, M_2, M_3, and M_4 steer the laser output beam around to a spatial filter assembly comprised of A_1, L_1, and A_2. Lens L_1 and pinhole A_2 form the conventional spatial filter arrangement, while the pre-aperture A_1 serves to block high-angle-beam trash, such as multiple reflections in the laser resonator output mirror. The long path length through M_1–M_4 and aperture A_1 also provides a significant reduction in laser tube discharge light that would otherwise pass through the system.

The spherically spreading wave coming from A_2 is recollimated off-axis by M_5. The collimated beam leaving M_5 has a diameter* $D(1/e)$ of approximately 1.65 cm. The wave-front planarity of this beam is measured and adjusted using a wave-front shearing interferometer aligned to give a 7-mm shear in the tangential plane. The tangential direction wave-front curvature is reduced to less than $\lambda/8$ over the beam aperture by translating the spatial filter assembly along the laser beam ($\hat{z}$) axis. It is important to note that the use of this off-axis collimation scheme produces a large amount of astigmatism and tangential plane coma. As a result, *it is not possible to make the probe-beam wave fronts straight in both the $\hat{x}$ and $\hat{y}$ directions simultaneously.*† The alignment procedure just described is intended to give diffraction-limited angular resolution in the $\hat{\theta}$ plane with some sacrifice in $\hat{\varphi}$ direction resolution.

The collimated probe beam is now sent to the scattering object at SV via the flat mirror M_7. Flat mirror M_8 collects the transmitted beam and small-angle scattered light and directs them to M_6.

In the tangential focal plane of spherical mirror M_6, the directly transmitted beam is brought to a vertical ($\hat{y}$) line focus at a position we define as $x_{KE} \equiv 0$. Light that has been scattered by some angle θ is brought to line focus in the same plane, but at a displaced transverse position

$$x_{KE}(\theta) = f_{M6} \tan \theta \approx f_{M6}\theta, \tag{2}$$

where $f_{M6} = 100$ cm is the focal length of M_6. Therefore, for sufficiently small values of θ, where $\tan \theta \approx \theta$, angular deflection maps linearly into lateral displacement at the focus with a position-angle dispersion (PAD) constant given by

$$PAD(KE) = \frac{x_{KE}(\theta)}{\theta} = f_{M6} = 1\ \mu\text{m}/\mu\text{rad}. \tag{3}$$

The knife-edge KE is located in this tangential focal plane with its edge vertical and can be set to intercept the transmitted beam at

---

* See Section 3.1 for the definition of these quantities.
† See Section 3.2.

$\theta = 0$ to prevent it from entering the remaining portion of the optical system. In normal practice, KE is adjusted to occult all light for which $\theta \leqq 30$ to 50 μrad. The position and orientation of KE relative to M_6 is fixed with diffraction-limited accuracy using the standard Foucault knife-edge test procedure.

Lens pair L_2 re-images the focal plane of M_6 onto the vertically oriented main receiving slit SL with a magnification of about (2.54). Therefore, the dispersion constant in the slit plane has the value

$$PAD(SL) = 2.54\ \mu\text{m}/\mu\text{rad} \tag{4}$$

or

$$PAD(SL) = 0.0001\ \text{in.}/\mu\text{rad}. \tag{5}$$

The magnification by L_2 allows the scattering angle θ to be read directly on the "english-units" micrometer that positions the slit. More importantly, it relaxes the stability and accuracy requirements that must be imposed on the slit scan mechanism. Since SL has a minimum opening setting of roughly 3 μm, the slit-limited angular resolution is about 1 μrad.

The proper locations and orientations for the main slit SL and lens L_2 are determined by an iterative procedure in which one of the jaws of SL and the *image* of the knife-edge formed by L_2 at the slit plane are positioned to form an apparent two-jawed slit. The absence of distortion in the Fraunhofer diffraction pattern formed when this "slit" is illuminated by a collimated beam becomes a diffraction-limited test for correct lens and slit alignment.

The scattered light passed by the main slit is collected by L_3 and sent to the photodiode DET. The focal length and position of L_3 are chosen such that the real image of the limiting aperture of M_6 formed at the plane M_6' by lens L_2 is re-imaged onto the detectors active area.

2.5 Observed angular-resolution performance

We can assess the θ direction angular resolution of the MK VI apparatus from measurements of intensity as a function of slit position (x) in the absence of a scattering object. Two such "instrumental profiles" are shown in Fig. 3. The ordinate scale is logarithmic in the detector photocurrent with a rough correspondence of 200 μA/mW of optical power. Curves A and B were taken under identical conditions except for the position of the knife-edge KE. For curve A, the knife-edge was withdrawn to allow the direct probe beam to reach the scanning slit, while for curve B, it was positioned to occult all light in the region $\theta \gtrsim 50$ μrad. Note that the use of the knife-edge provides a significant decrease in observed stray-light level, the reduction amounting to about an order of magnitude improvement for $\theta \gtrsim 600$ μrad (see Section IV).

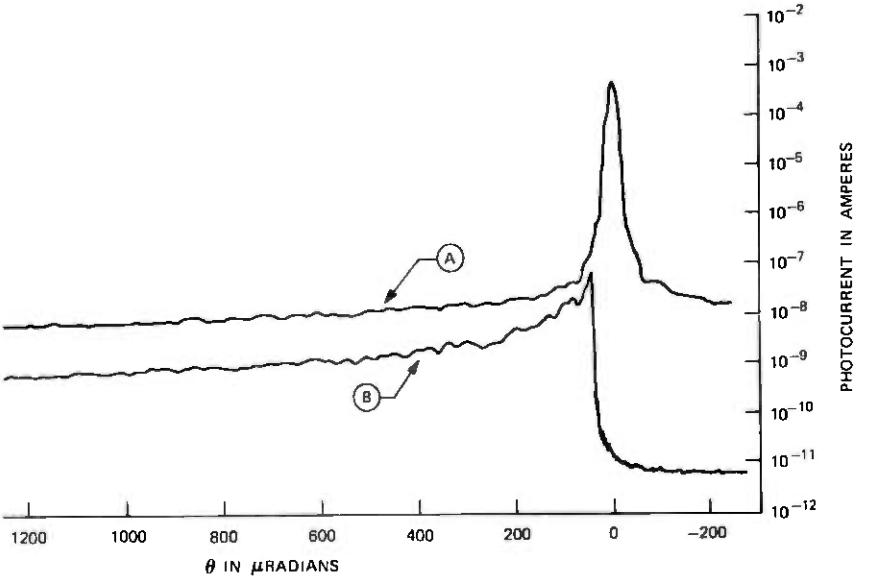


Fig. 3—Observed instrumental profiles for the MK VI apparatus plotted on a logarithmic intensity scale. Curve B was obtained with the knife-edge, KE, occulting the direct beam, while curve A was measured with KE retracted.

An expanded view of the $\theta \cong 0$ region of Fig. 3 is shown in Fig. 4. The dashed curve represents a best fit of the gaussian $\exp - [\theta^2/\delta\theta^2(1/e)]$ to the instrumental line shape, as detailed in Section 3.1. The full width at half-maximum of the fitted curve is

$$\Delta\theta(\tfrac{1}{2}) = 16\ \mu\text{rad}. \tag{6}$$

For the traces shown in Figs. 3 and 4, the main slit width was set at 5 μm which, from eq. (5), is equivalent to a 2-μrad acceptance angle. Under these conditions, the effect of artificial slit broadening on the line shape may be neglected, as outlined in Appendix A.

Deriving a value for the sagittal, or $\hat{\varphi}$ direction, resolution is a more complicated procedure because of the large instrumental astigmatism (see Section 3.2). However, a pragmatic number can be given using the following operational definition. If the *sagittal* resolution were measured in the *tangential* focal plane of M_6, the location of the main slit, the instrumental profile would have a full width at half maximum given by

$$\Delta\varphi(\tfrac{1}{2}) = 93\ \mu\text{rad}. \tag{7}$$

(See Section 3.2, especially Figs. 16 and 17.)

The overall angular resolution characteristics of the instrument are illustrated in Fig. 5. This sketch shows various fraction-of-maximum-intensity contours for the instrumental profile as determined at the main slit plane.

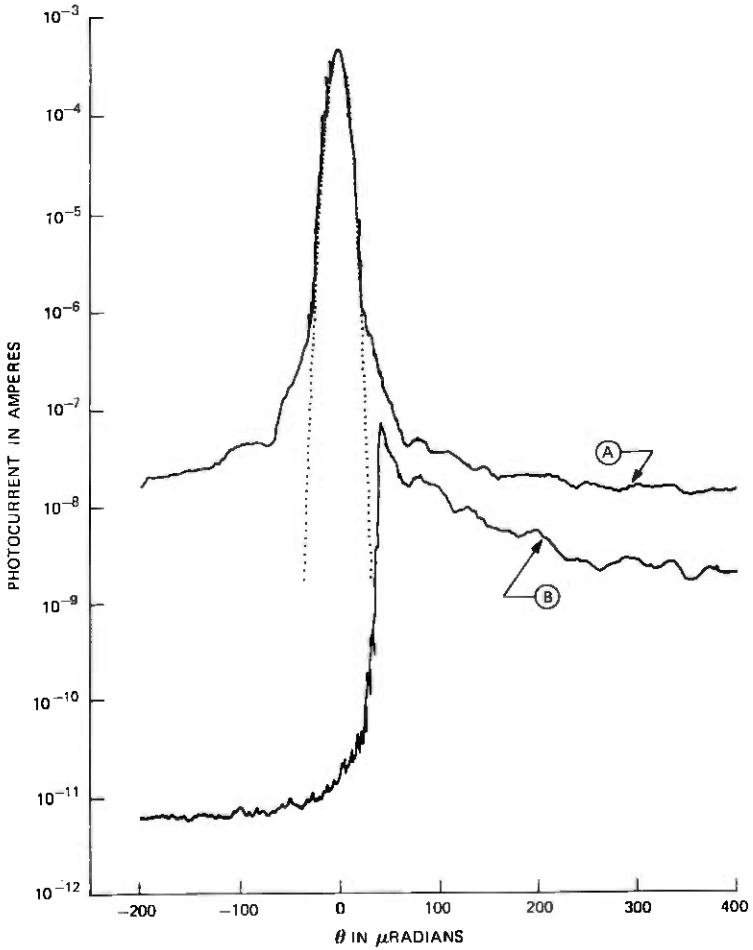


Fig. 4—Measured instrumental profiles for the MK VI apparatus in the region around $\theta = 0$. Curve B was obtained with the knife-edge, KE, occulting the direct beam, while curve A was measured with KE retracted. Dotted curve is best fit of the function $\exp[-\theta^2/\delta\theta^2(1/e)]$ to the transmitted beam profile.

The measured profiles presented in Figs. 3 and 4 and the corresponding contours of Fig. 5 were obtained using the full ϑ-axis beam height of the instrument, that is, in the absence of aperturing of the probe beam by slit S of Fig. 1. As such, the quoted $\Delta\varphi$ resolution does not include any diffraction broadening associated with ϑ direction vignetting of the main beam. At full aperture, the instrument's θ resolution is essentially diffraction limited, while the φ resolution is dominated by astigmatic blurring. However, as the beam height is stopped down, diffraction spreading will eventually override the astigmatism and the instrument will be solely diffraction limited. For

the MK VI, this crossover point occurs at a beam height of about 0.5 cm or roughly $\frac{1}{10}$ of the full design aperture. Therefore, the most advantageous use can be made of the present apparatus when the desired probe-beam geometry consists of a collimated "sheet" or ribbon illumination.

2.6 *Analysis of stray-light performance*

A second crucial performance characteristic of any light-scattering instrument is its stray-light level in relation to the scattering efficiency of the sample under investigation. For the MK VI instrument, the ratio of recorded stray-light photocurrent to the photocurrent observed at the peak of the transmitted beam, say, can be read directly from Figs. 3b and 4b; however, this ratio is not of immediate physical

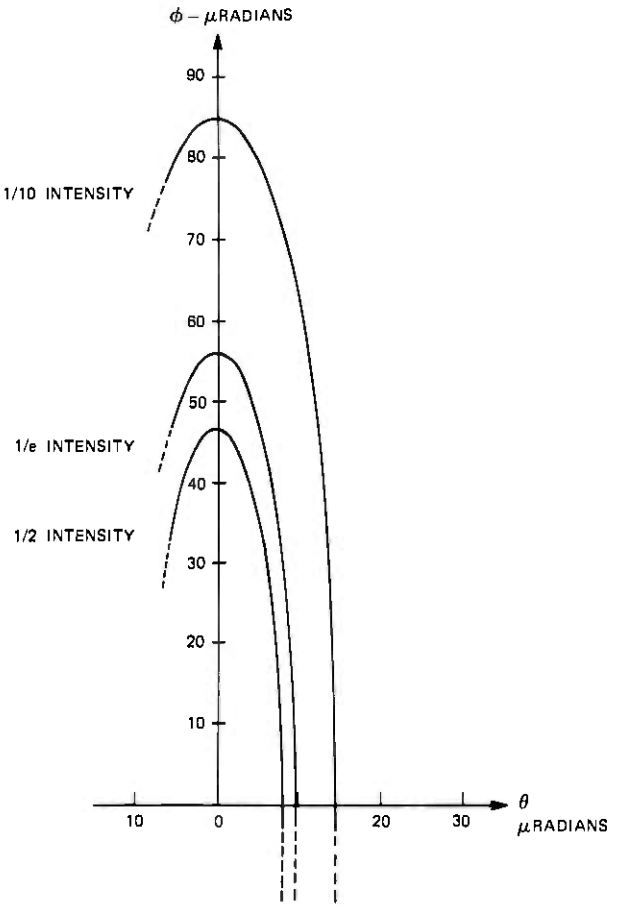


Fig. 5—Contours of constant intensity for the instrumental line shape. Each contour is labelled in terms of a fraction of the peak intensity, $I(\theta = 0, \varphi = 0)$.

significance because of the effects of residual instrumental astigmatism. The directly measured $i(\theta)/i(0)$ ratio requires a certain amount of mathematical interpretation to provide the stray-light ratio values that will be relevant in signal-to-noise or observability calculations.[1–3,27] In general, the quantities that are most important in this regard are:

(*i*) The scattered power per unit solid angle divided by the incident probe-beam power:

$$\frac{1}{\mathcal{P}_0}\frac{d\mathcal{P}_s(\theta,\ \varphi)}{d\theta d\varphi} \equiv \frac{d\mathcal{R}_s(\theta,\ \varphi)}{d\Omega}. \tag{8}$$

The ratio $[d\mathcal{R}_s(\theta,\ \varphi)/d\Omega]$ is a frequently used measure of the scattering power of an object; calculating a value of this ratio appropriate to the *stray* light $[d\mathcal{R}_{st}(\theta,\ \varphi)/d\Omega]$ provides a basis for estimating the observability of a particular scattering feature. This quantity can be extracted more or less directly from $i(\theta)/i(0)$ given (1) the effective solid angle subtended by the main slit and detection optics, and (2) the instrumental profile contours of Fig. 5.

(*ii*) The normalized scattered power per scattering normal mode:

$$\frac{\mathcal{P}_s(\mathbf{K}_j)}{\mathcal{P}_0}. \tag{9}$$

This latter quantity appears in scattered intensity calculations in which the index-of-refraction perturbations in the illuminated scattering volume are represented in terms of an orthonormal plane wave Fourier expansion.[3,27] The mean-square amplitude of these modes and their scattering efficiency are, in general, easily calculated from the known physical properties of the sample. Characterizing the stray-light via a ratio $\mathcal{P}_{st}(\mathbf{K}_j)/\mathcal{P}_0$ provides another convenient way of determining the observability of the scattering from a particular sample object.

(*iii*) The normalized scattered power per coherence solid angle in the scattered field:

$$\frac{1}{\mathcal{P}_0}\frac{d\mathcal{P}_s(\theta,\ \varphi)}{d\Omega_{COH}} = \frac{d\mathcal{R}_s(\theta,\ \varphi)}{d\Omega}\Omega_{COH}. \tag{10}$$

This quantity appears in signal-to-noise ratio calculations relevant to determining the spectrum of the scattered light from measurements of the temporal autocorrelation function or spectrum of the detected photocurrent.[2,3,27] In this case, it is useful to also characterize the stray-light level in terms of the quantity $[d\mathcal{R}_{st}(\theta,\ \varphi)/d\Omega_{COH}]$. The ratio $[d\mathcal{R}_s/d\Omega_{COH}]$ differs from that defined in eq. (8) in that the solid angle is specified as being the solid angle of spatial coherence in the scattered field, Ω_{COH}. The coherence solid angle is a measure of the range in θ and φ about some arbitrary reference direction $(\theta,\ \varphi)$ over which the amplitude and/or phase of the scattered electric field exhibits statistically correlated behavior. In the typical light-scattering experi-

ment, the extent of spatial coherence in the scattered field is controlled by the geometry of the scattering sample and wave diffraction. In a first approximation,[27] the coherence solid angle is just the diffraction solid angle of the scattering source; that is,

$$\Omega_{COH} = \Delta\theta_{COH} \times \Delta\varphi_{COH} \cong \left(\frac{\lambda}{b_\theta}\right)\left(\frac{\lambda}{b_\varphi}\right), \tag{11}$$

where b_θ and b_φ are extremal dimensions of the illuminated sample volume as viewed from a direction specified by (θ, φ). However, for the MK VI instrument, the extent of the spatial coherence is partly determined by residual aberration effects, and an evaluation of the ratio $[d\mathcal{R}_{sl}(\theta, \varphi)/d\Omega_{COH}]$ requires a specific calculation of the spatial-coherence properties of the optical field at the main slit plane.

Obtaining expressions for the various stray-light ratios when aberrations are present requires a rather lengthy detailed analysis, as is carried out in Section 3.3. For our purposes here, we merely quote those results that lead to the numerical ratios appropriate to the MK VI instrument. In each case, the procedure is to treat the observed stray-light level *as if it originated from a fictitious "sample"* placed at the normal position of the scattering volume. After deriving the expressions that relate slit-plane intensity to a real sample's scattering cross section, expressed for example as $[d\mathcal{R}_s(\theta, \varphi)/d\Omega]$, we utilize these results in reverse fashion to calculate the effective cross section of our fictitious stray-light sample. Of course, these expressions derive from the main slit-plane imaging characteristics of the instrument; therefore, in succeeding paragraphs, whenever angles or solid angles are specified, they are to be interpreted as slit-plane coordinates or areas converted to angular units via eq. (5). Consider first the quantity

$$\begin{aligned}\frac{d\mathcal{R}_{sl}(\theta, \varphi)}{d\Omega_{COH}} &= \frac{1}{\mathcal{P}_0}\frac{d\mathcal{P}_{sl}(\theta, \varphi)}{d\theta d\varphi}\Omega_{COH} \\ &= \frac{1}{\mathcal{P}_0}\frac{d\mathcal{P}_{sl}(\theta, \varphi)}{d\theta d\varphi}\overline{\Delta\theta}_{COH}\overline{\Delta\varphi}_{COH},\end{aligned} \tag{12}$$

where $\overline{\Delta\theta}_{COH}$ and $\overline{\Delta\varphi}_{COH}$ are the full-width coherence angles in the θ and φ directions. When the main-slit acceptance angles $\Delta\theta_{SL}$ and $\Delta\varphi_{SL}$ satisfy the inequalities

$$\begin{aligned}\Delta\theta_{SL} &\ll \overline{\Delta\theta}_{COH} = 24.1\ \mu\text{rad} \\ \Delta\varphi_{SL} &\gg \overline{\Delta\varphi}_{COH} = 140.4\ \mu\text{rad},\end{aligned} \tag{13}$$

as they do for the profiles of interest here, the right-hand side of eq. (12) can be expressed in terms of the measured photocurrent, $i(\theta)$, as

$$\frac{d\mathcal{R}_{sl}(\theta, \varphi = 0)}{d\Omega_{COH}} = \frac{i(\theta)}{i(0)} \times \sqrt{2}\,\frac{\overline{\Delta\varphi}_{COH}}{\Delta\varphi_{SL}}, \tag{14}$$

where $i(0)$ is the photocurrent observed at the peak of the direct

transmitted beam. In writing eq. (14), we have assumed that $\Delta\varphi_{SL}$ is symmetrically placed around $\varphi = 0$. The ratio $[\Delta\varphi_{SL}/\Delta\varphi_{COH}]$ is essentially the number of slit-plane coherence areas sampled by the detection optics. For the instrumental profiles shown in Figs. (3) and (4), $\Delta\varphi_{SL}$ was limited solely by the free aperture of lens L_3. Using the proper lens free aperture diameter and the linear dispersion constant given in eq. (5), we find an effective slit-acceptance angle

$$\Delta\varphi_{SL} \cong 1.4 \times 10^4\ \mu\text{rad} = 0.79°.$$

This value of $\Delta\varphi_{SL}$ corresponds to a slit height that samples approximately 100 coherence areas. The ratio $[\Delta\varphi_{SL}/\Delta\varphi_{COH}]$, eq. (14), and the data of Fig. 3 combine to give the $[d\mathcal{R}_{st}(\theta, 0)/d\Omega_{COH}]$ values listed in Table II.

The normalized stray-light power per mode can be found from $[d\mathcal{R}_{st}(\theta, 0)/d\Omega_{COH}]$ by the methods detailed in Section 3.3. The basic procedure involves calculating both the scattered power per coherence area and $\mathcal{P}_s(\mathbf{K}_j)$ from a common starting point to obtain the correction term that relates them. In the present case, the required relationship has the form

$$\frac{d\mathcal{R}_{st}(\theta, 0)}{d\Omega_{COH}} = \frac{\mathcal{P}_{st}(\mathbf{K}_j)}{\mathcal{P}_0} \times \frac{\overline{\Delta\varphi_{COH}}}{(\lambda/b_y)} \frac{\overline{\Delta\theta_{COH}}}{(\lambda/b_x)}, \tag{15}$$

where b_x and b_y are the clear aperture width and height of the instrument and λ is the optical wavelength. The product of the ratios

$$\frac{\overline{\Delta\varphi_{COH}}}{(\lambda/b_y)} \quad \text{and} \quad \frac{\overline{\Delta\theta_{COH}}}{(\lambda/b_x)}$$

is a weighted measure of the number of $\mathbf{K}_j$ modes contributing to the power observed in a single coherence area at the main slit. At the full aperture of the MK VI instrument, $b_x = 5$ cm and $b_y = 5$ cm, the correction factor has the value

$$\frac{\overline{\Delta\varphi_{COH}}}{(\lambda/b_y)} \frac{\overline{\Delta\theta_{COH}}}{(\lambda/b_x)} = 33.6. \tag{16}$$

Table II — Numerical values of various stray-light ratios for the MK VI instrument at selected scattering angles

θ-μrad	$i(\theta)/i(0)$	$d\mathcal{R}_{st}(\theta, 0)/d\Omega_{COH}$	$\mathcal{P}_{st}(\mathbf{K}_j)/\mathcal{P}_0$
50	1.07×10^{-4}	1.5×10^{-6}	4.5×10^{-8}
100	2.82×10^{-5}	4.0×10^{-7}	1.2×10^{-8}
200	9.55×10^{-6}	1.4×10^{-7}	4.0×10^{-9}
500	2.82×10^{-6}	4.0×10^{-8}	1.2×10^{-9}
1000	1.62×10^{-6}	2.3×10^{-8}	6.8×10^{-10}

$\Delta\theta_{SL} = 2$ μrad, $\Delta\varphi_{SL} = 14{,}000$ μrad, $\Delta\theta_{COH} = 24.1$ μrad, and $\Delta\varphi_{COH} = 140.4$ μrad.

Combining eqs. (15) and (16) with the values of $[d\mathcal{R}_{st}(\theta, 0)/d\Omega_{COH}]$ already calculated gives the stray-light-per-mode ratios to be found in Table II.

The numerical values of the various stray-light ratios may be put into perspective by calculating the amplitude of some physical perturbations that would generate a scattered intensity equal to the observed stray-light level. Based on a theoretical analysis of the scattering problem, it may be shown that, for sufficiently small scattering angles, the actual three-dimensional scattering volume can be taken to be equivalent to a two-dimensional phase-object placed normal to the incoming probe beam.[28] In this two-dimensional phase-plate equivalent, the scattering disturbances appear in the form of a spatially varying phase thickness $\psi(x, y)$, which is the line integral of the instantaneous index of refraction encountered by a ray traversing the actual sample at the lateral position (x, y). If $n(x, y, z)$ is the local index of refraction in the actual three-dimensional scattering problem then $\psi(x, y)$ is given by

$$\psi(x, y) = \frac{2\pi}{\lambda_0} \int_0^{L_z} n(x, y, z)dz, \tag{17}$$

where L_z is the length of the illuminated volume along the direction of the incident beam. The phase perturbation $\psi(x, y)$ may be represented in terms of a two-dimensional plane-wave Fourier expansion

$$\psi(x, y) = \sum_{K_x} \sum_{K_y} \tilde{\psi}(\mathbf{K}_j) e^{iK_x x} e^{iK_y y}, \tag{18}$$

with the $\mathbf{K}_j = (K_x, K_y)$ chosen to make the expansion functions orthonormal over the instrument's full aperture. In this formulation of the problem, the normalized scattered power per $\mathbf{K}_j$ mode has the simple form

$$\frac{\mathcal{P}_s(\mathbf{K}_j)}{\mathcal{P}_0} = \tfrac{1}{4}\langle |\tilde{\psi}(\mathbf{K}_j)|^2 \rangle, \tag{19}$$

where the angular brackets denote an appropriate time or ensemble average.

The expression for the scattered power given in eq. (19) may be used to interpret the stray-light levels observed in the MK VI instrument in terms of a minimum detectable amplitude for a specific physical scattering mechanism. In succeeding paragraphs, we consider three such scattering processes: (*i*) static index of refraction modulation in a transparent slab, (*ii*) surface height modulation on a reflecting mirror, and (*iii*) temperature modulation in an otherwise homogeneous liquid.

2.6.1 Refractive modulation in a slab

The scattering from a static sinusoidal refractive-index modulation in a plate is an interesting model problem relevant to holographic

memories and general phase-grating problems. We take the sample object to be a nominally homogeneous plate of thickness L_z and refractive index n_0 in which a small sinusoidal index disturbance

$$\delta n(x, y, z) = \delta n e^{i\mathbf{Q}\cdot\mathbf{r}} \tag{20}$$

has been created. The index perturbation is assumed to be uniform in the $\hat{z}$ direction such that $\mathbf{Q}$ lies in the (x, y) plane. The local index in the plate is

$$n(x, y, z) = n_0\left[1 + \frac{\delta n}{n_0} e^{i\mathbf{Q}\cdot\mathbf{r}}\right] \tag{21}$$

which, when inserted in eq. (17), gives the phase function $\psi(x, y)$ as

$$\psi(x, y) = \frac{2\pi}{\lambda_0}\left[n_0 L_z + \delta n L_z e^{i\mathbf{Q}\cdot\mathbf{r}}\right]. \tag{22}$$

The required Fourier amplitude, $\langle|\bar{\psi}(\mathbf{K}_j)|^2\rangle$, is obtained by inspection from eq. (22) as

$$\langle|\psi(Q)|^2\rangle = \left[\left(\frac{2\pi}{\lambda_0}\right)\left(\frac{\delta n}{n_0}\right) n_0 L_z\right]^2. \tag{23}$$

This result may be used together with eq. (19) to obtain the perturbation amplitude $(\delta n/n_0)$ necessary to produce a given scattered power per mode. For example, taking $L_z = 1$ cm, $n_0 = 1.5$, and $\lambda_0 = 5000$ Å, we find that a refractive-index amplitude $\delta n/n_0 = 1.2 \times 10^{-9}$ yields a normalized scattered power per mode equal to the MK VI's observed stray-light value at $\theta = 100$ μrad. To produce scattering at this angle, the wavelength of the perturbation $\Lambda = 2\pi/|\mathbf{Q}|$ would have to be $\Lambda = (\lambda_0/\theta) = 0.5$ cm. Table III lists the "background equivalent" $\delta n/n_0$ values corresponding to other values of $\theta(\Lambda)$.

2.6.2 Height modulation on a reflecting surface

Another interesting example from the viewpoint of stray-light level comparison is the scattering from a surface height disturbance on an otherwise perfect reflecting mirror. Clearly, this problem can also serve as a model for *calculating* the instrumental background when mirror surface roughness (see Section IV) is the dominant source of stray light.

Since the primary effect of a surface height deviation is to produce a phase perturbation on the reflected wave-front, the phase function $\psi(x, y)$ can be written down immediately as

$$\psi(x, y) = \frac{(2)(2\pi)}{\lambda_0} h(x, y), \tag{24}$$

where $h(x, y)$ gives the local physical height displacement from the nominally perfect geometric surface.

Table III — Amplitudes of three scattering perturbations necessary to scatter an amount of light equal to the instrument's stray-light level

θ (μrad)	Λ (cm)	$\mathcal{P}_s(\mathbf{K}_j)/\mathcal{P}_0$	Refractive Perturbation ($\delta n/n_0$)	Surface Corrugation (δh-Å)	Temperature Perturbation (δT-°C)
50	1.0	4.5×10^{-8}	2.2×10^{-9}	0.16	6.8×10^{-6}
100	0.5	1.2×10^{-8}	1.2×10^{-9}	0.084	3.5×10^{-6}
200	0.25	4.0×10^{-9}	6.7×10^{-10}	0.049	2.0×10^{-6}
500	0.1	1.2×10^{-9}	3.6×10^{-10}	0.027	1.1×10^{-6}
1000	0.05	6.8×10^{-10}	2.8×10^{-10}	0.021	8.4×10^{-7}

We will take $h(x, y)$ to be a small, static, sinusoidal corrugation

$$h(x, y) = \delta h\, e^{i\mathbf{Q}\cdot\mathbf{r}} \tag{25}$$

for which the phase function is just

$$\psi(x, y) = \frac{4\pi\delta h}{\lambda_0} e^{i\mathbf{Q}\cdot\mathbf{r}}. \tag{26}$$

The Fourier amplitude $\langle|\tilde{\psi}(\mathbf{K}_j)|^2\rangle$ follows trivially as

$$\langle|\tilde{\psi}(\mathbf{K}_j)|^2\rangle = \left[\frac{4\pi}{\lambda_0}\delta h\right]^2. \tag{27}$$

Combining eqs. (27) and (19) with the data of Table II gives the "background equivalent" surface corrugation amplitudes listed in Table III. Again, these are the surface amplitudes necessary to yield a normalized scattered power-per-mode equal to the MK VI's stray-light level. For example, when $\Lambda = 2\pi/|Q| = 0.5$ cm, the "background equivalent" corrugation has an amplitude of $\delta h = 0.084$ Å or, in the usual surface-figure parlance,

$$\delta h \cong \lambda/60{,}000.$$

2.6.3 Temperature modulation in a liquid

As a final example, we consider an otherwise homogeneous slab of liquid of thickness L_z on which is impressed a small sinusoidal temperature disturbance,

$$\delta T(x, y, z) = \delta T e^{i\mathbf{Q}\cdot\mathbf{r}}, \tag{28}$$

with Q lying in the (x, y) plane. The calculation of the scattering from such an object is really just a simple extension of the result obtained above for refractive index modulation. The temperature perturbation produces an associated index disturbance that is responsible for the scattering. If the temperature perturbation in eq. (28) is impressed

isobarically, the associated index modulation is simply

$$\delta n(x, y, z) = \left(\frac{\partial n}{\partial T}\right)_P \delta T(x, y, z)$$
$$= \left(\frac{\partial n}{\partial T}\right)_P \delta T e^{i\mathbf{Q}\cdot\mathbf{r}}. \tag{29}$$

Equations (20) through (23) may now be used to obtain the relevant mean-square phase amplitude, namely,

$$\langle |\psi(\mathbf{Q})|^2 \rangle = \left[\frac{2\pi}{\lambda_0}\left(\frac{\partial n}{\partial T}\right)_P L_z\, \delta T\right]^2. \tag{30}$$

Taking $L_z = 1$ cm and $\lambda_0 = 5000$ A and using a typical value for $(\partial n/\partial T)_P$ in liquids, $(\partial n/\partial T)_P = -5 \times 10^{-4}/°\text{C}$, we find the background-equivalent temperature amplitudes listed in Table III.

2.7 Conclusion

In this section, we described the basic features of an optical instrument capable of extending conventional light-scattering measurements to an angular range (50 μrad to 3 mrad) not previously accessible. In addition to a diffraction-limited angular resolution of a few seconds of arc, the MK VI instrument exhibits an exceptionally low stray-light background making it an effective tool for probing small-amplitude-scattering processes. Besides its primary purpose of studying long-wavelength (0.01 cm to 1 cm) thermal fluctuations, the present type of apparatus should prove quite useful in other areas where long-wavelength perturbations must be probed, such as,

(*i*) Holographic and optical memory imaging.
(*ii*) Surface roughness testing.
(*iii*) Index of refraction profiling.

In general, the MK VI offers a sensitivity improvement of a factor of about 1000 over the instrumentation normally used for such measurements.

While we have given a rather broad overview of the apparatus in the present section, we have not attempted to present the fundamental considerations on which the design is based. We refer the reader who is interested in these questions to the remaining sections of this paper.

III. THEORETICAL CONSIDERATIONS IN THE DESIGN OF A VERY-SMALL-ANGLE LIGHT-SCATTERING APPARATUS

3.1 Aperture apodization

In the simplest analysis, the ultimate angular resolution of any optical instrument is limited solely by diffraction. The expression most

widely used to estimate the limiting resolution is the so-called Rayleigh criterion,[29]

$$(\Delta\theta)(\Delta d) \cong \lambda_0. \tag{31}$$

The quantities $\Delta\theta$ and Δd can be interpreted in two ways:

(*i*) If Δd is the diameter of a "collimated" beam, then $\Delta\theta$ is the actual angular spread of the beam imposed by diffraction.

(*ii*) If $\Delta\theta$ is the collection angle for light emanating from an object, then Δd is the smallest spatial detail that can be resolved on that object.

Neglecting for the moment the off-axis features of the actual MK VI instrument, we can duplicate its basic function with the two-lens system sketched in Fig. 6. Applying the Rayleigh criterion to this particular optical arrangement for a collimated beam diameter $b = \Delta d = 5$ cm, and with $\lambda_0 = 5000$ Å, predicts an instrumental angular spread

$$\Delta\theta_{\text{RAYLEIGH}} = 10\ \mu\text{rad}. \tag{32}$$

Unfortunately, taken by itself, this value for $\Delta\theta_{\text{RAYLEIGH}}$ contributes little in the way of a quantitative understanding of the instrument's small-angle performance. In fact, the Rayleigh criterion can be misleading in a number of ways. First, it does not indicate how much light an object would have to scatter to be "visible" when the scattering angle approaches $\Delta\theta_{\text{RAYLEIGH}}$. Second, it implies that the angular diffraction spread can be decreased to an arbitrarily small value by simply increasing the beam diameter $b = \Delta d$. In reality, the presence of unavoidable optical aberrations will always limit the attainable angular resolution. In designing an instrument which is to attain a resolution approaching the diffraction limit, a quantitative approach to the problem is mandatory.

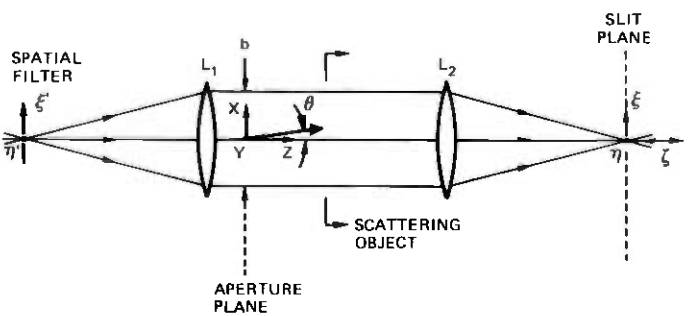


Fig. 6—A simplified "lens equivalent" version of the MK VI optical system.

An exact expression for the diffraction-limit resolution characteristics of an optical system may be obtained as follows. Let us assume we know the spatial dependence of the incident beam intensity on some surface in the optical system, say the aperture plane of Fig. 6. In our case, the electric field on this surface is of the form

$$E_0(x, y)e^{i(k_0 z_0 - \omega_0 t)}. \tag{33}$$

This aperture plane field can be decomposed into a set of infinitely extended plane waves,

$$E_D(\theta, \varphi) = E_D^0(\theta, \varphi)e^{-i(\mathbf{Q}\cdot\mathbf{r} - \omega_0 t)}, \tag{34}$$

propagating toward L_2 at various angles with respect to the $\hat{z}$ axis.* The plane-wave amplitudes, $E_D^0(\theta, \varphi)$, are found from the Fourier integral

$$E_D^0(\theta, \varphi) = \frac{1}{\lambda_0}\int\int_S dxdyE_0(x, y)e^{i(Q_x x + Q_y y)} \tag{35}$$

together with the relations,

$$\begin{aligned} |\mathbf{Q}| &= k_0 = 2\pi/\lambda_0 \\ Q_x &= \mathbf{Q}\cdot\hat{x} = k_0 \sin\theta \cong k_0\theta \\ Q_y &= \mathbf{Q}\cdot\hat{y} = k_0 \sin\varphi \cong k_0\varphi. \end{aligned} \tag{36}$$

Equation (35) is just a slightly modified form of the usual scalar diffraction theory result which utilizes spherically spreading waves as basis functions.[30]

Assume for the moment that lens L_2 in Fig. 6 is infinitely large and free of aberrations. Then each of the plane waves, $E_D(\theta, \varphi)$, is brought to a point focus in the slit plane at a position

$$\begin{aligned} \xi &= f\tan\theta \cong f\theta \\ \eta &= f\tan\varphi \cong f\varphi, \end{aligned} \tag{37}$$

where f is the focal length of L_2, and the approximate signs hold for small angles. Combining (35), (36), and (37) gives the field in the slit plane as

$$E_0(\xi, \eta) = \frac{1}{f\lambda_0}\int\int_S E_0(x, y)e^{i(2\pi/f\lambda_0)(\xi x + \eta y)}dxdy. \tag{38}$$

As eq. (38) shows, the field at the slit plane and the field at the aperture plane are related as Fourier transform pairs. It should be evident that eq. (38) can also be applied "backwards" in Fig. 6 to relate $E_0(x, y)$ to the field at the spatial filter aperture, $E_0(\xi', \eta')$. We will

* The cartesian coordinate and angle notation follows that adopted in Section II.

refer to the field $E_0(x, y)$ or its intensity

$$I_0(x, y) = \tfrac{1}{2}\sqrt{\epsilon\epsilon_0/\mu_0}E_0(x, y)E_0^*(x, y)$$

as the illumination or aperture function. The slit-plane field $E_0(\xi, \eta)$ or its intensity, $I_0(\xi, \eta)$, is the corresponding instrumental profile.

The procedure of aperture apodizing may be described simply as follows. The basic problem is to find and implement an instrumental illumination function, such that both the function itself and its transform have minimum spatial extent. The goal in a loose sense is to optimize the angular resolution per unit aperture opening. Of course, one of the general properties of Fourier transform pairs is that the second moments or "widths" of the pair members have an approximate inverse relationship. The Rayleigh criterion, in fact, is a simplified statement of this property. Even within the confines of this inverse relationship, however, there is still wide latitude for aperture apodizing, i.e., shaping the instrumental profile to obtain particularly desirable angular or spatial characteristics. Although the Fourier transform relationship between the illumination function and the instrumental profile in coherently illuminated optical systems is well known,[31–33] aperture apodizing schemes are not often applied in optical instrument design. Apodizing schemes are, however, extensively employed in high-frequency and microwave antenna design,[34,35] where they are used to create antenna systems exhibiting an angular directivity pattern that satisfies a particular objective.

In designing an apparatus for very-small-angle light scattering, the principal objective is the ability to observe the weak scattered light in close angular proximity to the unscattered beam. The goal, then, is an instrumental profile that not only has small angular half-power points but, more importantly, continues down rapidly to the 10^{-5} to 10^{-6} level. The proper shaping of the illumination function, $E_0(x, y)$ is absolutely crucial in obtaining this desired "steep-skirt" behavior.

In treating the question of aperture apodization for the MK VI instrument, we consider the optical system in the simplified form shown in Fig. 6. Therefore, the calculated instrumental profiles that are obtained below represent the instrument's ideal, diffraction-limited performance in the absence of all aberrations. The ways in which the residual aberrations of the actual off-axis configuration modify these results are taken up in detail in Section 3.2.

Given the idealized geometry of Fig. 6, the process of evaluating various illumination function/instrumental profile combinations can be further simplified by the following considerations. First, most of the interesting illumination functions and, therefore, their Fourier

transforms can be factored to the form

$$\begin{aligned} E_0(x, y) &= E_0(x)E_0(y) \\ E_0(\xi, \eta) &= E_0(\xi)E_0(\eta). \end{aligned} \tag{39}$$

Second, in the MK VI apparatus, the open height of the main scanning slit guarantees that the measured slit-plane profile is the integral over all η of the slit-plane intensity $I_0(\xi, \eta)$. Under either of these circumstances, we need consider only a one-dimensional form of eq. (38), namely,

$$E_0(\xi) = \frac{1}{(f\lambda_0)^{\frac{1}{2}}} \int E_0(x) e^{i(2\pi/f\lambda_0)\xi x} dx, \tag{40}$$

where $\xi = f\theta$. The corresponding aperture and slit-plane intensities are

$$\begin{aligned} I_0(x) &= \frac{1}{2}\sqrt{\frac{\epsilon\epsilon_0}{\mu_0}}\, E_0(x)E_0^*(x) \\ I_0(\xi) &= \frac{1}{2}\sqrt{\frac{\epsilon\epsilon_0}{\mu_0}}\,\frac{1}{f\lambda_0}\left| \int E_0(x) e^{i(2\pi/f\lambda_0)\xi x} dx \right|^2. \end{aligned} \tag{41}$$

In presenting the results of calculations based on eq. (41), it is convenient to adopt a concise terminology to describe the spatial and angular widths of the functions involved. We use the following notation:

$D_x(\frac{1}{2})$—Full width at half-maximum intensity for $I_0(x)$
$d_x(\frac{1}{2})$—Half width at half-maximum intensity for $I_0(x)$
$D_\xi(\frac{1}{2})$—Full width at half-maximum intensity for $I_0(\xi)$ or $I_0(\xi')$
$d_\xi(\frac{1}{2})$—Half width at half-maximum intensity for $I_0(\xi)$ or $I_0(\xi')$
$\Delta\theta(\frac{1}{2})$—Full width at half-maximum intensity for $I_0(\xi)$ or $I_0(\xi')$, expressed as an angular equivalent via eq. (37), $\Delta\theta(\frac{1}{2}) = (1/f)D_\xi(\frac{1}{2})$
$\delta\theta(\frac{1}{2})$—Half width at half-maximum intensity for $I_0(\xi)$ or $I_0(\xi')$, expressed as an angular equivalent.

For arguments other than $(\frac{1}{2})$ these quantities, give the width at the specified fraction of the peak intensity. Since the absolute normalization of the various intensity functions depends only on the total beam power, we present all results in terms of the ratio quantities:

$$I_0(x)/I_0(0),\ I_0(\xi)/I_0(0),\ I_0(\theta)/I_0(0),$$

where $I_0(\theta)$ describes the slit-plane intensity with position ξ given in terms of the equivalent angular deflection $\theta = \xi/f$.

Figure 7 shows the instrumental profiles calculated for two interesting illumination functions. The first is the uniform field

$$E_0(x) = \begin{cases} E_0; & -b/2 \leqq x \leqq b/2 \\ 0; & \text{otherwise} \end{cases}$$

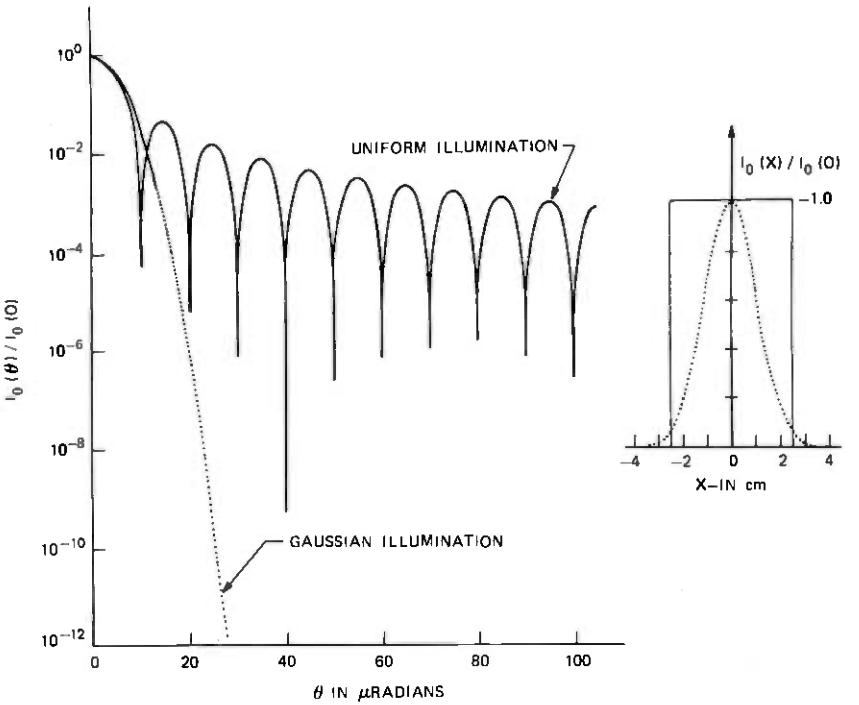


Fig. 7—Calculated instrumental profiles for uniform and gaussian illumination with $b = 5.0$ cm and σ chosen to give equal values of $\delta\theta(\frac{1}{2})$ for both profiles. The normalized illumination functions for each case are shown inset on a linear intensity scale.

for which one can easily calculate the following intensity ratios:

$$\begin{aligned} I_0(x)/I_0(0) &= \begin{cases} 1; & -b/2 \leqq x \leqq b/2 \\ 0; & \text{otherwise} \end{cases} \\ I_0(\xi)/I_0(0) &= \frac{\sin^2 (k_0 b\xi/2f)}{(k_0 b\xi/2f)^2} \\ I_0(\theta)/I_0(0) &= \frac{\sin^2 (k_0 b\theta/2)}{(k_0 b\theta/2)^2}, \end{aligned} \tag{42}$$

where $k_0 = (2\pi/\lambda_0)$. The second profile results from a gaussian aperture illumination,

$$E_0(x) = E_0\, e^{-x^2/2\sigma^2}$$

for which the relevant intensity ratios are

$$\begin{aligned} I_0(x)/I_0(0) &= \exp\,(-x^2/\sigma^2) \\ I_0(\xi)/I_0(0) &= \exp\,(-\sigma^2 k_0^2 \xi^2/f^2) \\ I_0(\theta)/I_0(0) &= \exp\,(-\sigma^2 k_0^2 \theta^2). \end{aligned} \tag{43}$$

The numerical parameters used in obtaining the curves plotted in Fig. 7 were

$$\lambda_0 = 5000 \text{ A}, \quad b = 5.0 \text{ cm}, \quad \sigma = 1.496 \text{ cm}. \tag{44}$$

The b value is representative of the maximum clear aperture of the MK VI apparatus; the value of σ was arbitrarily chosen to give equal $\delta\theta(\frac{1}{2})$ for both instrumental profiles. Also shown inset in Fig. 7 are the two aperture-plane intensity ratios, $I_0(x)/I_0(0)$. Note that the latter are plotted using a linear ordinate scale.

As is evident from these two calculated instrumental line shapes, the use of gaussian apodization is vastly superior to uniform illumination in regard to the observability of weak small-angle features even though the two $I_0(\theta)/I_0(0)$ profiles have identical half-widths. For example, at the 10^{-6} level we find

$$\delta\theta(10^{-6}) = 19.8\ \mu\text{rad} \quad \text{GAUSSIAN } I(x)$$
$$\delta\theta(10^{-6}) = 3183\ \mu\text{rad} \quad \text{UNIFORM } I(x).$$

In fact, from a theoretical standpoint, the gaussian is the ideal form of aperture functional. Among the families of possible illumination functions, it possesses a unique combination of two properties: (*i*) it has an extremely rapid shirt fall-off, and (*ii*) it goes over into itself under the Fourier transform operation. In a general situation where the available aperture illumination has some arbitrary (x, y) behavior, gaussian apodization would have to be accomplished by interposing a suitable neutral density mask at the aperture plane. Fortunately, laser sources with a reasonable cavity configuration and oscillating only on TEM_{00} modes have an output beam intensity pattern which is accurately gaussian, except in the extreme tails of the profile. The availability of such a source represents a crucial factor in the feasibility of constructing an instrument having the resolution and stray-light performance of the MK VI apparatus.

In the actual MK VI instrument, the ratio of the focal length of lens L_1 to that of mirror M_5 was chosen to generate a gaussian illumination function with an effective width

$$\sigma^* = d_x(1/e) = 0.826 \text{ cm} \tag{45}$$

in the collimated beam portion of the apparatus (see Fig. 1). The instrumental profile calculated via eq. (41) for this value of σ is plotted in Fig. 8. Also shown for comparison purposes is the profile to be expected if we uniformly illuminated the instrument's maximum design aperture

$$b^* = 5.0 \text{ cm}.$$

It is clear that for these specific values of σ and b, gaussian apodizing no longer exhibits an absolute superiority over uniform illumination. Although the "gaussian" profile still reaches the 10^{-6} level much more rapidly, it does sacrifice resolving power to the "uniform" profile down to approximately the 10^{-2} level. What this means to VSA scattering performance is the following. The "gaussian" instrument will excel in its ability to detect small amounts of light at very small scattering angles; however, it will not resolve approximately equal intensity features with as much detail as would the "uniform" instrument. As we see in the following paragraphs, the tradeoff, roughly speaking, involves paying for small-angle weak-intensity performance by sacrificing some ability to resolve the angular dependent features of the scattered light. This comparison can be made more quantitative by reference to Table IV, which gives various half-width angles for the profiles of Fig. 8.

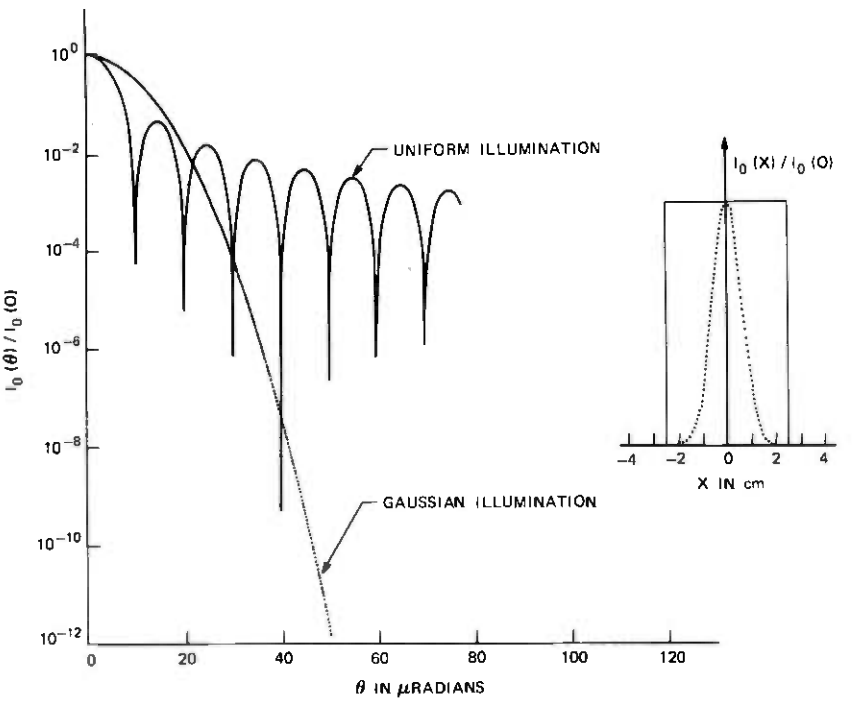


Fig. 8—Calculated instrumental profiles for uniform and gaussian illumination with $b = 5.0$ cm and $\sigma = 0.826$ cm. The b value corresponds to the maximum clear aperture of the MK VI instrument, while the σ value corresponds to the width of the gaussian illumination actually used in the present apparatus. Normalized illumination functions for each case are shown inset on a linear intensity scale.

Table IV—Calculated instrumental profile half-widths at various fractions of peak intensity for three different illumination functions

$\lambda_0 = 5000$ Å	Gaussian Illumination $d_x(1/e) = \sigma^*$ $\sigma^* = 0.826$ cm (μrad)	Uniform Illumination $d_x(1/e) = b^*/2$ $b^* = 5.0$ cm (μrad)	Truncated Gaussian Illumination $b = b^*$ $\sigma = \sigma^*$ $b/\sigma = 6.05$ (μrad)
$\delta\theta(1/2)$	8.0	4.43	8.09
$\delta\theta(1/e)$	9.6	5.23	9.69
$\delta\theta(10^{-1})$	14.6	7.36	14.62
$\delta\theta(10^{-2})$	20.7	31.8	20.6
$\delta\theta(10^{-3})$	25.3	100.6	25.5
$\delta\theta(10^{-4})$	29.2	318	29.7
$\delta\theta(10^{-5})$	32.7	1006	32.6
$\delta\theta(10^{-6})$	35.8	3183	76.8
$\delta\theta(10^{-7})$	38.7	10,060	243
$\delta\theta(10^{-8})$	41.3	31,830	768

From the inset plots of $I(x)/I(0)$ shown in Fig. 8, it may seem that the gaussian illumination profile used in the present apparatus was unnecessarily narrowed relative to the instrumental full aperture. This is, in fact, not the case. One crucial detail which has been omitted in obtaining the results presented in Figs. 7 and 8 is the possible vignetting effect of the instrument's maximum aperture. In calculating $E_0(\xi)$ for the gaussian $E_0(x)$, for example, the integral in eq. (40) was taken over all x, thereby neglecting any aperturing effects that might occur.

For the actual VSA scattering instrument, which has a fixed maximum aperture, b, $E_0(\xi)$, and $E_0(x)$ are related via the finite domain transform

$$E_0(\xi) = \frac{1}{(f\lambda_0)^{\frac{1}{2}}} \int_{-b/2}^{b/2} E_0(x) e^{i(2\pi/f\lambda_0)\xi x} dx. \tag{46}$$

Except for a few special cases, an analytical evaluation of this integral is not possible, and one must resort to a numerical approach to investigate various apodizing schemes. For the experimentally relevant case of gaussian illumination, eq. (46) becomes

$$E_0(\xi) = \frac{E_0}{(f\lambda_0)^{\frac{1}{2}}} \int_{-b/2}^{b/2} e^{-x^2/2\sigma^2 + i(2\pi/f\lambda_0)\xi x} dx. \tag{47}$$

On completing the square in the exponential and a change of variable, we can rewrite this expression in the form

$$E_0(\xi) = \frac{E_0}{(f\lambda_0)^{\frac{1}{2}}} \sqrt{2}\sigma e^{-(\sigma^2 K^2/2)} \int_{w_-}^{w_+} e^{-w^2} dw, \tag{48}$$

where K, w, w_+, and w_- are defined as follows:

$$\begin{aligned} w &= \frac{x}{\sqrt{2}\sigma} - i\frac{\sigma K}{\sqrt{2}} \\ K &= (2\pi/\lambda_0)\frac{\xi}{f} \\ w_+ &= +\frac{b}{2\sqrt{2}\sigma} - i\frac{\sigma K}{\sqrt{2}} \\ w_- &= -\frac{b}{2\sqrt{2}\sigma} - i\frac{\sigma K}{\sqrt{2}}. \end{aligned} \tag{49}$$

The analyticity of $\exp(-w^2)$ near $w = 0$ allows the complex plane w integral to be split into two terms, each having the form of an error function of complex argument. Tabulated values of this function are available in the literature[36] for a restricted range of the parameters (b/σ) and (σK).

In searching out an optimum configuration for the MK VI instrument, it was decidely more convenient to adopt a fully numerical approach in evaluating eq. (47). Appendix B outlines the methods that were used. The modified instrumental profile calculations were carried out for a range of values of the ratio (b/σ) with the aperture opening, b, held fixed at $b = b^* = 5.0$ cm.

Figure 9 shows four such profiles plotted in terms of the normalized intensity ratio $I_0(\theta)/I_0(0)$. Also shown are the corresponding aperture ratios $I_0(x)/I_0(0)$. The curve for $(b/\sigma) = 0.01$ is essentially equivalent to the result obtained above for uniform aperture illumination. The most striking feature of the remaining three $I_0(\theta)/I_0(0)$ curves is the presence of an effective background or floor contribution to the profile caused by edge diffraction at the aperture. This "shelf" or wing on the profile has the slow oscillatory decay of a $(\sin^2 x)/x^2$ functional dependence. In each case, however, the $\theta \approx 0$ portion of the curves closely approximates the gaussian profile expected from unapertured gaussian illumination.

The results given in Fig. 9 clearly illustrate the tradeoff involved in selecting a value of σ. In circumstances requiring an instrumental line shape with a very low background level, we are forced to accept a moderate increase in $\delta\theta(\frac{1}{2})$ and, therefore, a loss in angular resolving power. The curve given in Fig. 9 for $(b/\sigma) = 6.05$ corresponds to the choice that was made for the MK VI apparatus. Various half-width values for this profile have been included in Table IV for comparison with the results for unapertured gaussian and uniform illumination. In the actual instrument, this choice for (b/σ) guarantees that the calculated edge-diffraction "floor" constitutes less than 10 percent of the overall stray-light level. This point is illustrated in Fig. 10 which

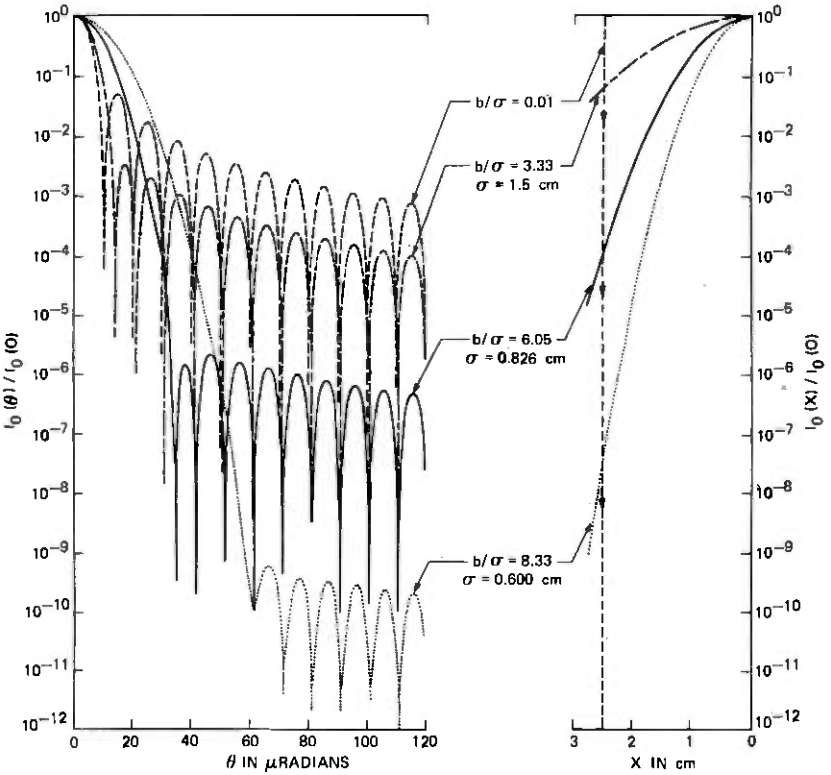


Fig. 9—Calculated instrumental profiles for truncated gaussian illumination and various values of (b/σ). Also shown, on a logarithmic intensity scale, are the corresponding normalized illumination functions $[I_0(x)/I_0(0)]$. The curves labelled $(b/\sigma) = 6.05$ are appropriate to the b and σ values used in the present apparatus.

shows the theoretical profile for $(b/\sigma) = 6.05$ superimposed on the measured profiles of the MK IV instrument.

At this point, it is crucial to realize that the truncated transform results apply not only downstream from the aperture plane of Fig. 6 but also upstream toward the spatial filter. There are, in fact, two other possible sources of beam vignetting in the system. The most obvious is the spatial filter itself. Since the field at the spatial filter and the field at the aperture plane are Fourier transform pairs, the same considerations involved in choosing (b/σ) also apply to the choice of spatial filter pinhole size. If edge-diffraction effects at the aperture plane are to dominate the system profile, then the ratio of pinhole diameter, b_{PH}, to the gaussian focal width at the pinhole, $d_{\ell'}(1/e)$, must exceed (b/σ). Specifically, for the present instrument, b_{PH} must satisfy the inequality

$$\frac{b_{PH}}{9.6\,\mu\text{m}} > \frac{b}{\sigma} = 6.05$$
$$b_{PH} > 58\,\mu\text{m}. \tag{50}$$

The pinhole diameter actually used is $b_{PH} = 100\ \mu m$. We note that a $[b_{PH}/d_{\xi'}(1/e)]$ ratio this large is contrary to usual spatial filtering practices.

A much more subtle source of possible beam aperturing is the internal cavity configuration of the laser source itself. Clearly, the ratio of laser tube inside diameter, b_{LASER}, to the mode $(1/e)$ radius, σ_{LASER}, must also satisfy the inequality

$$\frac{b_{LASER}}{\sigma_{LASER}} > \frac{b}{\sigma}. \tag{51}$$

The laser used in the MK VI has a gaussian mode diameter given as

$$D_{LASER}(1/e^2) = 1.4\ \text{mm}, \tag{52}$$

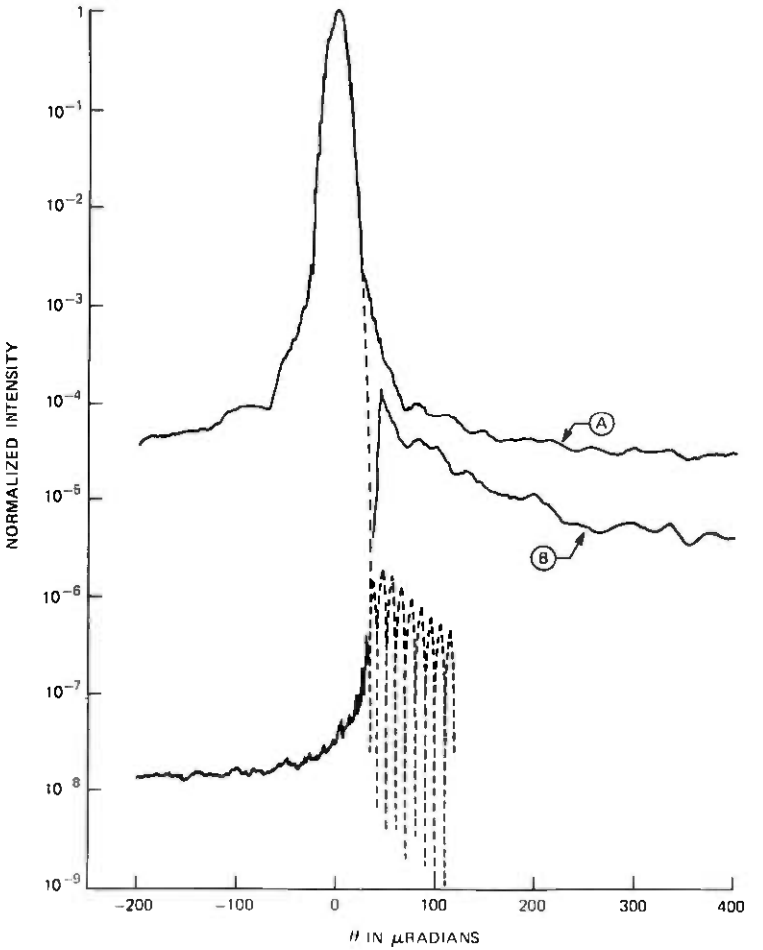


Fig. 10—Calculated instrumental profile for truncated gaussian illumination (dashed curve) superimposed on the measured profiles of the MK VI instrument. Curves A and B correspond to the two measurements described in Fig. 4.

which is equivalent to

$$\sigma_{\text{LASER}} = \frac{D_{\text{LASER}}(1/e^2)}{2\sqrt{2}} \cong 0.5 \text{ mm}. \tag{53}$$

The plasma tube ID for the laser is $b_{\text{LASER}} = 5.0$ mm yielding a b/σ ratio

$$\frac{b_{\text{LASER}}}{\sigma_{\text{LASER}}} \cong 10. \tag{54}$$

Therefore, in the MK VI instrument, the maximum design aperture available to the collimated beam is, in fact, the principal source of truncation effects.

In situations where it is advantageous to alter this resolution-background tradeoff by varying (b/σ), it soon becomes apparent that the numerical profile calculations of Appendix B are a rather unwieldy design tool. Instead, based on an examination of the results shown in Fig. 9, it seemed tempting to fit the profile tails to the form

$$\frac{I_T(\theta)}{I(0)} = A_1 \frac{\sin^2 (k_0 b\theta/2)}{(k_0 b\theta/2)^2} \tag{55}$$

and look for an interpolation formula relating the amplitude A_1 to the ratio (b/σ). By a trial-and-error procedure, the following relation was found to reproduce the best-fit A_1 values to within 10-percent error:

$$A_1^* = \left(1 + \frac{b^2}{8\sigma^2}\right) e^{-b^2/4\sigma^2}. \tag{56}$$

Table V gives the fitted and interpolated values of A_1 corresponding to the four (b/σ) ratios of Fig. 9. It is interesting to note that the exponential factor $\exp(-b^2/4\sigma^2)$, which dominates the (b/σ) dependence, is just the normalized aperture illumination at the aperture edge.

3.2 Optical aberrations

The fundamental diffraction limitations set out in Section 3.1 are really only a prediction regarding the ideal performance of an optical system. In the final analysis, the inherent optical aberrations of any particular apparatus design determine how close one will come to achieving the ideal of diffraction-limited performance. In this section, we give a brief summary of those aspects of optical aberration theory[37] that are relevant to the design of the MK VI apparatus. From a qualitative understanding of and analytical expressions for each of the various aberrations, we then determine the extent to which aberrations modify the ideal diffraction-limited characteristics of the instrument.

Table V — Fractional amplitude of the edge diffraction contribution to the instrumental profile of truncated gaussian illumination. The first column gives the A_1 values obtained by fitting eq. (55) to the tails of the profiles shown in Fig. 9; the second column gives the A_1 value predicted by the interpolation formula in eq. (56)

(b/σ)	A_1-Fitted	A_1^*-Interpolated
0	1.0	1.0
3.333	1.33×10^{-1}	1.49×10^{-1}
6.05	5.823×10^{-4}	5.92×10^{-4}
8.333	2.782×10^{-7}	2.79×10^{-7}

Finally, a number of measurements taken on the MK VI instrument are compared to the quantitative predictions of the aberration theory.

Because the present instrument is illuminated with monochromatic light, the various chromatic aberrations are absent, and the lowest-order non-zero distortions come from the third-order or primary aberrations. Here we follow the order-naming convention associated with the Taylor expansion of the function $\sin \psi$, i.e.,

$$\sin \psi = \psi - \frac{\psi^3}{3!} + \frac{\psi^5}{5!}, \tag{57}$$

where ψ is the angle of incidence of a ray on a reflecting or refracting surface. The approximation $\sin \psi = \psi$ leads to the usual paraxial optics formulae. The next term in the expansion, proportional to ψ^3, describes the primary aberrations.

The principle aberration-producing elements of the MK VI apparatus are the off-axis spherical mirrors M_5 and M_6 (see Fig. 1). Figure 11 shows the basic optical configuration in which the mirrors are used. The labelled geometrical parameters are:

$$\begin{aligned} u_p = \frac{\gamma}{2} &\text{—The half-field angle or off-axis angle} \\ \mathcal{a} &\text{—The semi-aperture} \\ R &\text{—The mirror radius of curvature.} \end{aligned} \tag{58}$$

The primary aberrations for an off-axis spherical mirror depend parametrically on two angles: u_p, the half-field angle and $(\mathcal{a}/R)$, the semi-aperture angle. The aberrations associated with the various

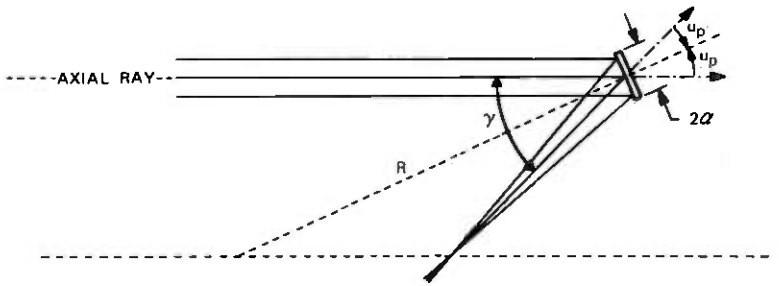


Fig. 11—Off-axis mirror configuration used in the MK VI instrument showing the important geometrical parameters.

third-order products of these angles are:

$$\begin{aligned} (\alpha/R)^3&\text{—Spherical aberration}\\ (\alpha/R)^2u_p&\text{—Coma}\\ (\alpha/R)u_p^2&\text{—Astigmatism}\\ u_p^3&\text{—Distortion.} \end{aligned} \tag{59}$$

3.2.1 Spherical aberration

Spherical aberration is a longitudinal focussing defect that is present even when the off-axis angle goes to zero. Figure 12 sketches the basic ray geometry for a spherical mirror exhibiting pure spherical aberration. As illustrated in the enlarged detail of the sketch, the marginal rays of an incoming parallel bundle are brought to a focus at a point closer to the mirror's surface than those lying nearer the axial ray. This constantly changing longitudinal focal position results in a transversely smeared focal spot rather than a focal point.

One common measure of the amount of spherical aberration is the minimum beam waist size produced in the focal region. For a spherical reflector, the diameter of this blur spot is given by

$$2TSC^* = \alpha^3/R^2. \tag{60}$$

Since transverse displacement at the focus is equivalent to an angular deviation in the parallel bundle, we can also express the spherical aberration in terms of a full-width angular blur,

$$\Delta\theta^*_{SC} = \frac{2TSC^*}{f} = 2(\alpha/R)^3, \tag{61}$$

where $f = R/2$ is the focal length of the reflector.

3.2.2 Astigmatism

Astigmatism, like spherical aberration, is the result of a longitudinal focussing defect. In contrast to the spherical aberration defect, how-

ever, the longitudinal focussing error depends not on aperture diameter, $2\mathcal{Q}$, but on the off-axis angle, u_p. Figure 13 illustrates the ray geometry of pure astigmatism for a spherical reflector.

One of the fundamental characteristics of the u_p dependent aberrations is the loss of rotational symmetry in the focal region. A non-zero off-axis angle destroys this symmetry and establishes two unique directions or planes of transverse blurring. The plane defined by the incident and reflected axial ray is the tangential plane. Cartesian or angular displacements perpendicular to the axial ray and lying in this plane are referred to as tangential displacements. The two planes orthogonal to this tangential surface and containing either the incident or reflected axial ray are called the sagittal planes. Cartesian or angular displacements from the axial ray in these planes are sagittal displacements.

For a spherical reflector exhibiting pure astigmatism, a fan of parallel tangential plane rays are brought to a focus closer to the mirror surface than an identical sagittal fan. The focal region pattern found by decomposing the entire illuminated aperture into such ray fans consists of the two longitudinally separated focal lines depicted in Fig. 13a. The longitudinal ($\hat{z}$) separation of the S and T foci ($2AC^*$)

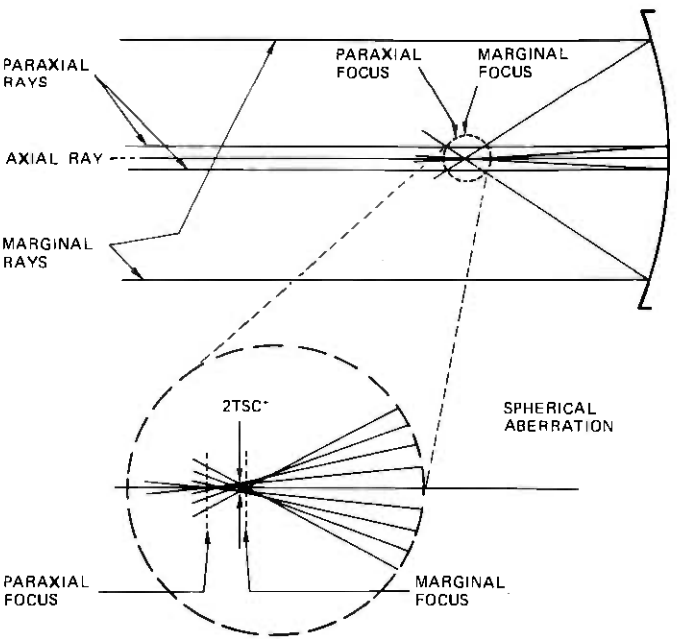


FIG. 12—Ray diagram for a single spherical mirror exhibiting a pure spherical aberration defect.

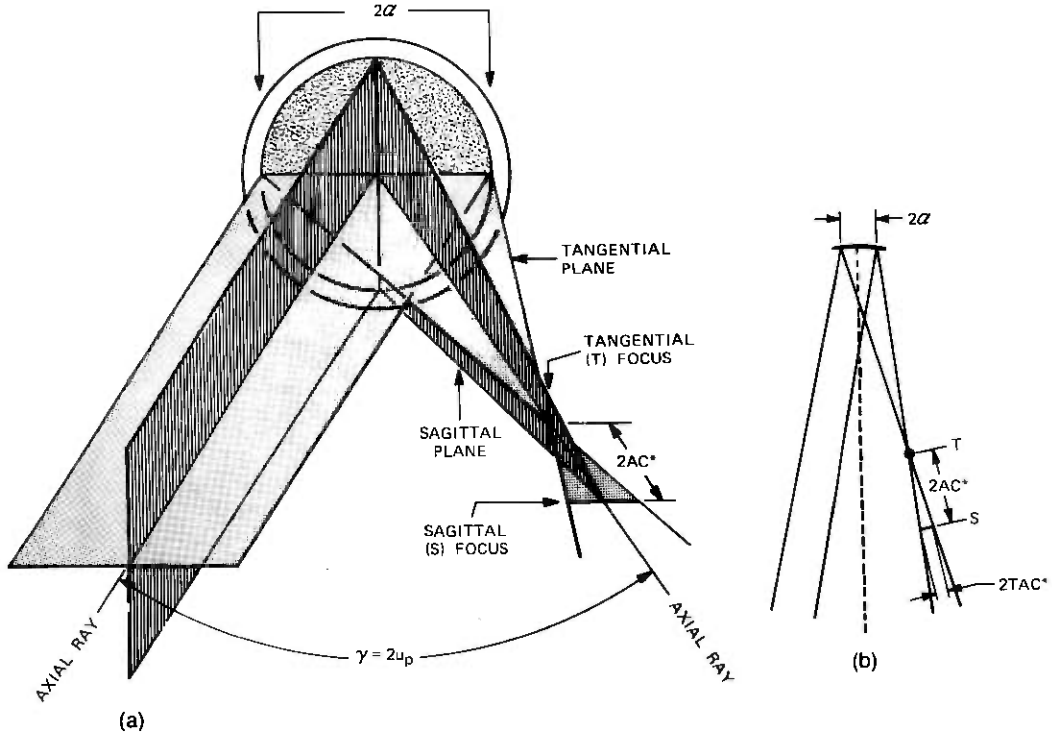


Fig. 13—Partial ray diagram illustrating the focal plane characteristics of an off-axis spherical mirror exhibiting the pure astigmatism defect.

can be calculated from the relation

$$AC^* = \frac{Ru_p^2}{4} \tag{62}$$

and depends only on the semi-field angle.

In the absence of other aberrations (and the effects of diffraction), the S and T focal lines are infinitely narrow in their respective planes. This means that the sagittal height of the tangential focus and the tangential width of the sagittal focus can be obtained from simple extremal ray geometry. For example, Fig. 13b shows the extremal rays seen in a tangential plane projection. Normally $2AC^*$ is small compared to the reflectors focal length $(R/2)$. It follows that the lengths of the two focal lines are identical and given by

$$2TAC^* = 2AC^*\left(\frac{2\mathcal{A}}{R/2}\right) = 2\mathcal{A}u_p^2. \tag{63}$$

The full-width angular spread equivalent to this spatial blur is

$$\Delta\theta_{TAC^*} = 4(\mathcal{A}/R)u_p^2. \tag{64}$$

3.2.3 Coma

When the off-axis angle is non-zero, the longitudinal focussing error that produces spherical aberration also gives rise to an asymmetric transverse blurring called coma. Figure 14a sketches the basic elements of the focal region pattern for a spherical reflector exhibiting a pure coma defect. Rays in the paraxial region are brought to a focus at the axial focus, P, while rays from larger-diameter annular zones on the mirror's surface form focal circles whose centers are tangentially displaced from P. The radius of a particular focal circle increases as the square of the radius of the zone producing it.

Figure 14b gives a qualitative representation of the characteristics of the focal pattern as found by dividing up the illuminated aperture into these annular zones. Mathematically speaking, the focal circles are not sharp unless the radial thickness of the corresponding zones vanish; however, the sketch does predict quite nicely the overall exterior outline of the coma blur patch.

The continuum of focal circles nest into a 60° wedge extending out from the axial focus forming a pattern commonly called comatic flare. The largest-diameter focal circle, produced by the annular zone at the edge of the illuminated aperture, has a radius CC^* given by

$$CC^* = \frac{\mathcal{A}^2 u_p}{2R}. \tag{65}$$

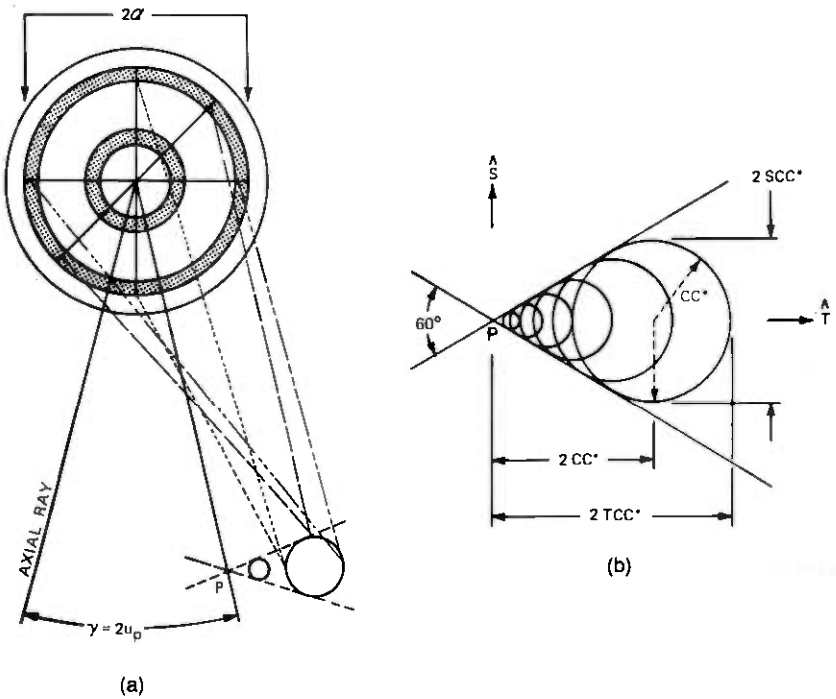


Fig. 14—Partial ray diagram, (A), of an off-axis spherical mirror exhibiting a pure coma defect. The details of the focal plane pattern are illustrated in (B).

Its center is tangentially displaced from the axial focus by an amount $2CC^*$.

Although coma is a highly asymmetric aberration, it is still convenient to specify its effect in terms of transverse and/or angular blur sizes. The numbers ordinarily quoted for coma correspond to the extremal dimensions of the coma patch in the tangential and sagittal directions. It follows easily from Fig. 14b that the full-width transverse spatial blurs are

$$\begin{aligned} \hat{T} \text{ DIRECTION} \quad & 2TCC^* = 3CC^* = \frac{3\mathcal{A}^2 u_p}{2R} \\ \hat{S} \text{ DIRECTION} \quad & 2SCC^* = 2CC^* = \frac{\mathcal{A}^2 u_p}{R}. \end{aligned} \tag{66}$$

The equivalent full-width blur angles are

$$\begin{aligned} \hat{T} \text{ DIRECTION} \quad & \Delta\theta^*_{TCC} = 3(\mathcal{A}/R)^2 u_p \\ \hat{S} \text{ DIRECTION} \quad & \Delta\theta^*_{SCC} = 2(\mathcal{A}/R)^2 u_p. \end{aligned} \tag{67}$$

3.2.4 Distortion

The last of the primary aberrations is a defect of off-axis magnification called distortion. This particular aberration is associated principally with optical systems that form a real image of an extended object at finite magnification. For example, the image of a rectangular grid of regularly spaced points will exhibit the classic "barrel" or "pincushion" appearance in an optical system involving pure distortion. For the situation of interest here, namely a spherical reflector with the image at infinity or at the focus, the amplitude of the pure distortion aberration vanishes identically.

3.2.5 Application of the aberration results to the MK IV instrument

The third-order aberration theory results for the off-axis spherical reflector are summarized in Table VI which gives the expressions for the various transverse and angular blurs. It must be emphasized that the aberration theory results outlined above are derived from purely geometric ray tracing. In no sense does this theory predict the actual intensity distribution in the image plane for a specific aperture illumination. The transverse and angular blur patterns define the outlines of a boundary between focal illumination and strict geometric shadow in the absence of all wave interference and diffraction effects. However, in certain situations, it is possible to combine the geometric aberration results with aberration-free diffraction calculations to obtain useful instrumental profile information. The approach works well when one or more of the following conditions are satisfied:

(*i*) Diffraction blurring is large or small compared to spherical aberration.
(*ii*) One primary aberration is dominant.
(*iii*) The ideal diffraction-limited system profile is free of large interference maxima and minima.

Table VI — Summary of the analytic expressions for the transverse and angular aberration blurs for a single off-axis spherical mirror

Aberration	Full-Width Transverse Blur	Full-Width Angular Blur
Spherical	$\alpha^3/R^2 \quad (2\,TSC^*)$	$2(\alpha/R)^3$
Coma	$\hat{T} \quad -3\alpha^2 u_p/2R \quad (3\,CC^*)$ $\hat{S} \quad -2\alpha^2 u_p/2R \quad (2\,CC^*)$	$-3(\alpha/R)^2 u_p$ $-2(\alpha/R)^2 u_p$
Astigmatism	$2\alpha u_p^2 \quad (2\,TAC^*)$	$4(\alpha/R)u_p^2$
Distortion	$Bu_p^3 R$	$2\,Bu_p^3$

In these cases, one can intuitively construct the aberration-affected profiles with a fair degree of accuracy using the one-to-one geometric mapping of regions of the aperture onto the focal plane. For example, it can be argued from Fig. 13b that for pure astigmatism the tangential direction intensity profile at the sagittal focus should be a demagnified replica of the aperture illumination. This type of analysis is essential in obtaining quantitative results from the aberration expressions.

In the MK VI instrument, the off-axis angle and radii of mirrors M_5 and M_6 are

$$\begin{aligned} \gamma &= 2|u_p| = 0.116 \text{ rad} \\ R &= 200 \text{ cm}. \end{aligned} \tag{68}$$

Assigning a value to be used for the semi-aperture $\mathfrak{a}$ is a more subtle question, especially since we are interested in gaussian rather than uniform aperture illumination. However, in the spirit of the astigmatism example given in the preceding paragraph, we take

$$\mathfrak{a} = d_x(1/e) = \sigma^* = 0.826 \text{ cm} \tag{69}$$

and *assume* that the focal plane profiles will also be gaussian. On the basis of geometrical imaging, the blur values calculated via Table VI should then be (with the exception of coma) the full-width to the $(1/e)$ points of a focal plane gaussian profile. We show in succeeding paragraphs that these assumptions lead to a selfconsistent picture of the experimentally observed aberration effects in the MK VI instrument.

Table VII gives the transverse and angular blurs for a single spherical reflector in the MK VI configuration. The table also includes the longitudinal separation of the S and T focal planes as well as the full angular width of the focus imposed by diffraction. From the viewpoint of small-angle-scattering performance, the most serious of the aber-

Table VII — Numerical values of the aberration blurs for a single spherical mirror used in the MK VI configuration

Aberration	Full-Width Transverse Blur (μm)	Full-Width Angular Blur (μrad)
Spherical	0.141	0.141
Coma	$\hat{T}$ 2.96	2.96
	$\hat{S}$ 1.97	1.97
Astigmatism	55.6	55.6
Distortion	0	0

$(S\text{-}T)_{\text{separation}} = 2\,AC^* = 0.336$ cm. $\Delta\theta(1/e)_{\text{diffraction}} = 19.3$ μrad.

Table VIII — Numerical values of the aberration blurs for the system of two off-axis spherical mirrors used in the MK VI apparatus

Aberration	Full-Width Transverse Blur (μm)	Full-Width Angular Blur (μrad)
Spherical	0.282	0.282
Coma	T 0 S 0	0 0
Astigmatism	111.2	111.2
Distortion	0	0

$(S\text{-}T)_{\text{separation}} = 2(2\,AC^*) = 0.672$ cm. $\Delta\theta(1/e)_{\text{diffraction}} = 19.3\ \mu$rad.

rations is coma. Even though the calculated comatic blurs are numerically small compared to the diffraction spread, the presence of coma can result in a distinctly asymmetric instrumental profile. Moreover, in a coherently illuminated system, the coma flare is criss-crossed by interference patterns whose tails extend far beyond the calculated geometric limits. This latter effect can significantly raise the effective "floor" level of the instrumental profile.

Fortunately, in a symmetric two-mirror system like the MK VI, the geometry may be chosen such that the total coma vanishes identically. In fact, all aberrations that depend on an odd power of the half-field angle u_p disappear if the field angles at the two elements are made equal and of opposite sign. By convention, u_p is defined as the angle through which the incoming axial ray must be rotated to bring it into coincidence with the local radius vector of the element's spherical surface (see Fig. 8). An inspection of Fig. 1 shows that in the MK VI apparatus the field angle rotations at M_5 and M_6 are of opposite sense. In this case, when the two off-axis angles are made equal in magnitude the coma and distortion aberrations vanish while the spherical aberration and astigmatism double.

The total calculated aberration blurs for the instrument are summarized in Table VIII. Since the spherical aberration is small compared to the diffraction spread, it is reasonable to expect that the interpretation of these blur values as $(1/e)$ full-widths of a gaussian blur profile should work quite well. This is in fact the case.

The tangential and sagittal foci of the MK VI instrument were located using a modified Foucault knife-edge procedure, and $\hat{T}$ direction scans of the intensity profiles were taken in each case. The measured S-T separation was

$$(S\text{-}T)\ \text{separation} = 0.660 \pm 0.013\ \text{cm}. \tag{70}$$

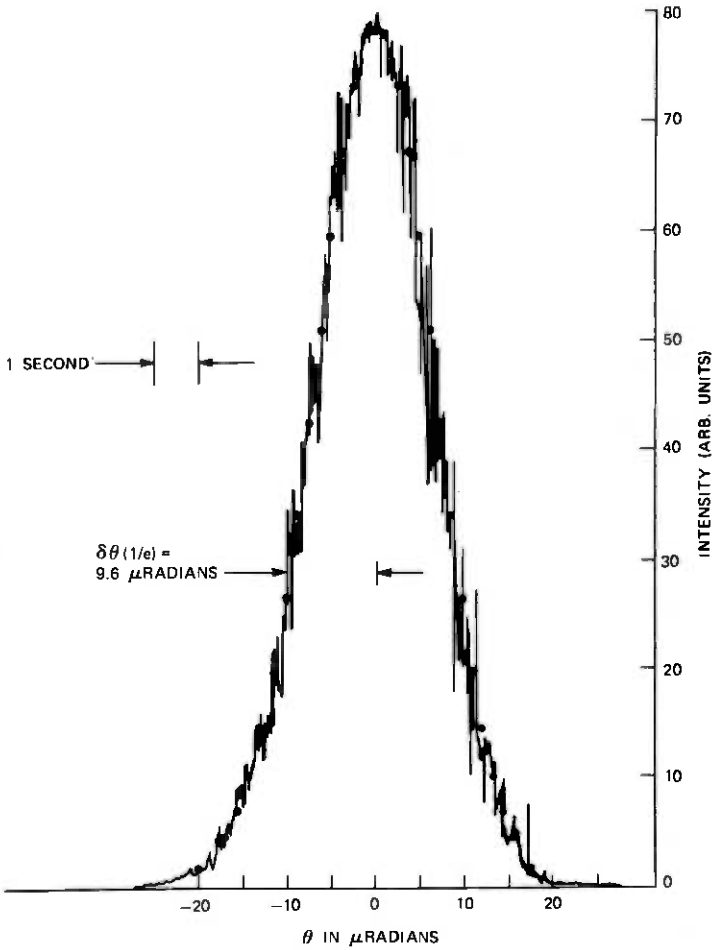


Fig. 15—Observed instrumental profile as measured by a tangential direction scan in the instrument's tangential focal plane. The heavy dots are a best fit to the function $\exp[-\theta^2/\delta\theta^2(1/e)]$.

At the T focus, only diffraction and spherical aberration contribute to the profile width. From the standard gaussian convolution formula, we can calculate the expected $\Delta\theta(1/e)$:

$$\begin{matrix} T \text{ PLANE} \\ \hat{T} \text{ SCAN} \end{matrix} \quad \Delta\theta(1/e) = \sqrt{(19.30)^2 + (0.282)^2}\ \mu\text{rad} = 19.302\ \mu\text{rad}. \tag{71}$$

Clearly the spherical aberration has a negligible effect on the ideal diffraction-limited broadening. Figure 15 shows a typical high-resolution T-plane scan for the instrument. The large apparent noise in this

trace is produced by residual air currents and vibration in the apparatus and corresponds to a peak-to-peak beam wander of roughly 0.5 μrad (0.7 arc second). The results of fitting the observed profile with a gaussian shape are indicated by the points in Fig. 15 and give an experimental full width

$$\Delta\theta(1/e) = 2\delta\theta(1/e) = 19.2\,\mu\text{rad}. \tag{72}$$

In the S focal plane, diffraction, astigmatism, and spherical aberration all contribute to the instrumental line shape. The full-width calculated from Table VIII is

$$\begin{matrix} S \text{ PLANE} \\ \hat{T} \text{ SCAN} \end{matrix} \quad \Delta\theta(1/e) = \sqrt{(19.3)^2 + (0.282)^2 + (111.2)^2}\ \mu\text{rad}$$
$$= 113\ \mu\text{rad}. \tag{73}$$

The predicted width comes predominately from the astigmatic blurring. Figures 16 and 17 give two experimental S-plane profiles recorded with logarithmic and linear intensity scales, respectively. A gaussian

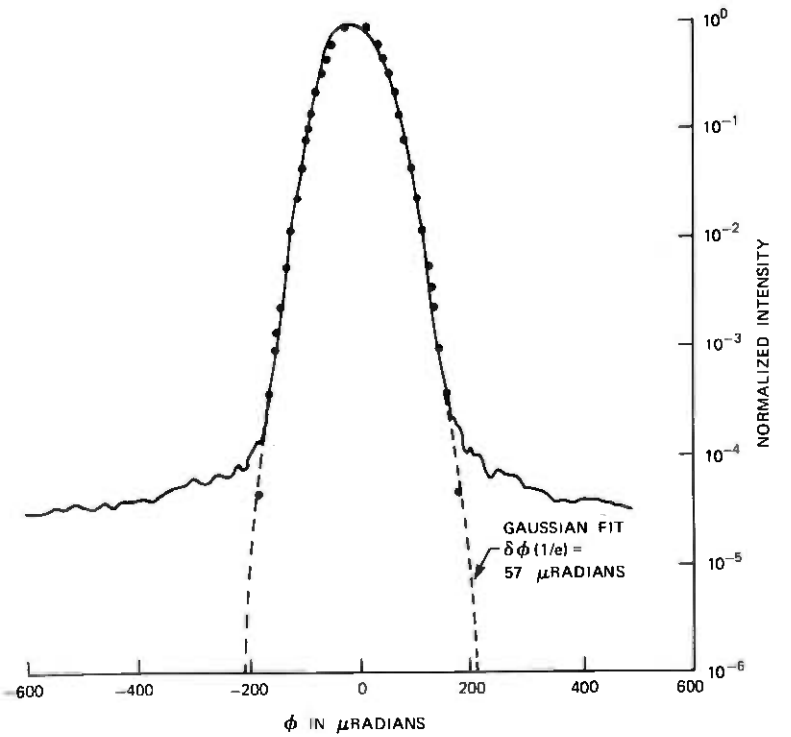


Fig. 16—Observed instrumental profile as measured by a tangential direction scan in the instrument's sagittal focal plane. The heavy dots and dashed curve a are best fit to the function exp $[-\varphi^2/\delta\varphi^2(1/e)]$.

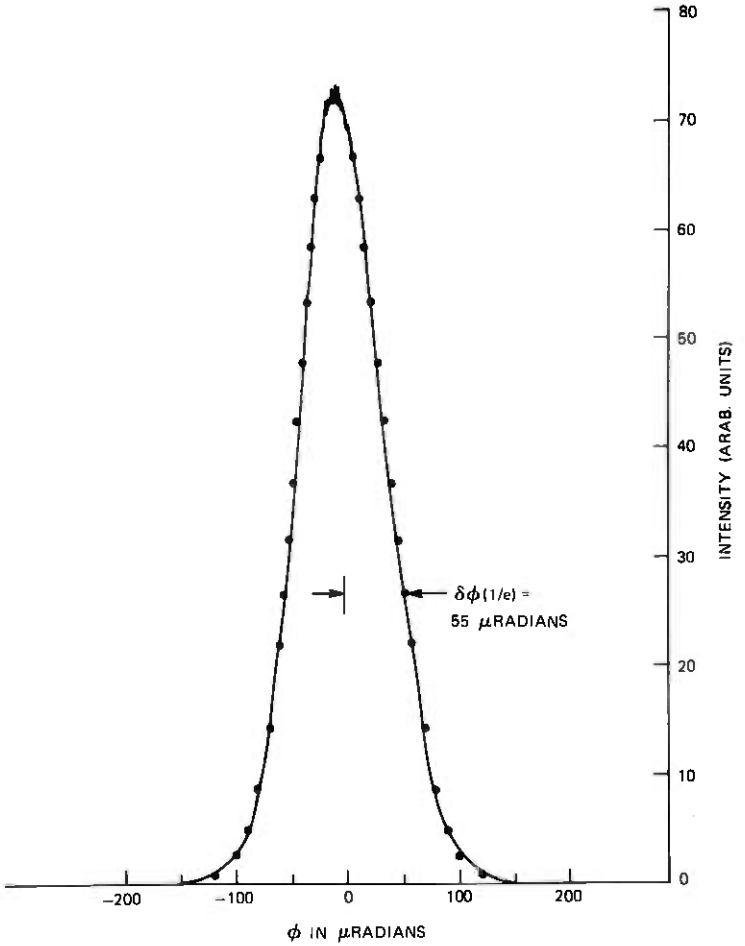


Fig. 17—Observed instrumental profile as measured by a tangential direction scan in the instrument's sagittal focal plane. The heavy dots are a best fit to the function $\exp[-\varphi^2/\delta\varphi^2(1/e)]$.

fit to the logarithmic curve, indicated by the points in Fig. 16, gives very good agreement with the observed line shape over roughly four orders of magnitude in intensity. The range and precision of the fit provide strong support to our assumptions regarding the interpretation of the aberration blur values. The best-fit half-width values for the logarithmic and linear scans are $\delta\theta(1/e) = 57$ μrad and $\delta\theta(1/e) = 55$ μrad, respectively. The mean observed full-width

$$\Delta\theta(1/e) = 112\ \mu\text{rad}$$

is in excellent agreement with the calculated value given in eq. (73).

3.2.6 Conclusions

In summarizing the discussion of aberrations, a number of general points deserve to be made and reiterated regarding the relationship between aberrations and small-angle-scattering performance.

(*i*) The cancellation of the asymmetric aberrations, coma especially, is crucial in obtaining an instrumental profile that has symmetry, the necessary steep skirt fall-off, and a low background value.

(*ii*) Spherical aberration, although it has a negligible effect relative to diffraction in the present instrument, can rapidly grow to serious proportions with increasing aperture size, $\mathfrak{a} = d_x(1/e)$. The angular blur of this aberration, $\Delta\theta^*_{SC}$, increases as the cube of the aperture size while the diffraction spread varies inversely with $d_x(1/e)$. The relative contribution of spherical aberration to the profile width will, therefore, increase as $\mathfrak{a}^4$. Since $\Delta\theta^*_{SC}$ depends only on the reduced quantity $(\mathfrak{a}/R)$, however, a constant ratio of spherical blur to diffraction spread can always be obtained by scaling the mirror radius R to keep $\mathfrak{a}^4/R^3$ constant. For example, an instrument with 10 times better angular resolution than the MK VI might conceivably utilize 80-cm-diameter mirrors with a 21.5-meter focal length.

(*iii*) The presence of a large residual astigmatism need not be detrimental if one is satisfied with an instrumental performance that is diffraction limited in only a single angular direction. It might appear from Fig. 13 that simultaneous sagittal and tangential resolution could be achieved by placing separate slits at the S and T foci of the collecting mirror. This is true if the wavefronts of the incoming ray bundle are perfectly parallel. However, in a two-mirror symmetric apparatus, such a bundle cannot be produced because of the collimating mirror astigmatism. For example, with reference to Fig. 1, the spatial filter pinhole, A_2, can be placed at either the T or S focus of mirror M_5. In the first case, the wavefronts of the beam travelling toward M_7 are tangentially collimated but sagittally curved; in the second case, the converse is true. From the viewpoint of light-scattering kinematics, this "collimated" beam will be able to conserve momentum with a relatively broad range of scattering vectors lying in the plane containing the wavefront curvature. Thus, even though mirror M_6 forms S and T focal lines of equal sharpness, only a single high-resolution axis actually exists for either position of pinhole A_2.

In the MK VI instrument, the spatial filter pinhole is at the T focus of the collimating mirror so that the probe beam is tangentially collimated. The nature of the wavefront curvature in the sagittal plane may be calculated in a simple fashion from the known (S-T) separation $2AC^*$. Since the pinhole (at the T focus) is closer to the mirror's surface

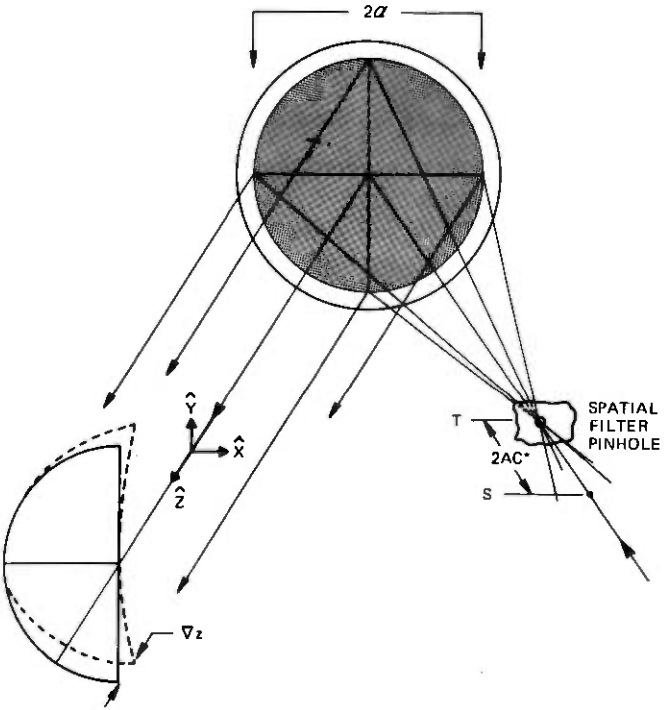


Fig. 18—Illustration of the wavefront curvature existing in the off-axis mirror collimating arm of the MK VI instrument.

than the S focus, the sagittal plane wavefronts appear to diverge from a point source lying behind the mirror. The object distance, q, between this virtual point source and the mirror, may be calculated via the usual paraxial formula. Since $(S\text{-}T) = 2AC^*$ is small compared to the focal length of M_5, we have

$$q \cong \frac{f^2}{2AC^*} = \frac{R^2}{4(2AC^*)}. \tag{74}$$

Inserting the appropriate numerical values, $R = 200$ cm and $2AC^* = 0.336$ cm, gives

$$q \cong 3.98 \times 10^4 \text{ cm}. \tag{75}$$

Figure 18 shows a sketch of the probe-beam constant-phase surfaces with the sagittal curvature greatly exaggerated for clarity. The magnitude of the curvature can be specified in terms of the longitudinal spatial separation between the wavefront and a reference plane which is tangent to the wavefront at the axial ray. Since the virtual sagittal source point lies so far behind the mirror, it does not matter exactly

where in the optical path we calculate this separation. The maximum deviation between the wavefront surface and the reference plane, ∇z, occurs at the sagittal extremes of the beam and is easily found to be given by

$$\nabla z = \frac{\mathcal{A}^2}{2q}. \tag{76}$$

If we take $\mathcal{A} = b^*/2$, where $b^* = 5.0$ cm is the maximum clear aperture diameter in the MK VI instrument, we have

$$\nabla z \cong 10.5 \times 10^{-5} \text{ cm},$$

or roughly 2λ of peak deviation from perfect collimation. In a situation where the full sagittal ($\hat{y}$) aperture height need not be used, it is possible to reduce this deviation substantially since ∇z is proportional to $\mathcal{A}^2$. For example, if we aperture the height of the probe beam to $b = 5$ mm, the peak wave-front deviation is reduced to $\nabla z \cong \lambda/50$, or essentially perfect collimation. Thus, the most desirable probe-beam configuration in the MK VI instrument corresponds to a "flat ribbon" or "sheet" type of illumination.

3.3 Scattered field intensity, spatial coherence, and scattering kinematics in the presence of aberrations

In a light-scattering optical system whose angular resolution capabilities are in some respect dominated by aberration effects—for example, the astigmatic φ blurring in the present instance—we find that other important properties of the observed scattered field are modified by the aberrations as well. In this section, we examine three aspects of normal light-scattering theory that are qualitatively altered by the presence of aberrations:

(*i*) The form of the spatial coherence function for the scattered field.

(*ii*) The application of the normal kinematic restrictions (or wave vector conservation conditions) in the scattering process.

(*iii*) The calculation of the amplitude of the perturbations that give rise to observed levels of scattered or stray light.

The effects of aberrations in all three cases have a straightforward physical interpretation connected with the fact that light scattered into a specific direction (θ, φ) is no longer brought to a diffraction-limited spot focus at the observation plane.

In typical calculations of the scattered field, in which an incoming plane wave is assumed to impinge on the sample, the far-field angular distribution of scattered intensity is shown to be simply the spatial Fourier transform of the refractive-index perturbations in the illumi-

nated volume.[1,3] The light observed at a particular angular position (θ, φ) is contributed by a Fourier component of the refractive index $\mathbf{K}$ given by

$$\mathbf{K} = \mathbf{k}_s - \mathbf{k}_0, \tag{77}$$

where $\mathbf{k}_0$ is the wave vector of the incident beam and $\mathbf{k}_s$ points in the direction of observation (θ, φ) and has a magnitude $|\mathbf{k}_s| = |\mathbf{k}_0| = 2\pi/\lambda_0$. For a finite illuminated volume, the refractive-index perturbations are most usefully represented in terms of a plane-wave Fourier expansion

$$\sum_j \delta n_j \exp(i\mathbf{K}_j \cdot \mathbf{r}'), \tag{78}$$

with the $\mathbf{K}_j$ chosen to make the expansion functions orthonormal over the scattering volume.[27] The scattering of a collimated incident beam by this assembly of plane waves consists of a family of diffracted beams that originate from the $\mathbf{K}_j$'s satisfying the Bragg condition, eq. (77). On the surface of a sphere in the far field, these diffracted beams form a contiguous but essentially nonoverlapping series of diffraction "spots," each associated with a particular $\mathbf{K}_j$. In the usual situation, where the amplitudes of the individual $\mathbf{K}_j$ disturbances are statistically independent, these patterns also delineate areas or solid angles of statistical field correlation. If the far-field scattered radiation is focussed onto the observation plane by an ideal lens or mirror, this contiguous angular distribution of "spots" is imaged one-for-one onto the focal plane. A ray penetrating the reference sphere at an angular position (θ, φ) is imaged onto the focal plane at a transverse position (ξ, η), where, in the small angle limit,

$$\xi = f\theta, \qquad \eta = f\varphi. \tag{79}$$

In this ideal situation, the intensity observed at some (ξ, η) is scattered essentially by a single $\mathbf{K}_j$ plane-wave mode. The measured intensity may, in theory, be used to calculate the mean-square-amplitude of the mode, or vice-versa. Furthermore, the spatial coherence properties of the field at the observation plane are determined uniquely by the angular distribution of intensity within *one* of the diffracted beams.

Formally speaking, the presence of aberrations in the imaging of the reference sphere scattered field produces qualitatively the same effects as any other imperfect focussing of the far-field pattern. The pattern of diffraction spots will be formed with a degree of spot broadening and overlap that depends on the nature and extent of the focussing defect. The scattered light reaching a specific (ξ, η) point at the observation plane is no longer associated with a single $\mathbf{K}_j$ disturbance, but is an appropriately weighted sum of contributions

from a number of modes. As a result, the Bragg condition, eq. (77), is not strictly applicable in relating a particular (ξ, η) observation point to a specific plane-wave disturbance. The aberration or defocussing effect must be understood in detail before the measured intensity, or its time evolution may be used to infer the physical behavior of modes responsible for the scattering. Under defocussed conditions the spatial correlation function is also modified, though its functional form does retain a close resemblance to the intensity pattern associated with a single $\mathbf{K}_j$ diffraction "spot." In the following paragraphs we consider how the residual astigmatic blurring in the MK VI instrument affects the three slit-plane field properties enumerated in the opening paragraph.

3.3.1 Kinematic relations

To understand how the wave vector conservation criterion is to be applied at the slit plane of the present apparatus, we need to know (*i*) the slit-plane intensity pattern formed by scattering from a single plane-wave disturbance, and (*ii*) the relative positioning of the spots from the various allowed $\mathbf{K}_j$. In the MK VI instrument, the intensity pattern associated with a single $\mathbf{K}_j$ is identical to the diffraction- and aberration-affected instrumental profile whose properties were treated in detail in Sections 3.1 and 3.2. At the slit plane, therefore, the single $\mathbf{K}_j$ diffraction spots have the elongated gaussian shape depicted in Fig. 5. Expressed in terms of angular coordinates via eq. (79), the normalized intensity distribution within a "spot" is simply

$$\frac{I(\theta, \varphi)}{I(\theta_j, \varphi_j)} = \exp\left\{-\frac{(\theta - \theta_j)^2}{\delta\theta^2(1/e)}\right\} \exp\left\{-\frac{(\varphi - \varphi_j)^2}{\delta\varphi^2(1/e)}\right\}. \tag{80}$$

The reference point (θ_j, φ_j) specifies the angular position of the spot center, which, in the present case, is correctly predicted by the Bragg condition, eq. (77).

Given a correct form for the intensity distribution within a single $\mathbf{K}_j$ pattern, we must still determine the slit-plane spacing of the spots associated with the family of allowed $\mathbf{K}_j$. Clearly, this spacing depends on the reciprocal lattice of the orthonormal expansion functions $\exp(i\mathbf{K}_j \cdot \mathbf{r}')$ which, in turn, is fixed by the geometry of the scattering volume. For sufficiently small scattering angles, the actual three-dimensional scattering sample can be taken to be equivalent to a two-dimensional phase object placed normal to the incoming probe beam.[28] The scattering disturbances in this "phase-sheet" may be represented in terms of a two-dimensional plane-wave Fourier expansion

$$\sum_{K_x} \sum_{K_y} \cdots \exp(iK_x x) \exp(iK_y y),$$

with the (K_x, K_y) chosen to make the expansion functions orthonormal over the instrument's full aperture. For a rectangular aperture with full-width x and y dimensions b_x and b_y, the allowed $\mathbf{K}_j$ can be obtained from the cyclic boundary condition relations

$$K_x(m) = \frac{2\pi m}{b_x}, \qquad K_y(n) = \frac{2\pi n}{b_y}, \tag{81}$$

where m and n are the integers

$$m, n = 0, \pm 1, \pm 2, \cdots. \tag{82}$$

The scattering angles (θ_j, φ_j) for the central ray of the diffracted beam produced by a particular $\mathbf{K}(m, n)$ are then given by the small-angle Bragg conditions,

$$\begin{aligned} K_x(m) &= k_0\theta_m = \frac{2\pi m}{b_x} \\ K_y(n) &= k_0\varphi_n = \frac{2\pi n}{b_y}. \end{aligned} \tag{83}$$

It follows that the family of diffracted beams are brought to a focus at the slit plane on the vertices of a rectangular mesh whose grid spacings are given by

$$\xi_{SP} = f\theta_{SP} = f\left(\frac{\lambda_0}{b_x}\right) \tag{84}$$

$$\eta_{SP} = f\varphi_{SP} = f\left(\frac{\lambda_0}{b_y}\right). \tag{85}$$

For the MK VI instrument at full aperture ($b_x = 5.0$ cm, $b_y = 5.0$ cm), the equivalent angular mesh spacings are

$$\theta_{SP} = \varphi_{SP} = 10\ \mu\text{rad}. \tag{86}$$

If we imagine the instrumental profile contours of Fig. 5 arranged on such a mesh, there will be little overlap in the $\hat{\theta}$ direction but considerable overlap along $\hat{\varphi}$. The light received at some (θ, φ) point in the slit plane will contain contributions from roughly 10 distinct $\mathbf{K}(m, n)$, each having the same $m(K_x)$ index but differing $n(K_y)$ components. Because the MK VI instrument is capable of probing θ values so close to the diffraction limit, corresponding to a very small m index, $m = 5$–300, these multiple contributions can prove a serious problem. This point is illustrated in Fig. 19, which shows the $(1/e)$ contour of a single $\mathbf{K}(m, n)$ intensity pattern centered at $\theta = 80$ μrad, $\varphi = 0$ superimposed on the slit-plane mesh of the (θ_m, φ_n). The θ and φ axes of the figure can also be labelled in terms of the wave vector components K_x and K_y to which the angles are directly proportional as

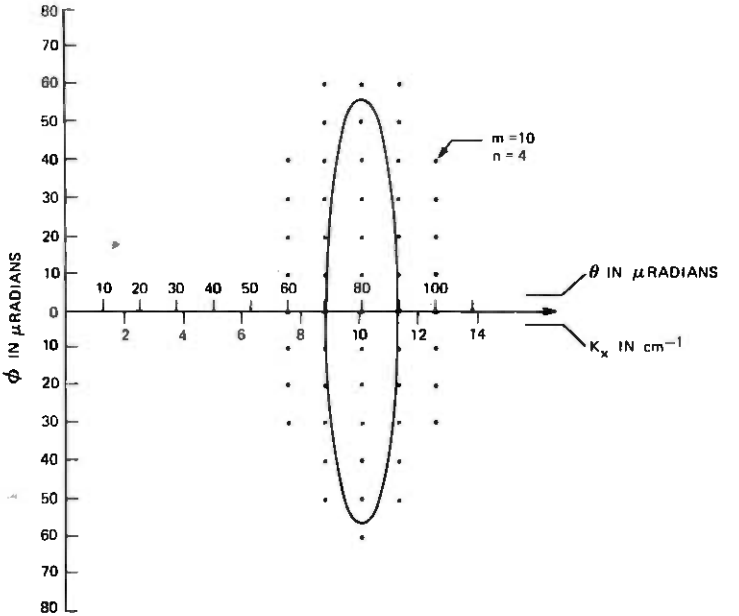


Fig. 19—Resolution function of the MK VI instrument superimposed on the slit-plane mesh points corresponding to the allowed scattering vectors $\mathbf{K}_j$.

indicated in eq. (83). For the eleven mesh points falling within the contour, the magnitude of the associated $\mathbf{K}(m, n)$ varies considerably. For the point at the center of the contour, we have ($m = 8$, $n = 0$) and

$$|\mathbf{K}| = 2\pi[(m/b_x)^2 + (n/b_y)^2]^{\frac{1}{2}} = 10.0 \text{ cm}^{-1},$$

while for the points at the φ extremes ($m = 8$, $n = 5$), we find

$$|\mathbf{K}| = 11.9 \text{ cm}^{-1}.$$

When the physical properties of the modes are strongly $|\mathbf{K}|$ dependent, this overlap can lead to a difficult task in the interpretation of the measured intensity and/or its time dependence.

Clearly, the $|\mathbf{K}|$ smearing effect becomes less significant as θ increases. Less obvious is the fact that the problem of multiple $\mathbf{K}_j$ contributions can be alleviated by stopping down the beam height of the instrument, b_y. As was pointed out in the conclusion of Section 3.2, the cylindrical distortion of the probe-beam wave fronts, which is a manifestation of the collimating mirror's astigmatism, can be made negligibly small by reducing b_y. With $b_y \leqq 0.5$ cm, for example, the beam incident on the sample can be considered as collimated to within the diffraction limit. In this case, the usual kinematic conditions apply in relation to the far-field scattered light; that is, the far-field array of

"spots" now form a contiguous and nonoverlapping pattern. Of course, each spot is elongated along $\hat{\varphi}$ by diffraction-spreading because of the imposed asymmetry of the probe-beam dimensions, i.e, $b_x = 5.0$ cm and $b_y \leqq 0.5$ cm, but the angular spot spacings $\theta_{SP} = (\lambda/b_x)$ and $\varphi_{SP} = (\lambda/b_y)$ are correspondingly asymmetric. Given this particular situation in the far field, we must still consider the effect of the astigmatism associated with the light-collecting mirror. As b_y is *decreased*, the $\hat{\varphi}$ direction astigmatic blurring at the slit plane is *decreased* proportionally,* while $\hat{\varphi}$ blurring due to diffraction increases. At some point, a crossover occurs beyond which diffraction spread dominates the slit-plane imaging. In this limit, the collecting mirror appears aberration free and the far-field pattern of $\mathbf{K}_j$ spots undergoes the normal one-to-one, no-overlap mapping onto the slit plane. For the MK VI configuration, the crossover occurs at $b_y \approx 0.5$ cm or roughly $\frac{1}{10}$ of the full design aperture height. Of course, it should be noted that under fully diffraction-limited conditions ($b_x = 5.0$ cm, $b_y < 0.5$ cm), the instrument retains its very asymmetric resolution profile. What we have done is to make the instrument appear to be in "good focus" by introducing a sufficient amount of φ direction diffraction spreading to swamp the aberration defocussing. The price paid for this is that the instrument becomes incapable of probing scattering disturbances having as long a $\hat{y}$ direction wavelength (that is as small a value of K_y) as can be resolved in the $\hat{x}$ direction.

3.3.2 Relation between the slit-plane intensity and the scattering cross-section of the sample

In the absence of aberrations or other defocussing problems, the far-field or observation plane scattered intensity can be related easily to the mean-square amplitude of the scattering perturbations using the standard integral expression for the scattered field.[1,3,27] In the presence of imaging aberrations, the total scattered power per plane-wave mode is unchanged; however, now the calculation of the slit-plane intensity is complicated by the overlap of the various $\mathbf{K}_j$ diffraction spots. For the MK VI instrument, a relation between the scattered power per mode and the observed slit-plane intensity may be obtained as follows.

The slit-plane intensity corresponding to a single $\mathbf{K}_j$ disturbance can be written down formally as

$$|E_{nm}(\theta, \varphi)|^2 = |E_{nm}(\theta_m, \varphi_n)|^2 \exp\left[-\frac{(\theta - \theta_m)^2}{\delta\theta^2(1/e)}\right] \exp\left[-\frac{(\varphi - \varphi_n)^2}{\delta\varphi^2(1/e)}\right] \quad (87)$$

with slit-plane position specified in angular units. The position of the

* See Figure 13 and the discussion pertaining to this figure.

central ray of the pattern (θ_m, φ_n) is given by eq. (83). The total scattered power in the single $\mathbf{K}_j$ pattern is just

$$P_s(\mathbf{K}_j) = \int\int d\theta d\varphi |E_{nm}(\theta, \varphi)|^2$$
$$= \pi\delta\theta(1/e)\delta\varphi(1/e)|E_{nm}(\theta_m, \varphi_n)|^2. \quad (88)$$

The problem now is to relate the observed total slit-plane intensity to the peak mode intensities, $|E_{nm}(\theta_m, \varphi_n)|^2$. The intensity observed at the position (θ, φ) is the sum over all possible mode contributions:

$$|E_s(\theta, \varphi)|^2$$
$$= \sum_{n,m} |E_{nm}(\theta_m, \varphi_n)|^2 \exp\left[-\frac{(\theta-\theta_m)^2}{\delta\theta^2(1/e)}\right] \exp\left[-\frac{(\varphi-\varphi_n)^2}{\delta\varphi^2(1/e)}\right]. \quad (89)$$

If $|E_{nm}(\theta_m, \varphi_n)|^2$ is independent of (n, m) over the range where the gaussian terms are nonvanishing, eq. (89) can be simplified to give

$$|E_s(\theta, \varphi)|^2 = |E_{nm}(\theta_m, \varphi_n)|^2 S_m S_n, \quad (90)$$

where S_m and S_n are the factored sums

$$S_m = \sum_m \exp - \{[\theta - m(\lambda_0/b_x)]^2/\delta\theta^2(1/e)\} \quad (91)$$

$$S_n = \sum_n \exp - \{[\varphi - n(\lambda_0/b_y)]^2/\delta\varphi^2(1/e)\}. \quad (92)$$

Consider the S_m sum expressed in the following dimensionless form

$$S_m = \sum_m \exp - \left\{\frac{\theta}{\delta\theta(1/e)} - m\left(\frac{\lambda}{b_x}\right)\frac{1}{\delta\theta(1/e)}\right\}^2$$
$$= \sum_m \exp(-[u - ma]^2), \quad (93)$$

where

$$u \equiv \frac{\theta}{\delta\theta(1/e)}, \qquad a \equiv \frac{1}{\delta\theta(1/e)}\frac{\lambda_0}{b_x}. \quad (94)$$

Although the indicated summation cannot be carried out explicitly, it can be expressed in a more useful form via the identity

$$\sum_m \exp[-(u - ma)^2] = \frac{\sqrt{\pi}}{a}\sum_m \exp\left(-\frac{m^2\pi^2}{a^2}\right)\cos(2m\pi u/a), \quad (95)$$

which is an immediate corollary of Poisson's formula.[38] Applying this identity to the S_m summation yields

$$S_m = \frac{\sqrt{\pi}\delta\theta(1/e)}{(\lambda/b_x)}\sum_m \exp\left[-\frac{m^2\pi^2\delta\theta^2(1/e)}{(\lambda/b_x)^2}\right]\cos\left[\frac{2m\pi\theta}{(\lambda/b_x)}\right]. \quad (96)$$

For the numerical parameters relevant to the MK VI instrument, $\delta\theta(1/e) = 9.6\ \mu\text{rad}$, $(\lambda/b_x) = 10\ \mu\text{rad}$, we need retain only the $m = 0$ term in the right-hand side of eq. (96). With an error less than 10^{-4}, we may take

$$S_m = \frac{\sqrt{\pi}\delta\theta(1/e)}{(\lambda/b_x)}. \tag{97}$$

A similar result follows easily for S_n. We then have for the total slit-plane intensity

$$|E_s(\theta, \varphi)|^2 = |E_{nm}(\theta_m, \varphi_n)|^2 \frac{\sqrt{\pi}\,\delta\theta(1/e)}{(\lambda/b_x)} \frac{\sqrt{\pi}\,\delta\varphi(1/e)}{(\lambda/b_y)}. \tag{98}$$

As this result shows, the various elongated gaussian patterns associated with the individual $\mathbf{K}_j$ modes overlap in the slit plane to produce an essentially uniform illumination.

The unknown intensity factors $|E_{nm}(\theta_m, \varphi_n)|^2$ can now be eliminated between eqs. (88) and (98) to give the desired relationship between the overall slit-plane intensity and the scattered power per mode, namely,

$$P_s(\mathbf{K}_j) = |E_s(\theta_j, \varphi_j)|^2(\lambda/b_x)(\lambda/b_y). \tag{99}$$

Equation (99) is the basic result which allows the measured intensity to be related quantitatively to the amplitudes of the individual scattering perturbations.*

Note that the power actually contained within the angular area of a single $\mathbf{K}_j$ pattern

$$|E_s(\theta_j, \varphi_j)|^2\delta\theta(1/e)\delta\varphi(1/e)$$

is larger than the scattered power per mode by the factor

$$\frac{\delta\theta(1/e)}{(\lambda/b_x)} \frac{\delta\varphi(1/e)}{(\lambda/b_y)}.$$

This ratio gives a rough gaussian weighted measure of the number of modes that contribute to the intensity reaching a particular (θ, φ).

3.3.3. The spatial coherence function of the slit-plane field

In light-scattering experiments designed to extract spectral information from the scattered field using photocurrent correlation techniques, the feasibility of a particular measurement is critically dependent on the range of transverse spatial correlation that characterizes the observation plane field.[2,3,27] The extent of the correlation is de-

* See the discussion which follows eq. (113) and leads to eq. (124).

scribed quantitatively by the normalized mutual coherence function[39,40]

$$T(\xi, \eta; \Delta\xi, \Delta\eta) = \frac{\langle \mathbf{E}_s(\xi, \eta; t) \cdot \mathbf{E}_s^*(\xi + \Delta\xi, \eta + \Delta\eta; t) \rangle}{\{\langle |\mathbf{E}_s(\xi, \eta; t)|^2 \rangle \langle |\mathbf{E}_s(\xi + \Delta\xi, \eta + \Delta\eta; t)|^2 \rangle\}^{\frac{1}{2}}}, \quad (100)$$

where (ξ, η) and $(\xi + \Delta\xi, \eta + \Delta\eta)$ are two arbitrary points in the observation plane. The angular brackets denote an appropriate ensemble or time average. The function $T(\cdots)$ reaches its maximum value, $T(\cdots) = 1$, for $\Delta\xi = \Delta\eta = 0$ and, in general, decreases smoothly to zero as $\Delta\xi$ and/or $\Delta\eta$ increase. The contour in $\Delta\xi$ and $\Delta\eta$ around (ξ, η) on which the coherence function reaches some specified numerical value may be taken as a measure of the area over which there is correlated temporal behavior of the two field amplitudes.

When the main probe beam is derived from a source having perfect transverse spatial coherence, as is the case here, then the presence of the spatial incoherence in the scattered field is totally attributable to the scattering processes taking place in the illuminated volume. The spatial coherence properties of the scattered field are uniquely determined at the exit face of the sample and are most easily specified analytically by calculating the mutual coherence function on a far-field reference sphere, 0, centered on the scattering volume. In purely formal terms, we can write

$$T_0(\mathbf{r}, \boldsymbol{\varrho}) = \frac{\langle \mathbf{E}_s(\mathbf{r}, t) \cdot \mathbf{E}_s^*(\mathbf{r} + \boldsymbol{\varrho}, t) \rangle}{\{\langle |\mathbf{E}_s(\mathbf{r}, t)|^2 \rangle \langle |\mathbf{E}_s(\mathbf{r} + \boldsymbol{\varrho}, t)|^2 \rangle\}^{\frac{1}{2}}}, \quad (101)$$

where $\mathbf{r}$ and $\mathbf{r} + \boldsymbol{\varrho}$ both terminate on the surface of the far-field sphere, 0. Generally speaking, $T_0(\mathbf{r}, \boldsymbol{\varrho})$ can be calculated in a straightforward fashion once it is assumed that the scattering perturbations satisfy certain basic stochastic criteria.

The relationship between the observation plane coherence function $T(\xi, \eta; \Delta\xi, \Delta\eta)$ and the far-field function $T_0(\mathbf{r}, \boldsymbol{\varrho})$ depends, of course, on the detailed characteristics of the optical system which collects and images the scattered light, and must include the effects of aberrations. There are two alternative procedures that may be used to obtain this relationship. The first involves the use of the plane wave $\mathbf{K}_j$ expansion of the scattering perturbations that was introduced in the beginning of this section. For the MK VI instrument, we have already calculated the slit-plane field produced by the scattering from the individual $\mathbf{K}_j$. In the notation of eq. (89), we have

$$\mathbf{E}_s(\xi, \eta; \mathbf{K}_j) = \mathbf{E}_{nm}(\xi_m, \eta_n) \exp\left[-\frac{(\xi - \xi_m)^2}{2f^2\delta\theta^2(1/e)} \right] \exp\left[-\frac{(\eta - \eta_n)^2}{2f^2\delta\varphi^2(1/e)} \right], \quad (102)$$

where f is the effective focal length of the light collection system. In

theory, therefore, we could calculate $T(\xi, \eta; \Delta\xi, \Delta\eta)$ directly by expressing the total slit-plane field as a sum over the $\mathbf{E}_s(\xi, \eta; \mathbf{K}_j)$ and then performing the statistical average indicated in eq. (100). In the absence of aberrations or other imaging defects, this direct method represents the simplest approach. For perfect imaging, the individual plane-wave scattered field patterns are essentially nonoverlapping at the observation plane and the coherence function is effectively dominated by the contribution of a single $\mathbf{K}_j$ term. However, when imaging errors produce a significant overlap of the $\mathbf{E}_s(\xi, \eta; \mathbf{K}_j)$ at the observation plane, as is the case for the present apparatus, then obtaining the analytical form of $T(\xi, \eta; \Delta\xi, \Delta\eta)$ by the direct method becomes a difficult mathematical problem.

The second alternative approach involves a direct calculation of the far-field coherence function $T_0(\mathbf{r}, \boldsymbol{\rho})$ from which $T(\xi, \eta; \Delta\xi, \Delta\eta)$ is obtained by using the fundamental laws that govern the "propagation" of mutual coherence in an optical system. This latter method is generally the more useful when the light-collection system departs significantly from ideal imaging.

For the scattering angles relevant to the MK VI instrument, the two-dimensional "phase sheet" model of the scattering sample may be used to simplify the calculation of $T_0(\mathbf{r}, \boldsymbol{\rho})$. For this two-dimensional model object, the reference sphere coherence function is given by the van Cittert-Zernike theorem[39] as

$$T_0(\theta, \varphi; \theta - \theta'; \varphi - \varphi') = \frac{2\pi \int_S \int dxdy\, |E_0(x, y)|^2 \exp\{ik_0[(\theta - \theta')x + (\varphi - \varphi')y]\}}{2\pi \int_S \int dxdy\, |E_0(x, y)|^2}, \quad (103)$$

where both $\mathbf{r}$ and $\boldsymbol{\rho}$ have been expressed in the cartesian angular coordinates θ and φ. In eq. (103), the factor $|E_0(x, y)|^2$ is the illumination function of the object, in our case the "phase-sheet" sample. The surface integral is to be taken over the entire (x, y) plane or over the open aperture of the object, as appropriate. It should be noted that the van Cittert-Zernike theorem will hold as long as the perturbations in the scattering "phase sheet" have a correlation distance, which is short compared to the characteristic spatial dimensions of $|E_0(x, y)|^2$. This condition is, in general, well satisfied in the typical scattering experiment.

At small angles, where eq. (103) is valid, the far-field spatial coherence function is independent of the absolute angular position of either observation point and depends only on the separations $(\theta - \theta')$ and $(\varphi - \varphi')$. In terms of these difference variables, $T_0(\cdots)$ is just

the normalized Fourier transform of the source *intensity*. As such, it bears an extremely close resemblance to the instrumental profile calculated in Sections 3.1 and 3.2.

For the MK VI instrument, the illumination function is the gaussian

$$|E_0(x, y)|^2 = E_0^2 \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right) \tag{104}$$

and we have from eq. (103)

$$T_0(\theta, \varphi; \theta - \theta', \varphi - \varphi') = \frac{\int_{-b_x/2}^{b_x/2} \exp(-x^2/\sigma^2) \exp[ik_0(\theta - \theta')x]dx}{\int_{-b_x/2}^{b_x/2} \exp(-x^2/\sigma^2)dx} \times \frac{\int_{-b_y/2}^{b_y/2} \exp(-y^2/\sigma^2) \exp[ik_0(\varphi - \varphi')y]dy}{\int_{-b_y/2}^{b_y/2} \exp(-y^2/\sigma^2)dy}, \tag{105}$$

where b_x and b_y are the aperture dimensions at the scattering sample. As is evident from eq. (105), the coherence function factors for the case of gaussian illumination and we can write

$$T_0(\theta, \varphi; \theta - \theta', \varphi - \varphi') = T_0'(\Delta\theta)T_0'(\Delta\varphi),$$

where $\Delta\theta = \theta - \theta'$ and $\Delta\varphi = \varphi - \varphi'$. The functions $T_0'(\Delta\theta)$ and $T_0'(\Delta\varphi)$ are given by the appropriate integrals in eq. (105). Each of these integrals is a finite domain Fourier transform of a gaussian kernel of the type considered in detail in Section 3.1 with respect to aperture apodization and vignetting. The only difference is that in eq. (105), the "intensity," $\exp(-x^2/\sigma^2)$, replaces the "field," $\exp(-x^2/2\sigma^2)$, which appeared in the diffraction calculations. It is not hard to show that the factored coherence functions $T_0'(\Delta\theta)$ and $T_0'(\Delta\phi)$ are identical to the normalized intensity profiles of Fig. 9 if one uses the correspondence

$$|T_0'(\psi)| = \frac{I(\psi/2)}{I(0)}. \tag{106}$$

Given the form of the far-field reference sphere function T_0, we must now determine the relationship between T_0 and the desired slit-plane correlation function.

One of the fundamental results of coherence theory is that second-order mutual coherence functions, such as $T(\cdots)$, propagate according to the wave equations as "field" variables. That is, once $T(\cdots)$ is specified on any surface in an optical system, its form on any other surface in the system may be found by treating the coherence function

as one would any electric field distribution. Therefore, all of the usual wave-diffraction and/or geometrical-optics approaches used to analyze wave-front propagation in an optical system are directly applicable to the coherence function.

For the MK VI apparatus, the coherence "field" described by T_0 is identical to the far-field electric field that describes the instrument's directly transmitted probe beam, except for a numerical change in the beam-width parameter σ. The effective beam width which characterizes the mutual coherence "field," σ_T, is related to the actual beam width of the instrument, σ^*, by the result

$$\sigma_T^2 = \frac{(\sigma^*)^2}{2}. \tag{107}$$

Except for this numerical change, the diffraction and aberration results of Sections 3.1 and 3.2 may be used intact to describe the slit-plane coherence function. In terms of angular coordinates at the slit and the widths $\delta\theta(1/e)$ and $\delta\varphi(1/e)$, which were used to characterize the instrumental profile, we have easily

$$\begin{aligned} T(\theta, \varphi; \theta - \theta', \varphi - \varphi') &= T'(\Delta\theta)T'(\Delta\varphi) \\ &= \exp\left\{-\frac{(\theta - \theta')^2}{4\delta\theta^2(1/e)}\right\} \exp\left\{-\frac{(\varphi - \varphi')^2}{4\delta\varphi^2(1/e)}\right\}. \end{aligned} \tag{108}$$

In the slit plane, as was the case on the surface of the far-field reference sphere, the slit-plane coherence functions are related to the intensity profile of the transmitted beam by the transformation

$$|T'(\psi)| = \frac{I(\psi/2)}{I(0)}.$$

Equation (108) is the basic result which may be used to evaluate the scattered or stray-light power-per-coherence region or estimate the number of coherence regions encompassed by a particular choice of main-slit size. For example, given the slit-plane scattered intensity $|E_s(\theta, \varphi)|^2$, we can form the weighted integral

$$\frac{dP_s(\theta, \varphi)}{d\Omega_{COH}} = \int\int d\theta' d\varphi' \, |E_s(\theta', \varphi')|^2 [T'(\theta - \theta')T'(\varphi - \varphi')]^2, \tag{109}$$

which is a useful measure of the power-per-coherence solid angle as measured at the slit.[2,3,27] In general, $|E_s(\theta', \varphi')|^2$ is slowly varying over the angular range where $[T'(\theta - \theta')T'(\varphi - \varphi')]^2$ is nonvanishing and can be removed from the integral to give

$$\frac{dP_s(\theta, \varphi)}{d\Omega_{COH}} = |E_s(\theta, \varphi)|^2 \overline{\Delta\theta}_{COH} \overline{\Delta\varphi}_{COH}, \tag{110}$$

where the mean full-width coherence angles, $\overline{\Delta\theta}_{COH}$ and $\overline{\Delta\varphi}_{COH}$, are defined by the integrals

$$\begin{aligned}\Omega_{COH} &= \overline{\Delta\theta}_{COH}\overline{\Delta\varphi}_{COH} \\ &= \int [T'(\Delta\theta)]^2 d(\Delta\theta) \int [T'(\Delta\varphi)]^2 d(\Delta\varphi).\end{aligned} \tag{111}$$

Combining eqs. (108) and (109) gives for the MK VI instrument

$$\begin{aligned}\overline{\Delta\theta}_{COH} &= \sqrt{2\pi}\delta\theta(1/e) = 24.1\ \mu\text{rad} \\ \overline{\Delta\varphi}_{COH} &= \sqrt{2\pi}\delta\varphi(1/e) = 140.4\ \mu\text{rad}\end{aligned} \tag{112}$$

and eq. (110) becomes

$$\frac{dP_s(\theta,\ \varphi)}{d\Omega_{COH}} = |E_s(\theta,\ \varphi)|^2 \sqrt{2\pi}\delta\theta(1/e)\sqrt{2\pi}\delta\varphi(1/e). \tag{113}$$

Earlier in this section, we obtained an expression for $|E_s(\theta, \varphi)|^2$ based on a plane-wave-mode expansion of the scattering perturbations. That result may be used in eq. (113) to yield a relationship between the observed scattered power-per-coherence solid angle and the scattered power-per-$\mathbf{K}_j$ mode. From eqs. (99) and (113), we find

$$\frac{dP_s(\theta_j,\ \varphi_j)}{d\Omega_{COH}} = \frac{\sqrt{2\pi}\delta\theta(1/e)}{(\lambda/b_x)} \times \frac{\sqrt{2\pi}\delta\varphi(1/e)}{(\lambda/b_y)} P_s(\mathbf{K}_j). \tag{114}$$

The product of the correction factors

$$\frac{\sqrt{2\pi}\delta\theta(1/e)}{(\lambda/b_x)} \times \frac{\sqrt{2\pi}\delta\varphi(1/e)}{(\lambda/b_y)}$$

is a rough measure of the number of modes that contribute to the power observed in a single coherence region at the slit plane, while the individual terms indicate the extent of the multiple mode contribution in the $\hat{\theta}$ and $\hat{\varphi}$ directions. For the MK VI instrument at full aperture, the numerical values of the correction factors are

$$\begin{aligned}\frac{\sqrt{2\pi}\ \delta\theta(1/e)}{(\lambda/b_x)} &= 2.4 \\ \frac{\sqrt{2\pi}\ \delta\varphi(1/e)}{(\lambda/b_y)} &= 14.0.\end{aligned} \tag{115}$$

The results given in eqs. (99), (113), and (114) together with the known form of the instrumental profile may be combined in various ways to calculate normalized scattering cross sections from measured slit-plane intensities. One important calculation of this type is to express the observed stray-light levels in the MK VI apparatus in

terms of an equivalent scattering cross section. For the experimental profile curves shown in Fig. 4, we may write down an analytical expression for the measured stray-light photocurrent, $i(\theta)$, as

$$i(\theta) = \alpha \int_{-\Delta\theta_{SL}/2}^{\Delta\theta_{SL}/2} \int_{-\Delta\varphi_{SL}/2}^{\Delta\varphi_{SL}/2} |E_{st}(\theta, \varphi)|^2 d\theta d\varphi, \tag{116}$$

where $|E_{st}(\theta, \varphi)|^2$ is the stray-light intensity at the slit plane and $\Delta\theta_{SL}$ and $\Delta\varphi_{SL}$ specify the full-width slit dimensions in angular units. The proportionality factor α relates the photocurrent to the optical power passed by the slit and includes the detector quantum efficiency, light-collection losses, etc. If the intensity $|E_{st}(\theta, \varphi)|^2$ is relatively constant over the slit aperture, we have simply

$$i(\theta) = \alpha |E_{st}(\theta, 0)|^2 \Delta\theta_{SL} \Delta\varphi_{SL}, \tag{117}$$

where we have assumed that $\Delta\varphi_{SL}$ is situated symmetrically around $\varphi = 0$. Combining this result with eq. (110) gives the relation between the measured photocurrent and the stray-light power-per-coherence solid angle as

$$i(\theta) = \alpha \frac{dP_{st}(\theta, 0)}{d\Omega_{COH}} \frac{\Delta\theta_{SL}}{\overline{\Delta\theta}_{COH}} \frac{\Delta\varphi_{SL}}{\overline{\Delta\varphi}_{COH}}. \tag{118}$$

To eliminate the unknown proportionality constant α, we make use of photocurrent observed at $\theta = 0$, the peak of the directly transmitted beam. Given the normalized slit-plane intensity profile of the direct beam, $I(\theta, \varphi)/I(0, 0)$, we can calculate the fraction of the total beam power, P_0, passed by the slit at $\theta = 0$ as

$$\frac{\int_{-\Delta\theta_{SL}/2}^{\Delta\theta_{SL}/2} \int_{-\Delta\varphi_{SL}/2}^{\Delta\varphi_{SL}/2} \frac{I(\theta, \varphi)}{I(0, 0)} d\theta d\varphi}{\iint_{\text{all }\Omega} \frac{I(\theta, \varphi)}{I(0, 0)} d\theta d\varphi} \equiv \gamma, \tag{119}$$

where for the MK VI apparatus we have

$$\frac{I(\theta, \varphi)}{I(0, 0)} = \exp\left[- \frac{\theta^2}{\delta\theta^2(1/e)} \right] \exp\left[- \frac{\varphi^2}{\delta\varphi^2(1/e)} \right].$$

The numerical value of the error function integrals in eq. (119) could be obtained from tabulated results for particular values of $\Delta\theta_{SL}$ and $\Delta\varphi_{SL}$; however, in the present case where the slit dimensions satisfy the inequalities

$$\begin{aligned} \Delta\theta_{SL} &\ll \delta\theta(1/e) \\ \Delta\varphi_{SL} &\gg \delta\varphi(1/e), \end{aligned} \tag{120}$$

we have the more useful analytical result

$$\gamma \cong \frac{\Delta\theta_{SL}}{\sqrt{\pi}\delta\theta(1/e)}. \tag{121}$$

The measured peak photocurrent, $i(0)$, is then

$$i(0) = \alpha P_0 \frac{\Delta\theta_{SL}}{\sqrt{\pi}\delta\theta(1/e)}. \tag{122}$$

Dividing eq. (118) by eq. (122) gives the useful result

$$\frac{i(\theta)}{i(0)} = \frac{1}{P_0}\frac{dP_{sl}(\theta, 0)}{d\Omega_{COH}}\frac{1}{\sqrt{2}}\frac{\Delta\varphi_{SL}}{\overline{\Delta\varphi_{COH}}}. \tag{123}$$

If desired, the quantity $dP_{sl}(\theta, 0)/d\Omega_{COH}$ can be replaced with the stray-light power per mode, $P_{sl}(K_j)$, by using eq. (114). This gives the very useful relationship

$$\frac{i(\theta)}{i(0)} = \frac{\mathcal{P}_{sl}(\mathbf{K}_j)}{\mathcal{P}_0}\frac{\sqrt{2\pi}\,\delta\theta(1/e)}{(\lambda/b_x)}\frac{\Delta\varphi_{SL}}{(\lambda/b_y)}. \tag{124}$$

IV. EMPIRICAL OBSERVATIONS ON THE STRAY-LIGHT BEHAVIOR OF OPTICAL ELEMENTS AT VERY SMALL ANGLES

Very little information of a quantitative nature is available concerning the imperfection scattering of optical elements at very small angles. As a result, the design and testing process leading to the present VSA instrument involved a significant amount of trial and error evaluation of various optical systems in a search for the desired stray-light performance. During this process, a certain amount of empirical information was obtained relating to the imperfection-scattering question. This section presents a brief discussion of these observations and their influence on the configuration adopted for the MK VI instrument.

4.1 Reflecting versus refracting optics

It is clear from a comparison of Figs. 1 and 6 that the implementation of a VSA scattering instrument using lenses would be significantly less involved than the MK VI off-axis mirror arrangement. The refracting system also has the advantage of strictly zero off-axis aberrations (coma, astigmatism, and distortion), although a "best form" single-element lens does have eight times the spherical aberration of an equivalent spherical mirror.[37] In fact, the earliest version of the present apparatus utilized precisely the kind of "straight-through" lens system illustrated in Fig. 6. This arrangement was abandoned because of

two problems:

(*i*) The presence of Newton's interference fringes crossing the illuminated field.
(*ii*) An excessive stray-light background.

The first problem arises because of the partial reflectivity of the two lens surfaces and can be solved to some extent through the use of antireflection (AR) coatings. However, even the best antireflection coated lens will form far-field Newton fringes with an integrated intensity of about $\frac{1}{4}$ percent of the incident beam power. This fact makes the refracting components generally unacceptable in a VSA system. The presence of these extraneous reflections and their associated interference fringes creates an intense fixed-pattern nuisance background which can make it impossible to observe the angular dependence of the sample scattered light. The stray-light background problem is a manifestation of small-angle scattering at the lens which may originate from three possible sources:

(*i*) Lens surface "roughness" or nonconformity (at least two surfaces).
(*ii*) Index of refraction inhomogeneity in the lens bulk material.
(*iii*) AR coating thickness nonuniformity (at least two surfaces).

By way of comparison, the possible sources of imperfection scattering from a first-surface reflector are

(*i*) Mirror surface "roughness" or nonconformity (one surface).
(*ii*) Reflective coating(s) thickness nonuniformity.
(*iii*) Reflective coating reflectivity nonuniformity.

From a theoretical standpoint, one should be able to evaluate the seriousness of each of these defects *a priori* by calculating the surface and/or bulk inhomogeneity scattering. This calculation is straightforward if one has available the spatial form of the roughness in terms of the spatial correlation function and the rms roughness amplitude. The effect of roughness or inhomogeneity is to impose a spatially random-phase perturbation in the optical path. The scattering that takes place as a result of this perturbation can be calculated via the same "phase-object" approach which is used for the primary scattering sample (see Section 2.6). The stray-light intensity observed at some specified scattering angles θ and φ is given by the Fourier transform of the roughness correlation function at a wave vector $|\mathbf{K}| = 2\pi/\Lambda$ satisfying the appropriate small-angle kinematic conditions. Unfortunately, the roughness wavelengths corresponding to the angular range of interest here (10^{-3} cm $\lesssim \Lambda \lesssim 1$ cm) are determined by a

spatial region of the roughness correlation function about which very little is presently known. This wavelength regime presents difficult measurement problems and is generally not probed by conventional roughness-testing techniques. The data that is available comes from two measurement techniques that tend to flank this regime on the short and long wavelength sides:

(*i*) The FECO* interferometer and allied methods[41,42] that exhibit good surface deviation resolution, 1 Å to 10 Å, but are useful only at short wavelengths ($\Lambda \lesssim 1000$ Å).

(*ii*) Conventional "surface-conformity" techniques such as the Foucault knife-edge and Twyman-Green interferometer tests that are useful primarily at longer roughness wavelengths (0.1 cm to 100 cm) and which exhibit relatively poor surface deviation resolution (50 Å → 2000 Å).

The stray-light measurements that were made during the course of the evolution of the present instrument provided the most sensitive roughness and inhomogeneity test for this awkward wavelength range. It was found experimentally that, for lenses and mirrors of the same fraction of the "state-of-the-art," the stray-light level of a refracting instrument was roughly 20 times that of its reflecting counterpart. In neither case did the VSA stray-light level correlate well with known short wavelength roughness and inhomogeneity data. Both types of components exhibited a spatial roughness spectrum that was strongly enhanced at long wavelengths. This enhancement did not appear to depend as strongly on the "surface-figure" of the component as one might be led to expect by qualitative theoretical arguments.

Comparisons were also made between mirror components having multilayer dielectric coatings and those with a conventional SiO-protected aluminized surface. The aluminized coatings can suffer from a spatially varying reflectivity caused by surface oxidation while high-reflectivity dielectric films tend to have a significantly smaller reflectivity modulation. However, the stray-light measurements showed no significant difference between the two types of coatings on similar "quality" substrates. Apparently the cumulative roughness of the greater number of dielectric layers offsets the dielectric coating's potential advantage.

4.2 Main scanning slit

Another major contributor to the stray-light level in earlier versions of the MK VI apparatus was the main angle-scanning slit. The slit selected for this application is a commercial Spex unit normally used

* Fringes of Equal Chromatic Order.

as an intermediate or exit slit on a double-grating spectrometer known for its low stray-light background. This fact not withstanding, severe stray-light problems were encountered in predecessors of the MK VI that had this slit located directly at the focal plane of the collecting mirror (mirror M_6 shown in Fig. 1). The origin of this problem was traced to scattering of the direct beam by the beveled surfaces of the slit jaws and to quasi-specular reflection from the slightly flattened and rounded jaw edges. This source of background by itself was of sufficient intensity to completely swamp the sum total of all other stray-light sources in the instrument.

This problem was solved in the MK VI apparatus by occulting the directly transmitted beam, before it reached the main scanning slit, with a precision knife-edge fabricated of highly attenuating black glass plate. The use of glass instead of metal permits the edge defining surfaces to be optically polished without cold flaw and rounding. In addition, the included angle formed by the edge surfaces is made obtuse, rather than the acute angle normally used, to avoid the feathering problems and surface irregularity enhancement associated with small included angles. The salient geometrical features of the knife-edge are illustrated in Fig. 20. The actual occulting edge is formed by a single beveling operation on polished flat stock and is oriented in use such that an incoming ray strikes the beveled face at the quasi-Brewster angle. The beveling angle is chosen so that the ray which is refracted into the plate travels parallel to the plate surfaces and is totally absorbed.

The improvement in stray-light level obtained by using the knife-edge to occult the direct beam, rather than relying solely on the main scanning slit, can be seen in Figs. 3 and 4. The improvement amounts to roughly an order of magnitude over the angular range of interest.

4.3 Aberration corrections and stray light

The reader familiar with optical system design will realize that the aberrations present in the MK VI instrument could be "corrected" using well-known techniques. However, the application of these correction methods has two drawbacks: cost and reduced stray-light performance. The simplest corrective measures, those which add the fewest number of optical elements to the basic apparatus, entail the use of off-axis fabricated, aspheric reflecting and/or refracting elements. These types of elements are, in general, exceedingly costly to fabricate. More sophisticated aberration-corrective designs, utilizing only spherical optics, require a larger number of additional elements. In either case, of course, the presence of additional optical surfaces means degraded stray-light performance. Furthermore, any corrective design

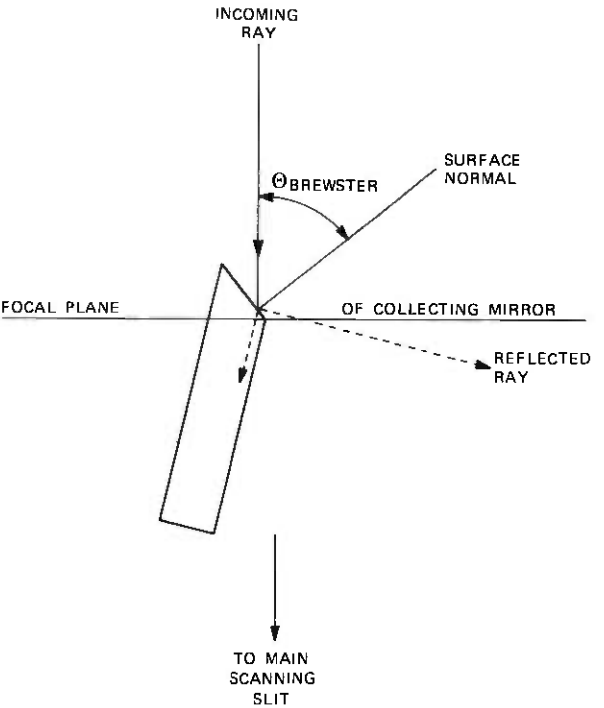


Fig. 20—Geometrical features of the occulting knife-edge used in the MK VI apparatus.

relying on the use of refracting elements will be further penalized by the excessive small-angle stray light which these elements generate.

APPENDIX A

Finite Slit-Width Effects in the Scanning of Gaussian Intensity Profiles

When a gaussian focal-plane profile is scanned by a finite-width slit, the transmitted power is proportional to the integral

$$J(\xi_0, \Delta) = \int_{\xi_0-\Delta}^{\xi_0+\Delta} \exp\left[-\xi^2/\delta\xi^2(1/e)\right]d\xi, \tag{125}$$

where ξ_0 gives the position of the center of a slit whose width is 2Δ. By writing the spatial coordinate ξ as

$$\xi = \xi_0 + \zeta \tag{126}$$

and making a change of variable, we may put eq. (125) into the form

$$J(\xi_0, \Delta) = \exp\left[-\xi_0^2/\delta\xi^2(1/e)\right] \times \int_{-\Delta}^{\Delta} \exp\left[-(2\xi_0 + \zeta^2)/\delta\xi^2(1/e)\right]d\zeta. \tag{127}$$

Table IX—Angular displacement, θ_0, at which an unbroadened gaussian and a slit-broadened gaussian reach specified fractions of peak intensity

$\frac{J(\theta_0, \Delta)}{J(0, \Delta)}$	θ_0 Unbroadened $\Delta_s = 0$ μrad (μrad)	θ_0 Slit Broadened $\Delta_s = 1$ μrad (μrad)	% Increase
1	0	0	—
1/2	7.993	8.022	0.36
$1/e$	9.600	9.635	0.36
10^{-1}	14.567	14.620	0.36
10^{-2}	20.601	20.675	0.36
10^{-3}	25.231	25.322	0.36
10^{-4}	29.135	29.239	0.36
10^{-5}	32.573	32.690	0.36
10^{-6}	35.682	35.810	0.36
10^{-7}	38.541	38.678	0.36
10^{-8}	41.203	41.348	0.36

For reasonably small values of the ratio $[\Delta/\delta\xi(1/e)]$, the gaussian term in the integrand of eq. (127) may be approximated by the leading term in its Taylor's series expansion

$$\exp\left[-\zeta^2/\delta\xi^2(1/e)\right] = 1 - \frac{\zeta^2}{\delta\xi^2(1/e)} + \cdots,$$

with a maximum error $\exp\left[-\Delta^2/\delta\xi^2(1/e)\right]$. Within this approximation, the remaining integral can be calculated in a straightforward manner to give

$$J(\xi_0, \Delta) = (2\Delta)\exp\left[-\xi_0^2/2\delta\xi^2(1/e)\right]\left\{\frac{\sinh\left[2\xi_0\Delta/\delta\xi^2(1/e)\right]}{\left[2\xi_0\Delta/\delta\xi^2(1/e)\right]}\right\}. \quad (128)$$

Since the function $(\sinh x)/x$ tends to unity as x goes to zero, the normalized slit-broadened profile is

$$\frac{J(\xi_0, \Delta)}{J(0, \Delta)} = \exp\left[-\xi_0^2/\delta\xi^2(1/e)\right]\left\{\frac{\sinh\left[2\xi_0\Delta/\delta\xi^2(1/e)\right]}{\left[2\xi_0\Delta/\delta\xi^2(1/e)\right]}\right\} \quad (129)$$

or its equivalent written in terms of the scattering angle $\theta = \xi/f$. Clearly, in the limit $\Delta \to 0$, eq. (129) describes the correct unbroadened gaussian. For $\Delta \neq 0$, the principal effect of the $(\sinh x)/x$ correction term is to push up the tails of the profile while leaving the peak of the gaussian relatively unaffected. A good quantitative feeling for the nature of this correction may be obtained by solving for the off-zero displacements, ξ_0, at which the broadened and unbroadened profiles reach specified fractions of their peak intensity. These ξ_0 values then specify the profile half-widths at the corresponding intensity level.

For the curves presented in Section II, the relevant numerical parameters, expressed in angular units, are:

$$\delta\theta(1/e) = (1/f)\delta\xi(1/e) = 9.6\ \mu\text{rad}$$
$$\Delta_\theta = (1/f)\Delta \qquad = 1.0\ \mu\text{rad}.$$

Table IX gives the calculated half-width values $\theta_0 = (\xi_0/f)$ obtained from eq. (129) for various choices of the ratio $J(\theta_0, \Delta)/J(0, \Delta)$. For comparison, the table also lists the corresponding half-widths of the uncorrected gaussian, and the percentage of line-width increase caused by the slit-width correction. As is evident from these results, the effect of the $(1/x)$ sinh x correction term is to alter the gaussian profile in such a way that the observed half-widths are an essentially constant *percentage* larger than the true values.

APPENDIX B

Numerical Evaluation of the Diffraction Profile of Apertured Gaussian Illumination

Equation (47) gives the basic integral for the truncated gaussian diffraction profile as

$$E(\xi) = \frac{E_0}{(f\lambda_0)^{\frac{1}{2}}} \int_{-b/2}^{b/2} \exp(-x^2/\sigma^2) \exp[i(2\pi/f\lambda_0)\xi x]dx. \quad (130)$$

This expression may be put into a form more suited to numerical computation as follows. We write the exp $i(\cdots)$ term as

$$\exp[i(2\pi/f\lambda_0)\xi x] = \cos Kx + i \sin Kx$$

with

$$K \equiv \frac{2\pi\xi}{f\lambda_0} = \frac{2\pi\theta}{\lambda_0} \quad (131)$$

and note that the sin Kx integral vanishes by symmetry. Next by a change of variable

$$x = \frac{bw}{2}, \quad (132)$$

we obtain

$$E(\xi) = \frac{bE_0}{(f\lambda_0)^{\frac{1}{2}}} \int_{w=0}^{w=1} \cos cw \exp(-a^2w^2)dw, \quad (133)$$

where a and c are defined as

$$c \equiv \frac{Kb}{2} = \left(\frac{2\pi\xi}{f\lambda_0}\right)\frac{b}{2} = \frac{\pi b\theta}{\lambda_0}, \qquad a^2 = \frac{b^2}{8\sigma^2}. \quad (134)$$

The gaussian in the integrand is now expressed in terms of its Taylor's

series expansion

$$\exp(-a^2x^2) = \sum_{n=0}^{\infty} \frac{(-1)^n(a^2x^2)^n}{n!}$$

to give $E(\xi)$ as

$$E(\xi) = \frac{bE_0}{(f\lambda_0)^{\frac{1}{2}}} \sum_{n=0}^{\infty} \frac{(-1)^n(a^2)^n}{n!} \int_{w=0}^{1} w^{2n} \cos(cw)\, dw. \quad (135)$$

Equation (135) forms the basis for the numerical computation of the profiles.

A simple closed-function form for the w integrals in eq. (135) does not exist; however, recursive relations among these integrals can be found from the standard integrals

$$\int_0^1 x^m \cos cx\, dx = \frac{\sin c}{c} - \frac{m}{c} \int_0^1 x^{m-1} \sin cx\, dx \quad (136)$$

and

$$\int_0^1 x^{m-1} \sin cx\, dx = -\frac{\cos c}{c} + \frac{(m+1)}{c} \int_0^1 x^{m-2} \cos cx\, dx. \quad (137)$$

Defining

$$L_m(c) \equiv \int_0^1 x^m \cos cx\, dx, \quad (138)$$

we easily obtain the following recursion formulae from eqs. (136) and (137):

$$(m+3)L_{m+2}(c) = \left(\frac{m+3}{c}\right) \sin c + \frac{(m+3)(m+2)}{c^2} \cos c - \frac{(m+3)(m+2)}{c^2} [(m+1)L_m(c)] \quad (139)$$

$$(m-1)L_{m-2}(c) = \left(\frac{c}{m}\right) \sin c + \cos c - \frac{c^2}{m(m+1)} [(m+1)L_m(c)]. \quad (140)$$

From eq. (136), we also have for $m = 0$

$$L_0(c) = \frac{\sin c}{c}. \quad (141)$$

In terms of the $L_m(c)$, the expression for the diffracted field takes the series form

$$E(\xi) = \frac{bE_0}{(f\lambda_0)^{\frac{1}{2}}} \sum_{n=0}^{\infty} \frac{(-1)^n(a^2)^n L_{2n}(c)}{n!} \quad (142)$$

and the desired normalized intensity profile is

$$\frac{I(\xi)}{I(0)} = \frac{\left\{ \sum_{n=0}^{\infty} \frac{(-1)^n (a^2)^n L_{2n}(c)}{n!} \right\}^2}{\left\{ \sum_{n=0}^{\infty} \frac{(-1)^n (a^2)^n L_{2n}(0)}{n!} \right\}^2}. \tag{143}$$

The zero argument L_m's can be written down explicitly from eq. (138), *viz.*

$$L_m(0) = \frac{1}{m+1}. \tag{144}$$

For the numerical results reported here, the series in eq. (143) were truncated at some $n = n_{MAX}$ by testing the value of $(1/n!)(a^2)^n L_{2n}(0)$ and terminating when this quantity was smaller than some chosen convergence criterion, ϵ. In the present case, ϵ was set at $\epsilon = 10^{-12}$. For the largest (b/σ) value, $(b/\sigma) = 8.33$, where the gaussian kernel of eq. (133) is

$$\exp(-a^2 w^2) = \exp(-8.68 w^2),$$

49 terms in the series were required for convergence.

For each individual pair of values for c and n_{MAX}, the required string of L_m's are generated by two subroutine programs.

B.1 Subroutine No. 1, c < 1

When the quantity $c = (Kb/2)$ is less than one, the L_m's are obtained by the following procedure.

(*i*) Calculate $L_m(c)$ for $m = 2n_{MAX}$ directly from the defining equation (138), using the Taylor expansion for cos cx to write

$$\begin{aligned} L_m(c) &= \sum_{n=0}^{\infty} \frac{(-1)^n c^{2n}}{(2n)!} \int_0^1 W^{2n+m} dW \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n c^{2n}}{(2n)!(m+2n+1)}. \end{aligned} \tag{145}$$

(*ii*) Truncate the sum in eq. (145) when $c^{2n}/(2n)!(m+2n+1)$ is less than 10^{-12}.

(*iii*) Use this result for $L_{n_{MAX}}(c)$ to obtain the required L_m's *via* the backward recursion formula, eq. (140).

B.2 Subroutine No. 2, c > 1

When the quantity $c = (Kb/2)$ is greater than 1, the L_m's are found by a two-part procedure that depends on the value of n_{MAX}.

(*i*) For $m = 2n$ values for which the inequality $m = 2n < c$ is satisfied, use $L_1(c) = (\sin c)/c$ and the forward recursion rela-

tion, eq. (140). If $2n_{MAX}$ is less than c this first step gives all required L_m's.

(*ii*) If $2n_{MAX}$ is greater than c, set

$$L_m(c) \approx \frac{\cos c}{m+1}$$

for some $m \gg 2n_{MAX}$ and work backward using recursion relation eq. (140). The calculated string of L_m's is joined onto the forward recursion values from step (*i*) for some $m \approx c$ and then renormalized.

This rather elaborate procedure for calculating the L_m's is made necessary by the rapid accumulation of numerical round-off errors which arise in the repetitive application of the basic recursion formulae.

REFERENCES

1. I. L. Fabelinskii, *Molecular Scattering of Light*, New York: Plenum Press, 1968, Chapter III, pp. 155–246.
2. H. Z. Cummins and H. L. Swinney, "Light Beating Spectroscopy," in *Progress in Optics*, Vol. VIII, Emil Wolf, ed., Amsterdam, Netherlands: North Holland Publishing, 1970, pp. 135–200.
3. B. Chu, *Laser Light Scattering*, New York: Academic Press, 1974, Chapters IV–VII, IX, and X.
4. *Small Angle X-Ray Scattering*, H. Brumberger, ed., Proc. of Conf. at Syracuse University, June 24–26, 1965; sponsored by American Crystallographic Society, the Army Research Office, the National Science Foundation, and the University of Syracuse; New York: Gordon and Breach, 1967.
5. A. J. Renouprez, "Diffusion des Rayons X aux Petits Angle," *International Union of Crystallography, Commission on Crystallographic Apparatus*, Bibliography *4*, 1970, pp. 19–24.
6. W. H. Aughey and F. J. Baum, "Angular Dependence Light Scattering—A High Resolution Recording Instrument for the Angular Range 0.05°–140°," J. Opt. Soc. Amer. *44*, No. 11 (November 1954), pp. 833–837.
7. C. H. Henry and J. J. Hopfield, "Raman Scattering by Polaritons," Phys. Rev. Lett. *15*, No. 25 (December 1965), pp. 964–966.
8. S. P. S. Porto, B. Tell, and T. C. Damen, "Near Forward Raman Scattering in Zinc Oxide," Phys. Rev. Lett., *16*, No. 11 (March 1966), pp. 450–452.
9. J. B. Lastovka and G. B. Benedek, "Spectrum of Light Scattered Quasielastically from a Normal Liquid," Phys. Rev. Lett., *17*, No. 20 (November 1966), pp. 1039–1042.
10. J. B. Lastovka and G. B. Benedek, "Light Beating Techniques for the Study of the Rayleigh-Brillouin Spectrum," in *Physics of Quantum Electronics*, P. L. Kelly, B. Lax, and P. E. Tannenwald, eds., Proceedings of the Physics of Quantum Electronics Conference, San Juan, Puerto Rico, June 28–30, 1965, sponsored by the Office of Naval Research, New York: McGraw-Hill, 1966, pp. 231–240.
11. D. Eden and H. L. Swinney, "Optical Heterodyne Studies of Brillouin Scattering in Xenon Near the Critical Point," Opt. Commun., *10*, No. 2 (February 1974), pp. 191–194.
12. S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, London: Oxford University Press, 1961.
13. V. M. Zaitsev and M. I. Shliomis, "Hydrodynamic Fluctuations Near the Convection Threshold," Zh. Eksp. Teor. Fiz., *59*, No. 5 (November 1970), pp. 1583–1592 [Sov. Phys. JEPT, *32*, No. 5 (May 1971), pp. 866–870].
14. R. Graham, "Generalized Thermodynamic Potential for the Convection Instability," Phys. Rev. Lett., *31*, No. 25 (December 1973), pp. 1479–1482.
15. M. G. Velarde, in *Hydrodynamics*, Proc. of the 1973 session of the École d'été de Physique Théorique, Les Houches, R. Balian, ed., New York: Gordon and Breach, in press.

16. P. Bergé and M. Dubois, "Convective Velocity Field in the Rayleigh-Bénard Instability: Experimental Results," Phys. Rev. Lett., *32*, No. 19 (May 1974), pp. 1041–1044.
17. W. A. Smith, "Temporal Correlations Near the Convection Instability Threshold," Phys. Rev. Lett., *32*, No. 21 (May 1974), pp. 1164–1167.
18. R. Farhadieh and R. S. Tankin, "Interferometric Study of Two-Dimensional Bénard Convection Cells," J. Fluid Mech., *66*, No. 4 (December 1974), pp. 739–752.
19. H. N. W. Lekkerkerker and J.-P. Boon, "Hydrodynamic Modes and Light Scattering Near the Convective Instability," Phys. Rev., A*10*, No. 4 (October 1974), pp. 1355–1360.
20. G. Ahlers, "Low Temperature Studies of the Rayleigh-Bénard Instability and Turbulence," Phys. Rev. Lett., *33*, No. 20 (November 1974), pp. 1185–1188.
21. J. B. McLaughlin and P. C. Martin, "Transition to Turbulence of a Statically Stressed Fluid," Phys. Rev. Lett., *33*, No. 20 (November 1974), pp. 1189–1192.
22. J. P. Gollub and M. H. Freilich, "Optical Heterodyne Study of the Taylor Instability in a Rotating Fluid," Phys. Rev. Lett., *33*, No. 25 (December 1974), pp. 1465–1468.
23. R. Graham, "Hydrodynamic Fluctuations Near the Convection Instability," Phys. Rev., A*10*, No. 5 (November 1974), pp. 1762–1784.
24. E. Guyon and P. Pieranski, "Convective Instabilities in Nematic Liquid Crystals," Physica (Utrecht), *73*, No. 1 (April 1974), pp. 184–194.
25. H. B. Möller and T. Riste, "Neutron-Scattering Study of Transitions to Convection and Turbulence in Nematic Para-azoxyanisole," Phys. Rev. Lett., *34*, No. 16 (April 1975), pp. 996–999.
26. *Fluctuations, Instabilities, and Phase Transitions*, T. Riste, ed., Proceedings of the NATO Advanced Study Institute, Geilo, Norway, April 11–20, 1975, New York: Plenum Press, 1975.
27. J. B. Lastovka, "Light Mixing Spectroscopy and the Spectrum of Light Scattered by Thermal Fluctuations in Liquids," Ph.D. Thesis, Massachusetts Institute of Technology, 1967, Chapter III, pp. 156–357.
28. J. B. Lastovka, unpublished paper.
29. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 3rd ed., New York: McGraw-Hill, 1957, pp. 298ff.
30. J. D. Jackson, *Classical Electrodynamics*, New York: John Wiley, 1962, pp. 280–282.
31. M. Born and E. Wolf, *Principles of Optics*, 2nd ed., New York: MacMillan, 1964, pp. 414–418.
32. M. Françon, *Diffraction-Coherence in Optics*, Oxford: Pergamon Press, 1966, Chapter VI, Section 6.5.
33. P. Jacquinot and B. Roizen Dossier, "Apodization," in *Progress in Optics, Vol. III*, Emil Wolf, ed., Amsterdam, Netherlands: North Holland Publishing, 1964, pp. 30–186.
34. R. C. Hansen, "Aperture Theory," in *Microwave Scanning Antennas, Volume I: Apertures*, R. C. Hansen, ed., New York: Academic Press, 1964, pp. 47–101.
35. E. A. Wolff, *Antenna Analysis*, New York: John Wiley, 1966, pp. 109–135.
36. K. A. Karpov, *Tables of the Functions* $F(Z) = \int_0^Z e^{x^2} dx$ *in the Complex Domain*, New York: MacMillan, 1964.
37. W. J. Smith, *Modern Optical Engineering*, New York: McGraw-Hill, 1966, pp. 385–387.
38. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford: Clarendon, 1948, pp. 61–66.
39. Jan Peřina, *Coherence of Light*, London: Van Nostrand Reinhold, 1972, pp. 32–42.
40. M. J. Beran and G. B. Parrent, *Theory of Partial Coherence*, Englewood Cliffs, New Jersey: Prentice-Hall, 1964, pp. 27–44.
41. H. E. Bennett and J. M. Bennett, "Precision Measurements in Thin Film Optics," in *Physics of Thin Films, Vol. 4*, G. Hass, ed., New York: Academic Press, 1967, pp. 1–96.
42. S. Tolansky, *Multiple-Beam Interferometry of Surfaces and Films*, New York: Dover Publications, 1970, Chapter IX, pp. 104–108.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 9, November 1976
Printed in U.S.A.

Exact Theory of TE-Wave Scattering From Blazed Dielectric Gratings

By D. MARCUSE

(Manuscript received April 9, 1976)

We present an exact description of scattering of an incident plane wave with TE-polarization at an interface between two dielectric media that is deformed by a grating with triangularly shaped teeth. The theory employs an expansion in plane waves outside of the grating region and describes the field in the grating region as a double Fourier series expansion. The results of this theory are represented graphically. That blazing provides substantial discrimination of the scattering process in favor of beams scattered into one or the other of the two media is shown. The exact theory is used to check an approximation for the effective reflection plane that is useful for future applications of the theory to scattering by gratings of guided waves in thin-film waveguides.

I. INTRODUCTION

This study of dielectric sawtooth gratings with deep grooves serves several purposes. Its principal aim is to investigate a particular analytical method for describing deep gratings with the view of applying it (at a later time) to waveguide-grating couplers. However, even without the added complication of one more dielectric interface that characterizes the waveguide problem, an examination of the response of dielectric gratings with deep grooves to a plane wave, incident at an angle that would lead to total internal reflection at the corresponding smooth surface, can teach us much about the expected behavior of waveguide-grating couplers.

The literature on the electromagnetic theory of diffraction gratings is vast. However, most papers are limited to discussions of metallic gratings,[1,2] and only a few papers mention dielectric gratings with sawtooth-shaped grooves and plane waves incident at angles larger than the critical angle for total internal reflection.[3] When it comes to providing numerical information for a given particular case, each worker must write a computer program to solve the problem at hand, since no publication can cover all conceivable cases in graphical form.

Developing the computer program for the study of deep gratings was one of the aims of this work.

Our method for treating TE-wave interaction with deep dielectric sawtooth gratings is basically simple and exact. We express the field above and below the grating as a series of plane waves using the periodicity imposed by the grating. In the grating region, the field is expressed as a double Fourier series expansion whose terms are not individually solutions of the wave equation. The unknown coefficients entering the various series expansions are determined by the requirement that the field in the grating region must be a solution of the wave equation and by enforcing the proper boundary conditions along two mathematical planes just above and below the grating region. By varying the number of terms used in the series expansions, it was found that the series converge very well, and good accuracy is obtained with relatively few terms. However, the required number of terms increases with increasing depth of the sawtooth grating.

The simple grating problem described here has the advantage that only an inhomogeneous equation system needs to be solved. Since the problem does not contain unknown eigenvalues, no search for suitable eigenvalue conditions is required. The exact solution of the corresponding waveguide problem would lead to an eigenvalue equation. A very large determinant with complex coefficients would have to be forced to vanish by proper choice of the propagation constant of the leaky wave inside the guide, one of whose interfaces between core and cladding is formed by the grating. The simple grating furnishes important information about the phase shift suffered by the reflected plane waves. This information can be used to estimate the eigenvalues of the modes inside of the waveguide with a grating on one of its interfaces. This information is useful for finding approximate solutions of the waveguide grating problem without the need for solving a costly and time-consuming eigenvalue problem.

For shallow gratings, our theory is in complete agreement with perturbation theory. Some of the features of deeper gratings with groove depth on the order of the wavelength can be explained by geometrical optics coupled with simple grating conditions. The ray paths in deep gratings (groove depth larger than the wavelength) are so complicated that an explanation of maxima or minima in terms of geometrical optics fails.

The sawtooth-shaped interface deformation is a blazed grating. It has the advantage that its shape can be adjusted to enhance certain grating orders. In particular, it is possible to let a high grating order predominate over lower orders. Furthermore, the grating shape can be used to favor scattering into the air space above the grating or, cor-

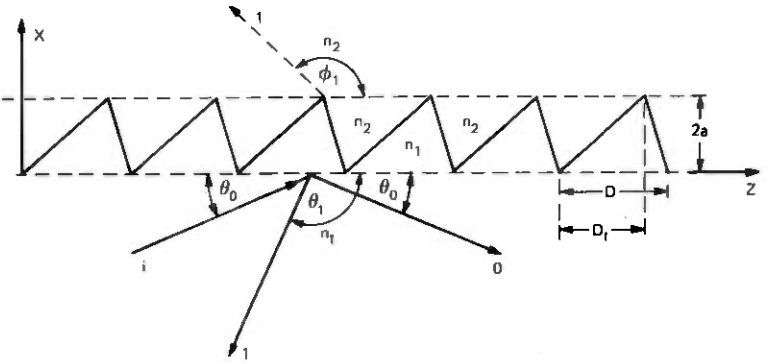


Fig. 1—Triangularly shaped dielectric grating as an interface between the two media with index n_1 and n_2. (This figure defines grating parameters and incident and scattered beams.)

respondingly, to favor scattering down into the higher dielectric region, the substrate, from which the incident wave impinged on the grating. This preferential scattering behavior is very useful for the construction of grating couplers. Additional grating responses, other than those used for the coupling beam, decrease the overall efficiency of a grating coupler. If unwanted grating lobes can be suppressed by properly shaping the grating teeth, higher coupling efficiencies are obtainable. Gratings that show strong asymmetry in favor of a certain grating order provide also high-reflection losses for the zero-order grating lobe (that would correspond to the guided mode field of a waveguide). The waveguide mode thus would decay rapidly over a few periods of the zig-zag path of the guided ray. This means that high-efficiency grating couplers based on this principle would have to be very short.

1.1 Theory of the dielectric sawtooth grating

Figure 1 shows the geometry of our sawtooth grating. A ray labeled i is incident from the medium with refractive index n_1 on the dielectric interface with the medium n_2 whose shape is a sequence of sawteeth. The specularly reflected beam is labeled 0. Also shown are two scattered beams labeled 1 which escape into the medium with index n_2 (subsequently to be called the air space) and into the medium with index n_1 (subsequently to be called the substrate). The grating period is D; D_1 is the distance along the base of each sawtooth from its beginning to the point underneath its peak. The grating amplitude is defined as $2a$.

We consider only TE-waves with the electric field component E_y and the magnetic field components[4]

$$H_x = -\frac{i}{\omega\mu_o}\frac{\partial E_y}{\partial z} \tag{1}$$

and

$$H_z = \frac{i}{\omega\mu_o}\frac{\partial E_z}{\partial x}. \tag{2}$$

(ω and μ_o are, respectively, the angular frequency of the wave and the magnetic permeability of vacuum.) The grating is infinitely extended in y direction, so that all y-derivatives vanish. The field components E_x, E_z, and H_y do not exist. The periodicity of the infinitely extended (in z direction) grating forces the electromagnetic field to be of the following form:

$$E_y = e^{-i\beta_i z}\left\{A_o^{(i)}e^{-i\sigma_o x} + \sum_{m=-\infty}^{\infty} A_m e^{i\sigma_m x}e^{i(2\pi/D)mz}\right\} \qquad \text{for } x \leqq 0, \tag{3}$$

and

$$E_y = e^{-i\beta_i z}\sum_{m=-\infty}^{\infty} C_m e^{-i\rho_m x}e^{i(2\pi/D)mz} \qquad \text{for } x \geqq 2a. \tag{4}$$

Since E_y must satisfy the wave equation

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} + n^2k^2E_y = 0 \tag{5}$$

with

$$k = \frac{2\pi}{\lambda_o} = \omega\sqrt{\epsilon_o\mu_o} \tag{6}$$

(λ_o = free space wavelength, ϵ_o = dielectric permittivity of vacuum, $n = n_1$ or n_2 refractive index of the dielectric medium), the parameters appearing in (3) and (4) must have the form,

$$\sigma_m = \left[n_1^2k^2 - \left(\beta_i - \frac{2\pi}{D}m\right)^2\right]^{\frac{1}{2}} \tag{7}$$

and

$$\rho_m = \left[n_2^2k^2 - \left(\beta_i - \frac{2\pi}{D}m\right)^2\right]^{\frac{1}{2}}. \tag{8}$$

A_i is the amplitude of the incident wave with propagation constant

$$\beta_i = n_1k\cos\theta. \tag{9}$$

The term propagation constant is used here in the same sense as in a waveguide; it is actually the z component of the plane wave propagation vector.

Note that the superposition of plane waves (3) and (4) was chosen so that the traveling parts of the wave move away from the grating with the exception of the incident wave of amplitude $A_o^{(i)}$ [the time dependence is understood to be $\exp(i\omega t)$]. It is clear that only a small

number of waves in the field expansions actually propagate in x direction, because almost all terms of the form (7) and (8) are imaginary. The signs of the imaginary quantities must be chosen, so that the evanescent fields decay in the direction away from the grating,

$$\sigma_m = -i|\sigma_m|, \qquad \rho_m = -i|\rho_m|. \tag{10}$$

Every term in the expansion (3) and (4) is a solution of the wave equation, but a similar expansion cannot be written down for the field in the grating region $0 \leqq x \leqq 2a$. Instead, we simply use a doubly infinite Fourier series.

$$E_y = e^{-i\beta_i z} \sum_{n,m=-\infty}^{\infty} B_{nm} e^{i(\pi/b)nx} e^{i(2\pi/D)mz} \qquad 0 \leqq x \leqq 2a. \tag{11}$$

Except for the phase factor $\exp(-i\beta_i z)$, the field solution is periodic in z with period D. This periodicity is a very important feature of the field solution and is imposed by the periodicity of the grating. The function (11) is also periodic in x direction with period length $2b$. This periodicity is quite arbitrary. It would appear natural to let $b = a$. However, this choice of b would force the field to have exactly the same values at $x = 0$ and $x = 2a$, which is physically unreasonable. For this reason, we must allow b to be arbitrary, but use $b > a$. As a practical matter $b = \sqrt{2}a$ has been used for the numerical calculations in the hope that this choice would facilitate the convergence of the series. Clearly, b should not be made too large and, of course, it must not be smaller than a.

It now remains to determine the expansion coefficients A_m, C_m, and B_{nm}. This is accomplished by substituting (11) into the wave equation (5), multiplying the resulting equation with $\exp(-i\pi n\, x/b)$ $\exp(-i2\pi m\, z/D)$ and integrating over z from 0 to D and over x from 0 to $2a$. Continuity of the fields at the planes $x = 0$ and $x = 2a$ requires us to force E_y and its x-derivative to be continuous at these planes. After elimination of A_m and C_m from the equation systems, we are left with the following three infinite simultaneous equations:

$$\sum_{n=-\infty}^{\infty} \left(\rho_m + \frac{\pi}{b} n\right) B_{nm} e^{i(2\pi/b)na} = 0. \tag{12}$$

$$\sum_{n=-\infty}^{\infty} \left(\sigma_m - \frac{\pi}{b} n\right) B_{nm} = 2\sigma_o A_o^{(i)} \delta_{mo}. \tag{13}$$

$$\sum_{n',m'=-\infty}^{\infty} \left\{ N_{n'-n,m'-m} - \left[\left(\frac{\pi n'}{b}\right)^2 + \beta_{m'}^2\right] M_{n,n'} \delta_{m,m'} \right\} B_{n'm'} = 0. \tag{14}$$

$\delta_{mm'}$ is Kronecker's delta symbol. In (12) and (13) m is allowed to be

any integer, and, similarly, n and m are allowed to be any integer in (14). The first two equations stem from the boundary conditions, while (14) expresses the requirement that the field expansion (11) satisfy the wave equation (5). The three sets of infinite equations (12) through (14) are used to express B_{nm} in terms of $A_0^{(i)}$. The coefficients $N_{n'-n,m'-m}$ and $M_{n,n'}$ are listed in the Appendix; β_m is defined as

$$\beta_m = \beta_i - \frac{2\pi}{D} m. \tag{15}$$

The amplitude coefficients A_m and C_m are obtained in terms of B_{nm} as follows:

$$A_m = \left(\sum_{n=-\infty}^{\infty} B_{nm} \right) - A_0^{(i)} \delta_{m0}. \tag{16}$$

$$C_m = \sum_{n=-\infty}^{\infty} B_{nm} e^{i[\rho_m + (\pi/b)n]2a}. \tag{17}$$

The power of the incident wave flowing through an element of unit area parallel to the x direction is given as

$$P_i = \frac{\sigma_0}{2\omega\mu_0} |A_0^{(i)}|^2. \tag{18}$$

It is convenient to express the power carried away by the scattered beams in terms of the power of the incident beam. For the grating orders carrying power into the air space, we obtain the relative power from

$$\frac{\Delta P_{ma}}{P_i} = \frac{\rho_m}{\sigma_0} \frac{|C_m|^2}{|A_0^{(i)}|^2}. \tag{19}$$

Similarly, we obtain the relative power carried into the substrate,

$$\frac{\Delta P_{ms}}{P_i} = \frac{\sigma_m}{\sigma_0} \frac{|A_m|^2}{|A_0^{(i)}|^2}. \tag{20}$$

Let us close this section with a few remarks about the numerical solution of the equation systems (12) through (14). As mentioned above, the terms in the series expansions (3) and (4) represent traveling as well as evanescent waves. It is clear that all terms corresponding to traveling waves must be included in the truncated series expansions used for approximate numerical solutions of the problem. According to (7), propagating grating orders are associated with m values in the interval

$$\left[\frac{D}{2\pi} (\beta_i - n_1 k) \right]_{\text{int}} < m < \left[\frac{D}{2\pi} (n_i k + \beta_i) \right]_{\text{int}}. \tag{21}$$

The label "int" is a reminder that the integer, whose absolute value is

just smaller than the value inside of the bracket, must be taken. Those terms in the series expansion (3) whose m-values lie outside the interval (21) belong to evanescent waves that do not carry away power. As a practical matter, we found that sufficient accuracy is obtained if just one—or at most a few—evanescent waves on each side of the interval (21) are included in the series expansions. The terms in the expansion (11) cannot be interpreted as traveling or evanescent waves. The sum over m is, of course, intimately related to the m-summations in (3) and (4), and an equal number of terms must be taken in all m-summations. We found that the n-summation in (11) converges more slowly, so that usually more terms are required in this series. In all numerical calculations whose discussions follow, we never used more than 11 terms in the n-summation, and often as few as 7 terms proved to be sufficient, if the grating amplitude remained below $2a/\lambda_o = 0.5$. The total number of unknowns B_{nm} in the equation system (12) through (14) is, of course, the product of the number of terms in both series expansions, n and m. For large values of $2a$, for example for $2a/\lambda_o = 2$, we used 66 unknowns B_{nm}, and for $2a/\lambda_o < 0.5$, 36 unknowns seemed to be sufficient.

The fact that the equations stemming from the boundary conditions (12) and (13) must be included in the equation system to be solved prevents us from using an equal number of terms in the n', m' summations of (14) and for the "free" indices n and m. Obviously, the number of m-values [the number of equations of the type (14)] that are used must be two less than the number of terms under the m-summation sign.

1.2 Geometrical optics considerations

Figure 1 shows the principal function of the diffraction grating. The incident plane wave breaks up into several components after striking the dielectric interface. The strongest wave leaving the grating region is usually the zero-order grating response that leaves in a direction corresponding to the specularly reflected beam at an ideal, smooth interface. Throughout this discussion, we assume that the incident wave strikes the interface at an angle θ_o that remains below the critical angle for total internal reflection at the unperturbed, smooth boundary. In addition to the incident and specularly reflected plane waves, a discrete number of scattered plane waves are generated. These waves emerge in directions that are defined by the condition that all scattered waves interfere constructively. The condition for such constructive interference is expressed by the relation

$$\beta_m = \beta_i - \frac{2\pi}{D} m. \tag{22}$$

The β_i and β_m are the z-components of the propagation vectors of the incident and scattered waves; m is a positive or negative integer. The angles θ_m of scattered waves in medium 1 are obtained from (9) and (22) as follows,

$$\theta_m = \arccos\left(\frac{\beta_m}{n_1 k}\right), \tag{23}$$

and the corresponding scattering angles in medium 2 are

$$\phi_m = \arccos\left(\frac{\beta_m}{n_2 k}\right). \tag{24}$$

The integer m defines the grating orders of the scattered beams. The direction of the specularly scattered plane-wave is obtained by using $m = 0$, and $m = +1$ gives the first grating orders, etc.

The intensities of the scattered beams decrease with increasing grating order, if the grating amplitude $2a$ is much smaller than the wavelength. However, for deep gratings whose amplitude is comparable to, or larger than, the wavelength, higher grating orders may well predominate over lower grating orders. In particular, it is possible to predict maxima of scattered waves based on geometric optics considerations. Such maxima occur when the direction of a grating lobe defined by (23) or (24) coincides with the condition of specular reflection of the incident beam on one of the facets of the grating teeth.

Consider the situation shown in Fig. 2. Geometrical optics allow us to calculate the angle θ_m of the reflected wave as

$$\theta_m = 2\alpha_2 + \theta_o. \tag{25}$$

If θ_m simultaneously satisfies (23), a strong grating response may be

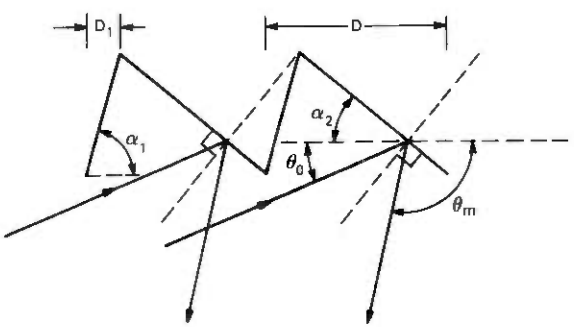


Fig. 2—Specular reflection from grating faces can be used to explain maxima of the grating lobes.

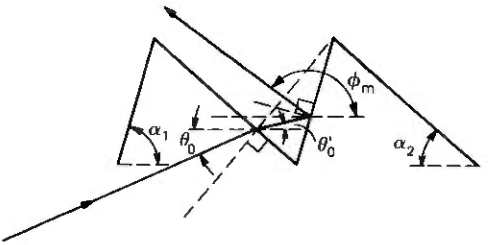


Fig. 3—Grating lobes in air can be enhanced by specular reflection from the grating teeth as shown.

expected. From (23) and (25), we find the following condition for α_2,

$$\alpha_2 = \frac{1}{2}\left[\arccos\left(\frac{\beta_i - \frac{2\pi}{D}m}{n_1 k}\right) - \arccos\left(\frac{\beta_i}{n_1 k}\right)\right]. \tag{26}$$

The grating angle α_2 is defined in terms of the other grating parameters as:

$$\alpha_2 = \arctan\left(\frac{2a}{D - D_1}\right). \tag{27}$$

Maxima for grating responses into the air space can occur in many different ways. One possibility is depicted in Fig. 3. The geometric optics condition for θ_m is computed in several steps. The refracted angle θ_o' follows from Snell's law,

$$\theta_o' = \arccos\left(\frac{n_1}{n_2}\cos(\alpha_2 + \theta_o)\right) - \alpha_2, \tag{28}$$

and the angle of the mth grating response in air follows from

$$\phi_m = 2\alpha_1 - \theta_o' = 2\arctan\left(\frac{2a}{D_1}\right) - \theta_o'. \tag{29}$$

The conditions that must be satisfied by the grating parameters to achieve equality of (24) and (29) can be found by an iterative calculation.

Geometric optics conditions leading to maxima of the grating response in air can be complicated in many ways. For example, the ray escaping into the air space shown in Fig. 3 may be intercepted by the grating tooth through which it just passed and may suffer further refraction. Another possibility is depicted in Fig. 4. The fact that many geometric optics conditions exist that may enhance the grating response in air makes it difficult to account for the maxima of the air lobes of deep gratings.

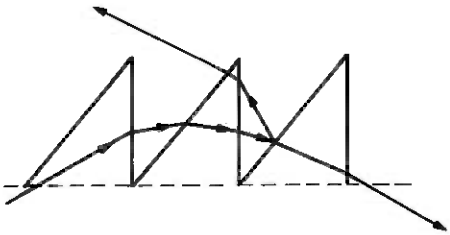


Fig. 4—The incident ray is bent back into the waveguide by successive refraction at the grating teeth. Scattering takes place at every tooth and may be enhanced if the specular reflection condition is satisfied.

Figure 4 not only shows how a ray may escape through the grating following a complicated path, but it also indicates that the unscattered portion of the ray is diffracted back into medium 1 following a path that takes it inside of the grating. This geometric optics picture suggests that the effective penetration depth of the reflected light field can be estimated by geometric optics methods. For this purpose, we assume that the grating acts on the refracted ray as a graded-index medium with an index distribution

$$\bar{n}^2(x) = \frac{2a - x}{2a} n_1^2 + \frac{x}{2a} n_2^2. \tag{30}$$

We can now use the WKB method[5] to determine the phase of the wave that penetrates into the grating region. It is well-known that the wave penetrates into the graded-index medium until it reaches the turning point of ray optics at $x = t$. The phase of the reflected wave taken at the reference plane $x = 0$ is given by[5]

$$\phi = 2 \int_o^t \sqrt{[\bar{n}(x)k]^2 - \beta_i^2}\, dx - \frac{\pi}{2} = \frac{8a\sigma_o^3}{3(n_1^2 - n_2^2)k^2} - \frac{\pi}{2}. \tag{31}$$

We define an effective reference plane by assuming that the medium with index n_1 reaches into the grating region to a depth $x = d_{app}$. The phase of a wave reflected at this reference plane (that is assumed to consist of the index discontinuity from n_1 to n_2) is

$$\phi = 2\sigma_o d_{app} - 2 \arctan \frac{\gamma}{\sigma_o} \tag{32}$$

with

$$\gamma = (\beta_i^2 - n_2^2 k^2)^{\frac{1}{2}}. \tag{33}$$

The first term in (32) accounts for the phase shift caused by the round trip from $x = 0$ to $x = d_{app}$, and the second term is the additional phase shift on reflection from the index discontinuity.[6] By equating (31) to (32), we obtain the following expression for the depth of the

effective reference surface inside of the grating:

$$d_{app} = \frac{4a\sigma_o^2}{3(n_1^2 - n_2^2)k^2} - \frac{1}{\sigma_o}\left(\frac{\pi}{4} - \arctan\frac{\gamma}{\sigma_o}\right). \tag{34}$$

If the tangent of the phase angle ψ of the reflected wave is known, for example, from the numerical solution of the grating problem, the effective reference plane can be calculated from the expression

$$d_e = \frac{2\arctan\dfrac{\gamma}{\sigma_o} - \psi \pm p\pi}{2\sigma_o}. \tag{35}$$

Note that the approximation (34) holds only for gratings that are sufficiently thick so that the turning point of the rays is located deep enough inside the grating, so that the evanescent field beyond the turning point has decayed to insignificant values by the time it reaches the top of the grating. Furthermore, (34) is certain to represent the effective reflection plane better as the grating period is short. Numerical comparisons of the two expressions (34) and (35) will be presented in the next section.

1.3 Examples and numerical evaluation

The boundary between the two media with index n_1 and n_2 is described by the function

$$f(z) = \begin{cases} \dfrac{2a}{D_1} z & 0 \leqq z \leqq D_1, \\ \dfrac{2a}{D - D_1}(D - z) & D_1 \leqq z \leqq D, \end{cases} \tag{36}$$

which is periodic in z with period D. Its Fourier coefficients are:

$$c_m = \frac{aD^2 e^{-i\pi m(D_1/D)}}{i\pi^2 m^2 D_1(D - D_1)} \sin\left(\pi m \frac{D_1}{D}\right). \tag{37}$$

This Fourier coefficient is important, because for the first order of perturbation theory[4,7] the grating responses are proportional to $|c_m|^2$.

In the remainder of this section we present the results of numerical evaluations of our theory in graphical form. Figure 5a shows the relative power that is scattered into the air space above the grating. The incident plane wave always arrives at an angle that is small enough (measured with respect to the plane interface) to ensure total internal reflection at the smooth interface between the two media. In Figs. 5 through 9, we use $\beta_i\lambda_o = 8.5$. The grating period was chosen as $D = 1.3\lambda_o$ resulting in three grating lobes labeled $m = 1$, 2, and 3.

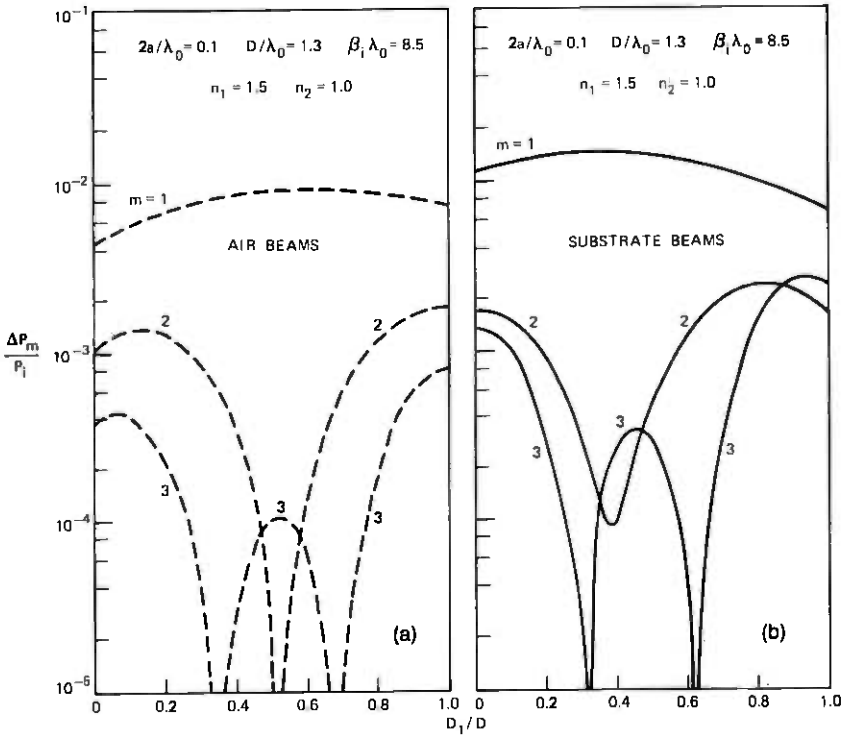


Fig. 5—Relative scattered power for the three grating lobes for $D/\lambda_o = 1.3$, $\beta_1\lambda_o = 8.5$, and $n_1 = 1.5$, $n_2 = 1.0$, and for a grating amplitude of $2a = 0.1\lambda_o$. (a) Shows the grating responses in air as functions of the grating shape factor D_1/D. (b) Shows the grating lobes in the substrate (index n_1).

The refractive index of the medium below the grating (called the substrate) is $n_1 = 1.5$, and the medium above the grating is assumed to be vacuum (or air) with $n_2 = 1.0$. The grating amplitude in Fig. 5a and b is $2a = 0.1\lambda_o$. This grating amplitude is already too large for perturbation theory to be accurate, but the zeros (or minima) and maxima of the grating responses can still be identified with the help of (37). Consider, for example, the second-order grating lobe with $m = 2$. First-order perturbation theory predicts that it has zero power at $D_1/D = 0.5$. Figure 5a for the grating responses in air shows that the zero of the second-order grating lobe is indeed very close to this value. The corresponding minimum (the power does not actually go to zero) for the substrate beam with $m = 2$ is, according to Fig. 5b, located at $D_1/D = 0.4$. Its position is shifted from the value that perturbation theory would predict, but the reason for the occurence of this minimum is still clearly discernible. The zeros for the third-order grating lobe, $m = 3$, would be located at $D_1/D = \frac{1}{3}$ and

$D_1/D = \frac{2}{3}$, if perturbation theory would apply. Correspondingly, Fig. 5a and b show that the zeros of the third-order grating responses are indeed close to these values.

Another interesting relationship results if we compare the power carried by the beams at $D_1/D = 1$. According to perturbation theory, we should find that the power ratio between the grating lobes $m = 1$ and $m = 2$ is 4, while the ratio between the lobes with $m = 1$ and $m = 3$ should be 9. According to Fig. 5a these power ratios are 4 and 8.8, respectively. The substrate beams shown in Fig. 5b give ratios of 3.8 and 1.8, respectively. The third-order grating response in the substrate is thus already considerably larger than perturbation theory would predict.

Figure 6a and b prove that all resemblance to perturbation theory is lost, if we increase the grating amplitude to $2a = 0.5\lambda_o$. According to perturbation theory, an increase of the grating amplitude by a factor of 5 should increase the scattered power by a factor of 25. No

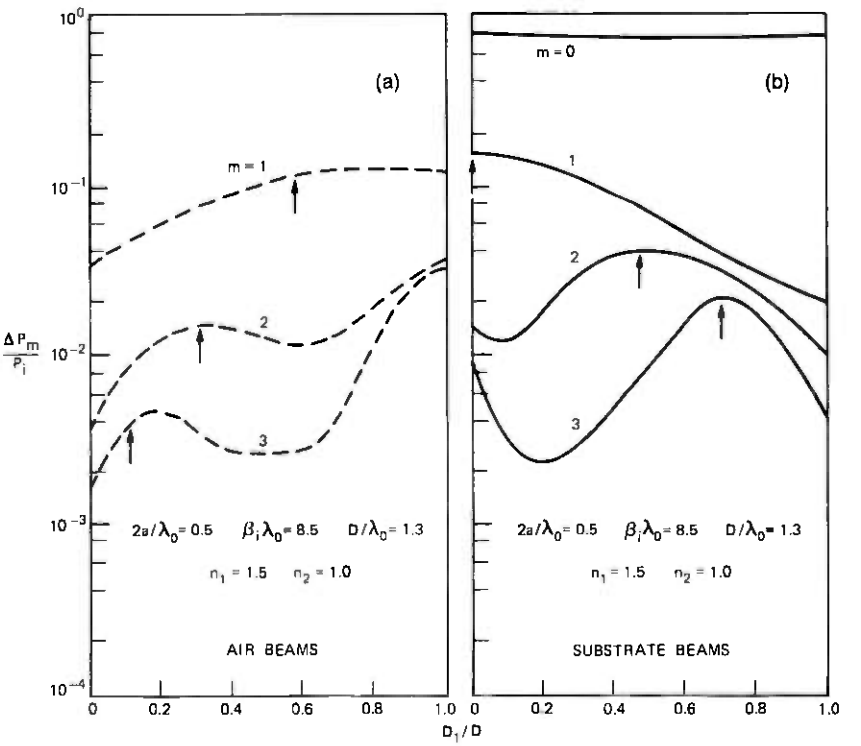


Fig. 6—Similar to Fig. 5a and b, except that the grating amplitude is $2a = 0.5\lambda_o$. (Arrows indicate the position of scattering enhancement by specular reflection from grating teeth.)

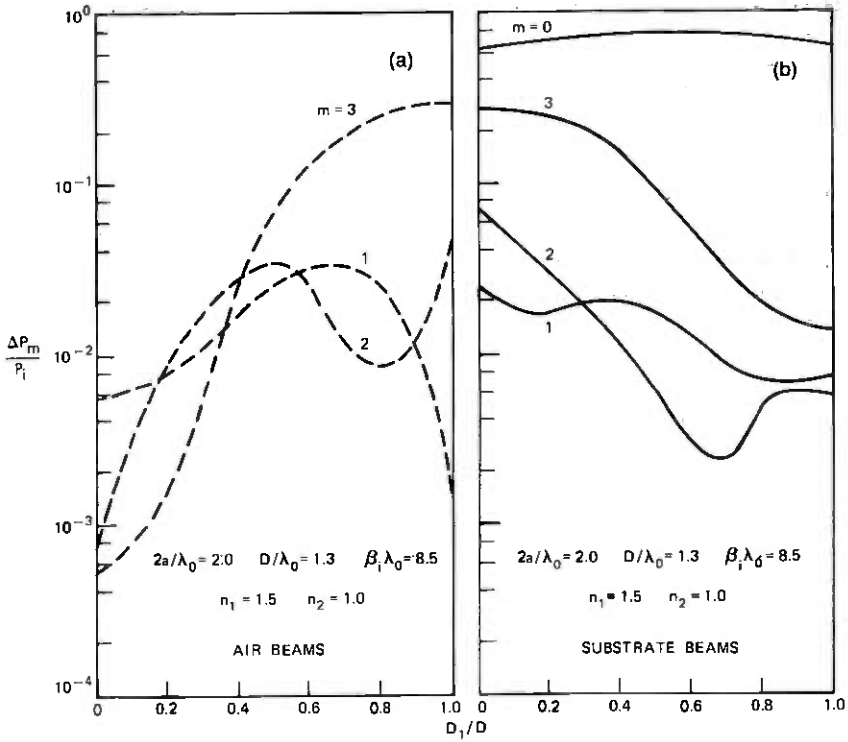


Fig. 7—Similar to Fig. 5a and b with $2a = 2.0\lambda_0$.

such increase is apparent for the first-grating order, nor is it indeed possible since the power in the scattered beams cannot exceed the input power. However, it is now possible to identify certain features of the curves by using geometrical optics. The arrows in Fig. 6a indicate the position where maxima of the grating lobes would be expected because of a coincidence of the direction of the grating lobes with specular reflection from the grating faces. The position of the arrows in Fig. 6a was computed from (24), (27), (28), and (29). Even though the agreement is not perfect, there is a strong indication that the maxima of the grating responses are indeed caused by specular reflection at the grating faces. The position of the arrows in Fig. 6b was computed from (23), (26), and (27). For the substrate lobes, the condition of specular reflection from the grating faces agrees very well with the actually observed grating maxima.

These figures show, furthermore, that very good discrimination between different grating responses can be obtained by a blazed grating. Consider a grating with $D_1/D = 1$. The first-order grating lobe in air carries 0.12 relative power while the corresponding substrate beam

carries only 0.02 relative power. The power of the higher-order grating lobes is less than one-third of the power in the first-order grating lobes. This observation has important consequences for grating couplers with blazed gratings, since loss of power to unwanted grating lobes can clearly be minimized. We shall see that even better results are obtainable with gratings that have only first-order grating lobes.

Finally, we let the grating amplitude grow to $2a = 2\lambda_o$ and show in Fig. 7a and b how the third-order grating lobe now dominates the grating response. The maxima and minima of these curves cannot easily be identified by ray tracing because of the many possible ray paths. However, the maximum at $D_1/D = 1$ of the curve with $m = 3$ in Fig. 7a seems to be caused by the specular reflection indicated in Fig. 4. The accuracy of the curves in Fig. 7a and b is not as high as that of the other figures. Whereas 42 simultaneous equations were sufficient to solve the problem with sufficient accuracy for $2a = 0.5\lambda_o$, 66 simultaneous equations were used to produce Fig. 7a and b. Computing time increases with the third power of the equation number. The accuracy of the curves in Fig. 6a and b is better than 10 percent, but the accuracy of the curves in Fig. 7a and b is poorer. However, these curves are certainly correct to order of magnitude and have the correct shapes.

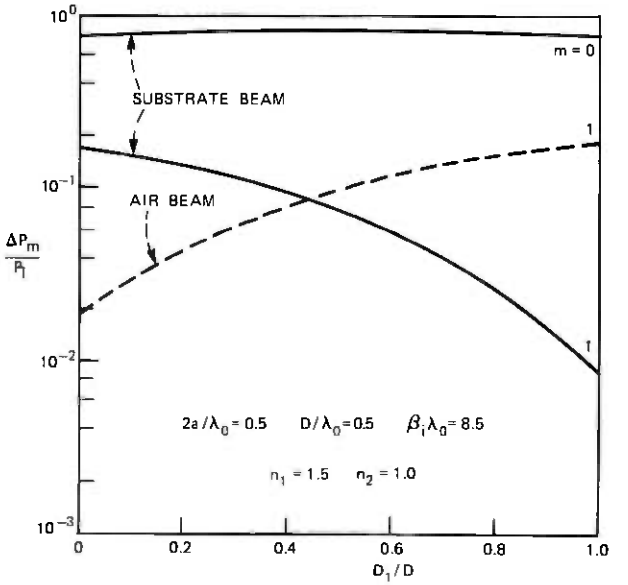


Fig. 8—Grating with only first order lobes. $D/\lambda_o = 0.5$, $\beta_i\lambda_o = 8.5$, $n_1 = 1.5$, $n_2 = 1.0$, and $2a/\lambda_o = 0.5$. The dotted line represents the relative scattered power in air; the solid lines represent the power in the substrate.

To gain insight in the beneficial effects of blazed gratings, a grating with only first-order lobes, $D/\lambda_o = 0.5$, $\beta_i\lambda_o = 8.5$, and an amplitude of $2a = 0.5\lambda_o$ was investigated. The results are plotted in Fig. 8. It is apparent that power is scattered predominantly into the substrate (solid line), if D_1/D is small and predominantly into air (dotted line) if D_1/D approaches unity. Figure 9 shows the ratio of air-to-substrate beam power for $D_1/D = 1$ as a function of the grating amplitude $2a$. Also shown in this figure is the power-reflection coefficient of the specularly reflected component; that is, the zero-order beam in the substrate. As the grating becomes deeper, the power discrimination between air and substrate beams becomes better, but the power reflection coefficient of the specular-beam component becomes lower. If we apply this situation to waveguide geometry, the incident plane wave and the reflected wave with $m = 0$ would both correspond to the guided

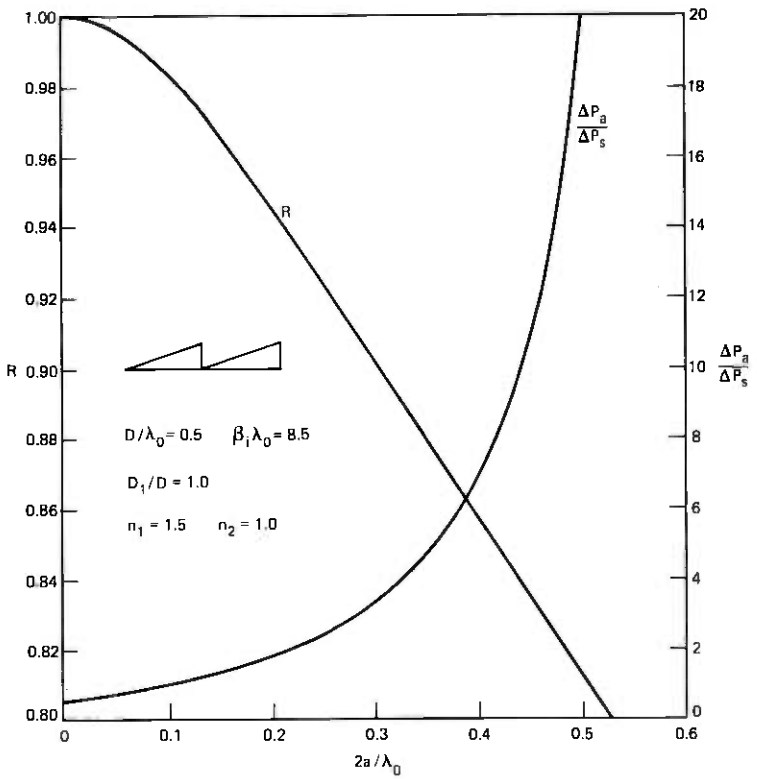


Fig. 9—Ratio of the powers $\Delta P_a/\Delta P_s$ that are scattered into air and substrate and also the power reflection coefficient R of the zero-order beam in the substrate as functions of the normalized grating depth $2a/\lambda_o$. It is $D/\lambda_o = 0.5$, $\beta_i\lambda_o = 8.5$, $D_1/D = 1.0$, $n_1 = 1.5$, and $n_2 = 1.0$.

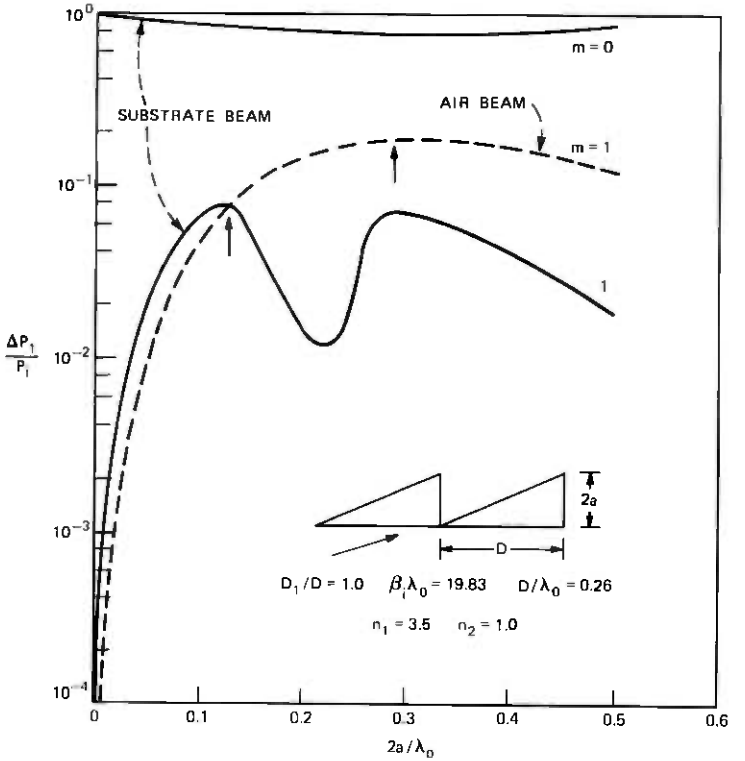


Fig. 10—Normalized scattered power into air (dotted line) and substrate (solid lines) for a grating with $n_1 = 3.5$, $n_2 = 1.0$, $D/\lambda_o = 0.26$, $D_1/D = 1.0$, and $\beta_i\lambda_o = 19.83$. The arrows indicate the position of points where specular reflection from the grating teeth coincides with the grating condition.

mode. The power loss on reflection expresses the mode attenuation per "bounce." A grating with good power discrimination between air and substrate beams suffers very high scattering losses.

Figures 10 and 11 complete our investigation of the scattering properties of blazed gratings with large-grating amplitudes. These figures apply to a substrate with high-refractive index, $n_1 = 3.5$. The gratings have $D_1/D = 1$ in Fig. 10 and $D_1/D = 0$ in Fig. 11. In both figures we used $D/\lambda_o = 0.26$ and $\beta_i\lambda_o = 19.83$. These figures show the scattered power as functions of the grating amplitude. It is obvious that the scattering levels off with increasing grating amplitude, so that it does not help to increase the grating depth beyond a certain point. However, the discrimination between air and substrate beams is affected by the grating depth. The arrows indicate points where specular reflection from the grating faces should enhance the scattered power. Except for obvious interference effects by some other ray path, it seems

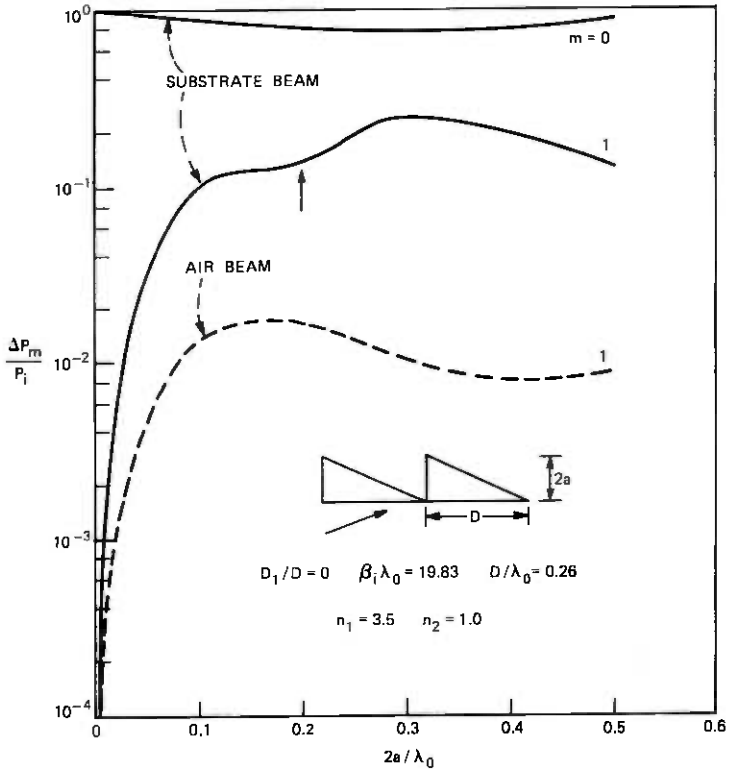


Fig. 11—Similar to Fig. 10 with $D_1/D = 0$.

that the maxima tend to be located where geometrical optics would predict.

We have extracted information about the effective reflection plane of the incident and specularly reflected plane waves. The theory of the effective reflection plane was presented in eqs. (34) and (35). Figure 12 shows the position d (normalized with respect to λ_o) of the effective reflection plane measured from the lower edge of the grating at $x = 0$. We assume that the phase of the zero-order beam in the substrate can be accounted for by reflection from an effective plane interface of the two media with index n_1 and n_2 located at $x = d$. The solid lines in Figs. 12 and 13 are obtained from our exact theory. The dotted curves represent the results of applying the WKB approximation to a continuous refractive index distribution as explained in connection with eq. (34). Figure 12 applies to a long grating period of $D/\lambda_o = 1.3$ and $n_1 = 1.5$, while Fig. 13 was drawn for $D/\lambda_o = 0.26$ and $n_1 = 3.5$. For large grating amplitudes the agreement with the approximate theory is apparently better for shorter grating periods. It might be expected

that the approximation would become very good for $D/\lambda \to 0$. Our use of the WKB approximation becomes inapplicable to $2a \to 0$. In this limit the effective reflection plane is better approximated by $d = a$. However, it is clear that the WKB approximation provides a useful estimate of the position of the effective reflection plane for deep gratings. This information is very important for an application of our theory to an approximate description of scattering by gratings on dielectric film waveguides.

The final figure, Fig. 14, shows a comparison between first-order perturbation theory and the exact grating theory. This figure represents the relative scattered powers in the first-order grating lobe (the only lobe that propagates in this case) as a function of the grating amplitude $2a$ for $D/\lambda_o = 0.5$, $\beta_i\lambda_o = 8.5$, $n_1 = 1.5$, and $n_2 = 1.0$. It is interesting to observe that the air beam is actually stronger than first-order perturbation theory would predict, while the substrate beam is considerably weaker. It is furthermore of interest that the relative strength of air to substrate-scattered power is predicted in reverse order by perturbation theory for large grating amplitudes. Whereas perturbation theory predicts that more power is scattered into the substrate than into air, the exact theory predicts just the opposite. At

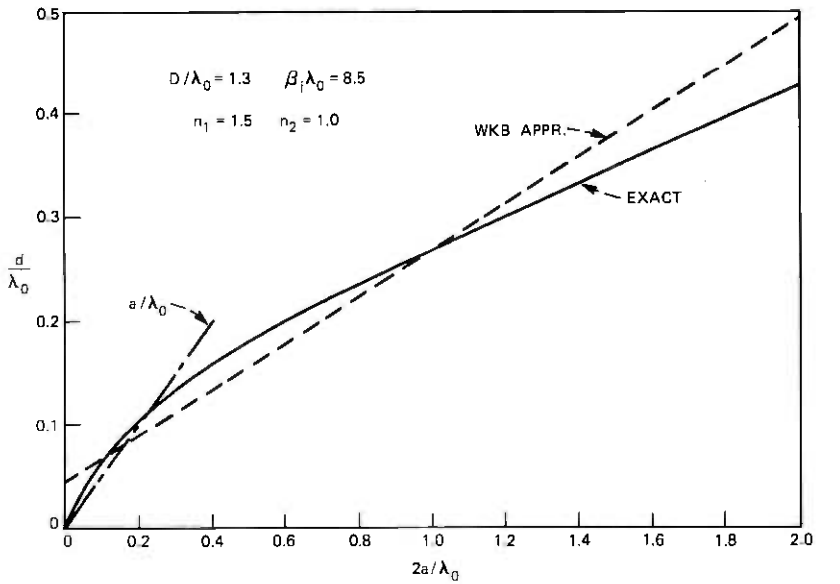


Fig. 12—Position d of the effective reflection plane as a function of grating amplitude $2a$ for $D/\lambda_o = 1.3$, $\beta_i\lambda_o = 8.5$, $n_1 = 1.5$, and $n_2 = 1.0$. The effective reflection plane is practically independent of D_1. The solid line is obtained from the exact theory while the dotted line was computed from the WKB approximation.

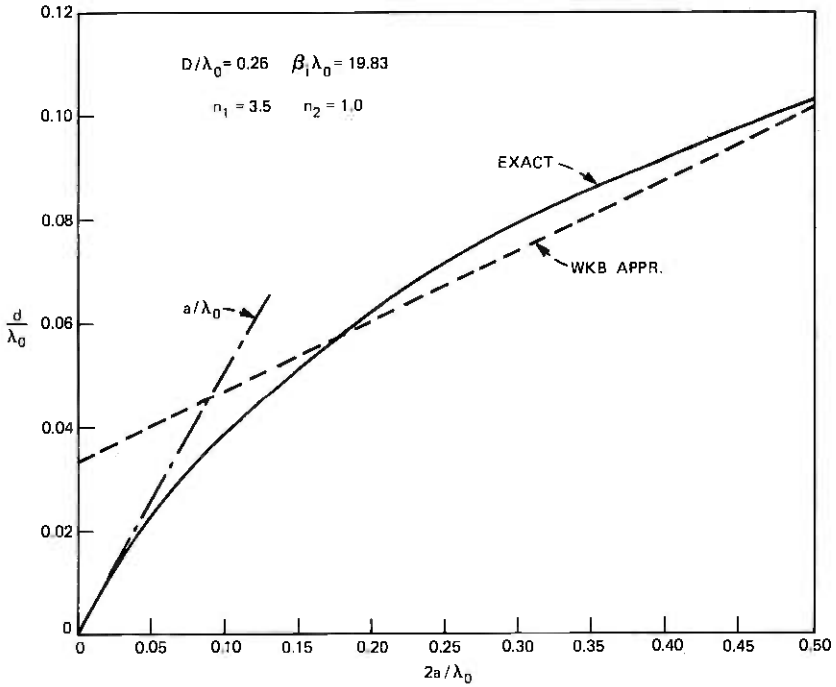


Fig. 13—Similar to Fig. 12 with $D/\lambda_o = 0.26$, $\beta_i\lambda_o = 19.83$, $n_1 = 3.5$, and $n_2 = 1.0$.

this point our theory is at variance with claims made by Tamir[3] whose results are in qualitative agreement with perturbation theory but disagree with our theory.

The arrow in Fig. 14 indicates the point at which the specular reflection condition from the faces of the grating teeth is satisfied for the substrate beam as shown in Fig. 2. This point is in good agreement with the maximum predicted by the exact theory. Perfect agreement cannot be expected for such small grating amplitudes and short grating periods, because geometrical optics cannot be expected to hold under these conditions.

Figure 14 shows that for this type of grating first-order perturbation theory is reasonably accurate for grating amplitudes below $2a = 0.05\lambda_o$.

II. CONCLUSION

We have found that our exact treatment of deep dielectric gratings with triangularly shaped teeth provides a satisfactory method for computing the scattering problem. Our theory has only been applied to incident plane waves of TE polarization. The field outside of the grating region was expanded in a series of plane waves, while the field

in the grating region was expressed as a double Fourier series expansion. Numerical evaluation of this scattering theory requires a modest computational effort. The required number of simultaneous equations that must be solved increases with increasing grating amplitude. The computer time increases with the third power of the number of equations used. For gratings with an amplitude of $2a/\lambda_o = 0.1$, 28 simultaneous equations were used, 42 equations were necessary for $2a/\lambda = 0.5$, and 66 equations were used for $2a/\lambda_o = 2.0$; however, somewhat greater accuracy seems desirable for accurate results in this latter case.

We found that blazed dielectric gratings are able to provide good discrimination of one grating lobe at the expense of other grating responses. The position of maxima and minima of the grating lobes as functions of the grating shape can be accounted for by perturbation theory for small grating amplitudes and by geometrical optics for larger grating amplitudes. However, multiple ray paths make the geometrical optics interpretation difficult for very deep gratings.

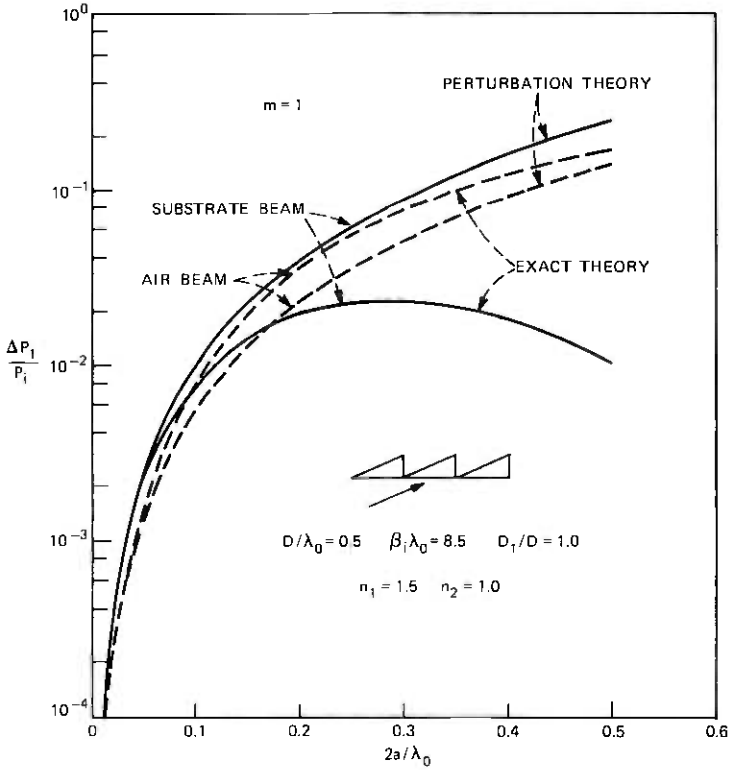


Fig. 14—Comparison with first-order perturbation theory. $D/\lambda_o = 0.5$, $D_1/D = 1$, $\beta_1\lambda_o = 8.5$, $n_1 = 1.5$, and $n_2 = 1$.

It is possible to define an effective reflection plane for the zero-order reflected grating lobe. This concept is useful for an approximate description of grating scattering of guided modes in thin dielectric films. Once the effective width of the film is known for the guided modes, scattering losses can approximately be calculated by accounting for the scattered power by means of a theory that is essentially no more complicated than the theory presented here. We have shown that the position of the effective reflection plane can be estimated by means of the WKB approximation.

III. ACKNOWLEDGMENT

I profited from fruitful discussions with W. W. Rigrod and from his considerable knowledge of the literature.

APPENDIX

We list here the coefficients that enter in the equation system (14).

$$N_{n'-n,m'-m} = \frac{(n_1^2 - n_2^2)k^2b}{2\pi^2(n' - n)} \left\{1 - \exp\left[i2\pi\left(\frac{D_1}{D}(m' - m) + \frac{a}{b}(n' - n)\right)\right]\right\} \times \left[\frac{D_1}{\frac{D_1}{D}(m' - m) + \frac{a}{b}(n' - n)} - \frac{D - D_1}{\frac{D - D_1}{D}(m' - m) - \frac{a}{b}(n' - n)}\right]$$

for $n' \neq n$ and $m' \neq m$;

$$N_{n'-n,0} = \frac{(n_1^2 - n_2^2)k^2b^2D}{2\pi^2(n' - n)^2a}\left\{1 - \exp\left[i2\pi\frac{a}{b}(n' - n)\right]\right\} + \frac{ik^2bD}{\pi(n' - n)}\left\{n_1^2 - n_2^2\exp\left[i2\pi\frac{a}{b}(n' - n)\right]\right\}$$

for $n' \neq n$;

$$N_{o,m'-m} = \frac{aD^3(n_1^2 - n_2^2)k^2}{2\pi^2(m' - m)^2D_1(D - D_1)}\left\{\exp\left[i2\pi\frac{D_1}{D}(m' - m)\right] - 1\right\}$$

for $m' \neq m$;

$$N_{0,0} = ak^2D(n_1^2 + n_2^2);$$

$$M_{n,n'} = \frac{2Dbe^{i\pi(a/b)(n'-n)}}{\pi(n' - n)}\sin\left[\pi\frac{a}{b}(n' - n)\right]$$

for $n' \neq n$;

$$M_{n,n} = 2aD.$$

REFERENCES

1. G. W. Stroke, "Diffraction Gratings," *Handbuch der Physik*, *29*, S. Fluegge, ed., Springer, New York (1967), pp. 426–754.
2. R. Petit, "Electromagnetic Grating Theories: Limitations and Successes," Nouv. Rev. Optique, *6*, No. 3 (1975), pp. 129–135.
3. T. Tamir, Beam Waveguide Coupler, In Integrated Optics, T. Tamir, ed., Springer Verlag, New York (1975), pp. 83–137.
4. D. Marcuse, *Theory of Dielectric Optical Waveguides*, New York: Academic Press, 1974.
5. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, II*, New York: McGraw-Hill, 1953.
6. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold, 1962, p. 18, Eq. (1.6–14).
7. W. W. Rigrod and D. Marcuse, "Radiation Loss Coefficients of Asymmetric Dielectric Waveguides with Shallow Sinusoidal Corrugations," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 9, November 1976
Printed in U.S.A.

Models for the Subjective Effects of Loss, Noise, and Talker Echo on Telephone Connections

By J. R. CAVANAUGH, R. W. HATCH, and J. L. SULLIVAN

(Manuscript received January 9, 1976)

Tests have been conducted at Bell Laboratories within the last 10 years to obtain subjective evaluations of the effects of loss, noise, and talker echo on telephone transmission quality. We use these subjective test results to formulate graphical and analytical models of subjective opinion that can be used in network planning studies to evaluate transmission performance of the network and to study the effects of network changes on performance. These models are based on the concept of a generalized transmission-rating scale. Separate opinion curves for each test take into account differences caused by factors such as subject group, type of test, and range of conditions. We also describe the methods of data analysis used in the formulation of the transmission-rating scale and opinion models, provide a comparison of the test results with the models, and discuss the models in sufficient detail to permit their application in transmission planning studies.

I. INTRODUCTION

In the last 10 years, several tests have been conducted at Bell Laboratories to determine the subjective evaluation of loss, noise, and talker echo in telephone connections. The purpose of these tests was to obtain information for use in network planning studies. Several hundred volunteers served as subjects and participated in several thousand test calls with various amounts of loss, noise, and talker echo. Some of these tests were conducted on normal business calls made by Bell Laboratories employees; others were conducted in a laboratory environment. At the end of each call, the subject was asked to indicate his or her opinion of the transmission quality on a five-category rating scale: excellent, good, fair, poor, and unsatisfactory.

The results from these tests were used to formulate the graphical and analytical models of opinion which are presented in this paper. These models have been used extensively in network planning studies

to evaluate present network performance and to study the effect of possible changes in the network. Although many of these studies are not yet published, a recent paper by Spang[1] provides an excellent example of the application of these models in toll transmission planning.

For many years the Bell System has used subjective tests to obtain information on the effects of transmission quality which could be used in network planning studies. For example, the present Via Net Loss design plan, which introduces loss to control talker echo in the Direct Distance Dialing (DDD) network, is based on the results of subjective tests for talker echo that are included in a 1953 paper by Huntley.[2] Subsequent talker-echo tests were described by Phillips in 1954.[3] Coolidge and Reier reported the results of tests of received telephone speech volume in 1959 and introduced the concept of volume grade-of-service.[4] Test results for message circuit noise were used in noise grade-of-service studies described by Lewinski in 1964.[5]

Most of the tests mentioned considered the effect of one transmission parameter at a time. Since transmission parameters appear in combinations and there are, in many instances, important interactions, a new series of conversational tests for the combined effects of connection loss and circuit noise was initiated in 1965. Subjective evaluations were obtained on normal business calls within Bell Laboratories. The results of these tests were reported by Sen in 1971.[6] Since then, the use of combined loss-noise grade-of-service based on these tests has largely replaced the use of the earlier noise and volume grade-of-service.

Another test to determine subjective reaction to loss and noise was conducted in 1969 on normal business calls within the Bell Laboratories location at Holmdel, New Jersey. There were two reasons for the new test. One was to obtain a larger number of subjects per test condition and thus reduce the experimental variability. The other reason was to include both symmetric loss conditions, as was done in the 1965 tests, and asymmetric loss conditions; i.e., unequal loss in the two directions of transmission. The test results for the symmetric conditions indicated a more critical assessment of quality than the 1965 tests, which could not be explained by known differences in the tests. Therefore, a third test was planned and conducted in 1972.

During this same period of time, new talker-echo tests were initiated. These echo tests used the same five-category rating scale as the loss-noise tests so that possible tradeoffs between loss-noise and echo grade-of-service could be studied. Some of the echo tests were conducted in the laboratory. Others were conducted on normal business conversations between Bell Laboratories employees, where values of loss, noise, echo-path loss, and echo-path delay could all be controlled.

As the results of the various tests became available, work was con-

tinued to modify and improve earlier loss-noise opinion models and to develop similar talker-echo opinion models. Systematic methods were formulated to analyze the results from individual tests and to combine the results from different tests into a composite loss-noise-echo opinion model. In addition, a transmission-rating scale was introduced that assigned a single numerical value to any specific combination of transmission conditions.

The concept of a generalized transmission-rating scale recognized that subjective test results can be affected by various factors such as the subject group, the type of test, and the range of conditions that are included in the test. These factors were found to cause changes in both the mean opinion score for a given condition and in the standard deviation.* Thus, there were difficulties in trying to establish a unique relationship between a given transmission condition and subjective opinion in terms of mean opinion score or other subjective measures of transmission quality. The introduction of a transmission-rating scale tended to reduce this difficulty by separating the relationship between transmission characteristics and opinion ratings into two parts. For the first part, the transmission rating as a function of the transmission characteristic, was anchored for two specific transmission conditions and thus tended to be much less dependent on individual tests. The second part, the relationship between transmission rating and subjective opinion ratings, could then be displayed for the individual test.

The essential features of the transmission-rating scale and opinion models are summarized in Section II of this paper in enough detail to permit their application in transmission planning studies. The remainder of the paper describes three subjective tests for connection loudness loss and circuit noise and four tests for talker echo, outlines the methods of analysis, and describes the formulation of the composite transmission-rating scale and opinion models. Comparisons of the individual test results and the final model are also presented.

II. SUMMARY OF TRANSMISSION-RATING AND OPINION MODELS

The models for transmission rating described in this section are based on the results of seven subjective tests. All of the tests were conducted with Western Electric 500-type telephone sets.[7] Loudness

* Results of subjective tests in terms of the number of votes in each of the several categories of a rating scale can be expressed in a number of ways. One way is to assign numerical values to each of the categories, e.g., excellent = 5, good = 4, fair = 3, poor = 2, and unsatisfactory = 1. Each of these numerics is then weighted by the proportion of votes in the corresponding category for a particular transmission condition, and the weighted values summed. The result is called the mean opinion score for that transmission condition.

Table I — Summary of tests

	Loss-Noise Tests			Talker Echo Tests			
	SIBYL 1965 (MH)	SIBYL 1969 (HO1)	SIBYL 1972 (HO2)	Lab 1966	Lab 1968	Lab 1970	SIBYL 1970
Number of subjects	66	78	74	29	30	100	45
Number of conditions	24	15	12	30	93	10	16
Number of ratings	685	1163	1684	870	2790	1000	752
Median ratings/condition	29	60	60	29	30	100	38
Connection loudness loss (dB)	5–30	10–30	5–30	*	*	18	10
Circuit noise (dBrnC)	21–44	22–45	25–42	28	18–38	33	30
Echo-path loudness loss (dB)	—	—	—	10–59	0–50	33–73	6–42
Echo-path delay (ms)	—	—	—	20–90	1.5–90	600, 1200	10–72
Sidetone-path loudness loss (dB)	13	12	12	9	12	12	12
Average room noise [dB(A)]	45	42	42	35	35	38	42

* These tests were not, strictly speaking, two-way conversation tests and, thus, connection loudness-loss values are not appropriate.

loss values used in the model describe the acoustic-to-acoustic transfer efficiency of overall telephone connections and are expressed in terms of the Electro-Acoustic Rating System (EARS) method.[8] Noise values used in the model are expressed at the line terminals of a telephone set with a reference receiving efficiency of 26 dB based on the EARS method.*

The major aspects of these tests are summarized in Table I. A more detailed description of the tests is presented in Sections III and IV.

The results of the subjective tests were used to derive transmission-rating models for (*i*) loss and noise, (*ii*) talker echo, and (*iii*) the combined effects of loss, noise, and talker echo. In addition, models were derived for the relationship between transmission rating and subjective opinion.

The procedures used in the analysis of the subjective-test results and the derivation of the transmission-rating scale are described in Sections V and VI. Although the procedures are somewhat complex for manual calculation, they are easily handled on a digital computer and have been found to provide a convenient and useful representation for a large variety of test data.

Mathematical expressions for the models are summarized in Table II. The derivations of these expressions are also given in Sections V

* The several subjective tests, results of which were used in deriving the model, were conducted with different circuit-noise values and with telephone sets operating at different EARS receiving efficiencies. To enable combination of results from the different tests to be made, it was necessary to express all circuit-noise values in terms of a telephone set with reference receiving sensitivity. The value of 26 dB was chosen because it is approximately the receiving efficiency of a customer loop consisting of a Western Electric 500-type telephone set,[7] a short-line facility, and a standard central office feeding bridge. Noise values are given in dBrnC.[9,10]

and VI. The remainder of the present section provides a general description and graphical presentation of the models.

2.1 Connection loudness loss and circuit-noise model

Transmission rating as a function of connection loudness loss and message-circuit noise is shown in Fig. 1. The curves were plotted using

Table II — Models for estimating subjective reaction to loss, noise, and echo

The models, in terms of a transmission-rating scale, for loss and noise (R_{LN}), echo (R_E), and loss, noise, and echo (R_{LNE}) are:

$$R_{LN} = 147.76 - 2.257\sqrt{(L_e - 7.2)^2 + 1} - 2.009N_F + 0.02037(L_e)N_F \quad (1)$$

$$R_E = 95.01 - 53.45 \log_{10}\{(1 + D)/\sqrt{1 + (D/480)^2}\} + 2.277E \quad (2)$$

$$R_{LNE} = \frac{R_{LN} + R_E}{2} - \sqrt{\left[\frac{R_{LN} - R_E}{2}\right]^2 + (10)^2}, \quad (3)$$

where L_e = Acoustic-to-acoustic loudness loss (in dB) of an overall telephone connection, determined using the Electro-Acoustic Rating System (EARS) method,
N = Circuit noise (in dBrnC) at the input to a set with a receiving-loudness rating of 26 dB, determined using the EARS method,
N_F = Total noise in dBrnC resulting from power addition of the circuit noise, N, and 27.37, both in dBrnC,
D = Round-trip echo-path delay (in milliseconds), and
E = Acoustic-to-acoustic loudness loss (in dB) of the echo path, determined using the EARS method.

The proportion of comments good or better (GoB) or poor or worse (PoW) are computed from R by:

$$GoB = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{A} e^{-t^2/2} dt$$

$$PoW = \frac{1}{\sqrt{2\pi}} \int_{B}^{\infty} e^{-t^2/2} dt,$$

where A and B are given in the table below for the various data bases.

Data Base	A	B
MH	$(R - 64.07)/17.57$	$(R - 51.87)/17.57$
HO1	$(R - 77.44)/17.07$	$(R - 60.70)/17.07$
HO2	$(R - 73.74)/15.68$	$(R - 58.03)/15.68$
Echo1	$(R - 75.05)/14.30$	$(R - 58.95)/14.30$
Echo2	$(R - 66.66)/11.84$	$(R - 53.33)/11.84$

The parameters A and B have been derived from opinion distributions, in terms of fit mean, μ, and fit standard deviation, σ, and then expressed as a function of R. Alternatively, the models can be expressed as follows:

$$\mu_{MH} = (R - 21.37)/12.20 \qquad \sigma_a = 1.44$$
$$\mu_{HO1} = (\mu_{MH} + 0.206)/1.372 = (R - 18.86)/16.74 \qquad \sigma_a = 1.02$$
$$\mu_{HO2} = (\mu_{MH} + 0.215)/1.288 = (R - 18.75)/15.71 \qquad \sigma_a = 0.998$$
$$\mu_{E1} = (R - 18.7)/16.1 \qquad \sigma_a = 0.888$$
$$\mu_{E2} = (R - 20)/13.33 \qquad \sigma_a = 0.888$$

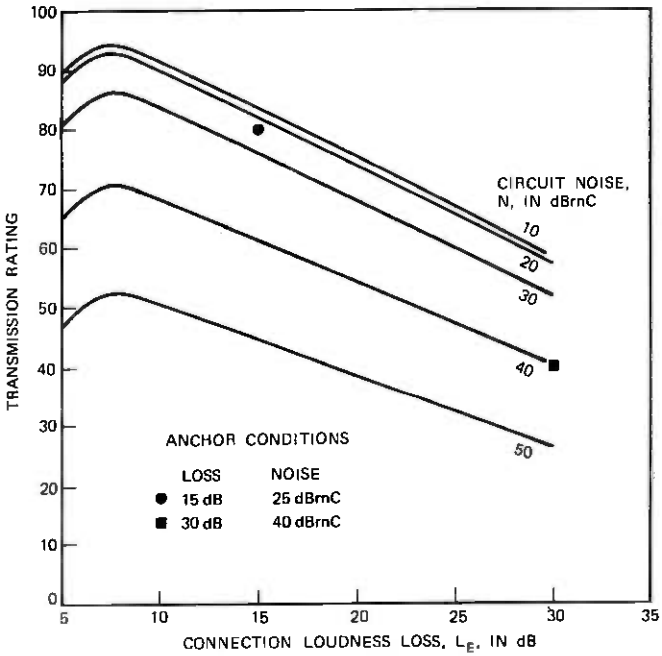


Fig. 1—Transmission rating for loss and noise.

eq. (1) below, which is also given in Table II.

$$R_{LN} = 147.76 - 2.257\sqrt{(L_e - 7.2)^2 + 1} - 2.009N_F + 0.02037(L_e)N_F, \quad (1)$$

where

L_e = Acoustic-to-acoustic loudness loss (in dB) of an overall telephone connection

N = Circuit noise (in dBrnC) at the input to a set with a receiving loudness rating of 26 dB

N_F = Power addition of the circuit noise, N, and 27.37 dBrnC.

The transmission-rating scale was derived, so that it is anchored at two points, as shown in Table III. These anchor points were selected to be well separated in quality, but within the range of conditions that are likely to be included in a test. Transmission ratings for other combinations of connection loudness loss and circuit noise are relative to those for the two anchor points. The rating values are such that most telephone connections will have positive ratings between 40 and 100, with the higher rating denoting higher quality. For most engineering applications sufficient accuracy can be achieved by the use of

Table III — Anchor conditions for the transmission-rating scale

Connection Loudness Loss (dB)	Circuit Noise (dBrnC)	Transmission Rating
15	25	80
30	40	40

whole numbers on the transmission-rating scale. For example, in the 1965 loss and noise tests, conditions with a transmission rating of approximately 80 were considered good or excellent by 80 percent of the subjects, while a transmission rating of 40 was considered good or excellent by only 10 percent.

2.2 Talker-echo model

Transmission rating as a function of talker-echo path loss and delay is shown in Fig. 2. The curves were plotted using eq. (2) below, which is also given in Table II. This equation was derived to exclude the effects of circuit noise and connection loudness loss.

$$R_E = 95.01 - 53.45 \log_{10} \{[1 + D]/\sqrt{1 + (D/480)^2}\} + 2.277E, \quad (2)$$

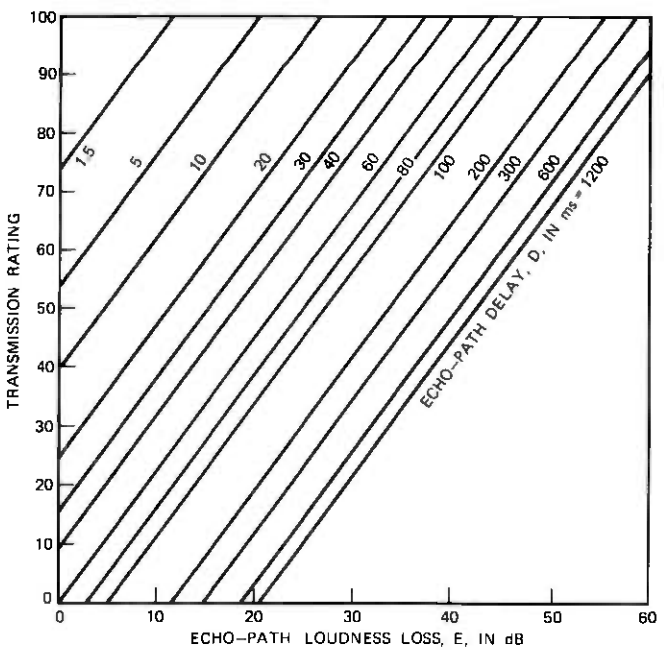


Fig. 2—Transmission rating for talker echo.

where

D = echo-path delay (in ms) and
E = acoustic-to-acoustic loudness loss (in dB) of the echo path.

The curves of Fig. 2 demonstrate the dependence of transmission quality on the two talker-echo path parameters, loss and delay, for connections where talker echo is important.

2.3 Connection loudness loss, circuit noise, and talker-echo model

Transmission ratings for the combined effects of connection loudness loss, circuit noise, echo-path loudness loss and echo-path delay are obtained from eq. (3) below, which is also given in Table II.

$$R_{LNE} = \frac{R_{LN} + R_E}{2} - \sqrt{\left(\frac{R_{LN} - R_E}{2}\right)^2 + (10)^2}, \qquad (3)$$

where

R_{LNE} = transmission rating for the combined effects of connection loudness loss, circuit noise, and talker echo,
R_{LN} = transmission rating for connection loudness loss and circuit noise
R_E = transmission rating for echo-path loudness loss and delay.

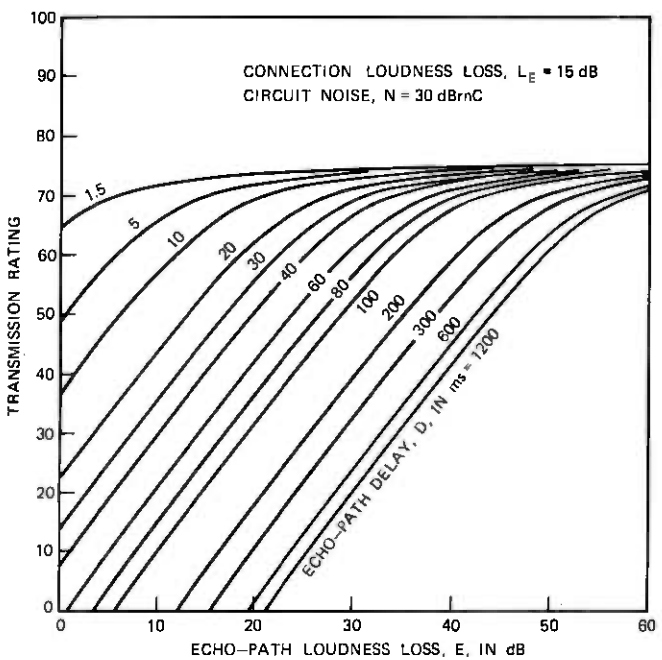


Fig. 3—Transmission rating for loss, noise, and talker echo.

Figure 3 illustrates curves generated by means of the above relationship for the transmission rating as a function of echo-path loudness loss and delay in a connection with a connection loudness loss of 15 dB and circuit noise of 30 dBrnC. For other values of connection loudness loss and circuit noise, the curves would become asymptotic to higher or lower values of R in accordance with the curves of Fig. 1.

2.4 *Subjective-opinion models*

Subjective opinion in terms of the proportion of ratings in each of the five categories (E, G, F, P, U) for a condition having a given transmission rating has been found to depend on various factors, such as the subject group, the range of conditions presented in a test, the year in which the test was conducted, and whether the test was conducted on conversations in a laboratory environment or on normal telephone calls. For the major tests on which the transmission-rating model is based, the observed relationship between subjective judgments and transmission rating can be represented as shown in Figs. 4 and 5 which are plotted from the equations for *GoB* and *PoW* of Table II.

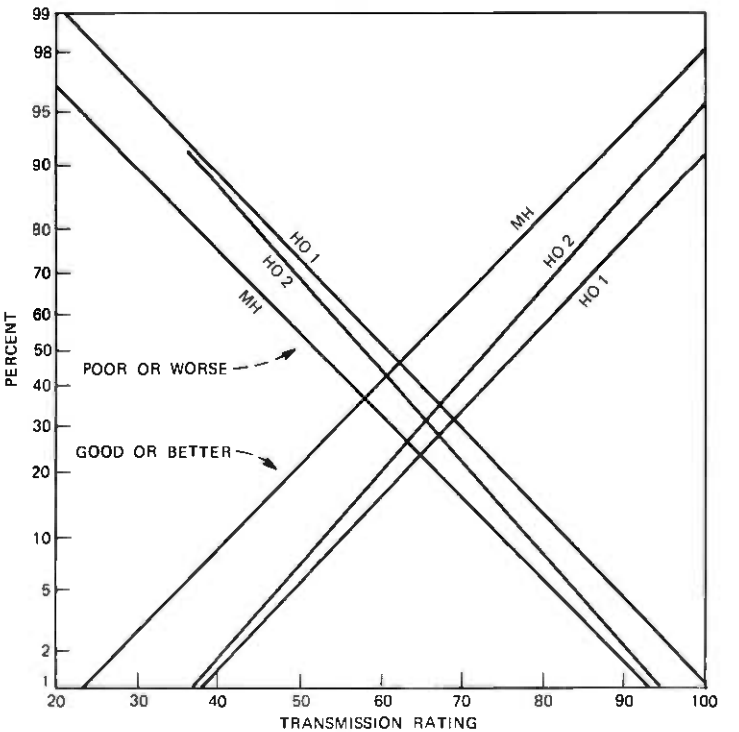


Fig. 4—Subjective opinion as a function of transmission rating for the connection loudness loss and circuit-noise tests.

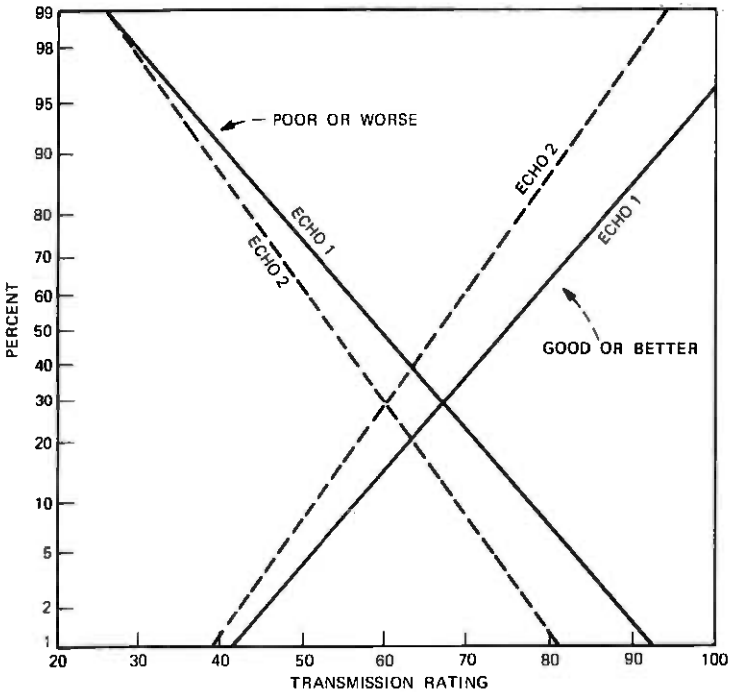


Fig. 5—Subjective opinion as a function of transmission rating for the echo tests.

Contours of constant percent good or better and percent poor or worse are given in Figs. 6 and 7, respectively, for the loss-noise results. These contours were computed from eq. (1) and the equations for *GoB* and *PoW* of Table II using the MH data base. Similar contours could be generated based on one of the other tests. However, the Murray Hill base is being used for current network planning studies for consistency with earlier studies. In addition, the opinion results from this test appear to be in close agreement with data obtained from customer interviews on typical DDD toll connections. Thus, at the time of publication, the Murray Hill base is recommended for conversion of the transmission ratings to subjective ratings. Eventually, we hope to determine other expressions for A and B (see Table II) that will provide even better agreement with customer interviews on various types of telephone connections.

2.5 Use of the models

The models summarized in preceding sections can be used to estimate transmission quality for telephone connections. The examples given below are based on representative 500-type telephone sets con-

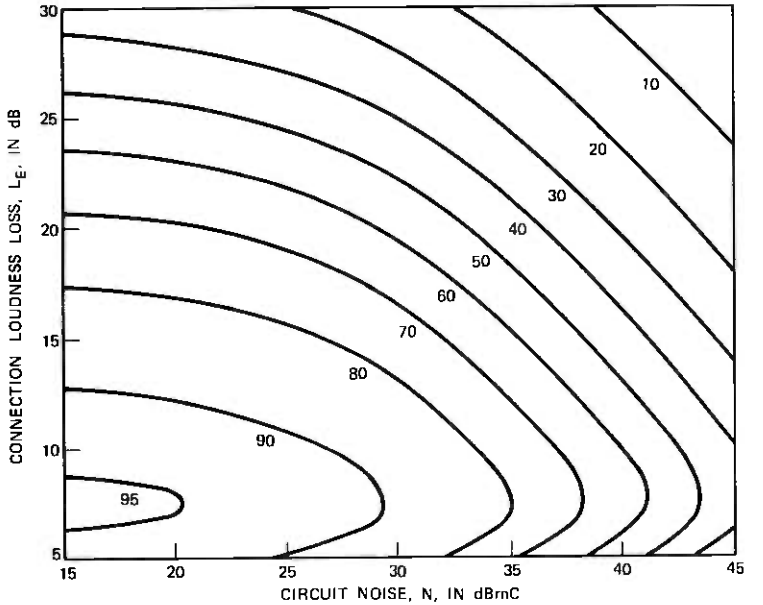


Fig. 6—Subjective-opinion contours of percent good or better at the MH base for loss and noise.

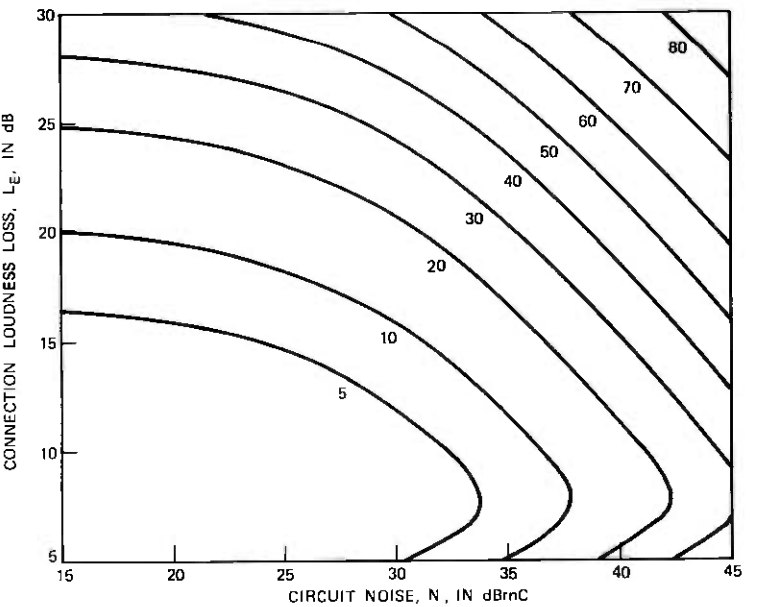


Fig. 7—Subjective-opinion contours of percent poor or worse at the MH base for loss and noise.

nected by specified lengths of 26-gauge nonloaded cable to the local Class 5 office.[1] The loss for such an arrangement is expressed in terms of:

TLR = Transmitting loop rating (in dB), which describes the loudness conversion efficiency in terms of an acoustic signal applied at the telephone set and the resulting electric voltage at the local Class 5 office.

RLR = Receiving loop rating (in dB), which describes the loudness-conversion efficiency in terms of an electric circuit voltage applied at the local Class 5 office and the resulting acoustic pressure received at the telephone set.[8]

The connection loudness loss, L_e, for a local call is the sum of TLR and RLR. For calls between customers served by different local Class 5 offices, L_e is the sum (TLR + RLR) plus the 1000-Hz loss between the offices.

The circuit noise, N (in dBrnC), as used in the model is in terms of a reference telephone set that has an RLR of 26 dB. Noise levels, as typically expressed at the line terminals of a telephone set, need to be corrected to the reference set. For the 500-type telephone set, the conversion factor is about 4 dB (increase in noise) and is nearly independent of loop length.

The effect of talker echo depends on the characteristics of the echo path. Generally, the dominating path is that from the talking customer to the distant Class 5 office and return, and is referred to as the far-end echo path.[11] The round-trip delay of this path, D, is taken to be the time required for a speech signal to go from the originating Class 5 office to the distant Class 5 office and return. (Delay in the customer's telephone set and loop is neglected as, with present plant, it is usually insignificant.) The loudness loss of the echo path, E, is the sum of TLR, RLR, and the echo-path loss from the originating Class 5 office to the distant Class 5 office and return.

Considerations of the preceding paragraphs provide the basis for several examples demonstrating use of the models. Details concerning computation for these examples are given in Appendix A. Results of these examples are summarized in Table IV. The examples are simplified representations of connections devised to illustrate application of the models and, thus, the results only approximately describe the performance of actual connections. (Methods of obtaining more accurate connection representations are covered in Ref. 1.)

Comparison of the results for Examples 1, 2, and 3 show that about optimum performance for local connections occurs for medium loops. The performance is below optimum for short and long loops, the former because the loss is lower, the latter because the loss is higher (see Figs. 1, 6, and 7).

Table IV — Examples of performance estimates obtained using the models

Example	L_E (dB)	N (dBrnC)	D (ms)	E (dB)	R_{LN}	R_e	R_{LNE}	GoB (%)	PoW (%)
1. Local connection, short loops	2.7	27	—	—	78.3	—	—	79.2	6.6
2. Local connection, medium loops	8.4	27	—	—	88.7	—	—	92	1.8
3. Local connection, long loops	16.9	27	—	—	75.5	—	—	74.2	8.9
4. Long-toll connection, medium loops, no talker echo	16.1	32.5	—	—	71	—	—	65.2	13.9
5. Long-toll connection, medium loops, with talker echo	16.1	32.5	37.3	31.7	71	82.6	65.2	52.6	22.4

Comparison of Examples 3 and 4 illustrates the effect of higher loss and noise typically encountered on toll connections.[12] Finally, comparison of Examples 4 and 5 indicates the effect of talker echo, demonstrating the need for echo control. (See Ref. 1 for detailed discussion of loss, noise, and talker echo for toll connections.)

III. DESCRIPTION OF LOSS-NOISE SUBJECTIVE TESTS

Three tests have been conducted using a special test facility called SIBYL to determine subjective reaction to loss and noise on telephone connections. This facility allowed control of transmission parameters during normal business calls of cooperating Bell Laboratories employees.[13–15]

SIBYL was first used at the Murray Hill Bell Laboratories location, and was moved to the Holmdel location in 1966.

The test subjects for the SIBYL studies were Bell Laboratories employees. Prior to the beginning of each of the tests, a list of employees reflecting a preselected makeup of age, sex, etc., was obtained, and the employees contacted to solicit their participation in the test. Upon obtaining agreement, their telephone lines were routed through SIBYL which could handle up to 100 subject lines.

An overall connection with SIBYL inserted is shown on Fig. 8. At the left is a subject's (participating employee's) telephone set. In close physical proximity (less than about 1500 feet of two-conductor cable) is SIBYL which converts the two-wire transmission path into a four-wire transmission path and separates signals transmitted from the subject's telephone set and signals received at the same telephone set. Separation of the signal paths in this manner permits (*i*) inserting different impairment values for the two directions of transmission and (*ii*) independent measurement of signals transmitted from and received at the subject's telephone set.

Proceeding from left to right in Fig. 8, the four-wire path is reconverted to a two-wire path in SIBYL, and connected to the serving

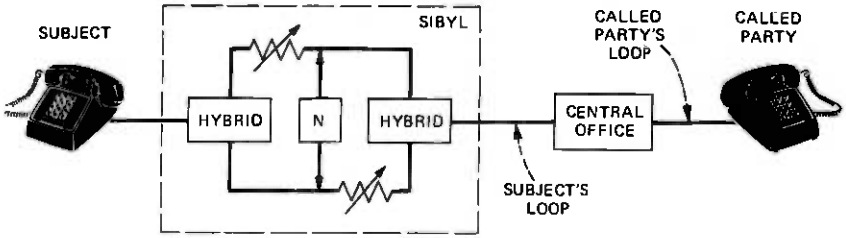


Fig. 8—Diagram of a telephone connection for the SIBYL tests.

central office over a two-wire cable pair which is about 1 mile in length for Murray Hill and about 3 miles in length for Holmdel. The central office switching machine connects the subject's line to the called line dialed by the subject. The called line (approximate lengths as given above for Murray Hill and Holmdel) is terminated with the telephone set of the called party, another Bell Laboratories employee at the same location who is usually not a subject in the experiment.

SIYBL recognized calls internal to the location, and only such calls were included in the experiments reported here. The major reason for this was to retain as much control as possible of the transmission parameters on test calls. Employees' telephone lines at any one location generally were of about the same physical makeup. Thus, variations in transmission parameter values between lines were small, and the values were considered to be identical for all lines. SIBYL then altered the normal parameter values to achieve the parameter values of interest in an experiment.

Restricting the calls to within the location had the further advantage that the subjects physical environment was reasonably uniform, consisting largely of offices with one to four desks and of electronic laboratories. Thus, room noise levels, which could affect the transmission quality of telephone calls, did not vary appreciably.

The subjects were provided with instructions at the beginning of the test, and thereafter, the procedure during a call was generally as shown in Fig. 9. (The procedure of Fig. 9 applies specifically to the Murray Hill test.) The subject initiated the procedure by lifting his handset from the telephone set cradle and dialing a number. If the call was *not* a test call, which was the case for about 85 percent of the calls from the subject group, the call was completed normally. That is, (*i*) the subject received a signal indicating the called line was busy and returned his handset to the cradle, (*ii*) the called party did not answer and the subject returned his handset to the cradle, or (*iii*) the called party answered and they conducted their conversation, after which the subject and called party returned their respective handsets to the cradles.

If the call was a test call and the called line was not busy, a test condition (e.g., a predetermined combination of loss and noise values) was inserted at the beginning of the call, and the conversation proceeded. If, during the call, the test condition was unacceptable to the subject (e.g., the noise was too loud) the subject could dial a digit (e.g., a five) on his telephone set dial. This signaled SIBYL to remove the test condition, after which the call completed normally. (This occurred on only about 1 to 2 percent of all test calls.)

In the more usual case, the subject was either not consciously aware of the test condition or did not find it unacceptable and completed the conversation, then returned the handset to the cradle. A short time thereafter, typically seconds, the subject received ringback—the telephone set ringer emitted a burst of sound—which alerted the subject to rate the transmission quality of the call just completed.

3.1 1965 Murray Hill SIBYL test

In 1965, a test to determine subjective reaction to loss and noise was conducted using the SIBYL facility, which was then located at Bell Laboratories, Murray Hill, New Jersey. The configuration of a typical connection incorporating SIBYL is shown on Fig. 8 as discussed earlier.

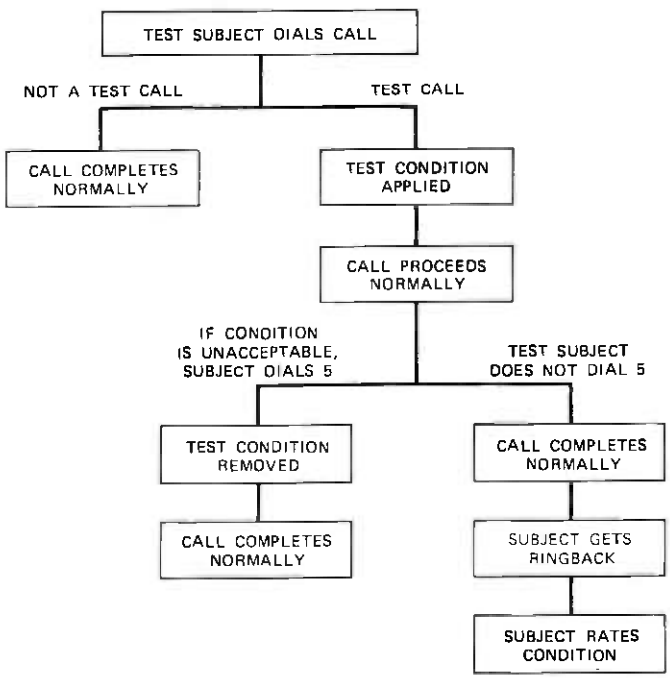


Fig. 9—Procedure for a SIBYL call.

Table V — 1965 Murray Hill SIBYL test of loss and noise: number of ratings for each test condition

Connection Loudness Loss (dB)	Circuit Noise (dBrnC)			
	21	28	36	44
5	35	12	24	20
10	61	36	34	34
15	33	33	27	28
20	29	43	23	35
25	35	32	29	24
30	12	12	23	11

The 24 combinations of loss and noise values that were tested are given in Table V together with the number of subjective ratings obtained for each combination. The loss values of Table V represent acoustic-to-acoustic loudness loss (in dB) of the connections, and are numerically equivalent to the electrical losses in terms of which results were published earlier for this test.[6] Circuit noise levels (in dBrnC) were originally reported in terms of (*i*) added noise, and did not include allowance for noise normally present on the connections, and (*ii*) the average receiving sensitivity of Western Electric 500-type telephone sets at the Murray Hill location. The noise levels of Table V represent total noise from all sources and are expressed at the line terminals of a telephone set with reference receiving sensitivity. The condition with loudness loss = 5 dB and noise = 21 dBrnC represents average transmission normally experienced by employees on calls within the Murray Hill location.

The lines connecting employees' telephone sets to the central office were all of about the same length. Thus, the loudness loss of the telephone side tone path,[16,17] strongly dependent on line length, was expected to be about the same for all employees. Side tone path loudness loss was measured on a sample of lines and telephone sets, and found to be about 13 dB. Circuitry was incorporated into SIBYL to ensure that the side tone path loudness loss would also be about 13 dB for subjects' telephone sets when connected to SIBYL.

Room noise was measured at a sample of subject locations using a sound-level meter with A-weighting.[18] The average value was found to be 45 dB(A) with a range of ±2 dB.

Sixty-six employees (subjects) participated in the test. The predetermined combinations of loss and noise were randomly introduced into about 15 percent of all within-location, normal-business calls placed by the subject group during the 3-month test interval. A subject had no prior information that any particular call was being placed

over a test connection. The call procedure summarized in Fig. 9 was followed for each call. At the end of each test call, a subject judged overall transmission quality by dialing a 9 for *excellent*, 8 for *good*, 7 for *fair*, and 6 for *poor*. Connections rejected during the call were considered to be *unsatisfactory*.

Test results are given in Table VI for each test condition in terms of the percent of subjects' votes in each of the five rating categories. These results are used in Sections V and VI in deriving the subjective-opinion models.

3.2 1969 Holmdel 1 SIBYL test

As discussed earlier, a second test to determine subjective reaction to loss and noise was conducted in 1969. This test utilized the SIBYL facility which was moved to the Holmdel location of Bell Laboratories in 1966. A major reason for these newer tests was to examine the effect of asymmetric transmission conditions on subjects' ratings. However, the test included 15 symmetric conditions (out of a total of 47 test conditions) and these are considered in this paper.

Table VI — 1965 Murray Hill SIBYL test of loss and noise: test results

Circuit Noise (dBrnC)	Connection Loudness Loss (dB)	Percent of Subjects' Votes in Each Category				
		Excel.	Good	Fair	Poor	Unsat.
21	5	60.0	11.4	14.3	14.3	0
	10	63.9	26.2	3.3	6.6	0
	15	66.7	21.2	12.1	0	0
	20	41.3	27.6	24.1	3.5	3.5
	25	34.3	14.3	31.4	8.6	11.4
	30	8.3	25	33.3	16.7	16.7
28	5	50.0	33.3	16.7	0	0
	10	66.7	30.5	0	2.8	0
	15	54.5	30.3	9.1	6.1	0
	20	25.6	30.2	18.6	16.3	9.3
	25	25.0	15.6	37.5	12.5	9.4
	30	8.3	8.3	33.3	16.8	33.3
36	5	54.2	20.8	16.7	8.3	0
	10	64.7	23.6	5.9	2.9	2.9
	15	33.3	29.6	22.2	14.9	0
	20	21.7	8.7	34.8	21.7	13.1
	25	13.8	3.5	17.2	34.5	31.0
	30	0	13.0	26.1	34.8	26.1
44	5	15.0	30.0	25.0	20.0	10.0
	10	17.6	26.5	11.8	20.6	23.5
	15	17.8	25.0	28.6	14.3	14.3
	20	8.6	5.7	20.0	40.0	25.7
	25	0	4.2	25.0	50.0	20.8
	30	0	0	0	36.4	63.6

Table VII — 1969 Holmdel 1 SIBYL test of loss and noise: number of ratings for each test condition

Connection Loudness Loss (dB)	Circuit Noise (dBrnC)			
	22	35	40	45
10	335	48	41	38
15	93	—	—	—
20	65	57	77	77
25	33	34	47	—
30	76	61	—	81

The symmetric combinations of loss and noise values that were tested are given in Table VII together with the number of subjective ratings obtained for each condition. The condition with connection loudness loss = 10 dB and circuit noise = 25 dBrnC was the reference condition that was repeated frequently during the test.

Side tone path loudness loss was measured on a sample of lines and telephone sets, and found to average about 12 dB with a range of ±2 dB both for normal connections and SIBYL connections. Room noise, measured at 12 subject locations, averaged about 42 dB(A) with a range of about ±2 dB.

Seventy-eight employees (subjects) participated in this test. Subjects followed the same procedure during the $2\frac{1}{2}$ month test interval as already described for the Murray Hill test. Comparison of the Holmdel 1 test results of Table VIII with the Murray Hill test results of Table VI shows that in the former, subjects gave lower ratings for approximately equivalent combinations of loss and noise than is the case for the latter. The major differences between the two tests were (*i*) the subject groups, (*ii*) the time difference of about 4 years, and (*iii*) the location. While it is not clear how these differences contributed to differences in results, the comparison suggests that subjects' expectations were higher in the later Holmdel tests.

3.3 1972 Holmdel 2 SIBYL test

A third test to determine subjective reaction to loss and noise was conducted in 1972. This test also utilized the SIBYL facility at the Holmdel location of Bell Laboratories.

A major purpose of the test was to determine whether or not the more critical subjective evaluations found in the Holmdel 1 test would continue to hold. The test included combinations of crosstalk, loss, and noise as well as combinations of loss and noise. Only the latter are considered here because only the results for these conditions are used in the subjective-opinion model for loss and noise.

Table VIII — 1969 Holmdel 1 SIBYL test of loss and noise: test results

Circuit Noise (dBrnC)	Connection Loudness Loss (dB)	Percent of Subjects' Votes in Each Category				
		Excel.	Good	Fair	Poor	Unsat.
22	10	37.9	41.2	13.7	4.5	2.7
	15	29.0	40.9	21.5	5.4	3.2
	20	15.4	26.1	43.1	7.7	7.7
	25	3.0	15.1	27.3	27.3	27.3
	30	2.6	13.2	34.2	30.3	19.7
35	10	12.5	20.8	39.6	25.0	2.1
	20	1.8	10.5	45.6	26.3	15.8
	25	0	2.9	20.6	53.0	23.5
	30	1.6	3.3	16.4	47.5	31.2
40	10	4.9	34.1	24.4	26.8	9.8
	20	0	1.3	14.3	53.2	31.2
	25	2.1	0	12.8	51.1	34.0
45	10	0	7.9	21.0	47.4	23.7
	20	1.3	0	22.1	39.0	37.6
	30	0	0	1.2	33.3	65.5

The 12 symmetric combinations of loss and noise values tested are given in Table IX together with the number of subjective ratings obtained for each combination. These conditions covered about the same range as for the Holmdel 1 (HO1) SIBYL test, except that a condition of lower connection loudness loss of 5 dB was included in the Holmdel 2 (HO2) SIBYL test to match the range of loss in the Murray Hill (MH) test.

Room-noise levels and side tone path loudness loss values were assumed to be identical with those for the HO1 test.

Seventy-four employees (subjects) participated in the test. (None of these employees had been subjects in the HO1 test.) The subjects followed the same procedure as has already been described for the MH

Table IX — 1972 Holmdel 2 SIBYL test of loss and noise: number of ratings for each test condition

Connection Loudness Loss (dB)	Circuit Noise (dBrnC)		
	25	32	42
5	99	70	53
10	1029	58	97
20	50	64	46
30	62	52	4

Table X—1972 Holmdel 2 SIBYL test of loss and noise: test results

Circuit Noise (dBrnC)	Connection Loudness Loss (dB)	Percent of Subjects' Votes in Each Category				
		Excel.	Good	Fair	Poor	Unsat.
25	5	49.5	35.4	13.1	0	2.0
	10	47.2	40.6	10.2	1.3	0.7
	20	14.0	26.0	40.0	18.0	2.0
	30	3.2	3.2	32.3	35.5	25.8
32	5	27.1	35.7	31.4	5.8	0
	10	19.0	39.6	32.8	6.9	1.7
	20	4.7	6.3	48.3	34.4	6.3
	30	1.9	7.7	15.4	51.9	23.1
42	5	11.3	11.3	35.8	34.0	7.6
	10	17.5	17.5	33.1	24.7	7.2
	20	4.3	10.9	19.6	41.3	23.9
	30	0	0	25.0	0	75.0

and HO1 tests, except that the voting procedure was changed. In the HO2 test, subjects were instructed that at the end of each experimental call, when they received ringback, they should rate the overall transmission quality of the call by dialing 9 for *excellent*, 8 for *good*, 7 for *fair*, 6 for *poor*, and 5 for *unsatisfactory*. In addition, they could reject an unacceptable connection during a call by dialing 4. In the latter case, they still received ringback after the call and were asked to rate the quality according to the 5-point scale. (In the MH and HO1 tests, a 4-point scale was used for post-call rating, and the fifth point, unsatisfactory, was assumed for dialed-out calls.)

Results of the HO2 test are given in Table X. These results are in close agreement with those of the HO1 test. Comparison of the test results for the three loss-noise tests are dealt with further in later sections covering derivation of the subjective-opinion models.

IV. DESCRIPTION OF TALKER-ECHO SUBJECTIVE TESTS

Talker echo occurs on a telephone connection when a portion of the primary speech signal is reflected at an impedance mismatch at some point in the connection, and returned to the talker delayed in time. The returned signal, talker echo, is defined in terms of echo-path delay and echo-path loudness loss. The echo-path delay that occurs because of the finite propagation velocity of the speech signal over transmission facilities and equipments is the time it takes the speech signal to traverse the path from the talker's lips to the point of impedance mismatch, then back to the talker's ear. Echo-path loudness loss represents the amount by which the talker's speech signal is attenuated when traversing the same path.

Four tests were conducted to determine subjective reaction to talker echo. Three of these tests—identified as the 1966 Laboratory Echo Test, the 1968 Laboratory Echo Test, and the 1970 Laboratory Echo Test—were conducted under laboratory conditions where the experimenter could closely control conditions. The laboratories used in these tests included rooms that were acoustically designed to muffle both internal and external noise.[15] The fourth test, identified as the 1970 SIBYL Echo Test, was conducted using the SIBYL facility.

The 1966 and 1968 Laboratory Echo Tests and the 1970 SIBYL Echo Tests were designed to study subjective reaction at short echo delays (<100 ms) such as might be encountered on long terrestrial connections. The 1970 Laboratory Echo Test considered the effects of long delays that might be encountered on connections using one- and two-hop synchronous-orbit satellite connections.

4.1 1966 laboratory echo test

The test conducted in 1966 to determine subjective reaction to talker echo utilized the test system shown in block diagram form on Fig. 10. This system provided (*i*) a fixed sidetone path with a loudness loss of about 9 dB, (*ii*) an echo path by means of which the subject heard his own voice delayed in time and attenuated under control of the experimenter, and (*iii*) a transmission path from the test administrator to the subject that had a loudness loss of 14 dB. (Transmission from the subject to the test administrator was obtained by means of an intercom system.) The administrator and subject were located in separate rooms for which the ambient room noise was about 35 dB(A), presumably sufficiently low so as to not affect subjects' ratings. Circuit noise was held constant at 28 dBrnC.

The subject was first given four practice conditions to illustrate the range of transmission quality. Then the actual test conditions were presented. The test incorporated 30 conditions, five different values of

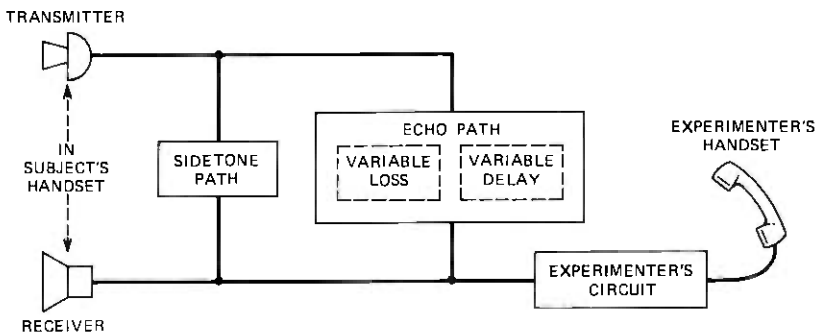


Fig. 10—Test system for the 1966 and 1968 laboratory echo tests.

echo-path delay each with six different values of echo-path loudness loss. These conditions were presented in random order and each subject evaluated each condition once.

At the beginning of each test session, a subject was seated in the test room and given general instructions for the test by the administrator speaking over the system of Fig. 10. The administrator's talking level was held constant. For each of the selected test conditions, the administrator spoke phonetically balanced sentences with the subject repeating each sentence immediately afterward. This continued until the subject arrived at a rating (excellent, good, fair, poor, unsatisfactory) for the test condition. Then the next condition was administered.

The test conditions and the results obtained for the 29-member subject group are given in Table XI. These results show that for any given echo-path delay, the transmission quality improves with increasing echo-path loudness loss. Also, the data indicate that for any given echo-path loudness loss, the transmission quality is degraded with increasing echo-path delay.

4.2 1968 laboratory echo test

This test, conducted in 1968, was designed on the basis of results obtained in the 1966 Laboratory Echo Test. The test system was the same as discussed in the preceding section.

Side tone path loudness loss was constant at 12 dB. Ambient room noise was about 35 dB(A).

The test incorporated 93 conditions. Three of these were base conditions at three different circuit-noise levels. The remaining 90 conditions represented various combinations of circuit-noise level, echo-path delay, and echo-path loudness loss. These conditions were arranged in random order and presented to each test subject in two sessions to avoid subject fatigue.

The procedure for each subject followed that discussed in Section 4.2 except that only three practice conditions were used.

The test conditions and the results obtained for the 30-member subject group are given in Table XII. These test results are reported in part 1 of Annex 4 to Question 6/XII in Ref. 19.

4.3 1970 laboratory echo test

Tests to determine subjective reaction to echo on circuits with echo-path delays of 0 ms, 65 ms, 600 ms, and 1200 ms were conducted in 1970. The tests consisted of a total of 25 conditions, many of which included echo suppressors. One condition at 0 ms, five conditions at 600 ms, and four conditions at 1200 ms did not employ echo sup-

Table XI — 1966 laboratory echo-test results

Echo Path		Percent of Subjects' Votes in Each Category				
Delay (ms)	Loudness Loss (dB)	Excel.	Good	Fair	Poor	Unsat.
20	10	3.5	3.5	41.3	37.9	13.8
	15	0	10.3	48.3	37.9	3.5
	20	6.9	27.6	51.7	13.8	0
	25	41.4	48.3	10.3	0	0
	30	44.8	48.3	6.9	0	0
	35	51.6	44.9	0	3.5	0
36	16.5	3.5	3.5	24.0	55.2	13.8
	21.5	3.5	20.7	20.7	51.6	3.5
	26.5	10.3	58.6	27.6	3.5	0
	31.5	20.7	58.6	17.2	3.5	0
	36.5	41.3	51.7	3.5	3.5	0
	41.5	62.0	31.0	3.5	3.5	0
56	26	3.5	3.5	55.1	34.4	3.5
	31	10.3	17.3	51.7	20.7	0
	36	10.3	34.5	51.7	3.5	0
	41	24.1	48.3	27.6	0	0
	46	51.7	41.3	3.5	0	3.5
	51	55.2	41.3	0	0	3.5
72	29	3.5	6.9	37.9	51.7	0
	34	0	31.0	51.7	17.2	0
	39	0	48.3	48.2	3.5	0
	44	20.7	51.7	27.6	0	0
	49	41.4	37.9	20.7	0	0
	54	44.7	48.3	3.5	3.5	0
90	34	0	10.3	48.3	37.9	3.5
	39	6.9	31.0	48.3	10.3	3.5
	44	10.3	48.3	41.4	0	0
	49	24.1	58.7	10.3	6.9	0
	54	41.4	44.8	13.8	0	0
	59	55.2	41.3	3.5	0	0

pressors. Results for these conditions were used in deriving the echo model.

The test system, shown on Fig. 11, provided for two-way conversation between pairs of subjects. The two ends of the test system were located in separate, acoustically treated rooms for which the ambient room noise was about 38 dB(A). Each subject was able to hear his own voice by means of (*i*) a side tone path with loudness loss of 12 dB and (*ii*) an echo path.

One hundred pairs of subjects participated in the tests; 50 of these evaluated the 600-ms delay case, the other 50 evaluated the 1200-ms delay case. The two members of a pair (they were acquainted) were located in the two separate test rooms. Prior to a test session, they were instructed that they should discuss a subject of mutual interest

Table XII — 1968 laboratory echo-test results—30 ratings (subjects) per condition

Circuit Noise (dBrnC)	Echo Path		Percent of Subjects' Votes in Each Category				
	Delay (ms)	Loudness Loss (dB)	Excel.	Good	Fair	Poor	Unsat.
18.0	—	100	63.4	30.0	3.3	0	3.3
18.0	1.5	0.4	3.3	33.3	36.7	23.4	3.3
		5.4	26.7	43.3	23.3	6.7	0
		10.4	33.3	46.7	10.0	10.0	0
		15.4	60.0	26.7	13.3	0	0
		20.4	33.3	46.7	13.3	6.7	0
		25.4	63.3	36.7	0	0	0
18.0	20.0	5.5	0	3.3	3.3	23.4	70.0
		10.5	3.3	0	13.3	40.0	43.4
		15.5	0	10.0	36.7	23.3	30.0
		20.5	33.3	30.0	16.7	13.3	6.7
		25.5	36.7	36.7	23.3	3.3	0
		30.5	26.7	56.7	6.6	10.0	0
18.0	56.0	15.6	0	0	6.7	33.3	60.0
		20.6	0	0	10.0	36.7	53.3
		25.6	3.3	6.7	26.7	36.6	26.7
		30.6	3.3	20.0	40.0	33.4	3.3
		35.6	20.0	43.3	26.7	10.0	0
		40.6	33.3	46.7	16.7	0	3.3
18.0	90.0	24.5	0	0	6.7	40.0	53.3
		29.5	3.3	10.0	20.0	36.7	30.0
		34.5	3.3	13.3	43.3	36.8	3.3
		39.5	23.3	26.7	16.7	30.0	3.3
		44.5	30.0	40.0	23.4	3.3	3.3
		49.5	50.0	40.0	10.0	0	0
28.0	—	100	26.6	60.0	6.7	6.7	0
28.0	1.5	0.4	3.3	26.8	43.3	13.3	13.3
		5.4	10.0	43.3	40.0	6.7	0
		10.4	16.7	56.7	20.0	6.6	0
		15.4	16.7	56.7	16.7	6.6	3.3
		20.4	6.7	56.7	30.0	3.3	3.3
		25.4	23.3	60.0	13.4	0	3.3
28.0	10.0	3.0	0	3.4	23.3	33.3	40.0
		8.0	0	3.3	43.3	36.7	16.7
		13.0	3.3	20.0	33.3	26.7	16.7
		18.0	20.0	56.7	13.3	3.3	6.7
		23.0	3.3	46.7	43.3	6.7	0
		28.0	10.0	60.0	26.7	3.3	0
28.0	20.0	5.5	0	0	13.3	33.3	53.4
		10.5	3.3	6.7	33.3	26.7	30.0
		15.5	10.0	3.3	23.3	43.4	20.0
		20.5	3.3	23.4	60.0	10.0	3.3
		25.5	13.3	56.7	23.3	0	6.7
		30.5	10.0	56.7	23.3	6.7	3.3

Table XII — Continued

Circuit Noise (dBrnC)	Echo Path		Percent of Subjects' Votes in Each Category				
	Delay (ms)	Loudness Loss (dB)	Excel.	Good	Fair	Poor	Unsat.
28.0	36.0	11.0	0	0	10.0	16.7	73.3
		16.0	0	0	6.7	33.3	60.0
		21.0	0	0	26.7	50.0	23.3
		26.0	3.3	20.0	60.0	16.7	0
		31.0	6.7	50.0	33.3	6.7	3.3
		36.0	13.3	60.0	20.0	6.7	0
	56.0	15.6	0	3.3	10.0	20.0	66.7
		20.6	3.3	0	6.7	36.7	53.3
		25.6	0	3.3	40.0	30.0	26.7
		30.6	0	13.3	50.0	33.4	3.3
		35.6	3.3	46.7	36.7	13.3	0
		40.6	16.7	50.0	33.3	0	0
	72.0	21.5	3.3	3.3	6.7	33.3	53.4
		26.5	0	0	13.3	63.4	23.3
		31.5	3.3	3.3	33.4	40.0	20.0
		36.5	6.7	36.6	43.3	6.7	6.7
		41.5	6.7	36.6	46.7	10.0	0
		46.5	23.3	56.7	13.3	6.7	0
	90.0	24.5	0	0	3.3	33.3	63.4
		29.5	3.3	0	26.7	50.0	20.0
		34.5	6.7	23.4	33.3	33.3	3.3
		39.5	6.7	30.0	46.6	16.7	0
		44.5	3.3	46.7	43.3	6.7	0
		49.5	10.0	50.0	36.7	0	3.3
38.0	—	100	0	26.7	53.3	13.3	6.7
38.0	1.5	0.4	3.3	0	26.7	46.7	23.3
		5.4	0	3.3	30.0	53.4	13.3
		10.4	0	26.7	50.0	13.3	10.0
		15.4	3.3	26.7	50.0	20.0	0
		20.4	0	13.3	56.7	20.0	10.0
		25.4	6.7	10.0	40.0	40.0	3.3
	20.0	5.5	0	0	3.3	20.0	76.7
		10.5	0	0	3.3	36.7	60.0
		15.5	0	3.3	20.0	43.4	33.3
		20.5	0	6.7	43.3	43.3	6.7
		25.5	6.7	13.3	46.7	23.3	10.0
		30.5	0	6.7	60.0	20.0	13.3
	56.0	15.6	0	0	3.3	26.7	70.0
		20.6	0	0	3.3	46.7	50.0
		25.6	0	0	20.0	53.3	26.7
		30.6	0	0	53.4	43.3	3.3
		35.6	6.7	16.7	40.0	33.3	3.3
		40.6	0	13.3	46.7	30.0	10.0
	90.0	24.5	0	0	13.3	36.7	50.0
		29.5	0	0	16.7	53.3	30.0
		34.5	0	13.3	36.7	43.3	6.7
		39.5	0	23.3	40.0	36.7	0
		44.5	6.7	26.7	40.0	23.3	3.3
		49.5	3.3	23.3	56.7	10.0	6.7

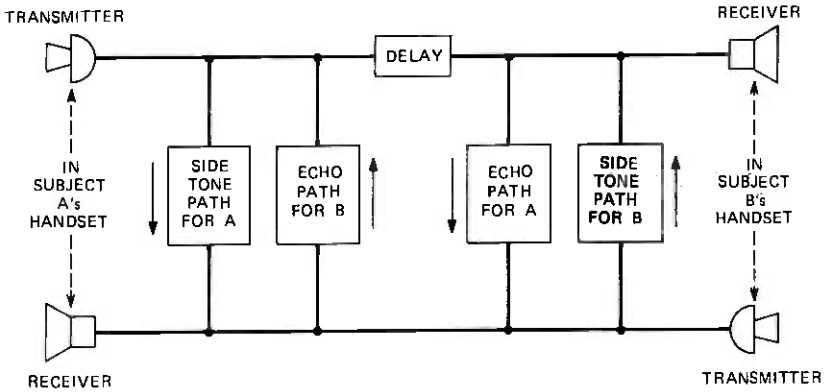


Fig. 11—Test system for the 1970 laboratory echo test.

over the system and told that each test condition would require about 4 minutes of conversation, at the end of which they should separately rate the condition on the 5-point scale: excellent, good, fair, poor, and bad. (These categories were further subdivided into undesignated thirds, resulting in a 15-point scale.)

The test conditions and results are given in Table XIII and are also reported in Part III of Annex 5 to Question 6/XII in Ref. 19. As with tests discussed earlier, these results show that (*i*) at a given echo-path

Table XIII — 1970 laboratory echo-test results—Approximately 100 ratings (subjects) per condition

Connection loudness loss = 18 dB
Side tone-path loudness loss = 12 dB
Circuit noise = 33 dBrnC

Echo Path		Percent of Subjects' Votes in Each Category				
Delay (ms)	Loudness Loss (dB)	Excel.	Good	Fair	Poor	Unsat.
—	33.0	9.0	49.0	34.0	8.0	0.0
600	33.0	1.0	1.0	2.0	17.6	78.4
	43.0	2.0	10.8	19.6	39.2	28.4
	53.0	3.9	36.3	44.1	11.8	3.9
	63.0	6.9	47.1	28.4	16.6	1.0
	73.0	6.0	44.0	42.0	6.0	2.0
1200	43.0	1.0	5.0	19.0	38.0	37.0
	53.0	6.0	29.0	40.0	19.0	6.0
	63.0	4.0	45.0	36.0	14.0	1.0
	73.0	7.0	39.0	39.0	13.0	2.0

delay, transmission quality improves with increasing echo-path loudness loss, and (*ii*) at a given echo-path loudness loss, transmission quality is degraded with increasing echo-path delay.

4.4 1970 SIBYL echo test

An echo test was conducted in early 1970 using SIBYL. The purpose of the test was to enable comparison of results obtained from subjects conversing on actual telephone calls to results obtained under laboratory conditions. The results of tests reported in Section 4.2 were used to guide selection of the conditions for the 1970 SIBYL echo test.

Forty-five subjects participated in these tests. Procedures followed by the subjects were the same as those for the loss-noise tests reported in Section 3.3.

Test variables were echo-path delay (three values) and echo-path loudness loss (five values for each delay). In addition, a condition without echo was included as a reference.

Connection loudness loss was 10 dB, the base condition for the Holmdel SIBYL tests reported in Sections 3.2 and 3.3. Circuit noise was 30 dBrnC. Side tone path loudness loss was about 12 dB and average room noise was estimated to be about 42 dB(A).

The test conditions and results are given in Table XIV. As with results of echo tests discussed in preceding sections, these results show that transmission quality is strongly dependent on echo-path delay and echo-path loudness loss.

V. ANALYSIS OF INDIVIDUAL TEST DATA

The raw test results from any individual test provide subjective-opinion information expressible in the form of percent of ratings in each of the five rating categories for each test condition. The utilization of these raw test results in this form for transmission planning is difficult because it is usually necessary to have ratings available for transmission parameter values not specifically included in the tests. Thus, some form of data analysis is frequently applied to obtain graphical or analytical representations of the data that are more convenient for use in transmission planning studies. This can involve simple curve fitting to the raw test data[5,6] or more elaborate models of subject ratings using binomial or other distributions.[20]

For example, Lewinski in Ref. 5 provides separate "smooth" fits for the percentage of responses in the categories excellent, good or better, fair or better, and poor or better as a function of circuit noise. Similarly Sen in Ref. 6 provides mathematical expressions and contours for the percent good or better and percent poor or worse as a function of connection loss and noise. A different approach is suggested by

Table XIV — 1970 SIBYL echo test

Connection loudness loss = 10 dB

Circuit Noise = 30 dBrnC

Echo Path		Number of Ratings	Percent of Subjects' Votes in Each Category				
Delay (ms)	Loudness Loss (dB)		Excel.	Good	Fair	Poor	Unsat.
—	∞	183	32.2	36.6	24.6	5.5	1.1
10.0	5.8	51	0	15.7	29.4	47.1	7.8
	10.8	23	0	8.7	30.4	47.8	13.1
	15.8	38	15.8	21.0	39.5	21.0	2.7
	20.8	45	8.9	53.3	20.0	13.3	4.5
	25.8	41	19.5	51.2	29.3	0	0
36.0	14.4	38	0	2.6	18.4	50.0	29.0
	19.4	37	2.7	0	29.7	37.9	29.7
	24.4	51	11.8	35.3	23.5	23.5	5.9
	29.4	31	6.4	35.6	25.8	25.8	6.4
	34.4	46	19.6	36.9	28.3	13.0	2.2
72.0	22.4	35	0	2.9	2.9	65.7	28.5
	27.4	20	10.0	0	15.0	60.0	15.0
	32.4	34	5.9	8.8	29.4	38.2	17.7
	37.4	33	12.1	24.2	30.4	24.2	9.1
	42.4	46	13.1	47.8	30.4	6.5	2.2

Prosser, Allnatt, and Lewis in Ref. 20 where they point out that five separate mathematical functions can be specified, one for each grade on the rating scale, but advocate the desirability of a more convenient and compact representation by means of a single mathematical model that embraces all five functions. They examined various models based on the binomial distributions as well as logistic and gaussian curves. They adopted the second-order binomial as the simplest adequate model to describe the opinion distribution found in their experiment.

We also recognized the advantages of a single mathematical model to represent the distribution of opinion in the five rating categories. The normal density curve was selected as a basis for the model described in the following sections because it provided somewhat greater flexibility in accommodating a variety of standard deviations. Because of the availability of digital computers for the data analysis, the additional computational complexity associated with the normal distribution was not judged to be a problem.

5.1 Analysis method

The subjective-test results for each test condition of connection loudness loss, L_e, and circuit noise, N, form a vote histogram containing

the proportion, P_i, of ratings for each of the rating categories, $i = 1$, 2, 3, 4, 5. Rating category 1 represents the unsatisfactory category, 2 represents poor, 3 represents fair, 4 represents good, and 5 represents excellent.

The P_i's for each test condition sum to unity.

The values of P_i for each test condition were used to calculate the mean opinion score (MOS) and sample standard deviation (SIGMOS) as in eqs. (4) and (5), respectively.

$$\text{MOS} = \sum_{i=1}^{5} iP_i \tag{4}$$

$$\text{SIGMOS} = \left[\sum_{i=1}^{5} i^2 P_i - (\text{MOS})^2 \right]^{\frac{1}{2}}. \tag{5}$$

The vote histogram was represented by a normal density curve with mean, μ, and standard deviation, σ. The area under this curve was divided into five regions, each with area $\hat{P}_i$. The areas were defined as follows: from minus infinity to 1.5 as $\hat{P}_1$, from 1.5 to 2.5 as $\hat{P}_2$, from 2.5 to 3.5 as $\hat{P}_3$, from 3.5 to 4.5 as $\hat{P}_4$, and from 4.5 to infinity as $\hat{P}_5$; the $\hat{P}_i$'s sum to unity. This quantization of the area under the normal curve into five discrete regions was the basis for using the normal curve

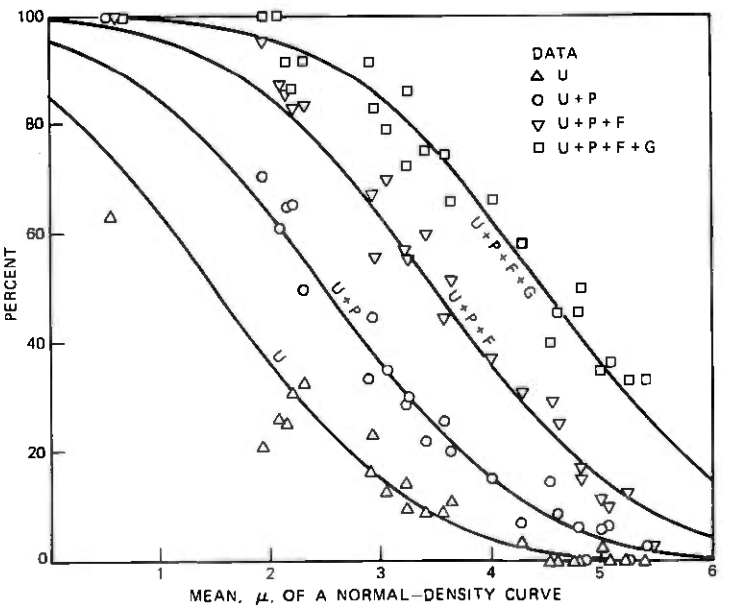


Fig. 12—MH loss-noise test data percentages as a function of the Step 3 normal-density means compared with percentages predicted using the MH standard deviation, 1.44.

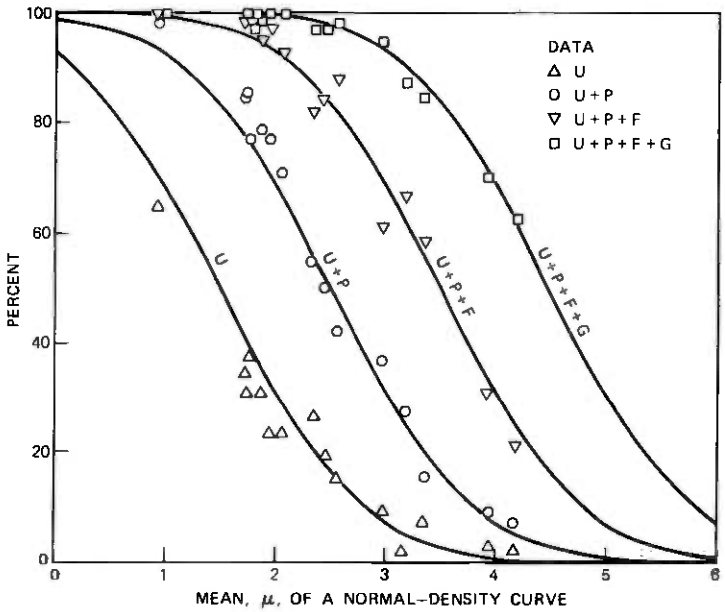


Fig. 13—HO1 loss-noise test data percentages as a function of the Step 3 normal-density means compared with percentages predicted using the HO1 standard deviation, 1.02.

to represent this type of data. These $\hat{P}_i$'s were used to compute the pseudo mean opinion scores (MOS_Q) and sample standard deviations (SIGMOS_Q) in terms of the mean, μ, and standard deviation, σ, of the normal curve as given in eqs. (6) and (7), respectively.

$$\text{MOS}_Q = \sum_{i=1}^{5} i\hat{P}_i = 5 - \sum_{j=1}^{4}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(j+0.5-\mu)/\sigma}\exp\left(-\frac{t^2}{2}\right)dt\right]. \quad (6)$$

$$\begin{aligned}\text{SIGMOS}_Q &= \left[\sum_{i=1}^{5} i^2\hat{P}_i - (\text{MOS}_Q)^2\right]^{\frac{1}{2}} \\ &= \left\{25 - \sum_{j=1}^{4}\left[\frac{2j+1}{\sqrt{2\pi}}\int_{-\infty}^{(j+0.5-\mu)/\sigma}\exp\left(-\frac{t^2}{2}\right)dt\right] - (\text{MOS}_Q)^2\right\}^{\frac{1}{2}}. \quad (7)\end{aligned}$$

Step 1 in the four-step analysis was to find a normal density curve for each test condition such that MOS = MOS_Q and SIGMOS = SIGMOS_Q. These two constraints were used in an iterative computer procedure to determine the values of the mean, μ, and standard deviation, σ, of the normal density curve used to represent the data for each test condition.

Three other criteria were considered in the determination of a normal density curve during the development of this analysis procedure. One criterion was to determine μ and σ such that the proportions good or better ($P_4 + P_5$) and poor or worse ($P_1 + P_2$) were the same for the data and the normal curve. Another was to determine μ and σ to minimize the sum of the squares of the differences between P_1 and $\hat{P}_1$, $P_1 + P_2$ and $\hat{P}_1 + \hat{P}_2$, $P_1 + P_2 + P_3$ and $\hat{P}_1 + \hat{P}_2 + \hat{P}_3$, and finally $P_1 + P_2 + P_3 + P_4$ and $\hat{P}_1 + \hat{P}_2 + \hat{P}_3 + \hat{P}_4$. The third criterion was to determine μ and σ to minimize the sum of the square of the differences $P_i - \hat{P}_i$, $i = 1, 2, \cdots, 5$. The selection of MOS = MOS_Q and SIGMOS = SIGMOS_Q was chosen as the general criterion of fit, since the other criteria were found to be more sensitive to the experimental variability inherent in subjective data of this type, particularly when the number of subjects used is small (less than about 100). This selection was made after applying the several analysis criteria to several hundred sets of test data generated by Monte Carlo simulation.

After all the test conditions were represented by a normal density curve with mean, μ, and standard deviation, σ, an average of the standard deviations for all of the test conditions was computed in Step 2. In determining this average, the individual σ's were weighted in accordance with the number of votes per condition and by the weight-

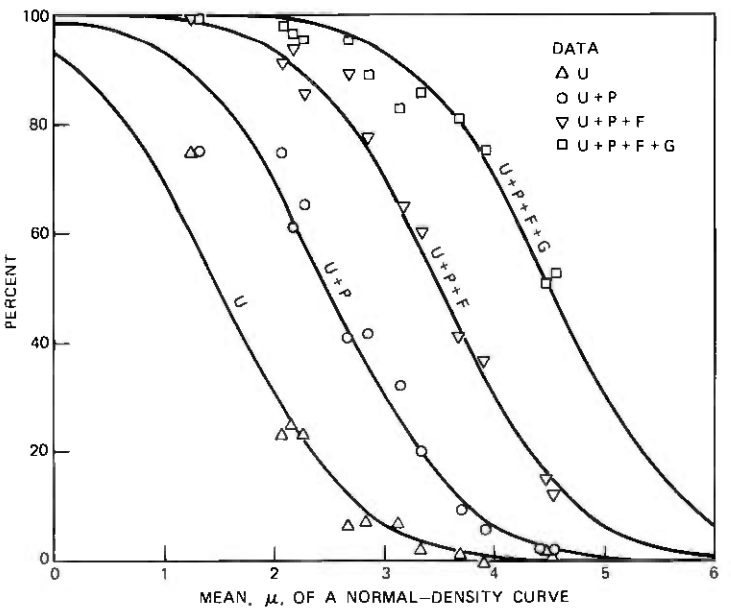


Fig. 14—HO2 loss-noise test percentage as a function of the Step 3 normal-density means compared with percentages predicted using HO2 standard deviation, 0.998.

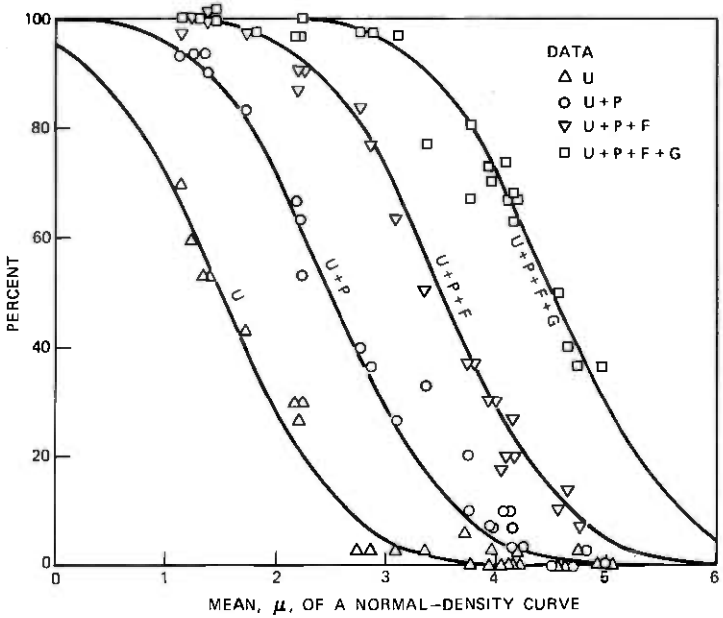


Fig. 15—1968 laboratory echo data percentages (noise=18 dBrnC) as a function of the Step 3 normal-density means compared with percentages predicted using the echo standard deviation, 0.888.

ing given in eq. (8). This latter weighting was based on the analyses of the Monte Carlo data mentioned previously by examining the variations in the standard deviations as a function of the fit mean, μ.

$$\text{Weighting} = \left[\frac{1}{1 + \frac{(\mu - 3)^2}{3}}\right]^2. \tag{8}$$

The weighted average standard deviation, σ_a, was then used in Step 3 as the standard deviation for all of the normal density curves. Using σ_a, a new mean, μ_a, was computed for each test condition subject to the constraint, MOS = MOS_Q. The end result was a family of normal density curves, all with the same standard deviation and with means as determined above.

Figures 12 to 19 illustrate the results from the first three steps in the procedure for several of the tests described in this paper. In these figures the cumulative percent of ratings in four categories—unsatisfactory, unsatisfactory plus poor, unsatisfactory plus poor plus fair, unsatisfactory plus poor plus fair plus good—are plotted against the fit mean, μ_a, determined in Step 3. The solid curves are plotted using the weighted average standard deviation, σ_a, from Step 2. Also shown

are the raw data plotted against the respective fit mean, μ_a, for each condition. These figures show that the normal density curves defined by the values, μ_a, and an average standard deviation, σ_a, provide a convenient and simple representation of the raw data for any single test.

In Step 4, the means of all the normal curves are fitted by a suitable analytical function of the test parameters using a least-squares-fit technique.

In summary, the steps involved in the analysis procedure can be described as follows:

Step 1. A normal density curve is used to represent the vote histogram for each test condition such that MOS = MOS_Q and SIGMOS = SIGMOS_Q.

Step 2. The standard deviations of the normal curves of Step 1 for all test conditions under consideration are weighted and averaged to obtain a single value for the standard deviation, σ_a.

Step 3. The single value of standard deviation from Step 2 is used for each test condition as the standard deviation of the

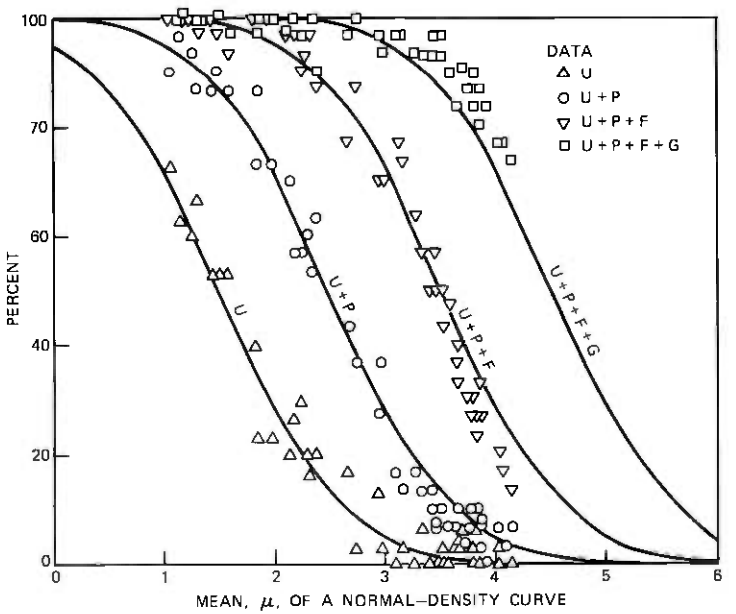


Fig. 16—1968 laboratory echo data percentages (noise=28 dBrnC) as a function of the Step 3 normal-density means compared with percentages predicted using the echo standard deviation, 0.888.

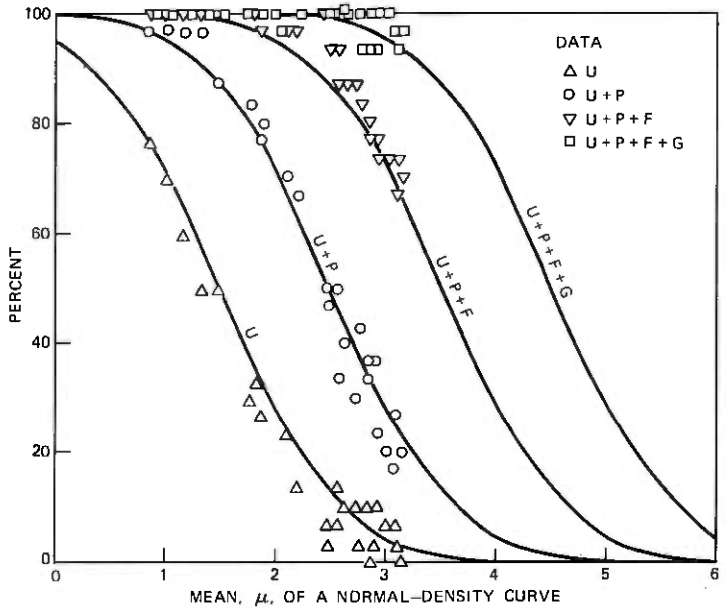


Fig. 17—1968 laboratory echo data percentages (noise=38 dBrnC) as a function of the Step 3 normal-density means compared with percentages predicted using the echo standard deviation, 0.888.

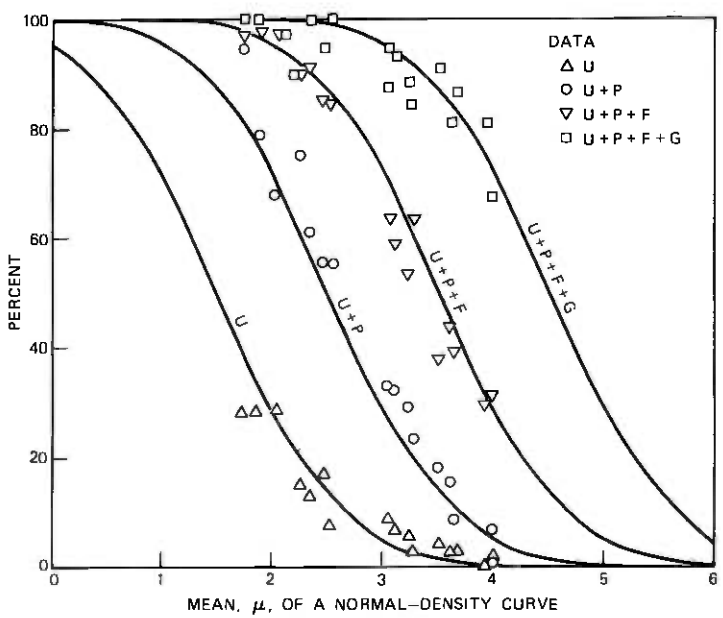


Fig. 18—SIBYL echo data percentages as a function of the Step 3 normal-density means compared with percentages predicted using the echo standard deviation, 0.888.

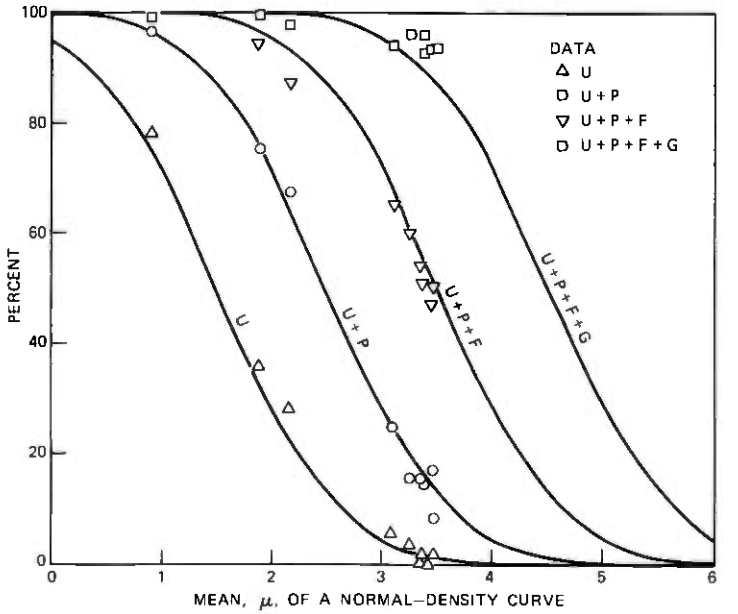


Fig. 19—1970 laboratory echo data percentages as a function of the Step 3 normal-density means compared with percentages predicted using the echo standard deviation, 0.888.

corresponding normal curve, and the mean of the normal curve, μ_a, is adjusted such that MOS = MOS$_Q$.

Step 4. The means of all normal curves of Step 3 are fitted, on a least-squares-error basis, to an appropriate function of the test parameters.

The results at each step of this analysis are summarized in Table XV for HO2 loss/noise data.

In this table, the results presented for Step 4 are based on the analytical function [eq. (11) in the next section] fitted in Step 4. Comparison of the entries for Step 4 with the entries for the previous steps illustrate the extent to which this analytical function provides a good fit to the test results. For this test, all of the fit means in Step 4 are within approximately 0.2 of the fit means in Step 3. The individual differences have a mean of 0.035 and a standard deviation of 0.13. This agreement is considered reasonable in view of the average standard deviation of 0.998 and about 50 to 70 ratings per condition. Somewhat larger differences were obtained in the MH and HO1 tests where the number of ratings per test condition was smaller.

5.2 Loss-noise analysis

The results obtained by applying the analysis method to the data from the three loss/noise tests are given in eqs. (9), (10), and (11),

Table XV—Summary of results for HO2 loss-noise test at each step in the analysis procedure

Step in the Fit Process	Percentage of Votes					Mean Opinion Score*	Standard Deviation*	Fit Mean	Fit Sigma
	Excel.	Good	Fair	Poor	Unsat.				
Test Condition 1, Loss = 5 dB, Noise = 25 dBrnC									
Raw Data	49.49	35.35	13.13	0.00	2.02	4.303	0.846	—	—
Step 1	51.54	30.94	14.15	3.05	0.32	4.303	0.846	4.543	1.117
Step 2	51.71	33.48	12.77	1.92	0.11	4.348	0.782	4.543	0.998
Step 3	49.16	34.51	13.96	2.23	0.14	4.303	0.801	4.479	0.998
Step 4	50.00	34.18	13.56	2.12	0.13	4.318	0.795	4.500	0.998
Test Condition 2, Loss = 10 dB, Noise = 25 dBrnC									
Raw Data	47.23	40.62	10.20	1.26	0.68	4.325	0.762	—	—
Step 1	48.79	36.59	12.98	1.57	0.06	4.325	0.762	4.472	0.923
Step 2	48.88	34.61	14.10	2.26	0.15	4.298	0.803	4.472	0.998
Step 3	50.40	34.02	13.38	2.07	0.13	4.325	0.792	4.510	0.998
Step 4	50.00	34.18	13.56	2.12	0.13	4.318	0.795	4.500	0.998
Test Condition 3, Loss = 20 dB, Noise = 25 dBrnC									
Raw Data	14.00	26.00	40.00	18.00	2.00	3.320	0.989	—	—
Step 1	11.81	31.46	36.85	16.75	3.14	3.320	0.989	3.333	0.985
Step 2	12.11	31.24	36.45	16.88	3.31	3.320	0.998	3.333	0.998
Step 3	12.11	31.24	36.45	16.88	3.31	3.320	0.998	3.333	0.998
Step 4	12.26	31.38	36.37	16.73	3.26	3.326	0.998	3.340	0.998
Test Condition 4, Loss = 30 dB, Noise = 25 dBrnC									
Raw Data	3.23	3.23	32.26	35.48	25.81	2.226	0.974	—	—
Step 1	1.26	8.78	27.39	36.40	26.16	2.226	0.974	2.166	1.043
Step 2	0.97	8.10	27.83	37.88	25.23	2.217	0.946	2.166	0.998
Step 3	0.99	8.24	28.04	37.82	24.91	2.226	0.947	2.176	0.998
Step 4	1.03	8.43	28.34	37.73	24.47	2.238	0.949	2.190	0.998
Test condition 5, Loss = 5 dB, Noise = 32 dBrnC									
Raw Data	27.14	35.71	31.43	5.71	0.00	3.843	0.889	—	—
Step 1	25.24	40.95	27.16	6.17	0.48	3.843	0.889	3.885	0.922
Step 2	26.89	38.13	26.72	7.42	0.84	3.828	0.938	3.885	0.998
Step 3	27.45	38.19	26.35	7.20	0.80	3.843	0.935	3.902	0.998
Step 4	25.42	37.91	27.70	8.02	0.95	3.788	0.945	3.840	0.998
Test condition 6, Loss = 10 dB, Noise = 32 dBrnC									
Raw Data	18.97	39.66	32.76	6.90	1.72	3.672	0.917	—	—
Step 1	19.46	39.06	31.65	8.93	0.89	3.672	0.917	3.700	0.929
Step 2	21.14	36.80	30.60	10.09	1.37	3.662	0.965	3.700	0.998
Step 3	21.46	36.91	30.38	9.91	1.34	3.672	0.963	3.711	0.998
Step 4	27.72	38.22	26.17	7.10	0.79	3.850	0.933	3.910	0.998
Test condition 7, Loss = 20 dB, Noise = 32 dBrnC									
Raw Data	4.69	6.25	48.44	34.38	6.25	2.688	0.864	—	—
Step 1	1.40	14.76	42.65	33.60	7.59	2.688	0.864	2.684	0.826
Step 2	3.44	17.24	36.64	30.91	11.77	2.697	0.999	2.684	0.998
Step 3	3.37	17.03	36.53	31.11	11.97	2.687	0.999	2.674	0.998
Step 4	5.44	21.94	38.19	26.39	8.03	2.904	1.007	2.900	0.998

Table XV — Continued

Step in the Fit Process	Percentage of Votes					Mean Opinion Score*	Standard Deviation*	Fit Mean	Fit Sigma
	Excel.	Good	Fair	Poor	Unsat.				
Test condition 8, Loss = 30 dB, Noise = 32 dBrnC									
Raw Data	1.92	7.69	15.38	51.92	23.08	2.135	0.920	—	—
Step 1	0.67	6.66	25.95	38.87	27.85	2.135	0.920	2.076	0.981
Step 2	0.76	6.92	25.87	38.26	28.19	2.138	0.931	2.076	0.998
Step 3	0.75	6.87	25.78	38.27	28.33	2.135	0.930	2.072	0.998
Step 4	0.45	4.89	21.72	38.15	34.80	1.980	0.895	1.890	0.998
Test condition 9, Loss = 5 dB, Noise = 42 dBrnC									
Raw Data	11.32	11.32	35.85	33.96	7.55	2.849	1.088	—	—
Step 1	6.83	20.85	34.25	26.55	11.51	2.849	1.088	2.839	1.116
Step 2	4.80	20.59	37.91	27.72	8.98	2.845	1.005	2.839	0.998
Step 3	4.84	20.67	37.93	27.63	8.92	2.849	1.005	2.843	0.998
Step 4	5.23	21.50	38.11	26.83	8.34	2.884	1.006	2.880	0.998
Test condition 10, Loss = 10 dB, Noise = 42 dBrnC									
Raw Data	17.53	17.53	32.99	24.74	7.22	3.134	1.181	—	—
Step 1	14.29	24.81	30.53	20.76	9.60	3.134	1.181	3.150	1.265
Step 2	8.81	27.48	37.97	20.83	4.91	3.144	1.005	3.150	0.998
Step 3	8.63	27.24	38.02	21.07	5.03	3.134	1.006	3.139	0.998
Step 4	7.45	25.51	38.30	22.83	5.90	3.058	1.007	3.060	0.998
Test condition 11, Loss = 20 dB, Noise = 42 dBrnC									
Raw Data	4.35	10.87	19.57	41.30	23.91	2.304	1.081	—	—
Step 1	2.92	11.45	26.35	31.67	27.61	2.304	1.081	2.217	1.206
Step 2	1.11	8.82	28.91	37.54	23.62	2.262	0.953	2.217	0.998
Step 3	1.25	9.51	29.86	37.16	22.23	2.304	0.960	2.263	0.998
Step 4	1.24	9.46	29.79	37.18	22.32	2.301	0.959	2.260	0.998
Test condition 12, Loss = 30 dB, Noise = 42 dBrnC									
Raw Data	0.00	0.00	25.00	0.00	75.00	1.500	0.866	—	—
Step 1	0.98	3.30	9.20	17.76	68.75	1.500	0.866	0.705	1.626
Step 2	0.01	0.25	3.35	17.68	78.71	1.252	0.521	0.705	0.998
Step 3	0.05	1.06	8.84	28.93	61.12	1.500	0.705	1.218	0.998
Step 4	0.12	1.98	13.00	33.70	51.20	1.661	0.786	1.470	0.998

* Mean opinion score and standard deviation are calculated from the percentage of votes given in each table for the corresponding step in the fit process.

respectively, for the MH, HO1, and HO2 tests.

$$\mu_{MH} = 11.54 - 0.1099|L_e - 11.7| - 0.168N_1 - 0.001059L_eN_1$$
$$\sigma = 1.44, \tag{9}$$

$$\mu_{HO1} = 7 - 0.1365|L_e - 10.31| - 0.1219N_2 + 0.001577L_eN_2$$
$$\sigma = 1.02, \tag{10}$$

$$\mu_{HO2} = 7.17 - 0.1681|L_e - 6.7| - 0.1058N_3 + 0.002106L_eN_3$$
$$\sigma = 0.998, \tag{11}$$

where

L_e = Acoustic-to-acoustic loudness loss (in dB) of an overall telephone connection, determined using the Electro-Acoustic Rating System (EARS) method.

N = Circuit noise (in dBrnC) at the input to a set with a receive-loudness rating of 26 dB, determined using the EARS method.

N_1 = Total noise in dBrnC resulting from power addition of the circuit noise, N, from the MH tests with 34.03 dBrnC.

N_2 = Total noise in dBrnC resulting from power addition of the circuit noise, N, from the HO1 tests with 23.76 dBrnC.

N_3 = Noise, N from the HO2 tests.

The values 34.03 and 23.76 were determined as fit parameters. The particular functional form was selected to provide as simple a model as possible of the systematic effects observed in the data.

The results represented by eqs. (9), (10), and (11) revealed two important differences between the MH and HO tests. First, the standard deviation, σ, was considerably larger for the MH tests than for either HO1 or HO2. Second, the subjective opinions, as represented by the means, μ, calculated from eqs. (9), (10), and (11) were considerably higher in the MH test compared with the HO tests. These differences occurred despite the similarities of the tests. A careful examination of either the raw data or the smooth results clearly shows that the subjects' ratings tended to be more critical in the two HO tests compared with the subjects in the MH test.

A clearer picture of the differences is obtained by selecting a set of loss (L_e) and circuit noise (N) values over a common range of the tests for MH, HO1, and HO2 and computing the corresponding values of the means (μ) for the three tests from eqs. (9), (10), and (11). If plots are made of the MH means versus both the HO1 and HO2 means and the appropriate linear regression made for both plots, then eqs. (12) and (13), respectively, represent the regression line between the MH and HO1 means and the MH and HO2 means. Such a plot is shown in Fig. 20 for MH and HO1.

$$\mu_{MH} = 1.372\mu_{HO1} - 0.206. \tag{12}$$

$$\mu_{MH} = 1.288\mu_{HO2} - 0.215. \tag{13}$$

The Pearson product moment coefficient of correlation was found to be 0.9586 and 0.9693 for MH with HO1 and HO2, respectively. Eqs. (12) and (13) clearly show the difference between the means for the MH and the two HO tests. These equations also show the close agreement between the two Holmdel tests.

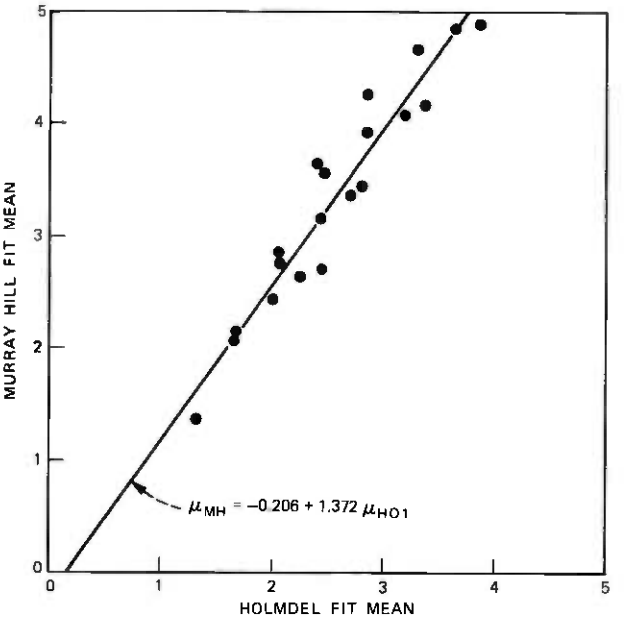


Fig. 20—Comparison of 1965 Murray Hill and 1969 Holmdel 1 loss-noise results.

Equations (12) and (13) are used to adjust for the difference between the HO results and the MH results in Section VI of this paper.

5.3 Echo analysis

The four echo subjective tests yielded data on the subjective effects of echo-path loudness loss, E, in dB and echo-path delay, D, in ms. Circuit noise, N, was not the same for all four tests, and needed to be considered as a test variable. Loudness loss, L_e, was a factor in only two of these tests (the other two were listening only tests) and, as a first approximation, it was decided to ignore L_e and concentrate on E, D, and N in the analysis.

Preliminary analyses of the individual test data indicated that there were only relatively small differences in the absolute ratings among the 1968–1970 tests. Thus, it was feasible to combine these data and use the analysis method described previously. The resulting equation relating the normal density means to the test variables, as realized at Step 4 in the analysis, was a function of E, D, and N, where N for this preliminary analysis was the actual noise at the telephone set terminals. The effect of noise was asymptotic. That is, the fit mean, μ, was determined largely by the values of E and D when echo was the predominant impairment, while, for any value of D, increasing the value of E gradually led to the mean being solely determined by the

value of N. This asymptotic effect of the noise made it relatively simple to separate the effects of noise and echo in the function depicting the mean as a function of the test variables. The resulting functions are given in eq. (14).

$$\begin{aligned} \mu_{E1} &= 4.74 - 3.32 \log_{10} [(1 + D)/\sqrt{1 + (D/480)^2}] + 0.1414E \\ \mu_N &= 6.38 - 0.094N \\ \mu_{NE} &= \frac{\mu_E + \mu_N}{2} - \sqrt{\left(\frac{\mu_E - \mu_N}{2}\right)^2 + (0.627)^2} \\ \sigma &= 0.888. \end{aligned} \tag{14}$$

Although these functions were a useful interim result, subsequent analysis described in the later sections of this paper indicated that when the effects of both loss and noise were included and the noise was referred to the input of a reference set, further modifications were necessary. These functions are included here because they provided a basis for the subsequent modifications.

VI. COMBINATION OF MODELS

The results of the three loss-noise tests as given by eqs. (9), (10), and (11) showed fundamental differences among the tests. Despite the similarity of the tests and the general character of the results, both the raw data and the smoothed results showed that the subjects' ratings in the two HO tests tended to be more critical in their evaluations than they were in the MH tests. This could have occurred because of one or more differences in the tests, such as room noise, sidetone path loss, year of test, or some fundamental difference in the attitude of the subject groups. The frequent repetition of high-quality conditions in the Holmdel tests may also have been a factor.

The exact reasons for the differences in the test results could not be determined. Because of these differences, direct pooling of results from the three tests did not appear to be justifiable. However, the test results were combined by adjusting the HO1 and HO2 results to a MH base using the linear transformation obtained from the linear regressions introduced previously in eqs. (12) and (13). In this way, the systematic differences among the test results were preserved, while achieving the advantages of a larger data base. The transformed means were then included with the MH means, and a new equation was obtained by applying Step 4 to this combined set of means. Table XVI shows the fit means at Step 3 for the MH, HO1, and HO2 test results and the adjusted fit means for the HO1 and the HO2 test results. Thus, the final fit was based on a total of 51 conditions. The final result

Table XVI — Data for combined fit—Step 3

Murray Hill			Holmdel 1				Holmdel 2			
Connection Loudness Loss (dB)	Circuit Noise (dBrnC)	Fit Mean (μ_a)	Connection Loudness Loss (dB)	Circuit Noise (dBrnC)	Fit Mean (μ_a)	Adjusted Fit Mean (μ_a')	Connection Loudness Loss (dB)	Circuit Noise (dBrnC)	Fit Mean (μ_a)	Adjusted Fit Mean (μ_a')
5	21	4.53	—	—	—	—	5	25	4.48	5.56
10	21	5.08	10	22	4.18	5.53	10	25	4.51	5.59
15	21	5.23	15	22	3.94	5.20	—	—	—	—
20	21	4.27	20	22	3.35	4.39	20	25	3.33	4.07
25	21	3.62	25	22	2.36	3.03	—	—	—	—
30	21	2.90	30	22	2.46	3.17	30	25	2.18	2.59
5	28	4.80					5	32	3.90	4.81
10	28	5.39					10	32	3.71	4.56
15	28	4.80					—	—	—	—
20	28	3.56					20	32	2.67	3.22
25	28	3.41					—	—	—	—
30	28	2.29					30	32	2.07	2.45
5	36	4.59	—	—	—	—				
10	36	5.01	10	35	3.17	4.14				
15	36	4.01	—	—	—	—				
20	36	3.05	20	35	2.54	3.28				
25	36	2.20	25	35	1.94	2.46				
30	36	2.09	30	35	1.86	2.35				
			10	40	2.97	3.87				
			20	40	1.73	2.17				
			25	40	1.72	2.15				
5	44	3.24	—	—	—	—	5	42	2.84	3.44
10	44	2.93	10	45	2.06	2.62	10	42	3.14	3.83
15	44	3.21	—	—	—	—	—	—	—	—
20	44	2.16	20	45	1.76	2.21	20	42	2.26	2.70
25	44	1.91	—	—	—	—	—	—	—	—
30	44	0.54	30	45	0.93	1.07	30	42	1.22	1.36

is eq. (15).

$$\mu_{MH} = 10.36 - 0.185\sqrt{(L_e - 7.2)^2 + 1} - 0.1647N_F + 0.00167L_E N_F, \quad (15)$$

where

N_F = Power addition of noise with 27.37 dBrnC.

6.1 General rating scale

Equation (15) above is calculated in terms of the MH base. This equation together with eqs. (12) and (13) can be used to express the representation of the subjective ratings in terms of any one of the other test bases, HO1 or HO2. Each test base has a standard deviation associated with it that can then be used in conjunction with the computed means to calculate predicted vote histograms from the normal density curves. However, if this is done, the fit means and vote histograms will be different for each test in accordance with the difference in absolute ratings obtained for each test. To eliminate the need for three separate equations, one for each test, a general transmission-rating scale was established.

This transmission rating scale, referred to as the R-scale, is simply a linear transformation of the normal density means, defined by eq. (15), with the constraints that two preselected transmission conditions are to be the anchor conditions for the transformation. R-scale values of 80 and 40, respectively, were selected for the transmission conditions $L_e = 15$, $N = 25$, and $L_e = 30$, $N = 40$. These two transmission conditions were selected to be well separated in quality. The first pair is typical of a short intertoll connection and the latter represents an extreme condition of loss and noise that should rarely occur even on long intertoll connections between long loops.

Using the above transmission conditions as anchors, R-scale values can be specified in terms of μ for each test through the linear transformation $R = a + b\mu$, with a and b determined from the anchor constraints.

6.2 Loss-noise model

From eq. (15), the transmission condition $L_E = 15$, $N = 25$ yields $\mu_{MH} = 4.806$ and $L_e = 30$, $N = 40$ yields $\mu_{MH} = 1.528$. Using these values of μ, respectively, with R-scale values of 80 and 40 determines the transformation to the R-scale from the μ scale as given in eq. (16).

$$R = 21.37 + 12.20\mu_{MH}. \quad (16)$$

Substituting eq. (15) for μ_{MH} into eq. (16) gives eq. (17), which is

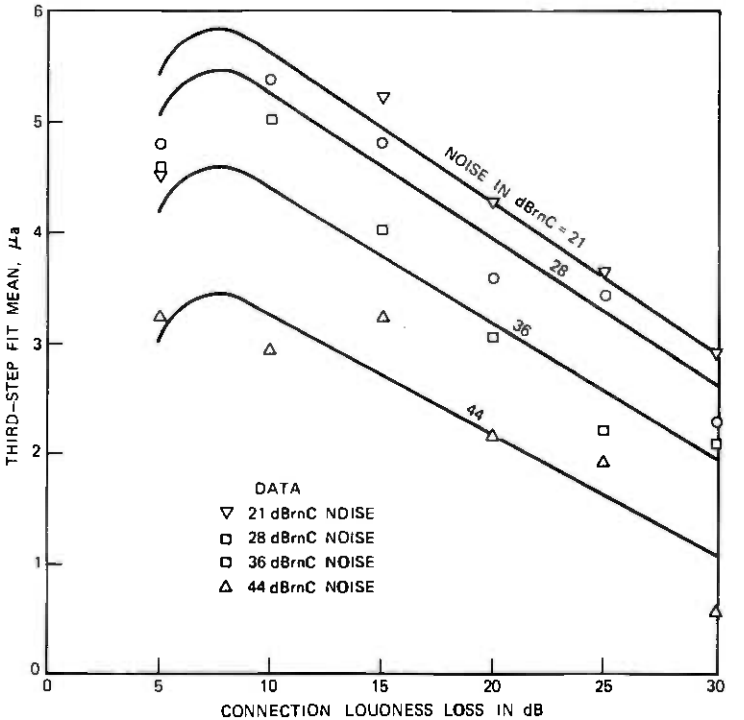


Fig. 21—MH loss-noise test, Step 3 means from data as a function of loudness loss and noise compared with means predicted from the loss-noise model at the MH base.

the R-scale representation of the subjective opinion for loss and noise.

$$R_{LN} = 147.76 - 2.257\sqrt{(L_e - 7.2)^2 + 1} - 2.009N_F + 0.02037L_eN_F. \tag{17}$$

Equation (17) is plotted in Fig. 1 as transmission rating versus L_e for a selected set of values of N. These curves represent the predicted transmission rating, in terms of the R-scale, for selected values of L_e and N.

The R-scale result of eq. (17) can also be used in conjunction with the appropriate standard deviations associated with eqs. (9), (10), and (11) and the appropriate inverse linear regression lines of eqs. (12) and (13) to calculate percent good or better, poor or worse, or other characteristics at the chosen test base.

For the μ-scale, the proportion of ratings good or better is the integral of the standard normal density curve from $(3.5 - \mu)/\sigma$ to infinity. In the R-scale, this corresponds to the integral of the standard normal density curve from minus infinity to $[R - (a + 3.5b)]/\sigma b$.

Similar computations can be made for proportion poor or worse, or for proportions of ratings in any of the five categories. The appropriate limits of integration to compute the proportion of ratings good or better and poor or worse are given in Table II for the three loss-noise test bases (MH, HO1, HO2).

The above discussion concerning the relationships between the proportions good or better and poor or worse and the R-scale lead to the plots of Fig. 4 which show these relationships for the three test bases.

Finally, the results summarized in Table II were also used to generate curves showing the tradeoff between L_e and N for selected values of percent good or better and percent poor or worse as shown in Figs. 6 and 7, respectively. As noted on the figures, these results correspond to the MH data base.

In Figs. 21 to 23, the third-step fit means for the individual tests are plotted as a function of loudness loss with circuit noise as a parameter. The solid lines in the figure correspond to the values of transmission rating in the final model transformed by the appropriate relation between μ and R for each test. These figures show a generally good fit to the individual test results for each of the tests.

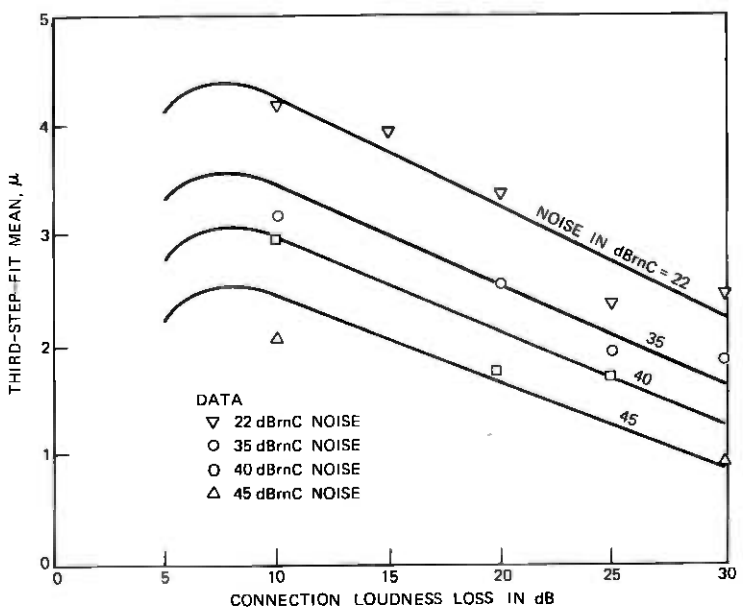


Fig. 22—HO1 loss-noise test, Step 3 means from data as a function of loudness loss and noise compared with means predicted from the loss-noise model at the HO1 base.

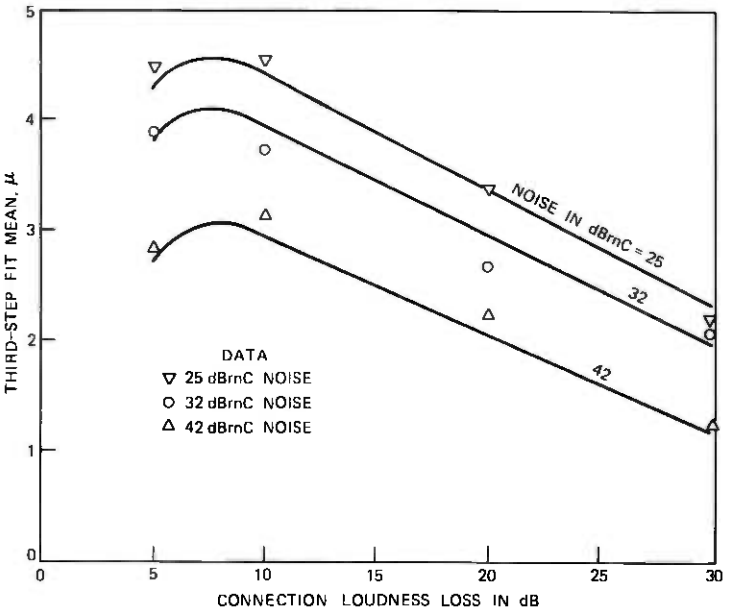


Fig. 23—HO2 loss-noise test, Step 3 means from data as a function of loudness loss and noise compared with means predicted from the loss-noise model at the HO2 base.

6.3 Talker-echo model

The R-scale was introduced previously in terms of L_e and N. Thus, it was necessary to use the conversational echo test results which included both loss and noise as parameters to establish an expression for the echo results on the R-scale. The 1970 SIBYL test was used as a basis for this conversion. As indicated in Table XIV, this test included a base condition with L_e = 10 dB and N = 30 dBrnC for which the fit mean, μ, was 4.01. The R-scale value corresponding to this combination of loss and noise is 83.47. Similarly, the results for the noise condition in the 1968 lab tests were taken into account but given less weight because they were obtained in a less realistic test environment. A further aid in deriving this conversion was the HO1 SIBYL test for loss and noise which preceded the 1970 SIBYL echo test and was conducted in the same manner with the same subjects. The final conversion is given in eq. (18) below.

$$R = 18.7 + 16.1\mu_{E1}. \tag{18}$$

This relationship is almost identical to that of the HO1 SIBYL tests at low values of μ. However, a slight correction was included for higher values of μ to take account of the base condition in the 1970 SIBYL

echo test. For the HO1 test the value of μ corresponding to the base condition ($R = 83.47$) was 3.86. The relationship in eq. (18) gives a value of 4.02 and provides good agreement with the actual value of μ which was 4.01 in the 1970 SIBYL echo tests.

Substituting eq. (14) for μ_{E1} into eq. (18) yields eq. (19) which is the transmission-rating model for echo.

$$R_E = 95.01 - 53.45 \log_{10} [(1 + D)/\sqrt{1 + (D/480)^2}] + 2.277E. \quad (19)$$

Equation (19) is plotted in Fig. 2.

In the analysis above, eq. (19) was derived from eqs. (14) and (18) to provide excellent agreement for the 1970 SIBYL echo tests. However, these same relationships did not provide good agreement with the 1970 lab tests which had a higher loss included in all conditions. The base condition for this test as given in Table XII, with $L_e = 18$ dB and $N = 33$ dBrnC ($R = 67.36$), had a fit mean, μ, of 3.59.

The equation,

$$R = 20 + 13.33\mu_{E2}, \quad (20)$$

provided a good match at this point and retained the relationship between μ and R at low values of μ obtained previously for the SIBYL echo tests. With the relationship defined by eq. (20), the transmission-rating model for echo given in eq. (19) provided an excellent representation of the results from the 1970 lab test for echo. The extent of the agreement is illustrated in Section 6.4 where the combined loss-noise-echo model is discussed.

6.4 Loss-noise-echo model

In the development of the echo results, it was noted that the degradations considered were echo-path-loudness loss (E), echo-path delay (D), loudness loss (L_e), and noise (N). Loss and noise were eliminated in the final echo result because it was felt that for any combined result, the loss and noise influence should be based on the larger data base available from the SIBYL tests.

The original analysis of the echo-loss-noise data showed that the loss and noise really only affected the circuit performance as an asymptote. That is, for very large E, the ratings were determined by L_e and N. Use was made of this fact when the echo result of eq. (14) was developed. The combination of R_{LN} and R_E was made such that this asymptotic behavior was retained in the final model.

The final result for loss, noise, and echo is presented in terms of the R-scale as R_{LNE}. This final result is shown in eq. (21).

$$R_{LNE} = \frac{R_{LN} + R_E}{2} - \sqrt{\left(\frac{R_{LN} - R_E}{2}\right)^2 + C^2}. \quad (21)$$

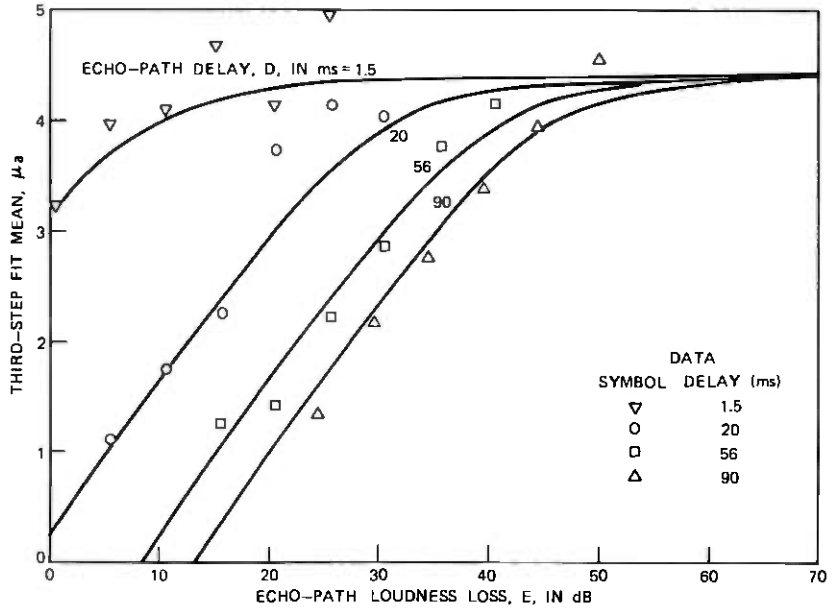


Fig. 24—1968 laboratory echo test (noise = 18 dBrnC, loss = 10 dB) Step 3 means from data as a function of echo-path loss and delay compared with means predicted from the loss-noise-echo model at the echo 1 base.

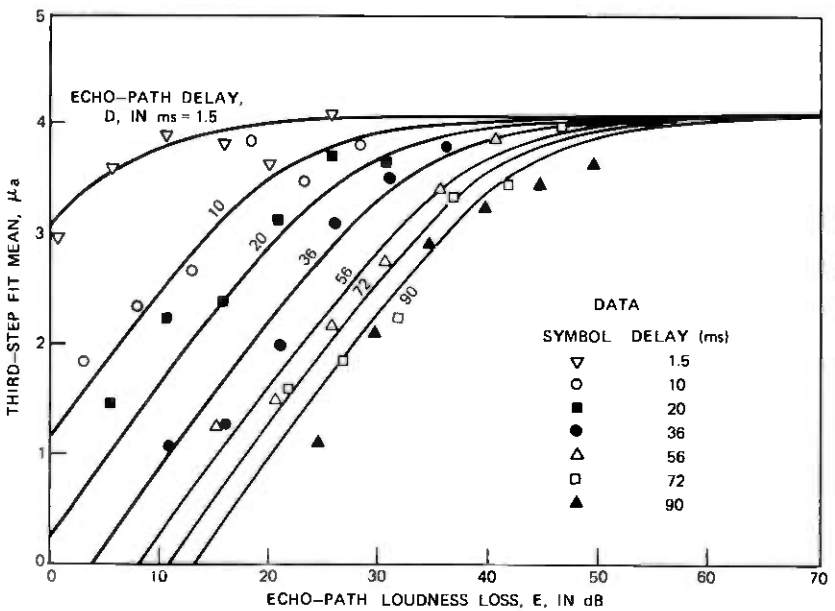


Fig. 25—1968 laboratory echo test (noise = 28 dBrnC, loss = 10 dB) Step 3 means from data as a function of echo-path loss and delay compared with means predicted from the loss-noise-echo model at the echo 1 base.

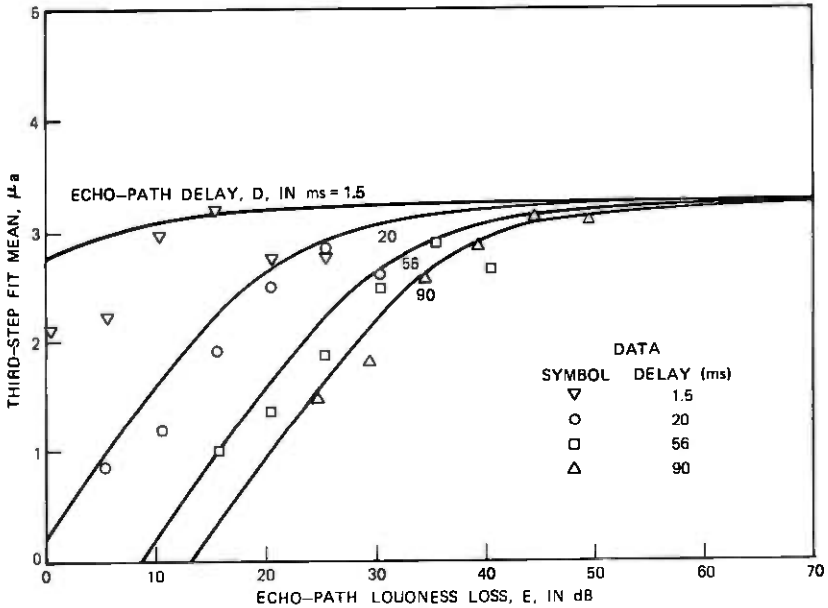


Fig. 26—1968 laboratory echo test (noise = 38 dBrnC, loss = 10 dB) Step 3 means from data as a function of echo-path loss and delay compared with means predicted from the loss-noise-echo model at the echo 1 base.

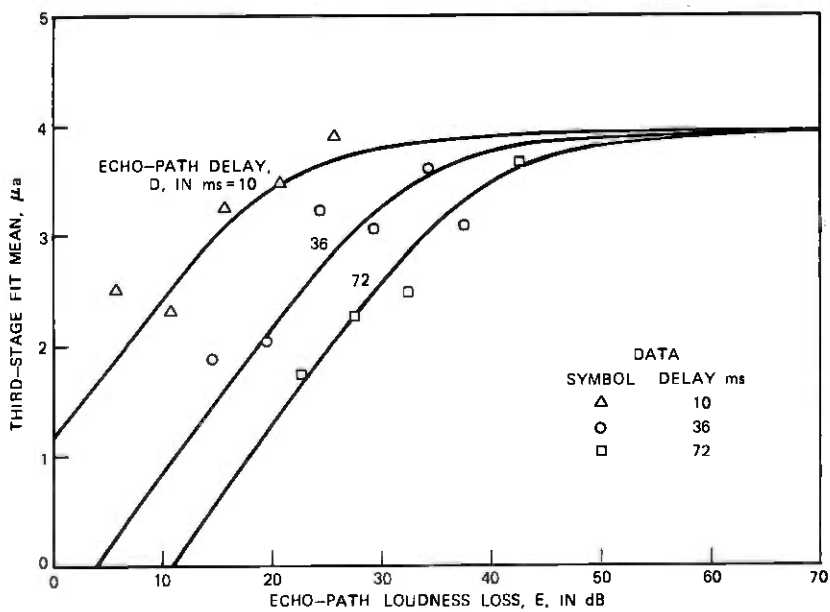


Fig. 27—SIBYL echo test, Step 3 means from data as a function of echo-path loss and delay compared with means predicted from the loss-noise-echo model at the echo 1 base.

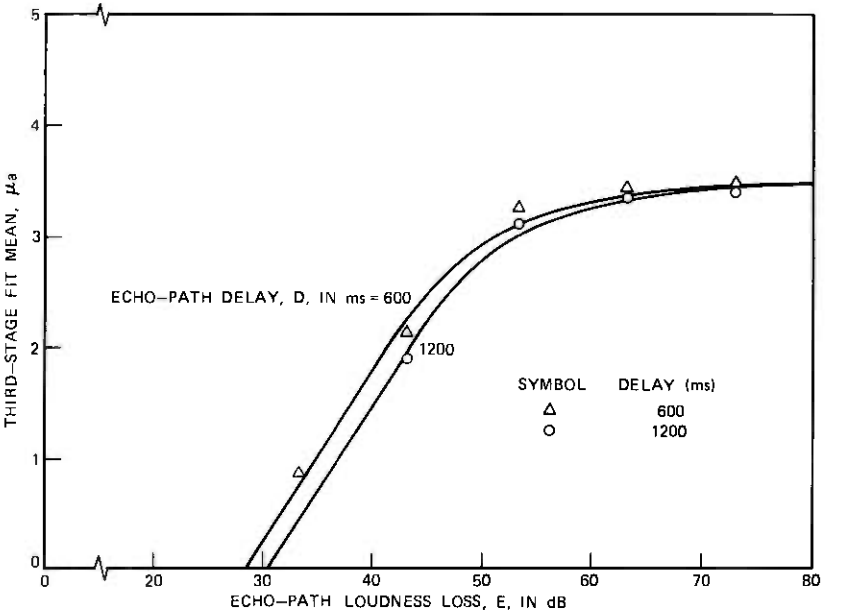


Fig. 28—1970 laboratory echo test, Step 3 means from data as a function of echo path loss and delay compared with means predicted from the loss-noise-echo model at the echo 2 base.

With $C = 0$, R_{LNE} is simply the lesser of R_{LN} and R_E. The factor C is selected to represent the additional degradation when R_{LN} and R_E are nearly equal. The value of $C = 10$ was based on echo tests that included echo, loss, and noise and was obtained as the product of the constant, 0.627, in eq. (14) and the slope of the line relating R and μ given in eq. (18).

The final result in terms of the R-scale is:

$$R_{LNE} = \frac{R_{LN} + R_E}{2} - \sqrt{\left(\frac{R_{LN} - R_E}{2}\right)^2 + (10)^2}. \qquad (22)$$

For high echo-path-loudness loss, the ratio is determined mainly by connection loudness loss and circuit noise and the result reduces to the R_{LN} result. Similarly, for connection loudness loss near optimum and low circuit noise, the rating is determined mainly by the echo, and the result effectively reduces to the R_E result.

Comparison of the final model for loss, noise, and echo with the third-step-fit means for the individual tests are shown in Figs. 24 to 28. As in the case of the loss-noise model, the final loss-noise-echo model provides good agreement with the test results from each of the individual tests.

VII. ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of a large number of colleagues who contributed to the studies described in this paper. Particular note is due Mr. T. K. Sen who was principally responsible for the loss-noise subjective tests and to S. H. Franz, J. T. Powers, D. J. Rhoads, and E. J. Thomas who conducted the talker-echo tests.

APPENDIX A

Examples Demonstrating Use of the Transmission-Rating Models

A.1 Example 1—local connection

500-type telephone sets on a calling and a called customer loop, each of which consists of 1 kilofoot of 26-gauge nonloaded cable:

Loss (L_e): TLR $= -23.6$ dB
RLR $= 26.3$ dB
$L_e =$ (TLR + RLR) dB
$= 2.7$ dB

Noise (N): $N_T =$ Total noise at each telephone set from two loops, each of which meet the 20 dBrnC loop-noise objective.[5]
$= 23$ dBrnC
$N =$ Total noise referred to a telephone set with RLR of 26 dB
$= 27$ dBrnC.

Talker Echo (assumed negligible):

Using Table II with $L_e = 2.7$ dB and $N = 27$ dBrnC,

$$R_{LN} = 78.3.$$

Using Table II for the MH data base,

$$GoB = 79.2\%, \qquad PoW = 6.6\%.$$

A.2 Example 2—local connection

500-type telephone sets and a calling and a called customer loop, each of which consists of 8 kilofeet of 26-gauge nonloaded cable:

Loss (L_e): TLR $= -18.9$ dB
RLR $= 27.3$ dB
$L_e = 8.4$ dB

Noise (N): Assumed to be the same as for Example 1:
$N = 27$ dBrnC.

Talker echo (assumed negligible):

Using Table II with $L_e = 8.4$ dB and $N = 27$ dBrnC,

$$R_{LN} = 88.7.$$

Using Table II for the MH data base,

$$GoB = 92\%, \qquad PoW = 1.8\%.$$

A.3 Example 3—local connections

500-type telephone sets on a calling and a called customer loop, each of which consists of 15 kilofeet of 26-gauge nonloaded cable:

Loss (L_e): TLR $= -13.2$ dB
RLR $= 30.1$ dB
$L_e = 16.9$ dB

Noise (N): Assumed to be the same as for Example 1.

$$N = 27 \text{ dBrnC}.$$

Talker echo (assumed negligible):

Using Table II with $L_e = 16.9$ dB and $N = 27$ dBrnC,

$$R_{LN} = 75.5.$$

Using Table II for the MH data base,

$$GoB = 74.2\%, \qquad PoW = 8.9\%.$$

A.4 Example 4—toll connection

500-type telephone sets on a calling and a called customer loop, each of which consists of 8 kilofeet of 26-gauge nonloaded cable:

Loss (L_e): TLR $= -18.9$ dB
RLR $= 27.3$ dB
L = Class 5 office-to-Class 5 office loss
$= 7.7$ dB (mean for the connection-length category, 775 to 2900 miles, Table VI of Ref. 12)
L_e = TLR + RLR + 7.7
$= 16.1$ dB.

Noise (N): Assume no noise from the loops:

N_T = Total noise from the Class 5 office-to-Class 5 office connection (33.8 dBrnC for the connection length category 775 to 2900 miles, Table III of Ref. 12 referred to the telephone set)
$= 28.5$ dBrnC
$N = 32.5$ dBrnC.

Talker echo (assumed negligible):

Using Table II with $L_e = 16.1$ dB and $N = 32.5$ dBrnC,

$$R_{LN} = 71.$$

Using Table II for the MH data base,

$$GoB = 65.2\%, \qquad PoW = 13.9\%.$$

A.5 Example 5—toll connection

Same as Example 4, Section A.5, except that it takes into account talker echo:

Echo-path delay (D)

D = Far-end echo-path loss
= 37.3 ms (mean for connection length 1450 to 2900 miles, Table III of Ref. 11).

Echo-path loudness loss (E)

E' = Loss of echo path from near-end Class 5 office to distant end and return,
= 23.3 dB (mean for connection length 1450 to 2900 miles, Table II of Ref. 11).

E = TLR + RLR + E'
= 31.7 dB.

Using Table II with $L_e = 16.1$ dB, $N = 32.5$ dBrnC, $D = 37.3$ ms, and $E = 31.7$ dB.

$$R_{LN} = 71$$
$$R_E = 82.6$$
$$R_{LNE} = 65.2.$$

Using Table II for the MH data base,

$$GoB = 52.6\%, \qquad PoW = 22.4\%.$$

REFERENCES

1. T. C. Spang, "Loss-Noise Echo Study of the Direct Distance Dialing Network," B.S.T.J., *55* (January 1976), pp. 1–36.
2. H. R. Huntley, "Transmission Design of the Intertoll Telephone Trunks," B.S.T.J., *32* (September 1953), pp. 1019–1036.
3. G. M. Phillips, "Echo and Its Effects on the Telephone User," Bell Laboratories Record (August 1954), pp. 281–284.
4. O. H. Coolidge and G. C. Reier, "An Appraisal of Received Telephone Speech Volume," B.S.T.J., *38* (May 1959), pp. 877–897.
5. D. A. Lewinski, "A New Objective for Message Circuit Noise," B.S.T.J., *43* (March 1964), pp. 719–740.
6. T. K. Sen, "Subjective Effects of Noise and Loss in Telephone Transmission," IEEE Trans. Commun. Technol., *COM-19* (December 1971), pp. 1229–1233.
7. A. H. Inglis and W. L. Tuffnell, "An Improved Telephone Set," B.S.T.J., *30* (April 1951), pp. 239–270.

8. J. L. Sullivan, "A Laboratory System for Measuring Loudness Loss of Telephone Connections," B.S.T.J., *50* (October 1971), pp. 2663–2739.
9. A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," B.S.T.J., *39* (July 1960), pp. 879–909.
10. W. T. Cochrane and D. A. Lewinski, "A New Measuring Set for Message Circuit Noise," B.S.T.J., *39* (July 1960), pp. 911–931.
11. F. P. Duffy et al., "Echo Performance of Toll Telephone Connections in the United States," B.S.T.J., *54* (February 1975), pp. 209–243.
12. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50* (April 1971), pp. 1311–1347.
13. H. D. Irvin, "Studying Tomorrow's Communications ··· Today," Bell Laboratories Record (November 1958), pp. 398–402.
14. H. D. Irvin, "SIBYL: A Laboratory for Simulation Studies of Man-Machine Systems," 1958 IRE Wescon Conv. Rec., Part 4, pp. 277–285.
15. J. L. Sullivan, "Is Transmission Satisfactory? Telephone Customers Help Us Decide," Bell Laboratories Record (March 1974), pp. 90–98.
16. A. M. Noll, "Subjective Effects of Sidetone During Telephone Conversation," Commun. Elec., *83* (May 1964), pp. 228–231.
17. A. Michael Noll, "Effects of Head and Air-Leakage Sidetone During Monaural-Telephone Speaking," J. Acoust. Soc. Amer., *36* (March 1964), pp. 598–599.
18. American National Standards Institute, *Specification for Sound Level Meters, S1.4-1971.*
19. International Telecommunication Union, *Telephone Transmission Quality, Local Networks and Telephone Sets*, CCITT Green Book, *5*, 1973.
20. R. D. Prosser, J. A. Allnatt, and N. W. Lewis, "Quality Grading of Impaired Television Pictures," Proc. IEE., *111*, No. 3 (March 1969), pp. 491–502.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 9, November 1976
Printed in U.S.A.

Speech Encryption by Manipulations of LPC Parameters

By M. R. SAMBUR and N. S. JAYANT

(Manuscript received May 14, 1976)

This paper discusses several manipulations of LPC (linear predictive coding) parameters for providing speech encryption. Specifically, the paper considers temporal rearrangement or scrambling of the LPC code sequence, as well as the alternative of perturbing individual samples in the sequence by means of pseudo-random additive or multiplicative noise. The latter approach is believed to have greater encryption potential than the temporal scrambling technique, in terms of the time needed to "break the secrecy code." The encryption techniques are assessed on the basis of perceptual experiments, as well as by means of a quantitative assessment of speech-spectrum distortion, as given by an appropriate "distance" measure.

I. INTRODUCTION

Encryption can be an important requirement in speech communication systems. Conventionally, encryption has largely been accomplished by signal manipulations in the frequency domain; for example, by means of spectrum inversion techniques.[1] With the increased popularity of digital codes for speech transmission, time-domain encryption techniques have received increased attention. Typically the time-domain encryption technique consists of temporal rearrangement of samples within a time block. For the scrambling of PCM bits in speech waveform coding, a block-length that is at least a pitch period long is usually adequate to provide a nonspeech-like output waveform. Similarly, the scrambling of differential PCM and delta-modulation bits can also produce a nonspeech-like output waveform provided that the time-block is sufficiently long. For example, in a 24-kb/s speech code, this constraint implies approximately a block length of 64 samples for an adequate scrambling of the coded bits.[2]

The temporal scrambling of speech samples within millisecond-length blocks generally provides what may be referred to as casual encryption. This means that a noncasual 'eavesdropper' can break the

speech secrecy code by the simple expedient of running through a finite number of possible rearrangements of the disarranged speech samples that are received. Greater degrees of encryption or secrecy can be achieved by employing much longer speech blocks for scrambling or, alternatively, by subjecting individual speech samples to pseudo-random additive or multiplicative perturbations whose undoing is typically more time-consuming than a simple temporal rearrangement of clean digits or bits.

The purpose of this paper is to point out that casual encryption as well as more formal secrecy can be achieved by appropriate manipulations of the linear predictive coding (LPC) parameters.[3,4] The use of an LPC code is by no means a necessary requirement for encryption; it can be achieved in conjunction with any kind of speech digitizers, such as the waveform codes[5] discussed above. However, when the channel capacity of communication systems dictates a low-bit-rate vocoder instead of a generally higher-bit-rate waveform code, the LPC parameter manipulations discussed in this paper may provide a naturally appropriate basis for speech encryption and/or secrecy. It shall also be seen that an efficient encryption of the LPC parameters can be achieved more readily than similar techniques used to encrypt waveform codes. For example, an adequate block length for scrambling the LPC parameters can be as short as 6 to 8 samples, while the block length for waveform scrambling is typically 16 to 64 samples.

In this paper, Section II provides a brief description of LPC encoding of speech, while Section III considers the use of temporal scrambling and pseudo-random sample perturbations for casual and formal encryption in the LPC domain. Section IV describes attempts to measure the efficacy of the encryption techniques. These measurements involved informal perceptual experiments (the results are usually unambiguous and one-dimensional enough not to require formal subjective testing), as well as a comparison of alternative techniques in terms of speech-spectrum distortions that they provide. The spectrum distortion was assessed by an appropriate distance measurement. This distance approach has the advantage of being quantitative; however, as discussed in Section IV, the distance criterion has to be invoked with caution because spectral distortion, as such, is not a definitive measure of speech encryption.

II. LINEAR PREDICTION SPEECH MODEL

The method of linear prediction has proved quite popular and successful for use in speech-compression systems.[3,6,7] In this method, speech is modeled as the output of an all-pole filter $H(z)$ that is excited by a sequence of pulses separated by the pitch period for voiced sounds

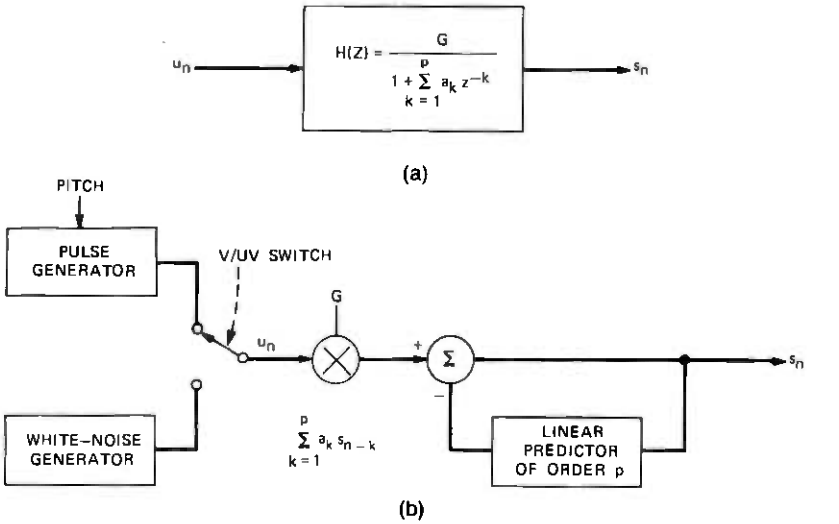


Fig. 1—Discrete model of speech production as employed in linear prediction. (a) Frequency-domain model. (b) Time-domain model.

or pseudo-random noise for unvoiced sounds. These assumptions imply that within a frame of speech, the output speech sequence is given by

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n,$$

where p is the number of modeled poles, u_n is the appropriate input excitation, G is the gain of the filter, and the a_k's are the coefficients characterizing the filter (linear prediction coefficients). Figure 1 illustrates the frequency-domain as well as the equivalent time-domain model of linear prediction speech production. To account for the nonstationary character of the speech waveform, the parameters a_k of the modeled filter are periodically updated during successive speech frames.* Generation of speech in this method requires a knowledge of the pitch, the filter parameters, and the gain of the filter (amplitude of excitation) in each speech frame.

The LPC coefficients model the combined effects of the vocal tract, glottal source, and radiation load in each frame of speech. Manipulations of the LPC coefficients can seriously perturb the frequency character of the speech signal and, hence, destroy the linguistic information present in the signal. In contrast, the measurements of pitch and gain represent the prosodic aspects of the speech and some characteristics

* A frame is a segment of speech thought adequate to assume stationarity of the speech process. Typical frame lengths employed range from 10 to 30 ms.

of the speaker. Manipulations of pitch and gain parameters will affect the prosody of the speech, but not seriously diminish the linguistic aspects of the waveform. In Section III, we consider several methods for efficiently manipulating the LPC coefficients so as to encrypt the speech signal.

Since the purpose of this paper is the consideration of encryption techniques for low-bit-rate vocoders (2.4 kb/s or less), the manipulation schemes discussed in Section III were not performed directly on the LPC coefficients, but rather on more desirable alternate representations of these coefficients. The stability of the linear-prediction filter, $H(z)$, is extremely sensitive to small perturbations in the LPC coefficients and, thus, it is not possible to achieve low-bit-rate coding by transmitting the LPC coefficients.[6] However, by transmitting either the log area coefficients or the parcor coefficients, a 2.4-kb/s vocoder is readily achieved.[6] The log area coefficients are nonlinearly related to the LPC coefficients by

$$g_i = \log \frac{1 + k_i}{1 - k_i},$$

where the k_i's are termed the parcor coefficients.[7] If we denote $a_i^{(j)}$ as the ith linear prediction coefficient for a jth-pole linear-prediction model, then

$$k_i = a_i^{(i)}.$$

The parcor coefficients have the very important property that if

$$|k_i| < 1, \qquad i = 1, \cdots, p,$$

then it is guaranteed that the linear prediction filter is stable.[4] Thus, small perturbations in the parcor coefficients or log-area coefficients will not affect the stability of the modeled filter.

III. ENCRYPTION TECHNIQUES

3.1 Temporal scrambling

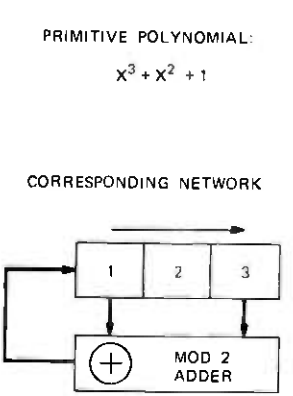
The rearrangement of samples within a block of length L is achieved by assigning to each sample a new address A ($A = 1$, or 2, or 3, $\cdots$, or L) as determined by the state of a maximal-length shift-register arrangement. The theory and design of maximal length sequences is well documented.[8,9] Here, we simply provide a constructive recapitulation for the purpose of this paper. The idea is to start with a shift register whose length is $D = \log_2 L$ (assume that the block length is a power of 2, and that elements in the register are either 1 or 0). The next step is to select a so-called primitive polynomial $P_D(x)$ of degree D, and to include stage $(D - S)$ in the register ($S = 0$ to $D - 1$) in

an exclusive OR (modulo 2 add) feedback arrangement, if the coefficient of x^S in $P(x)$ is nonzero. The resulting network now generates a succession of $2^D - 1 = L - 1$ nonzero states in the shift register at successive 'clock' times, after which the cycle repeats, starting once again with the original initial state of the shift register. The number of nonzero states in the cycle is identically equal to the repetition period $L - 1$ of the cycle. Consequently, the $L - 1$ states of the shift-register (specifically, their decimal equivalents) can be utilized as "pseudo-random" addresses for a block of $L - 1$ input samples in a one-to-one mapping of addresses. If the input block has L rather than $L - 1$ samples (because of the frequent requirement that L be a power of 2), the address of the Lth sample is usually left unaltered by the scrambler. Such simplicity is not, however, inevitable, and appropriate manipulations that scramble all L samples are quite conceivable.

Figure 2 illustrates the scrambler design for the example of $D = 3$ and $L = 7$, as defined by a primitive polynomial $P_3(X) = X^3 + X^2 + 1$. It is seen how input samples (1, 2, 3, 4, 5, 6, and 7) get scrambled into the pseudo-random positions (1, 4, 6, 7, 3, 5, and 2) in a reversible one-to-one mapping.

Figure 3 illustrates an alternative design, as defined by a second primitive polynomial of degree 3, $P_3(X) = X^3 + X + 1$. In this case, the output addresses of the input samples are the positions (1, 4, 2, 5, 6, 7, and 3).

It is clear that in each of the arrangements in Figs. 2 and 3, the use of a different initializing sequence (other than 001) can lead to a totally different mapping of sample addresses. There would be $L - 1$ nonzero initializations, corresponding to every given $P_3(X)$. Incidentally, the number of primitive polynomials of degree 3 is 2.



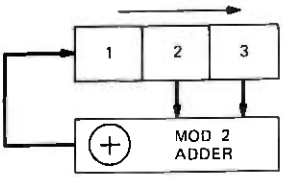
NUMBER	STATE OF SHIFT REGISTER	DECIMAL EQUIVALENT
1	001	1
2	100	4
3	110	6
4	111	7
5	011	3
6	101	5
7	010	2
↑	001	↑
INPUT ADDRESSES	REPETITION PERIOD = 7	OUTPUT ADDRESSES

Fig. 2—Scrambler design with a three-stage shift register.

PRIMITIVE POLYNOMIAL:

$X^3 + X + 1$

CORRESPONDING NETWORK



NUMBER	STATE OF SHIFT REGISTER	DECIMAL EQUIVALENT
1	001	1
2	100	4
3	010	2
4	101	5
5	110	6
6	111	7
7	011	3
↑	001	↑
INPUT ADDRESSES	REPETITION PERIOD = 7	OUTPUT ADDRESSES

Fig. 3—Alternate scrambler design with a three-stage shift register.

Table I lists for $D = 1$ to 12 a typical set of primitive polynomials and also the number of primitive polynomials for each D. Note, for example, that a 12-stage shift register with an exclusive OR feedback network involving stages 12, 11, 8, and 6 provides one of 144 possible bases for a scrambler that would operate on an input block of $2^{12} = 4096$ samples.

The possibility of alternate scrambler designs (as defined by different initializations and/or different primitive polynomials) is an important consideration from the point of view of the average descrambling time needed for an eavesdropping code-breaker.

3.2 LPC parameter scrambling

The effectiveness of any scrambling scheme in perturbing the sequence of samples is directly proportional to the lack of similarity or

Table I—List of primitive polynomials

Degree D	Typical Primitive Polynomial	Number of Primitive Polynomials of Degree D
1	$X + 1$	1
2	$X^2 + X + 1$	1
3	$X^3 + X + 1$	2
4	$X^4 + X + 1$	2
5	$X^5 + X^2 + 1$	6
6	$X^6 + X + 1$	6
7	$X^7 + X + 1$	18
8	$X^8 + X^4 + X^3 + X^2 + 1$	16
9	$X^9 + X^4 + 1$	48
10	$X^{10} + X^3 + 1$	60
11	$X^{11} + X^2 + 1$	176
12	$X^{12} + X^6 + X^4 + X + 1$	144

dynamic ranges of the samples to be scrambled. The greater the range of values assumed by the samples, the more effective the scrambling scheme.[1] For an efficient scrambling of the LPC parameters, let us begin by ordering the parameters in the following manner: the first sample in the first block is x_{11}, where x_{in} denotes the ith LPC parameter* in the nth analysis frame. The second sample is x_{21} and the third sample is x_{31}. The arrangement proceeds in this fashion until the $(p + 1)$th† sample, which is defined as x_{12}. Thus, the ordering of LPC parameters for purposes of scrambling is simply a concatenation of the p LPC parameters in each sequential analysis frame.

Using this particular arrangement, it can be seen that within a block of data there is a wide distribution of values assumed by the various samples. This observation follows from the fact that the measured LPC parameters for any given analysis frame will usually vary across the entire permissible range of values. For example, the p measured values of the parcor coefficients in any given frame will typically be somewhat uniformly spread across the permissible range of -1 to 1.[4] The particular arrangement of LPC parameters given above will thus be effective for scrambling purposes due to the large resulting dynamic range. In Section IV, we show that a block length as small as eight samples ($L = 8$) is sufficient to destroy the linguistic information in the synthetic signal produced by a 12th order analysis ($p = 12$).

3.3 Pseudo-random perturbations

For a more secure secrecy coding of the speech signal, the LPC parameters can be modified by a pseudo-random additive or multiplicative perturbation. Since the repetitive period of any typical pseudo-random number generator is extremely large, the process of undoing or breaking the encryption is quite difficult and time-consuming.

Since one of the goals of the present study was to perceptually assess the linguistic information in the synthesized speech generated by the encrypted LPC parameters, the pseudo-random number perturbation scheme was designed to retain the stability of the modified LPC filter. Thus, for the manipulation of the parcor coefficients, the pseudo-random number technique involved the transmission of the sequence of parameters

$$y_{in} = k_{in} \times r_{in},$$

where

k_{in} = ith parcor coefficient in nth frame

r_{in} = ith pseudo-random number in nth frame; $|r_{in}| \leq 1$.

* The LPC parameters considered in this paper are either the log-area coefficients or parcor coefficients.

† p = order of LPC analysis.

Since $|r_{in}| \leq 1$, $|y_{in}| < 1$ and the stability of the LPC filter is guaranteed. For the modification of the log-area coefficients, the technique is simply to transmit

$$y_{in} = g_{in} + r_{in}.$$

The stability of the resulting LPC filter is guaranteed regardless of the range of r_{in}. This result follows from the fact that any real value of y_{in} will lead to parcor parameters that are less than 1.

In viewing the pseudo-random number manipulation of the LPC parameters, it should be noted that the spectral characteristics of the LPC filter are more sensitive to changes in the parcor coefficients than to changes in the log-area coefficients.[10] Thus, manipulation of the parcor coefficients is a more direct and efficient technique for perturbing the spectral properties of the LPC filter. For this reason the pseudo-random techniques discussed in this paper were applied only to the parcor coefficients. If pseudo-random number manipulation is to be applied to the log-area coefficients, the manipulation can be made most effective if the probability distribution of the random number generator is nonuniform, in order to mimic that of the log-area coefficient.[10]

For the experimental examination of the pseudo-random number perturbation of the parcor coefficients, the following two probability distributions were used for generating r_{in}:

(*i*) r_{in} was uniformly distributed between -1 and 1, or

(*ii*) r_{in} was, with equal probability, set to -1 or 1.

The second distribution was studied because the resulting manipulation of the parcor coefficients is particularly easy to perform and, as we shall soon discuss, is effective in destroying the intelligibility of the encrypted speech. However, the "breaking" of the encryption coding using the second distribution is not difficult to achieve by using the available knowledge of the statistical range of the parameters. For example, it is well known that the first parcor coefficient is almost always positive.[4] Thus, a negative value of the first parcor coefficient indicates a manipulation of this parameter. If the listener knows that a $+1$ or -1 manipulation of the parameters is being employed, then a simple reversal of sign breaks the encryption.

IV. EXPERIMENTAL STUDY

In this section, we examine the effectiveness of the various encryption techniques in destroying the intelligibility of the output speech signal. For this purpose, an informal perceptual evaluation was conducted. To evaluate objectively the efficacy of the techniques, an LPC distance measure proposed by Itakura[11] was used to reinforce and supplement

the perceptual examination. Before discussing the LPC distance measure, we emphasize that *this measure may not be a definitive or complete description of encryption efficiency;* but it is a good measure of spectral distortion, which in turn turns out to be a useful (if not ideal) indicator of intelligibility loss.

4.1 Distance measure

The LPC distance measure is defined as

$$d_n = \ln (\mathbf{a}_n V \mathbf{a}_n^T / \mathbf{b}_n V \mathbf{b}_n^T),$$

where

$\mathbf{a}_n$ = Original LPC coefficient vector $(1, a_1, \cdots, a_p)$ measured in the nth frame of the speech signal.

$\mathbf{b}_n$ = LPC coefficient vector determined after manipulation of the original parameters in the nth frame

and

$$V = [v(|i - j|)], \qquad (i, j = 0, 1, \cdots, p),$$

where $v(i)$ are the normalized correlation coefficients that are computed directly from $\mathbf{b}_n$.[3,10]

The measure d_n has been very effectively applied in problems of speech recognition,[11] speaker recognition,[12] and variable frame-rate synthesis.[13,14] Gray and Markel[15] have recently demonstrated that the measure d_n is very closely related to the rms spectral distance measure. Sambur and Jayant[16] have also studied the significance of the measure, and a complete discussion of the utility of the measure for assessing spectral distortions can be found in their paper. For purposes of this paper, the important facts to appreciate about the measure d_n are

(*i*) The greater the value of d_n, the more pronounced the spectral distortions of the original sound.

(*ii*) A value of $d_n = 0.9$ is a "perceptually" significant boundary for evaluating spectral distortion.[13]

4.2 Experiment

For the experimental study, four sentences spoken by four different speakers were analyzed using a 12th order ($p = 12$) LPC autocorrelation analysis for each contiguous 20-ms frame. The sentences analyzed were:

(*i*) A lathe is a big tool.

(*ii*) May we all learn a yellow lion roar.

(*iii*) Few thieves are never sent to the jug.

(*iv*) It's time we rounded up that herd of Asian cattle.

The encryption schemes that were formally evaluated both perceptually and with the distance measure of Section 4.1 were:

a. Scrambling
   (1) Block length = 16
   (2) Block length = 8
b. Pseudo-random manipulation*
   (1) Uniform distribution of r_{in} for $i = 1$ and $r_{in} = 1$ for $i > 1$.
   (2) Uniform distribution of r_{in} for $i \leqq 6$ and $r_{in} = 1$ for $i > 6$.
   (3) Uniform distribution of r_{in} for all i $(1 \leqq i \leqq 12)$.
   (4) ± 1 distribution of r_{in} for $i = 1$ and $r_{in} = 1$ for $i > 1$.
   (5) ± 1 distribution of r_{in} for $i \leqq 6$ and $r_{in} = 1$ for $i > 6$.
   (6) ± 1 distribution of r_{in} for all i $(1 \leqq i \leqq 12)$.

Experiment b was performed to determine the number of parcor coefficients that must be altered to effectively encrypt the signal. Since the parcor coefficients are approximately ordered in terms of their spectral sensitivity,[4] these experiments were performed by sequentially removing from manipulation the less sensitive parameters.

4.3 Results

4.3.1 Distance evaluation

4.3.1.1 Uniform pseudo-random manipulation. Figure 4 illustrates the distance-evaluation of the sentence "May we all learn a yellow lion roar" for the uniform pseudo-random number manipulation of the parcor coefficients. Parts (a), (b), and (c) of Fig. 4 indicate, respectively, the results of experiments b(1), b(2), and b(3) of Section 4.2. The straight solid line in each part of the figure depicts the perceptually significant threshold for assessing spectral distortions ($d = 0.9$). Any frame with a distance larger than the threshold is perceptually different from the nonencrypted speech. To show just how dramatically the perturbation in the spectral character of the speech can be, Fig. 5 illustrates the calculated linear prediction spectrum (dotted line) for the nonencrypted speech frame and the corresponding linear prediction spectrum (solid line) for the same frame of encrypted speech. The measured LPC distance between the illustrated spectra is $d_n = 3.0$, or approximately the average value of distance for uniform pseudo-random manipulation of the first coefficient. From this figure, it can be expected that the character of the encrypted speech is completely different from that of the original speech.

* Remember r_{in} denotes the pseudo-random number multiplicative factor for the ith parcor coefficient in the nth frame.

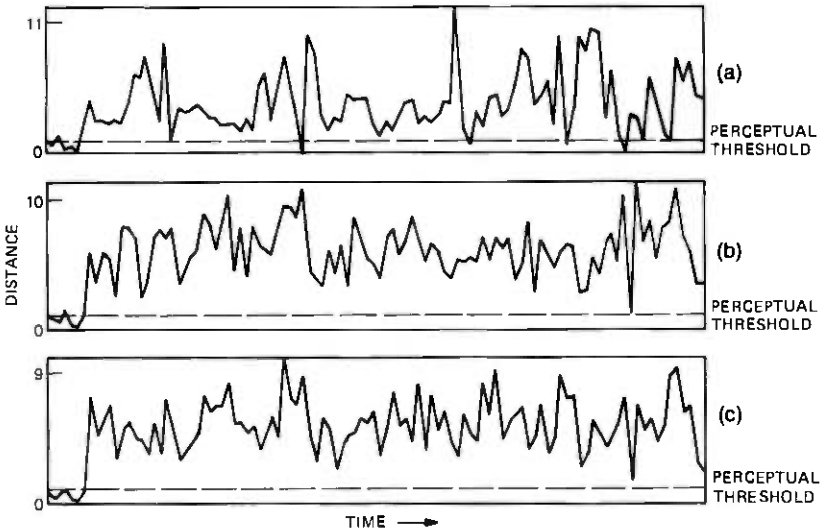


Fig. 4—LPC distance as a function of time across the utterance, "May we all learn a yellow lion roar," for uniform pseudo-random perturbation of the parcor parameters. (a) Manipulation of k_1; average distance = 3.4. (b) Manipulation of k_1 to k_6; average distance = 4.4. (c) Manipulation of all k_i; average distance = 4.4.

The results depicted in Fig. 4 are typical of the distance evaluation results for the uniform pseudo-random manipulation of the parcor coefficients determined for the other sentences examined. It is interesting to note that the average distance for an encryption scheme that manipulates the first six parameters is not significantly lower than the average distance obtained for the manipulation of all 12 parameters. This result can be anticipated from the fact that the higher-ordered parcor coefficients are much less sensitive than the lower-ordered parameters, and changes in these higher-ordered parameters do not significantly change the spectral character of the sound.[4] Thus, a less-expensive and equally effective encryption scheme can be obtained by manipulating only a few lower-ordered parameters. To determine the optimum number of parameters necessary for an efficient, uniform,

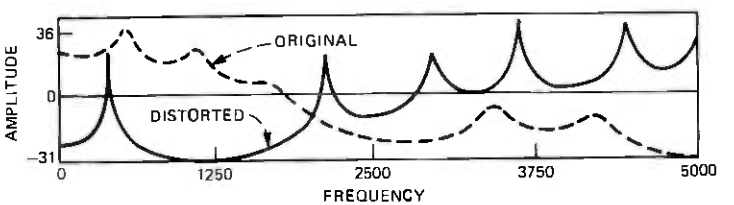


Fig. 5—Comparison of the distorted LPC spectra and the original LPC spectrum. Distance between the spectrum equals 3.0.

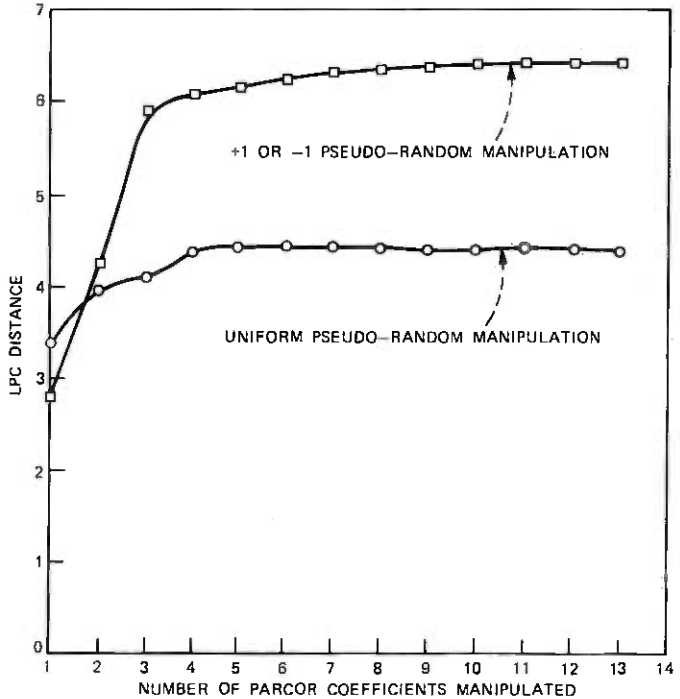


Fig. 6—Average LPC distance as a function of number of parcor coefficients manipulated by pseudo-random number techniques.

pseudo-random encryption, we sequentially increased the number of parcor parameters perturbed by uniform pseudo-random manipulation and measured the average LPC distance. Figure 6 illustrates the average LPC distance as a function of the number of parameters manipulated. From this figure, it can be seen that a scheme that perturbs only the first four parcor coefficients is quite efficient.

4.3.1.2 Pseudo-random manipulation of +1 or −1. Figure 7 shows the detailed distance-evaluation scores for the $+1$ or -1 pseudo-random perturbation of the sentence "May we all learn a yellow lion roar." Parts (a), (b), and (c) of the figure correspond to experiments b(4), b(5), and b(6), respectively. Figure 6 illustrates the average LPC distances obtained for encryption schemes that sequentially increase the number of parameters subjected to $+1$ or -1 pseudo-random manipulations. We note from Figs. 6 and 7 that again the perturbation of the higher-ordered parcor coefficients does not significantly add to the effectiveness of the encryption scheme. It can also be seen from these figures that $+1$ or -1 pseudo-random manipulation is generally superior (except for the manipulation of only k_1) to the uniform pseudo-random number scheme in distorting the speech signal. How-

ever, as noted previously, this form of encryption is easier to break than uniform pseudo-random number coding.

4.3.1.3 Scrambling. Figure 8 shows the frame-by-frame distance scores for the scrambling of the parcor coefficients for the sentence "May we all learn a yellow lion roar." The illustrated results are typical of the results obtained for the other analyzed sentences. A comparison of the distances results of the pseudo-random schemes (Fig. 6) shows that a scrambling encryption with a block length of only eight samples ($L = 8$) is at least as effective in distorting the spectral properties of the original signal as a pseudo-random manipulation of the first parcor coefficient. A scrambling scheme with a block length of 16 ($L = 16$) or more samples is superior to any of the pseudo-random schemes studied. It is interesting to note that the scrambling manipulation saturates in effectiveness for block length greater than 16. Since the range of the parcor coefficients is confined to $-1 \leqq k_i \leqq 1$, increasing the block length beyond 16 does not increase the dynamic range of the sample within the block and, thus, the effectiveness of the scrambling is not enhanced for $L > p$ (see Section 3.2).

4.3.2 Perceptual evaluation

To support the results of the distance study, the various encrypted utterances were presented to a group of listeners for an informal perceptual evaluation of the manipulation schemes. To avoid any

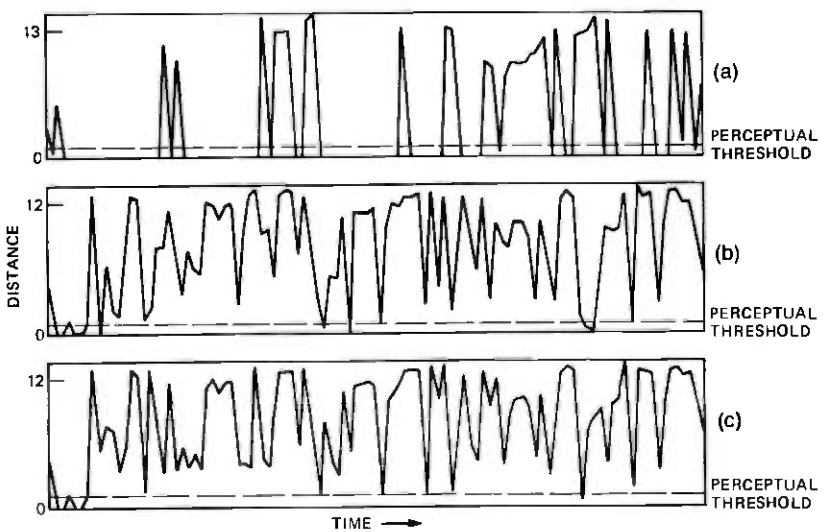


Fig. 7—LPC distance as a function of time across the utterance "May we all learn a yellow lion roar" for the +1 or −1 pseudo-random manipulation of the parcor coefficient. (a) Manipulation of k_1; average distance = 2.8. (b) Manipulation of k_1 to k_8; average distance = 6.2. (c) Manipulation of all k_i; average distance = 6.3.

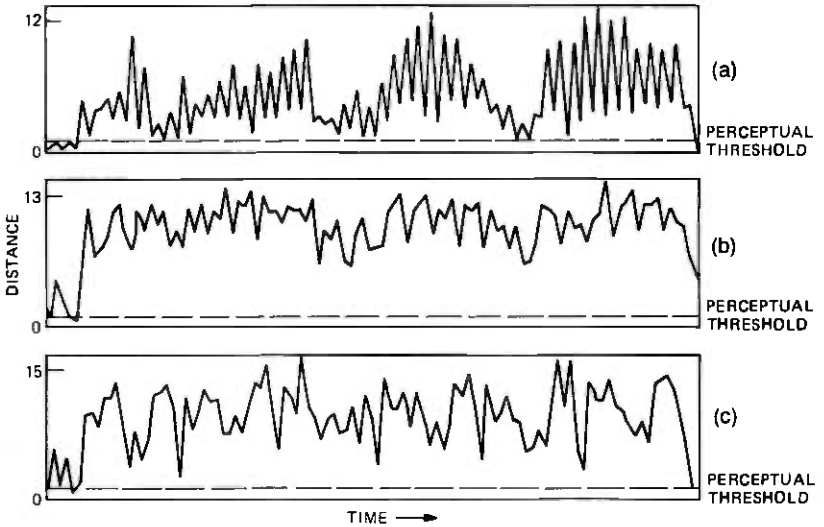


Fig. 8—LPC distance as a function of time across the utterance "May we all learn a yellow lion roar" for the scrambling of the parcor coefficient. (a) Block length = 8; average distance = 3.8. (b) Block length = 16; average distance = 7.7; (c) Block length = 64; average distance = 7.6.

problems posed by the awkward linguistic content of the analyzed sentences, the listeners in this study were all familiar with the sentences, and were also informed that the utterances to be heard were typical sentences used to evaluate vocoder systems.

The listeners in the experiment were asked to determine the intelligibility of the utterance and to rank-order the effectiveness of the encryption schemes. For all the techniques studied, except for the +1 or −1 manipulation of only k_1, the listeners unanimously agreed that the encrypted utterances were clearly nonspeech-like. However, for the uniform pseudo-random techniques manipulating only the first parcor coefficients, the listeners noted that, even though the complete utterances could not be understood, there were certain instances in the encrypted utterances that were somewhat speech-like and understandable. These instances probably correspond to points in the encrypted speech for which the LPC distances fall below the perceptual threshold. In characterizing the nonspeech-like quality of the encrypted utterances, the listeners termed the pseudo-random perturbed utterances as sounding like "one continuous buzz;" the scrambled utterances sounded like "water running through a pipe."

In rank-ordering the encryption schemes, the listeners were quite definite in characterizing the +1 or −1 psuedo-random manipulation of only the first parcor coefficient as least effective. The scrambling

with block length of 16 ($\bar{d} = 7.7$) was ranked about equal to the +1 or −1 pseudo-random manipulation of all 12 parcor coefficients ($\bar{d} = 6.3$), and also to the same manipulation of only the first six coefficients ($\bar{d} = 6.2$). The uniform pseudo-random scheme that altered all 12 coefficients ($\bar{d} = 4.4$) was ranked equal to the scheme that perturbed only the first six coefficients ($\bar{d} = 4.4$), and both techniques were ranked slightly less effective than the scrambling with block length of 16 ($\bar{d} = 7.7$) and the equivalent +1 or −1 pseudo-random schemes. The other techniques were ranked somewhere in the middle. The perceptual rank-ordering of the various manipulation schemes corresponded almost exactly to the distance evaluation and, thus, reinforced the conclusions in that evaluation.

V. CONCLUSIONS

There is great interest in low-bit-rate speech-transmission systems and in the "securing" of these transmission systems. The purpose of this paper is to investigate various methods for encrypting a low-bit-rate LPC transmission system. The methods chosen for investigation were schemes that either scrambled the string of input parcor coefficients or multiplied the coefficients by a pseudo-random number. The schemes were evaluated by an informal perceptual experiment and by the use of an LPC distance measure. The results of the evaluations suggest that all the schemes are somewhat successful in distorting the original signal. The most successful scheme was the scrambling technique with a block length of 16 samples. The pseudo-random manipulations were almost as effective.

In viewing the results of the evaluations, it is important to note that the distortion of the speech signal is only one consideration in designing an encryption system. Another consideration is the difficulty of "breaking" the security code. Of the codes examined, the uniform pseudo-random number manipulation is the most difficult to break. The scrambling scheme is the next most difficult and the +1 or −1 pseudo-random scheme is the easiest. Still another consideration is the transmitter-end complexity of the encryption scheme. Although this complexity is somewhat difficult to assess, it appears that the scrambling scheme is the least complex and the uniform pseudo-random manipulation is the most complex. In choosing any of these encryption schemes, a user would balance the various merits and liabilities of the techniques.

VI. ACKNOWLEDGMENT

The manipulation of LPC parameters was suggested by Professor B. S. Ramakrishna of the Electrical Communication Engineering

Department, Indian Institute of Science, Bangalore, during a visit by one of the authors (N. S. Jayant).

REFERENCES

1. D. Kahn, *The Code Breakers*, New York: Macmillan, 1967, pp. 551–554.
2. N. S. Jayant, "Step-size Transmitting Differential Coders for Mobile Telephony," B.S.T.J., *54*, No. 9 (November 1975), pp. 1557–1581.
3. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer., *50* (1971), pp. 637–655.
4. J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear Prediction of Speech-Theory and Practice," Speech Communication Research Laboratory, Inc., Santa Barbara, Calif., Monograph 10, 1973.
5. N. S. Jayant, "Digital Coding of Speech Waveforms—PCM, DPCM, and DM Quantizers," Proc. IEEE (May 1974), pp. 611–632.
6. M. R. Sambur, "An Efficient Linear Prediction Vocoder," B.S.T.J., *54*, No. 10 (December 1975), pp. 1693–1723.
7. F. Itakura et al., "An Audio Response Unit Based on Partial Autocorrelation," IEEE Trans. Commun., *COM-20*, No. 4 (August 1972), pp. 792–797.
8. S. Golomb, *Shift Register Sequences*, San Francisco: Holden Day, 1967.
9. R. G. Gallager, *Information Theory and Reliable Communications*, New York: John Wiley, 1968.
10. J. Makhoul and R. Viswanathan, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," Bolt Beranek and Newman, Inc., Report No. 2800, April 1974.
11. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech and Signal Proc., *ASSP-23*, No. 1 (February 1975), pp. 67–72.
12 H. Wakita, "On the Use of Linear Prediction Error Energy for Speech and Speaker Recognition," J.A.S.A., *57*, Supplement No. 1 (Spring 1975). (A)
13. D. T. Magill, "Adaptive Speech Compression for Packet Communication Systems," Telecommunications Conference Record, IEEE Publ. 73, CH0805-2, 29D 1–5.
14. J. R. Makhoul, L. Viswanathan, L. Cosel, and W. Russel, "Natural Communication with Computers: Speech Compression Research at BBN," BBN Report No. 2976, Vol. II, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts, December 1974.
15. A. H. Gray and J. D. Markel, "COSH Measure for Speech Processing," J.A.S.A., *58*, Supplement No. 1 (Fall 1975). (A)
16. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis From Speech Inputs Containing Noise or Additive White Noise," IEEE Trans. Acoust., Speech, and Signal Proc., *ASSP-24*, No. 6 (December 1976).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 9, November 1976
Printed in U.S.A.

Multiple Tone Parameter Estimation From Discrete-Time Observations

By D. C. RIFE and R. R. BOORSTYN*

(Manuscript received May 11, 1976)

In a previous paper, we discussed estimation of the parameters of a single tone from a finite number of noisy discrete-time observations. In this paper, we extend the discussion to include several tones. The Cramér-Rao bounds are derived and their properties examined. Estimation algorithms are discussed and characterized.

I. INTRODUCTION

In a previous paper,[1] we reported on the estimation of the parameters of tones from a finite number of noisy, discrete-time observations and described the case of a single complex tone. In this report, we discuss the situation when the signal consists of several, say k, tones, either real or complex. By *real signal* we mean

$$s(t) = \sum_{i=1}^{k} b_i \cos(\omega_i t + \theta_i).$$

The corresponding *complex signal* is of the form

$$s(t) + j\check{s}(t) = \sum_{i=1}^{k} b_i \exp[j(\omega_i t + \theta_i)],$$

where $\check{s}(t)$ is the Hilbert transform of $s(t)$.

A computer observes, through the A–D converters, noisy versions of the signal, $X(t)$, and possibly its Hilbert transform $Y(t)$. That is, samples are taken of

$$X(t) = s(t) + W(t), \tag{1}$$

and

$$Y(t) = \check{s}(t) + \check{W}(t), \tag{2}$$

where $W(t)$ and $\check{W}(t)$ are the noise and its Hilbert transform, respectively.

* Polytechnic Institute of New York.

The observations are made at times denoted t_n. The computer will process one or both sample vectors:

$$\mathbf{X} = [X_0, X_1, \cdots, X_{N-1}]^T \quad \text{and} \quad \mathbf{Y} = [Y_0, Y_1, \cdots, Y_{N-1}]^T,$$

where T denotes matrix transpose,

$$X_n = X(t_n), \tag{3}$$

and

$$Y_n = Y(t_n). \tag{4}$$

We assume the noise samples, W_n and $\breve{W}_n$, are independent, zero-mean, gaussian random variables with variance σ^2.

Let $\boldsymbol{\alpha}$ be the p-element vector of unknown signal parameters. We assume all signal parameters are unknown, so that $p = 3k$, and use the convention:

$$\begin{aligned} \alpha_{3i-2} &= \omega_i, \\ \alpha_{3i-1} &= b_i, \end{aligned} \tag{5}$$

and

$$\alpha_{3i} = \theta_i, \qquad i = 1 \text{ to } k.$$

This model describes several situations. The real signal may be received from a data set during a test or it could be a probe signal used to characterize a data-transmission channel. The real and imaginary parts of the complex signal could occur as the result of in-phase and quadrature modulation processes, as described by Palmer.[2] The imaginary part of the complex signal could be the output of a 90-degree phase-shift network (Hilbert transformer) through which the real signal is passed before the sampling is done. This is done in certain types of data sets that use all-digital means to demodulate received signals. Samples of the complex signal are easier to process because of the absence of negative frequency components, as we show below. The model also applies to certain mathematically equivalent, phased-array radar problems, such as the one described in Refs. 3, 4, 5, and 6.

There are two main aspects to the problem of estimating the parameters of the signal: lower bounds to estimation accuracy and algorithms for doing the estimation. In the next section, the properties of the Cramér-Rao (C-R) bounds are explored. There are many other bounds that could be applied but we have only examined the C-R bounds. Section III describes and evaluates some approximations to maximum-likelihood (ML) estimation. In Ref. 1, we found that when the signal consists of a single complex tone ($k = 1$), then ML estimates can be obtained with any desired accuracy. When several tones are present, ML estimation is sufficiently complicated that suboptimum alternatives are attractive.

II. CRAMÉR-RAO BOUNDS

2.1 General theory

Maximum-likelihood estimates of signal parameters are unbiased at high signal-to-noise ratio (s/n).[4,7] We will develop estimation algorithms that have very little bias, so we have only studied the C-R bounds to unbiased estimation accuracy. Even when an estimator has some bias, the unbiased bounds serve as useful goals for estimation accuracy. Since the accuracy of ML estimates approaches the unbiased C-R bounds at high s/n, the unbiased bounds also show what could be done if exact ML estimation algorithms were used.

We found in Ref. 1 that low s/n is that range of s/n where estimation anomalies occur. None of the known bounds seem to be very tight under these conditions.

The first property of the C-R bounds that we consider is one for which we need the following general notation.

Let $\mathbf{V}$ be a "signal" vector whose typical component is of the form

$$V_n = \sum_{i=1}^{K} b_i g_i(\omega_i, \theta_i, n). \tag{6}$$

Notice that each $g_i(\cdot)$ has an associated level, b_i, and is a function only of n and the ith set of unknown parameters. Time does not necessarily enter into the $g_i(\cdot)$ functions. Let $\mathbf{X}$ be a noisy observation of $\mathbf{V}$. Assume the noise is additive, multivariate normal with zero mean and correlation matrix $\mathbf{R}^{-1}$. If the noise vector is $\mathbf{W}$, then

$$\mathbf{X} = \mathbf{V} + \mathbf{W}, \tag{7}$$

and the probability density function of $\mathbf{X}$ given $\mathbf{V}$ is

$$f(\mathbf{X}/\mathbf{V}) = \frac{|\mathbf{R}|^{\frac{1}{2}}}{(2\pi)^{N/2}} \exp[-\tfrac{1}{2}(\mathbf{X} - \mathbf{V})^T \mathbf{R}(\mathbf{X} - \mathbf{V})], \tag{8}$$

where N is the dimension of $\mathbf{V}$ and the T denotes transpose (see Ref. 8, page 207).

The C-R bounds require certain regularity conditions on $\mathbf{V}$, which are satisfied by our model.[9] The bounds are the diagonal elements of the inverse of the Fisher information matrix, $\mathbf{J}$, whose typical element[10] is:

$$J_{ab} = -E\left\{\frac{\partial^2}{\partial\alpha_b \partial\alpha_a} \log f\right\}, \tag{9}$$

where $E\{\cdot\}$ denotes expected value of $\{\cdot\}$. The bounds are:

$$\text{Var}\,\{\hat{\alpha}_a - \alpha_a\} \geqq J^{aa}, \tag{10}$$

where J^{aa} is the ath diagonal element of $\mathbf{J}^{-1}$ and $\hat{\alpha}_a$ is an unbiased estimate of α_a.

It is easy to show[11] that

$$J_{ab} = \frac{\partial \mathbf{V}^T}{\partial \alpha_b} \mathbf{R} \frac{\partial \mathbf{V}}{\partial \alpha_a}. \tag{11}$$

We now present a few theorems that characterize the C-R bounds. We assume $\mathbf{J}$ is not singular, for reasons discussed below.

Theorem 1: *The C-R bounds to unbiased estimation of the parameters ω_i and θ_i of V are functions of b_i but are independent of the other levels, b_j: $j \neq i$. The bound to unbiased estimation of a level, b_i, is independent of all the levels.*

Proof: Equation (11) is equivalent to

$$J_{ab} = \sum_n \sum_m R_{nm} \frac{\partial V_n}{\partial \alpha_b} \frac{\partial V_m}{\partial \alpha_a}, \tag{12}$$

where R_{nm} is an element of $\mathbf{R}$.

The elements of $\mathbf{J}$ that are functions of the parameters of g_i and g_j, using the convention given by (5) and the notation $g_i(n) = g_i(\omega_i, \theta_i, n)$, are:

$$J_{3i-2,3j-2} = b_i b_j \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \omega_i} \frac{\partial g_j(m)}{\partial \omega_j}. \tag{13a}$$

$$J_{3i-2,3j-1} = b_i \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \omega_i} g_j(m). \tag{13b}$$

$$J_{3i-2,3j} = b_i b_j \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \omega_i} \frac{\partial g_j(m)}{\partial \theta_j}. \tag{13c}$$

$$J_{3i-1,3j-2} = b_j \sum_n \sum_m R_{nm} g_i(n) \frac{\partial g_j(m)}{\partial \omega_j}. \tag{13d}$$

$$J_{3i-1,3j-1} = \sum_n \sum_m R_{nm} g_i(n) g_j(m). \tag{13e}$$

$$J_{3i-1,3j} = b_j \sum_n \sum_m R_{nm} g_i(n) \frac{\partial g_j(m)}{\partial \theta_j}. \tag{13f}$$

$$J_{3i,3j-2} = b_i b_j \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \theta_i} \frac{\partial g_j(m)}{\partial \omega_j}. \tag{13g}$$

$$J_{3i,3j-1} = b_i \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \theta_i} g_j(m). \tag{13h}$$

$$J_{3i,3j} = b_i b_j \sum_n \sum_m R_{nm} \frac{\partial g_i(n)}{\partial \theta_i} \frac{\partial g_j(m)}{\partial \theta_j}. \tag{13i}$$

An examination of (13a) through (13i) shows that the submatrix of $\mathbf{J}$ has the form $\mathbf{D}_i\mathbf{Q}_{ij}\mathbf{D}_j$, where

$$\mathbf{D}_i = \begin{bmatrix} b_i & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & b_i \end{bmatrix} \tag{14}$$

and the matrix $\mathbf{Q}_{ij}$ is not a function of any b_i. It follows that $\mathbf{J}$ has the form

$$\mathbf{J} = \mathbf{DQD}, \tag{15}$$

where

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} & \cdots & \\ \mathbf{0} & \mathbf{D}_2 & & \\ \vdots & & \ddots & \\ & & & \mathbf{D}_k \end{bmatrix} \tag{16}$$

and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1k} \\ \vdots & \ddots & \\ \mathbf{Q}_{k1} & & \mathbf{Q}_{kk} \end{bmatrix}. \tag{17}$$

$\mathbf{0}$ is a matrix whose elements are all zeros. From (15),

$$\mathbf{J}^{-1} = \mathbf{D}^{-1}\mathbf{Q}^{-1}\mathbf{D}^{-1}, \tag{18}$$

from which the theorem follows. For example,

$$\text{Var } \{\hat{\omega}_1 - \omega_1\} \geqq Q^{11}/b_1^2. \tag{19}$$

This theorem is not entirely new. It is alluded to in Ref. 6. However, this form of the theorem shows that, contrary to Ref. 6 and popular opinion, precisely known sampling times (or antenna element spacing in the equivalent radar problem) are not necessary for the theorem to hold.

The theorem is true whether or not the noise samples are independent and regardless of the sampling times. Of course, if the sampling times are not known, then the C-R bounds cannot be accurately calculated, but that does not obviate the theorem.

It should also be clear that the number of unknown parameters is unimportant to the theorem. Clearly the theorem holds if, for example,

$$g_i(\omega_i, \theta_i, n) = \cos(\omega_i t_n + \theta_i),$$

and if

$$t_n = nT.$$

Theorem 2: *The bounds associated with the parameters of the first k tones, when there are $k + m$ tones, are not less than the bounds when there are only k tones.*

Proof: The matrix $\mathbf{J}$ is always positive semidefinite. Thus, if it is not singular, it is positive definite.

Suppose $\mathbf{J}$ is the Fisher information matrix for $k + m$ tones and is partitioned so that $\mathbf{J}_k$ is the $\mathbf{J}$ matrix associated with k of the tones. This partitioning is always possible. Then write

$$\mathbf{J} = \left[\begin{array}{c|c} \mathbf{J}_k & \mathbf{K} \\ \hline \mathbf{K}^T & \mathbf{J}_m \end{array}\right]. \tag{20}$$

Since $\mathbf{J}$ is positive definite, so are $\mathbf{J}_k$ and $\mathbf{J}_m$.

Write the inverse of $\mathbf{J}$ in the form

$$\mathbf{J}^{-1} = \left[\begin{array}{c|c} \mathbf{U} & \mathbf{W} \\ \hline \mathbf{W}^T & \mathbf{V} \end{array}\right], \tag{21}$$

where $\mathbf{J}_k$ and $\mathbf{U}$ are both $3k$ by $3k$ matrices. Theorem 2 is true if

$$\mathbf{U} \geqq \mathbf{J}_k^{-1}, \tag{22}$$

which means $\mathbf{U} - \mathbf{J}_k^{-1}$ is positive semidefinite and which we now prove.

Using the fact that

$$\mathbf{J}\mathbf{J}^{-1} = \mathbf{I}, \tag{23}$$

one can show that

$$\mathbf{U} = [\mathbf{J}_k - \mathbf{K}\mathbf{J}_m^{-1}\mathbf{K}^T]^{-1}. \tag{24}$$

Observe that $\mathbf{K}\mathbf{J}_m^{-1}\mathbf{K}^T$ is positive semidefinite. That is,

$$\mathbf{K}\mathbf{J}_m^{-1}\mathbf{K}^T \geqq 0. \tag{25}$$

Since $\mathbf{J}_k$ and hence $\mathbf{U}$ are positive definite, (24) and (25) imply that $\mathbf{U}^{-1} \leqq \mathbf{J}_k$, which implies (22).

Another implication of the proof of Theorem 2 is that the bounds for p of $p + m$ unknown parameters are not less than the bounds when only the first p parameters are unknown.

This theorem is also not entirely new, although we have not seen it stated before. A restricted version of the theorem is mentioned in Ref. 12, (page 33), and Problem 2.4.23 in Ref. 10 hints at this kind of result.

The theorem depends upon $\mathbf{J}$ being nonsingular. It is easy, but tedious, to show that if the signal vector is composed of samples of the real or complex signal described in the introduction and only two tones are involved, then $\mathbf{J}$ is singular only if the two tone frequencies are equal, modulo $2\pi/T$, where T is the intersample time. (Remember that a real tone has a component at $+\omega_i$ and another at $-\omega_i$.)

We have not been able to prove this result for an arbitrary number of tones, but all of our calculations of various J matrices support the hypothesis that J is singular only if two or more of the tone frequencies are equal, modulo $2\pi/T$ (assuming N is large enough).

When two of the tone frequencies are equal, the receiver is receiving one less tone than expected. In this paper, we assume that the correct number of tones, k, is known and that all of the frequencies are distinct.

2.2 *Equally spaced samples and independent noise*

We now concentrate on the problem described in the introduction. Assume all noise samples are independent with variance σ^2. That is,

$$\mathbf{R} = \frac{1}{\sigma^2}\mathbf{I}, \tag{26}$$

where **I** is an identity matrix.

Define

$$\mu_n = \sum_{i=1}^{K} b_i \cos(\omega_i t_n + \theta_i), \tag{27}$$

$$\nu_n = \sum_{i=1}^{K} b_i \sin(\omega_i t_n + \theta_i), \tag{28}$$

and

$$t_n = nT; \qquad n = 0, 1, \cdots, N-1. \tag{29}$$

As is mentioned in Ref. 1, the time of the first sample, t_0, has an effect upon bounds and estimation accuracy. We have ignored that problem in this paper and taken t_0 to be zero.

The signal vector is

$$V_n = \begin{cases} \mu_n & n = 0 \text{ to } N-1 \\ \nu_{n-N} & n = N \text{ to } 2N-1 \end{cases} \qquad \text{(complex signal)} \tag{30}$$

or

$$V_n = \mu_n; \qquad n = 0 \text{ to } N-1 \qquad \text{(real signal)}. \tag{31}$$

Then a typical element of **J** is

$$J_{ab} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left[\frac{\partial \mu_n}{\partial \alpha_a}\frac{\partial \mu_n}{\partial \alpha_b} + \frac{\partial \nu_n}{\partial \alpha_a}\frac{\partial \nu_n}{\partial \alpha_b}\right] \qquad \text{(complex signal)}. \tag{32}$$

The ν_n terms are dropped if the signal is real.

Let $\mathbf{M}(\omega, \theta)$ be the matrix defined by

$$\mathbf{M}(\omega, \theta) = \begin{bmatrix} T^2\sum n^2\cos\Delta_n & -T\sum n\sin\Delta_n & T\sum n\cos\Delta_n \\ T\sum n\sin\Delta_n & \sum\cos\Delta_n & \sum\sin\Delta_n \\ T\sum n\cos\Delta_n & -\sum\sin\Delta_n & \sum\cos\Delta_n \end{bmatrix}, \tag{33}$$

where

$$\Delta_n = n\omega T + \theta; \qquad n = 0 \text{ to } N-1. \tag{34}$$

Let $\mathbf{P}_{ij}$ be the matrix defined by

$$\mathbf{P}_{ij} = \mathbf{M}(\omega_i - \omega_j, \theta_i - \theta_j) \tag{35}$$

and let $\mathbf{P}$ be the p-by-p matrix defined by

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \cdots & \mathbf{P}_{1k} \\ \vdots & & \vdots \\ \mathbf{P}_{k1} & & \mathbf{P}_{kk} \end{bmatrix}. \tag{36}$$

Let

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{37}$$

Define a matrix $\mathbf{Q}_{ij}$ by

$$\mathbf{Q}_{ij} = \tfrac{1}{2}[\mathbf{M}(\omega_i - \omega_j, \theta_i - \theta_j) - \mathbf{M}(\omega_i + \omega_j, \theta_i + \theta_j)\mathbf{B}] \tag{38}$$

and a matrix $\mathbf{Q}$ by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1k} \\ \vdots & & \vdots \\ \mathbf{Q}_{k1} & \cdots & \mathbf{Q}_{kk} \end{bmatrix}. \tag{39}$$

Then it can be shown (13) that $\mathbf{J}$ is given by:

$$\mathbf{J} = \frac{1}{\sigma^2}\mathbf{DPD} \qquad \text{complex tones} \tag{40}$$

or

$$\mathbf{J} = \frac{1}{\sigma^2}\mathbf{DQD} \qquad \text{real tones.} \tag{41}$$

Theorem 3: *When the signal consists of two equal-level complex tones, the C-R bounds for the same parameters (e.g., the two frequencies) are equal. In other words, the mutual interference is reciprocal.*

Proof: The $\mathbf{J}$ matrix is

$$\mathbf{J} = \frac{1}{\sigma^2}\begin{bmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_2 \end{bmatrix}\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{11} \end{bmatrix}\begin{bmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_2 \end{bmatrix} \tag{42}$$

because $\mathbf{P}_{11} = \mathbf{P}_{22} = \mathbf{M}(0, 0)$. Observe that $\mathbf{P}_{12}^T = \mathbf{B}\mathbf{P}_{12}\mathbf{B}$ because $\mathbf{M}^T(\omega, \theta) = \mathbf{B}\mathbf{M}(\omega, \theta)\mathbf{B}$. Thus, $\mathbf{J}^{-1}$ has the form

$$\mathbf{J}^{-1} = \sigma^2\begin{bmatrix} \mathbf{D}_1^{-1} & 0 \\ 0 & \mathbf{D}_2^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{U} & \mathbf{W} \\ \mathbf{W}^T & \mathbf{V} \end{bmatrix}\begin{bmatrix} \mathbf{D}_1^{-1} & 0 \\ 0 & \mathbf{D}_2^{-1} \end{bmatrix}, \tag{43}$$

where

$$\mathbf{V} = \mathbf{BUB}. \tag{44}$$

Hence,

$$|U_{ij}| = |V_{ij}| \tag{45}$$

and $U_{ii} = V_{ii}$, which proves the theorem.

Theorem 4: *The bounds for two tones, real or complex, are periodic in* θ_1 *and* θ_2 *with period* π.

Proof: The theorem follows from the easily checked fact that $\mathbf{M}(\omega, \theta + \pi) = -\mathbf{M}(\omega, \theta)$.

Theorem 5: *The bounds for real or complex tones are periodic in each frequency with period* $2\pi/T$.

Proof: The theorem follows from the fact that $\mathbf{M}(\omega + 2\pi/T, \theta) = \mathbf{M}(\omega, \theta)$.

Theorem 6: *The bounds associated with complex tones depend upon the difference frequencies and phases but not upon the absolute values.*

Proof: The theorem follows from (35), (36), and (40).

It is, in general, tedious to invert J and obtain formulas for the bounds. However, it is a simple matter to have a computer calculate the elements of J and its inverse. We have done this to obtain a better understanding of the bounds.

A number of illustrative curves are given in Ref. 13. In the interest of brevity, we will present only two of the figures here.

The main thing we learned from the calculations is that there is a critical frequency separation, $4\pi/NT$, associated with multitone C-R bounds. In Ref. 1, it is shown that when a single complex tone is present, the bounds are independent of the frequency of the tone. When more than one complex tone is present, the bounds approach the single-complex-tone bounds when the minimum frequency separation (modulo $2\pi/T$) exceeds the critical frequency. The multitone bounds increase rapidly as the minimum frequency separation goes below this critical frequency.

This rule applies to a single real tone if it is considered to be two complex tones, one at a frequency, say, of ω_1, and one at $-\omega_1$. Thus, if the frequency of a single real tone is less than $2\pi/NT$, modulo π/T, then its C-R bounds are much larger than the corresponding single-complex-tone bounds.

In all cases, the multitone bounds depend upon the tone phases, as might be expected.

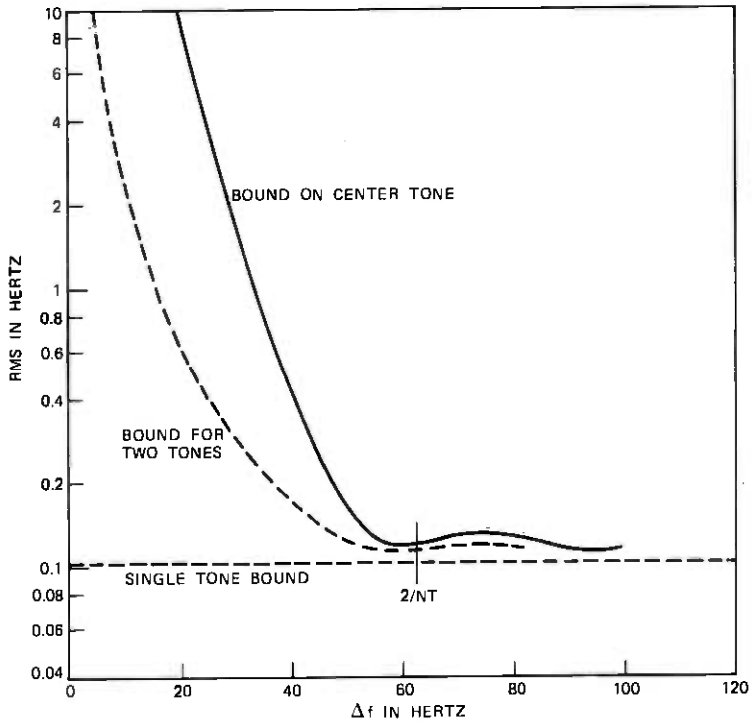


Fig. 1—Frequency estimation bound vs Δf for center of three equally spaced complex tones with worst relative phase and 20 dB s/n. N is 128. $1/T$ is 4000 Hz. Corresponding single and double tone bounds also shown.

Figure 1 illustrates the critical frequency for frequency-estimation bounds. The worst phase, i.e., the phase that gives the largest frequency estimation bound, was used at each difference frequency. Figure 2 shows the critical frequency effect upon the frequency-estimation bound for a single real tone.

To facilitate comparisons, in all figures we used a sampling frequency of 4000 Hz for complex tones and 8000 Hz for real tones. Thus, in both cases, the unknown tone frequencies are assumed to fall in the range of 0 to 4000 Hz.

III. ESTIMATION ALGORITHMS

3.1 General

The ML estimation procedure is conceptually simple. Given that a sample vector, $\mathbf{X}$, is received, the ML estimate of the parameter vector, $\hat{\boldsymbol{\alpha}}$, is the value of $\boldsymbol{\alpha}$ that maximizes the p.d.f. of $\mathbf{X}$. That is, $\hat{\boldsymbol{\alpha}}$ maximizes $f(\mathbf{X}/\mathbf{V})$. $\hat{\boldsymbol{\alpha}}$ may not be unique. Maximum likelihood estimation of the

parameters of a single complex tone has been shown to be relatively easy to implement.[1] It was shown in Ref. 1 that single complex-tone ML estimators have variances almost equal to the C-R bounds over a wide range of s/n. No other unbiased estimators could do significantly better over that range of s/n.

Maximum likelihood estimation when several tones are present is much more difficult to implement. However, we show below ways to approximate ML estimation. We start the discussion with complex tones and examine a practical approximation to ML estimation, the resulting bias effects, the use of window functions to reduce bias, and a time-saving interpolation algorithm. Then we briefly discuss how the ideas and results apply to real tones.

Recall from Ref. 1 that we seek to maximize the function

$$L = \frac{2}{N} \sum_n (X_n \mu_n + Y_n \nu_n) - \frac{1}{N} \sum_n (\mu_n^2 + \nu_n^2), \tag{46}$$

where X_n and Y_n are as defined in (3) and (4).

After carefully arranging the terms, we obtain the likelihood func-

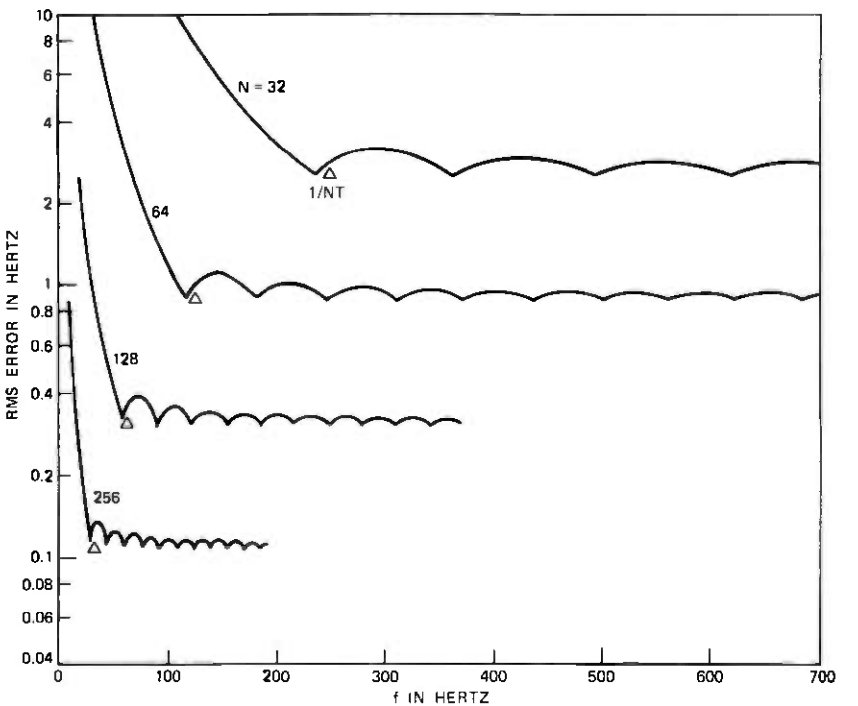


Fig. 2—Frequency estimation bounds vs frequency for single real tone at 20 dB s/n and worst phase.

tion in a form similar to that used in Ref. 1:

$$L = \sum_{i=1}^{k} \{2b_i \operatorname{Re}[e^{-j\theta_i} A(\omega_i)] - b_i^2\} - \frac{1}{N} \sum_{i \neq m} \sum_{m} b_i b_m \sum_{n} \cos(n\omega_i T - n\omega_m T + \theta_i - \theta_m), \quad (47)$$

where

$$A(\omega) = \frac{1}{N} \sum_{n=0}^{N-1} (X_n + jY_n) e^{-jn\omega T}. \quad (48)$$

L as given by (47) has two main terms and would be difficult to maximize by a simple program. It can be done, but a lot of work is involved. We notice, however, that when there is only one tone ($k = 1$), the second term of (47) vanishes. Also, when N is large and $k > 1$, the magnitude of the second term is still relatively small and does not involve the data. Thus, we are led to drop the second term in L and maximize the remainder. This, of course, will only give ML estimates when $k = 1$ and will give "almost ML" estimates otherwise.

3.2 An almost ML algorithm

Suppose the cross-product terms in (47) are dropped. Then to make estimates, we need to maximize

$$L_1 = \sum_{i=1}^{k} 2b_i \operatorname{Re}[e^{-j\theta_i} A(\omega_i)] - b_i^2. \quad (49)$$

From Ref. 1, each frequency estimate, $\hat{\omega}_i$ maximizes $|A(\omega)|$. Then the corresponding level and phase estimates are

$$\hat{b}_i = |A(\hat{\omega}_i)| \quad (50)$$

and

$$\hat{\theta}_i = \arg[A(\omega_i)]. \quad (51)$$

The function $|A(\omega)|$ has many maxima and large peaks near the frequency of each tone. Thus, the frequencies of these large peaks, as illustrated in Fig. 3, are taken to be the frequency estimates, $\hat{\omega}_i$. Due to the periodicity of $|A(\omega)|$, all the ω_i should be confined to a range no wider than $\omega_s = 2\pi/T$ to avoid ambiguous frequency estimates. Normally the range $(0, 2\pi/T)$ is used. When real tones are involved, the range should not exceed π/T.

3.3 Bias

Consider the case of only two tones. An example of $|A(\omega)|$ when the noise power is zero is shown on Fig. 3.

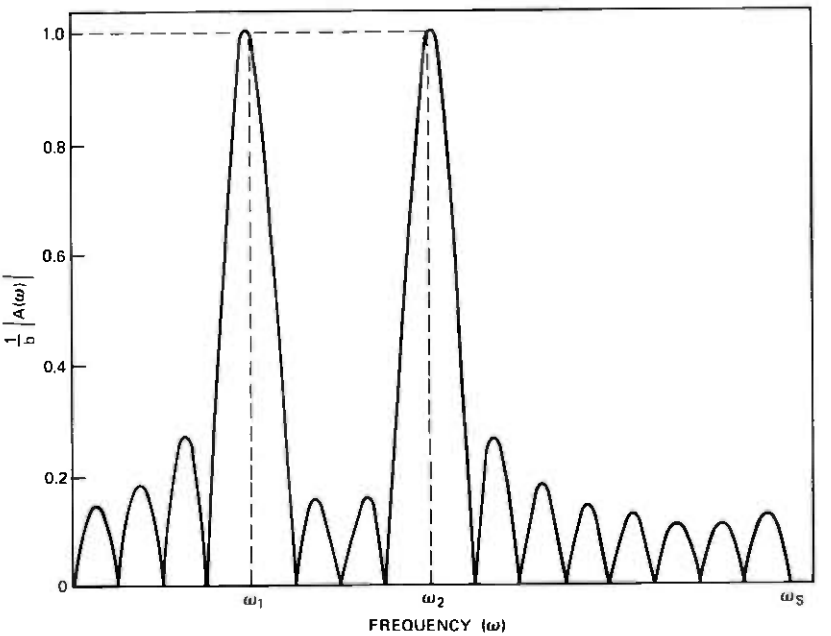


Fig. 3—Shape of $|A(\omega)|$ from two complex tones of equal phase and level, without noise. N is 16.

The figure has large peaks near ω_1 and ω_2. The peaks in the example are actually both displaced away from the average of the two frequencies. Thus, the penalty for neglecting the cross-product term in (47) is a bias in estimates of frequencies and levels.

The frequency and level bias in the zero-noise case is easily calculated. An example of such calculations is shown on Fig. 4. The figure shows the dependence of frequency estimation bias on the difference frequency (Δf) of the two tones. When two tones have almost the same frequency, the two large peaks merge into one at a frequency equal to the average of the two tone frequencies. This accounts for the negative slope of $-\frac{1}{2}$ at low Δf on Fig. 4. There is also a dependence upon the difference phase ($\Delta\theta$).

Figure 4 shows the bias for one of the two complex tones. The bias for the other has the same magnitude but opposite sign. In general, the magnitudes of the biases for two tones are not equal. However, they are equal when the two tones are equal-level complex tones.

3.4 Window functions

In discrete Fourier transform (DFT) work, window functions (also called weighting functions) are often used to minimize the effects of one tone upon another. The modification of the DFT of samples of one

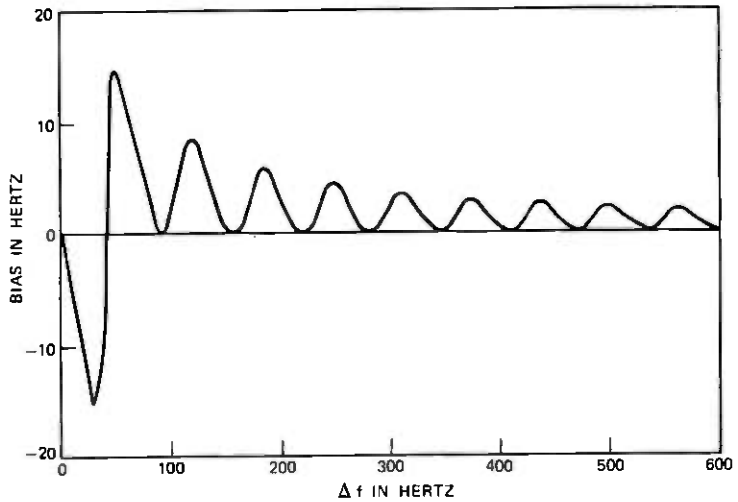


Fig. 4—Bias in peak frequency of $|A(\omega)|$ from two equal-level, equal-phase complex tones vs difference frequency, without noise. N is 64.

tone by the presence of samples of another tone is called *leakage*. See Rife and Vincent[14] for a discussion of leakage and how window function will reduce it.

In the time domain, a window function (or time window), say $h(t)$, is characterized by its samples, $h(nT)$. In use, each data sample, $X_n + jY_n = Z_n$, is multiplied by $h_n = h(nT)$ before $A(\omega)$ is computed. Thus, $A(\omega)$ becomes

$$A(\omega) = \frac{1}{N} \sum_n h_n Z_n e^{-jn\omega T}. \tag{52}$$

When a window function is used, the bias in the frequencies of the peaks of $|A(\omega)|$ is modified. If a good window is used, the bias can be greatly reduced. The penalty, as we see below, is an increase in the variance of $\hat{\omega}_i$ and $\hat{b}_i$. Palmer also reported this penalty in Ref. 2.

In the context of the DFT, window functions can be written in the form

$$h_n = 1 + \sum_{i=1}^{M} d_i \cos(2\pi in/N). \tag{53}$$

The number M, which can be assumed to be less than $N/2$, and the d_i define particular windows. With h_n in this form,

$$\frac{1}{N} \sum_{n=0}^{N-1} h_n = 1.$$

Table I—Values of d_i in ascending order of i for various windows

Hanning	Standard	Taylor
−1	−1.43596	−1.03538
	0.497536	0.0824936
	−0.061576	−0.00116197
		−0.00188862
		−0.00123387
		−0.000671595
		−0.000275885

A window that is better than many at reducing bias is the one identified by Rife and Vincent[14] as $g_3(t)$. We call this the *standard* window. Another useful window is one of the Taylor windows.[14] These windows are defined in Table I.

Figure 5 is an attempt to summarize the way window functions affect bias. The curves on the figure compare upper bounds to the bias associated with each of the previously defined window functions. The curves were obtained by computing at each frequency the bias at the worst phase (the phase that gave the largest bias). The resulting curves were flattened as indicated for the Taylor curve.

Figure 5 shows the Taylor window does the best job when the tone frequency separation is small. At large separations, however, the standard window does much better. The figure also shows how bad the bias is if no window is used.

Windowing increases the variance of frequency and level estimates. It can be shown[13] that the increase in variance is related to the function.

$$\eta = \frac{1}{N} \sum_{n=0}^{N-1} h_n^2. \tag{54}$$

It is easy to show that

$$\eta = 1 + \tfrac{1}{2} \sum_{i=1}^{M} d_i^2, \qquad \text{if} \qquad N > 2M. \tag{55}$$

Thus, η is not a function of N. Simulations verify that larger RMS errors are associated with larger values of η. Some values of η are tabulated below.

Window	η
None	1.00
Hanning	1.50
Taylor	1.54
Standard	2.16

Durrani et al. call the sum η a *dispersion factor*.[15] They have compared many windows and have tabulated their parameters, including dispersion factors. Other windows are mentioned in Blackman and Tukey.[16]

The data on Fig. 6 shows the general effects of windows on RMS errors when a single complex tone is present. The Hanning window produces almost the same RMS error as the Taylor window and is not shown on the figure.

Bias contributes to RMS errors more than variance does at high s/n,

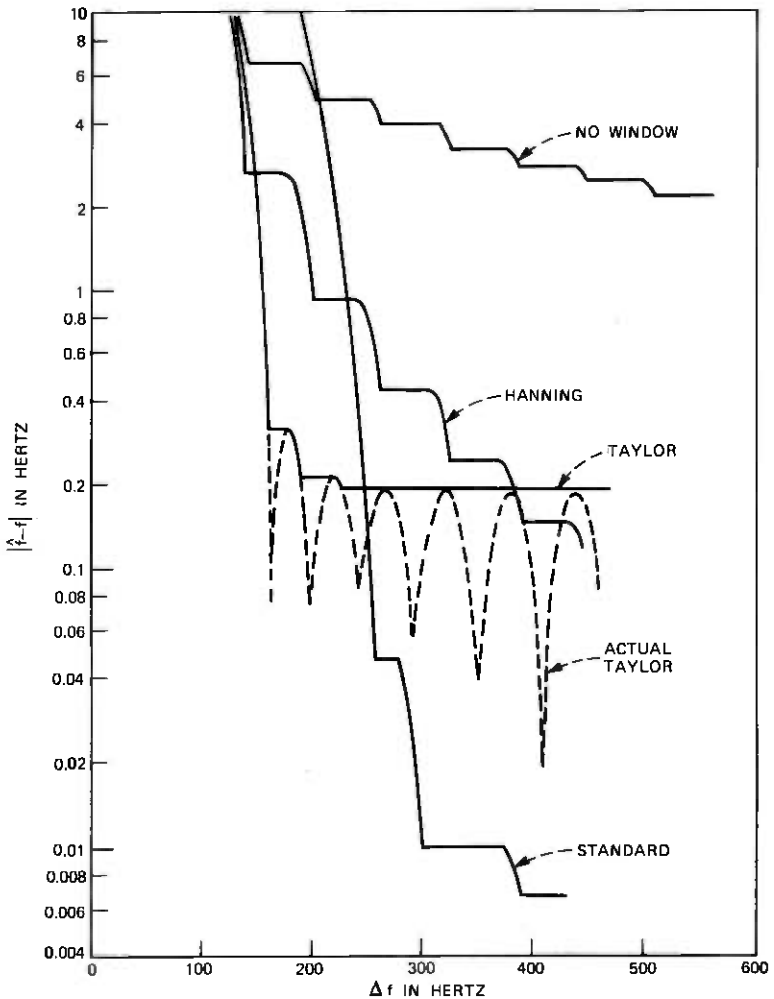


Fig. 5—Magnitude of frequency estimation bias for two equal-level complex tones using window functions. Curves are leveled as described in the text. Worst-phase was used at each frequency. N is 64.

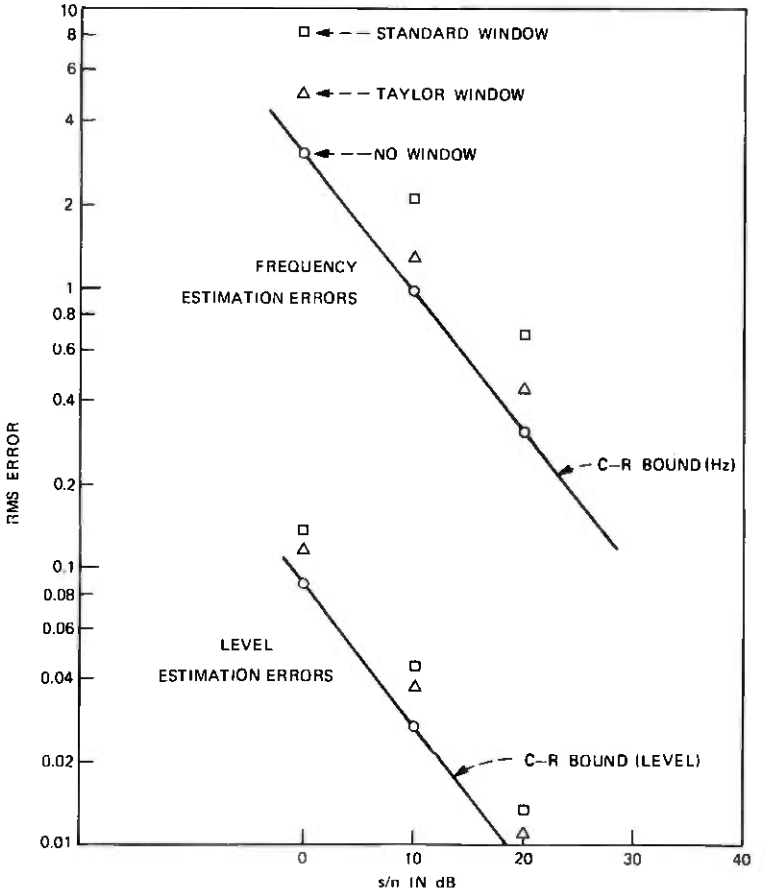


Fig. 6—Simulation results showing the effect of window functions on estimation variance with a single complex tone. N is 64.

while estimation variance controls RMS errors at low s/n. Thus, while a given window may produce lower RMS errors than another at high s/n, the roles may be reversed at low s/n. The "best" window for a given application will, therefore, depend upon the tone frequency spacings, the expected s/n, and possibly other factors. Figure 7 illustrates this point. On the figure, the Taylor window is best at 10 dB s/n, but the standard window is best at 40 dB s/n, where the bias associated with the Taylor window causes the RMS error curve to level off.

3.5 Interpolation

Maximization of $|A(\omega)|$ involves a search routine. A two-step algorithm that has a coarse search and a fine search was described in

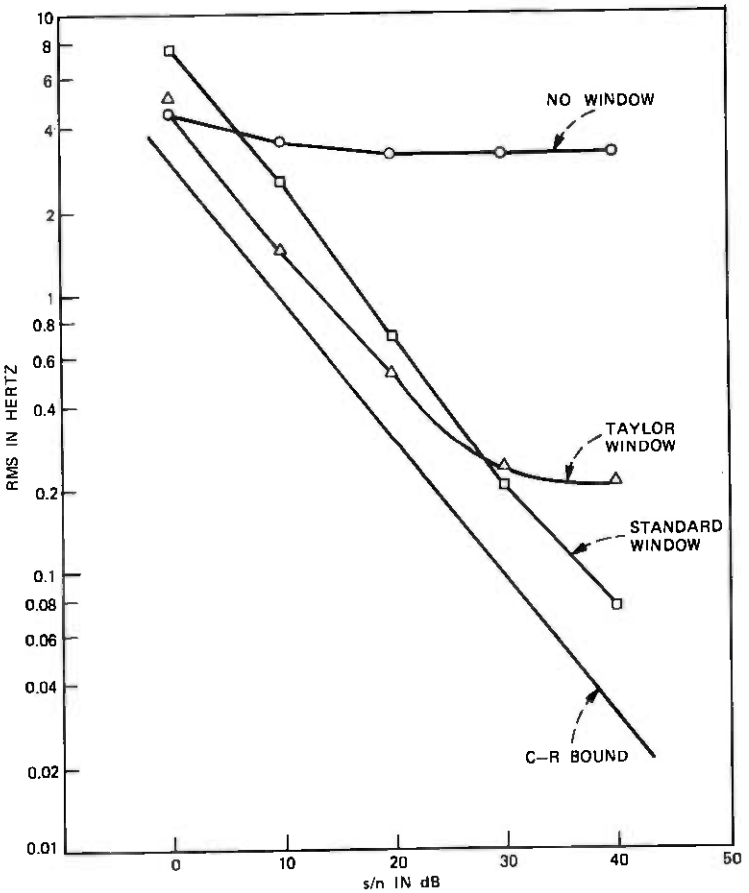


Fig. 7—Simulation results showing combined effects of bias and variance on frequency estimation for one of two complex tones. Frequency difference is 380 Hz; worst-phase was used for each window. N is 64.

Ref. 1. Fine searches are time consuming. This can be serious if computer time is important. One way to trade accuracy for speed is to use an interpolation algorithm on the DFT of the input data to arrive at frequency and level estimates.

Rife and Vincent developed several interpolation algorithms.[14] The one we investigate here is the following.

Assume the output of the FFT is the set:

$$A_k = \frac{1}{N}\sum_{n=0}^{N-1} h_n Z_n e^{-j2\pi nk/N}, \qquad k = 0 \text{ to } N-1. \tag{56}$$

Suppose a coarse search is conducted over $0 < k < N$. This results in locating $|A_l|$ which is the largest $|A_k|$ in the interval. Choose $\alpha = \pm 1$ such that $|A_{l+\alpha}| \geqq |A_{l-\alpha}|$

Let

$$a_1 = |A_l| \tag{57}$$

and

$$a_2 = |A_{l+\alpha}|. \tag{58}$$

Assume the sampling frequency is $\omega_s = 2\pi/T$.

The formulas from Rife and Vincent are:

$$\hat{\omega} = \frac{\omega_s}{N}(l + \alpha\delta) \tag{59}$$

and

$$\delta = \frac{2\pi a_1 X}{\sin(\pi X)\left[1 + \sum_{n=1}^{M} d_n \delta^2/(\delta^2 - n^2)\right]}, \tag{60}$$

where the d_n define a window and

$$\delta = \frac{C_1 a_2 - C_2 a_1}{C_3 a_2 + a_1}. \tag{61}$$

The numbers C_1, C_2, and C_3 are given by Rife and Vincent in Table II for several windows.

The interpolation formulas give estimates that are only a little worse than the fine search gives. RMS frequency errors are typically increased by about 30 percent when interpolation is used. RMS level errors increase less.

When many tones are present, window functions can provide a satisfactory reduction of leakage as long as the minimum frequency separation is no less than about $8\pi/NT$. The data on Fig. 8 illustrate this point. The tone phases were all made random for these simulations. Thus, the points indicate the RMS errors one might encounter in a working system. The bound shown on the figure is the (unbiased) C-R bound maximized over the possible phases of the center tone.

We consider a real-tone estimation system to be equivalent to a complex-tone system if the two systems have the same useful bandwidth and the same frequency resolution. This means (*i*) the real sampling frequency is twice the complex sampling frequency and (*ii*) the total sampling time, NT, is the same for the real tones as for the

Table II — Constants for computing delta in eq. (61)

Window	C_1	C_2	C_3
None	1	0	1
Hanning	2	1	1
Taylor	1.96339	1.01643	0.893534
Standard	3.6020	2.5862	1.0317

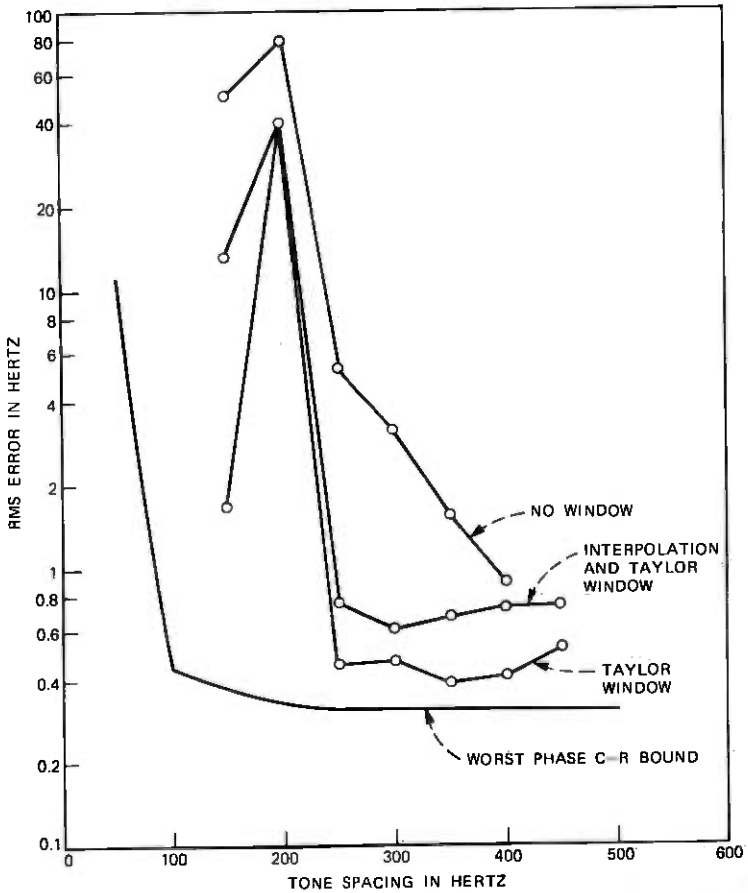


Fig. 8—Simulation results showing effects of Taylor window and interpolation algorithm upon frequency estimates of center tone of three equally spaced real tones. Center tone frequency is 2000 Hz. All have random phase. N is 128 and s/n is 20 dB.

complex. For example, a real-tone system using $1/T = 8000$ Hz and $N = 32$ is equivalent to a complex-tone system using $1/T = 4000$ Hz and $N = 16$.

The estimation algorithms described above for complex tones can be applied to real tones whose frequencies, in Hz, are in the range $(1/NT,\ 1/2T - 1/NT)$. The resulting accuracies are about the same as in the equivalent complex case.

IV. CONCLUSIONS

We have studied the problem of estimating the parameters, such as level and frequency, of several sinusoidal signals from a number of

noisy observations, taken at discrete-time instants. Gaussian noise and ideal analog-to-digital conversion were assumed. The nature of the problem led us to study the generalized Cramér-Rao lower bounds to estimation accuracy and maximum likelihood estimation. The complexity of maximum likelihood estimation algorithms led us to examine several algorithms that yield estimates that are almost, but not exactly, maximum likelihood estimates of the signal parameters.

We were able to obtain estimators that have negligible bias, at least at high s/n. Thus, we considered in detail only the generalized C-R bound for unbiased estimators. Even when the resulting numbers are not, strictly speaking, lower bounds (e.g., when an estimator is biased), the unbiased estimation bounds can be considered to be desirable objectives for estimators.

Several properties of the bounds were derived from the properties of the J matrix. Other properties, such as the existence of critical frequencies, were revealed from computations.

The J matrix in the real tone cases is more complicated than in the complex cases and does not have quite the same structural properties. Thus, for example, the lower bounds for a single complex tone are not also lower bounds for the equivalent single real tone. On the other hand, the bounds for the case of many real tones approach the bounds for the equivalent complex cases when none of the real tones have frequency differences less than $2/NT$, modulo $1/2T$ (in Hz).

The cases of many complex tones and of real tones present some difficulties. Maximum likelihood estimation is difficult to implement because of the presence of cross-product terms. To properly implement ML estimation, multidimensional search procedures over a nonconvex function would be necessary. We found that when the tone frequencies are separated far enough, the cross-product terms could be neglected, thereby permitting the use of a simple algorithm whose estimates are almost equal to ML estimates.

The penalty for dropping the cross-product terms is a bias in frequency and level estimates. We found that the use of a suitable window function will reduce the bias to the point where it can be neglected when the minimum frequency separation of the tones is $4/NT$. Three window functions were discussed and compared.

We found that the use of a window to reduce bias increased the variance of the frequency and level estimates. The RMS error of frequency estimates is increased by about 35 percent with Taylor window and by over 100 percent with standard window. The use of the interpolation formulas increases RMS frequency errors by another 30 percent or so. Level estimates are affected less by windows and interpolation. All of these figures apply when the s/n is above threshold.

REFERENCES

1. D. C. Rife and R. R. Boorstyn, "Single Tone Parameter Estimation from Discrete-Time Observations," IEEE Trans. Inform. Theory, *IT-20*, No. 5 (September 1974), pp. 591–598.
2. L. C. Palmer, "Coarse Frequency Estimation Using The Discrete Fourier Transform," IEEE Trans. Inform. Theory, *IT-20*, No. 1 (January 1974), pp. 104–109.
3. L. E. Brennan, "Angular Accuracy of a Phased Array Radar," IRE Trans. Ant. and Propag., *AP-9*, No. 3 (May 1961), pp. 268–275.
4. E. J. Kelly, I. S. Reid, and W. Root, "The Detection of Radar Echoes in Noise," Part II, J. Soc. Ind. Appl. Math., *8* (September 1960), pp. 481–507.
5. A. A. Ksienski and R. B. McGhee, "A Decision Theoretic Approach to the Angular Resolution and Parameter Estimation Problem for Multiple Targets," IEEE Trans. Aerosp. Electron. Syst., *4*, No. 3 (May 1968), pp. 443–455.
6. J. R. Sklar and F. C. Schweppes, "On the Angular Resolution of Multiple Targets," Proc. IEEE, *52*, No. 9 (September 1964), pp. 1044–45.
7. L. P. Seidman, *Design and Performance of Parameter Modulation Systems*, Ph.D. Dissertation, University of California, Berkley, 1966.
8. A. M. Wood and F. A. Graybill, *Introduction to the Theory of Statistics*, 2nd, ed., New York: McGraw-Hill, 1963.
9. S. Zacks, *The Theory of Statistical Inference*, New York: Wiley, 1971.
10. H. L. VanTrees, *Detection, Estimation, and Modulation Theory, Part I*, New York: Wiley, 1968.
11. D. Slepian, "Estimation of Signal Parameters in the Presence of Noise," Trans. IRE Prof. Group Inform. Theory, *PG IT-3*, *68* (March 1954), pp. 68–89.
12. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, New York: Hafner, 1961.
13. D. C. Rife, *Digital Tone Parameter Estimation in the Presence of Gaussian Noise*, Ph.D. Dissertation, Polytechnic Institute of Brooklyn, June 1973.
14. D. C. Rife and G. A. Vincent, "Use of the Discrete Fourier Transform in the Measurement of Frequencies and Levels of Tones," B.S.T.J., *49*, No. 2 (February 1970), pp. 197–228.
15. R. S. Durrani et al., "Data Windows for Digital Spectral Analysis," Proc. Inst. Elec. Eng., London, *119*, No. 3 (March 1972), pp. 343–352.
16. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, New York: Dover, 1959.

The Bell System Technical Journal
Vol. 55, No. 9, November 1976
Printed in U.S.A.

On the Rearrangeability of Some Multistage Connecting Networks

By F. K. HWANG

(Manuscript received July 24, 1975)

We generalize the concept of rearrangeability to a finer measure of the connecting power of a network, called the c-rearrangeability function. It can be interpreted as the proportion of calls a network guarantees to connect under a given traffic load. We study the c-rearrangeability function for many well-known rearrangeable networks, including one-sided rearrangeable, two-sided rearrangeable, as well as several other kinds. We also give constructions for some new classes of networks, study their c-rearrangeability functions, and describe conditions under which the networks are rearrangeable. We show that these newly constructed rearrangeable networks compare favorably with the well-known ones with respect to the number of crosspoints.

I. INTRODUCTION

A multistage connecting network can be described by the following (see Fig. 1 for a three-stage example):

(*i*) There are s ordered stages, where $s > 1$ is arbitrary. The ith stage, $i = 1, \cdots, s$, consists of r_i copies of a switch ν_i. The jth copy of ν_i is denoted by ν_{ij}.

(*ii*) Links can exist only between switches of adjacent stages or between $\nu_1(\nu_s)$ and input (output) terminals of the network. The set of links incident to a particular ν_i is partitioned into two subsets. Those which are linked to either ν_{i-1} or input terminals are called input links of ν_{ij} and those linked to either ν_{i+1} or output terminals are called output links.

(*iii*) The r_1 copies of ν_1 are called input switches of the network. Each ν_1 is connected to n_1 input terminals. The r_s copies of ν_s are called output switches of the network. Each ν_s is connected to n_s output terminals.

The three-stage Clos network is a special case of a multistage connecting network, satisfying the additional restrictions that $s = 3$ and

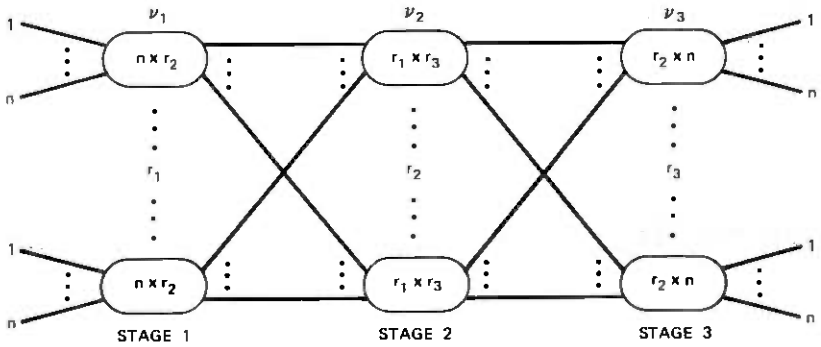


Fig. 1—Generalized three-step Clos network.

that there is exactly one link between every pair (ν_1, ν_2) and every pair (ν_2, ν_3). When ν_1, ν_2, ν_3 themselves are allowed to be multistage connecting networks, then a three-stage Clos network is called a generalized Clos network. For simplicity, we assume $n_1 = n_3 = n$ throughout this paper. Then, a generalized Clos network can be denoted by $C(\nu_1, \nu_2, \nu_3, r_1, r_2, r_3, n)$ (see Fig. 1).

Define a request to be a pair of idle terminals seeking connection. A request becomes a call once the two terminals are connected in the network. An assignment is a set of requests and the size of an assignment is the number of requests in it. An assignment is said to be realizable if every request in it can be simultaneously connected in the network without any link being used more than once. A network is said to be rearrangeable if it can realize every possible assignment.

Consider a multistage connecting network ν. Let $\mathcal{I}$ be the set of input terminals, $\mathcal{O}$ the set of output terminals of ν, and $\mathcal{I} + \mathcal{O} = T$. In many actual cases, not every possible pair in T will generate a request. In general, there could be two subsets $I, \Omega \subseteq T$ such that all requests are generated in the product space $I \times \Omega$. However, the four most important cases are:

(*i*) the one-sided case: $I = \Omega = T$.
(*ii*) the two-sided case: $I = \mathcal{I}$, $\Omega = \mathcal{O}$.
(*iii*) the input-mixed case: $I = \mathcal{I}$, $\Omega = T$.
(*iv*) the output-mixed case: $I = T$, $\Omega = \mathcal{O}$.

The last two are often combined and called the mixed case.

A network is said to be one-sided rearrangeable if it can realize every one-sided assignment. Similarly, we can define two-sided rearrangeable, input-mixed rearrangeable, and output-mixed rearrangeable. Thus, a one-sided rearrangeable network means that every set of pairs of terminals can be simultaneously connected, and a two-sided

rearrangeable network means that every set of (input link–output link) terminals can be simultaneously connected. It is clear that one-sided rearrangeability implies mixed rearrangeability, which, in turn, implies two-sided rearrangeability.

Rearrangeability is a strong condition which is manifested in two aspects. First, all the requests in an assignment must be simultaneously connected; i.e., if one request fails, the whole assignment fails. Second, every assignment must be realizable; i.e., if one assignment fails, the whole network fails. Even for a nonrearrangeable network, it is still of interest to know the degree of its nonrearrangeability. We introduce a new concept of rearrangeability in this direction. First, we score an assignment by the largest number of requests it can guarantee to connect simultaneously. Second, we partition the set of all assignments into classes according to the size of an assignment. We score a class by the lowest score achieved by any member in this class. Now, a bad assignment can still bring down the score of its class, but not of the other classes, and not to a score of zero. Thus, we define $R_\nu(c)$, the c-rearrangeability function, as the largest number of requests the network ν can guarantee to connect given any assignment of size c. Thus, $R_\nu(c)/c$ is the proportion of requests ν can guarantee to connect given that the traffic load is approximately $c/(\text{capacity of } \nu)$. When $R_\nu(c) = c$, we say ν is c-rearrangeable. If ν is c-rearrangeable for all c, then c is rearrangeable in the classical sense.

In this paper, we study the c-rearrangeability functions for some well-known rearrangeable networks. We also construct some new classes of networks, study their c-rearrangeability functions, and describe conditions under which the networks are rearrangeable. We show that these newly constructed rearrangeable networks can save a significant number of crosspoints over the well-known networks.

II. ANALYSES OF SOME WELL-KNOWN REARRANGEABLE NETWORKS

As switches are the basic components of a network, to understand the rearrangeable property of a network, we have first to know what the rearrangeable properties of its switches are (the switches mentioned in this paper are all cross-point grid switches). For a switch, the definition of rearrangeability is similar to that for networks, except that input links and output links replace the roles of input terminals and output terminals in a network.

Two links of a switch have direct access to each other if they intersect at a crosspoint. In many networks, the cost of crosspoints still dominates the other costs. Therefore, we would like to minimize the number of crosspoints in a network. A relevant question is, for a given rearrangeable property, which switch has the minimum number of

crosspoints? This problem has recently been solved by Chung.[1] However, in our networks, we will stick to the more traditional switches for their engineering feasibility and for ease of comparisons with existing networks. Consider a switch with n input links and m output links. It is called a triangular switch if there is a crosspoint between every pair of links, input or output. Therefore, a triangular switch has $[(n+m)(n+m-1)]/2$ crosspoints and is clearly one-sided rearrangeable. The switch is called a rectangular switch if there is a crosspoint between every input link and every output link.

A rectangular switch has $n \times m$ crosspoints and is two-sided rearrangeable. The switch is called a trapezoidal switch if there is a crosspoint between every pair of links with at least one of the links belonging to a fixed side. A trapezoidal switch either has $n(n-1)/2 + nm$ or $m(m-1)/2 + nm$ crosspoints and is either input-mixed or output-mixed rearrangeable, depending on which side is the fixed side. Note that an $n \times m$ rectangular switch is in fact input-mixed rearrangeable if $m \geqq n-1$. This is because any pair of input links can be connected through an output link, and there are always enough output links to do it. While the existing networks always use trapezoidal switches when mixed-rearrangeable switches are needed, we will use rectangular switches to save crosspoints when the condition $m \geqq n-1$ for input-mixed and $n \geqq m-1$ for output-mixed is met.

In every network we discuss in this paper, ν_1 and ν_s are always assumed to be two-sided rearrangeable (or stronger). Hence, two terminals from two distinct ν_1 and/or ν_s can be connected if and only if their corresponding switches $\nu_1(\nu_s)$ can be connected. Thus, we can redefine a request as a pair of ν_1 and/or ν_s and an assignment as a collection of requests where each ν_{1i} or ν_{sj} can appear at most n times. If a request is (ν_{1i}, ν_{1i}) or (ν_{sj}, ν_{sj}), then we have to discuss separately how they can be connected.

Consider $\nu = C(\nu_1, \nu_2, \nu_3, r_1, r_2, r_3, n)$ shown in Fig. 2. (In our figures, $\triangle$ represents a one-sided rearrangeable ν_i, $\square$ a two-sided rearrangeable

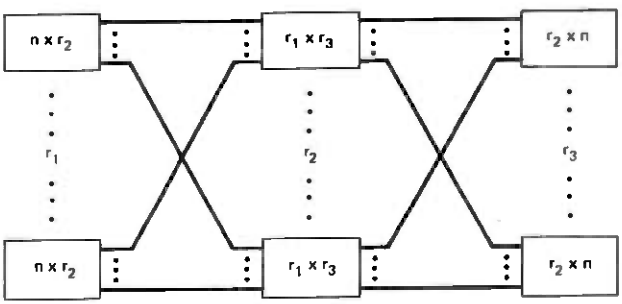


Fig. 2—Ordinary three-stage Clos network.

ν_i, ◿ an input-mixed rearrangeable ν_i, and ◺ an output-mixed rearrangeable ν_i. The number of input links and output links will be shown inside these figures.)

Theorem 1: (See Slepian,[2] Duguid,[3] and Beneš.[4]) *ν is two-sided rearrangeable if and only if $r_2 \geqq n$.*

Theorem 2:

$$R_\nu(c)/c \begin{cases} \cong \min\{r_2/n, 1\}, & \text{for } c > r_2; \\ = 1, & \text{for } c \leqq r_2. \end{cases}$$

Proof: Actually, we will prove an exact expression for $R_\nu(c)/c$. We need only consider the case $n > r_2$ and $c > r_2$ since the other cases are trivial. Consider any assignment of size c. Let $c = pn + q$ where $0 \leqq q < n$. We can assume that the r_2 real ν_2 are embedded in a set of n imaginary ν_2. Then by Theorem 1, all requests can be simultaneously connected by the n ν_2. Rank the n ν_2 according to the number of calls they carry and select the r_2 ν_2 with the highest ranks to be the r_2 real ones. They must carry a total number of calls not less than $\min\{r_2, q\} \times (p + 1) + \max\{r_2 - q, 0\} \times p$. On the other hand, when all the requests in an assignment involve only a few ν_1, say as few as possible, then every ν_2 carries essentially the same number of calls, differing at most by one. Hence,

$$R_\nu(c)/c = \begin{cases} (r_2p + q)/(np + q), & \text{if } r_2 \geqq q; \\ (r_2p + r_2)/(np + q), & \text{if } r_2 < q. \end{cases}$$

Theorem 2 gives a good approximation to this when c and r_2p are large relative to q.

To compute the number of crosspoints of a network, we always make the simplifying assumptions that $r_1 = r_3 = n$ and all ν_i are nonblocking switches so that we can easily compare the various networks. Under these assumptions, then, the current network for $r_2 = n$ has $3n^3$ crosspoints.

Next consider $\nu_2 = C(\nu_1, \nu_2, \nu_3, r_1, r_2, r_3, n)$ shown in Fig. 3a.

Theorem 3:[5–7] *ν is one-sided rearrangeable if and only if $r_2 \geqq \lfloor 3n/2 \rfloor$, where $\lfloor x \rfloor$ is, as usual, the integer part of x.*

Theorem 4:

$$R_\nu(c)/c \begin{cases} \cong \min\left\{r_2 \Big/ \left\lfloor \frac{3n}{2} \right\rfloor, 1\right\}, & \text{for } c > r_2; \\ = 1, & \text{for } c \leqq r_2. \end{cases}$$

Proof: Again, we need only consider the case $\lfloor 3n/2 \rfloor > r_2$ and $c > r_2$. The proof that $R_\nu(c)/c \geqq r_2/\lfloor 3n/2 \rfloor$ uses a similar argument to that

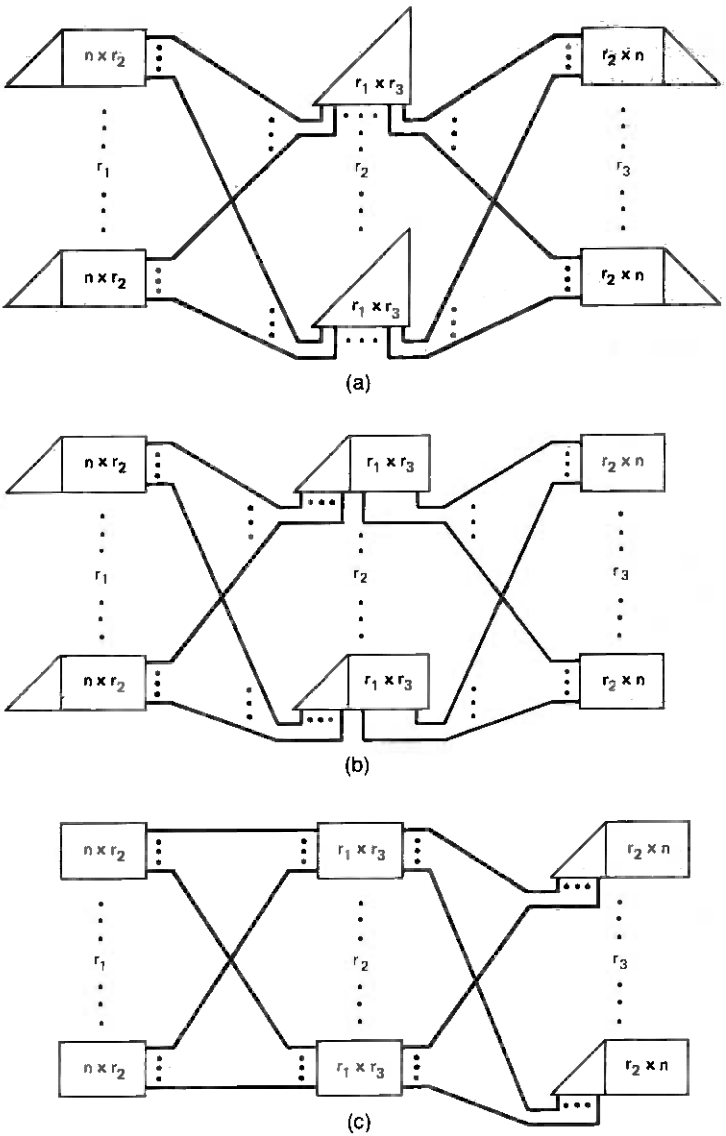


Fig. 3—Mixed three-stage Clos networks.

in the proof of Theorem 2. To prove the reverse inequality, label all the ν_1 and ν_3 by the numbers 1 to $r_1 + r_3$. Consider the $\lfloor 3n/2 \rfloor$ requests,

$$(1, 2), \cdots, (1, 2), \quad (\lfloor n/2 \rfloor \text{ of them}),$$
$$(1, 3), \cdots, (1, 3), \quad (\lfloor n/2 \rfloor \text{ of them}),$$
$$(2, 3), \cdots, (2, 3), \quad (\lfloor (n + 1)/2 \rfloor \text{ of them}).$$

No two of them can be carried by the same ν_2, since otherwise the two will share a link. Consider an assignment whose requests can be partitioned into sets of $\lfloor 3n/2 \rfloor$ requests of the above type and a remainder set which is a subset of the above type of $\lfloor 3n/2 \rfloor$ requests. Then the number of calls carried by each ν_2 can differ at most by one. Hence, $R_\nu(c)/c \leqq r_2/\lfloor 3n/2 \rfloor + (\text{a constant})/c \cong r_2/\lfloor 3n/2 \rfloor$. The proof is completed.

Under the simplifying assumptions previously stated, the number of crosspoints for this network for $r_2 = \lfloor 3n/2 \rfloor$ is $7n^3$. Note that this compares favorably with the one-sided network[8] obtained from a two-sided rearrangeable network where both sides are $\mathcal{I} + \mathcal{O}$. Such a network needs $12n^3$ crosspoints. It is also better than the one obtained by joining three two-sided networks together, the first having $\mathcal{I}$ in both sides, the second having $\mathcal{O}$ in both sides, and the third having $\mathcal{I}$ in one side and $\mathcal{O}$ the other. Such a network needs $9n^3$ crosspoints.

We can easily obtain an input-mixed rearrangeable network from the above one-sided rearrangeable network by changing ν_2 from one-sided rearrangeable to input-mixed rearrangeable and ν_3 from output-mixed rearrangeable to two-sided rearrangeable (Fig. 3b). Let ν be the network shown in Fig. 3b.

Theorem 5:

$$R_\nu(c)/c \begin{cases} \cong \min\left\{ r_2 \Big/ \left\lfloor \frac{3n}{2} \right\rfloor, 1 \right\}, & \text{for } c > r_2, \\ = 1, & \text{for } c \leqq r_2. \end{cases}$$

Proof: The proof is similar to the proofs for Theorems 3 and 4.

For $r_2 = 3n/2$, this network has $(23/4)n^3$ crosspoints. However, we can also obtain an input-mixed rearrangeable network by joining two two-sided rearrangeable networks together; one is $(\mathcal{I}, \mathcal{O})$-two-sided and the other has $\mathcal{I}$ in both sides. Such a network needs $6n^3$ crosspoints.

III. A NEW INPUT-MIXED REARRANGEABLE NETWORK

Consider $\nu = C(\nu_1, \nu_2, \nu_3, r_1, r_2, r_3, n)$ shown in Fig. 3c.

Theorem 6: *ν is input-mixed rearrangeable if*

(*i*) $r_2 \geqq n$,
(*ii*) $(r_2 - 1)r_3 \geqq nr_1$.

Proof: We first explain how a request is connected in this network. A (ν_{1i}, ν_{3j}) request is still connected through some ν_2 which has an idle link to ν_{1i} and an idle link to ν_{3j} just as is done in the networks of Section II. But a (ν_{1i}, ν_{1j}) request cannot be connected in this manner

since ν_2 cannot connect two input links. Instead, we will connect both ν_{1i} and ν_{1j} to some ν_{3k} and then use the input-mixed rearrangeable property of ν_3 to complete the connection. One question is whether there are enough input links of ν_3 to accommodate all (ν_1, ν_1) requests. Now each (ν_1, ν_3) request takes up one input link of ν_1 and one of ν_3, and each (ν_1, ν_1) request takes up two input links of ν_1 and two of ν_3. Hence, regardless of the distribution of (ν_1, ν_1) requests relative to the (ν_1, ν_3) requests, the maximum number of ν_3 input links needed is, except for a minor correction, the maximum number of ν_1 input links available, which is nr_1. The minor correction is because each (ν_1, ν_1) request takes up a pair of ν_3 input links from the same ν_3. Hence, occasionally, a ν_3 input link may be wasted since it has no partner. Discounting one input link from each ν_3, we obtain condition (*ii*).

If condition (*ii*) is satisfied, then each (ν_{1i}, ν_{1j}) request can be replaced by two requests (ν_{1i}, ν_{3k}) and (ν_{1j}, ν_{3k}). Hence, an input assignment is turned into a two-sided assignment. The requirement that each ν_{3k} must appear no more than n times is irrelevant here because the connection of (ν_{1i}, ν_{3k}) does not involve any output links of ν_{3k}. By Theorem 1, the derived two-sided assignment is rearrangeable if $r_2 \geqq n$. Theorem 6 is proved.

If $r_1 = r_3$, then $r_2 = n + 1$ satisfies both conditions of Theorem 6. Furthermore, since the size is right, we can use rectangular switches for ν_3 for mixed-rearrangeable property. This network has $3n^3 + 3n^2$ crosspoints (under the simplifying assumptions) as compared to $(23/4)n^3$ for the input-mixed rearrangeable network in Section II.

Since a (ν_{1i}, ν_{1j}) request takes twice as many links to connect as a (ν_{1i}, ν_{3j}) request, one might suspect that the blocking probability for the former request is much larger. This is not necessarily true, however, since there are only r_2 distinct connecting paths of two links for a (ν_{1i}, ν_{3j}) request but $\binom{r_2}{2} r_3$ paths of four links for a (ν_{1i}, ν_{1j}) request.

Theorem 7: *Let ν be the network in Theorem 6. Then*

$$R_\nu(c)/c \cong \min\left\{1, \frac{r_2}{n}\right\} \times \min\left\{1, \frac{\max\{2c - nr_1, 0\} + (r_2 - 1)r_3}{2c}\right\}.$$

Proof: For the time being, suppose $r_2 = n$. Consider any assignment of size c and let u be the number of (ν_1, ν_3)-type requests in it. Then $u \geqq \max\{2c - nr_1, 0\}$ since $u + 2(c - u) = 2c - u$ input links of ν_1 are required while only nr_1 are available. If there are not enough input links of ν_3 to take care of all c requests, then priority should be given to (ν_1, ν_3) type requests to maximize the number of requests connected. The priority is due to the fact that a (ν_1, ν_3) request needs only one input link of ν_3 while a (ν_1, ν_1) request needs two. For $u \geqq r_2 r_3$, the

maximum number of requests connectable is r_2r_3; for $u < r_2r_3$, the maximum is approximately

$$u + \frac{(r_2 - 1)r_3 - u}{2} = \frac{u + (r_2 - 1)r_3}{2} < r_2r_3.$$

The worst case occurs when u is at its minimum, i.e., $u = \max \times \{2c - nr_1, 0\}$. But still the network guarantees to connect at least $\frac{1}{2}$ $[\max (2c - nr_1, 0\} + (r_2 - 1)r_3]$ requests. Now look at the distribution of these calls in the ν_2. If $r_2 < n$, select the $r_2\nu_2$ that carry the most calls. In this way, we obtain Theorem 7.

Corollary: *ν is input-mixed c-rearrangeable if $r_2 \geqq n$ and either*

$$c \leqq \frac{(r_2 - 1)r_3}{2} \quad \text{or} \quad c \geqq \frac{(r_2 - 1)r_3}{2} \geqq \frac{nr_1}{2}.$$

IV. A NEW ONE-SIDED REARRANGEABLE NETWORK

Consider $\nu = C(\nu_1, \nu_2, \nu_3, r_1, r_2, r_3, n)$, where ν_1 and ν_3 are input- and output-mixed rearrangeable and ν_2 is one-sided rearrangeable. Also assume n and r_2 are even. We construct a ν' from ν by inserting something between the pair $(\nu_{2,2i-1}, \nu_{2,2i})$ for each $i = 1, \cdots, r_2/2$ (see Fig. 4) to provide some limited access between links of $\nu_{2,2i-1}$ and links of $\nu_{2,2i}$. One way to do this is to insert two two-sided rearrangeable networks μ_{i1} and μ_{i2} between $\nu_{2,2i-1}$ and $\nu_{2,2i}$. The input links of μ_{i1} are the extensions of the n links of $\nu_{2,2i-1}$ and the output links of μ_{i2} are the extensions of the n links of $\nu_{2,2i}$. μ_{i1} has $\frac{1}{3}(r_1 + r_3)$ output links that become the input links of μ_{i2}. Thus any link of $\nu_{2,2i-1}$ can seize an output link of μ_{i1} and then connect to any link of $\nu_{2,2i}$ in μ_{i2}. Of course, $\frac{1}{3}(r_1 + r_3)$ such connections can be made simultaneously.

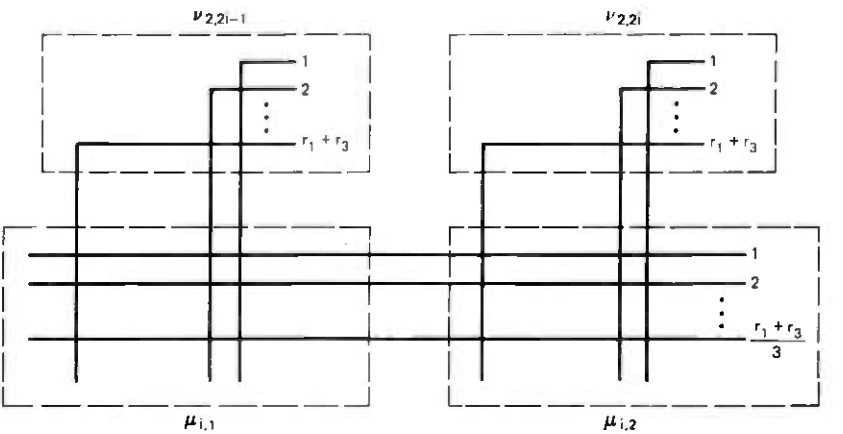


Fig. 4—One-sided rearrangeable networks.

Theorem 8: ν *is one-sided rearrangeable if and only if* $r_2 \geqq n$.

Proof: For a given assignment, define an assignment graph by taking all ν_1 and ν_3 as vertices and every request as an edge. We can augment the assignment graph to become a regular graph of degree n by adding suitable edges to it. By a theorem of Petersen (see Refs. 9 or 10), a regular graph of even degree is 2-factorable; i.e., the assignment graph can be decomposed into $(n/2)$ 2-factors, where a 2-factor is a subgraph in which every vertex is of degree 2. Hence, a 2-factor consists of a set of disjoint circuits. Now any circuit of length 1 represents a request from two terminals of the same switch. We can connect them within that switch because of its mixed-rearrangeable property. Aside from that, we can partition all edges in an odd circuit into three sets such that edges in the same set are all disjoint; and we can partition all edges in an even circuit into two such sets. Since all circuits in a 2-factor are disjoint, we can combine those sets into three large sets A_1, A_2, A_3, such that the edges in each large set are all disjoint. As each edge represents a request, all the requests in A_1 can be connected through, say, ν_{21}, since they are all disjoint (we can ignore those edges which are augmented to the assignment graph). Similarly, all the requests in A_2 can be connected through ν_{22}. For a request in A_3, say (x, y), if (x, y) is disjoint with every request in $A_1(A_2)$, then we can connect it through $\nu_{21}(\nu_{22})$. Otherwise, suppose x has appeared in A_1 and y in A_2. Then we connect x to ν_{22} and y to ν_{21} and then connect them through μ_{11} and μ_{12}. We do this for every request in A_3. Therefore, all requests in a 2-factor can be connected by a pair of ν_2. There are $n/2$ 2-factors; hence, $n/2$ pairs of ν_2 will suffice. If we have less than $n/2$ pairs of ν_2, then there is no way to handle the $3n/2$ requests given in the proof of Theorem 4. Hence, Theorem 8 is proved.

Theorem 9:

$$R_\nu(c)/c \begin{cases} \cong \min\left\{\frac{r_2}{n}, 1\right\}, & \text{for } c > \frac{3r_2}{2}, \\ = 1, & \text{for } c \leqq \frac{3r_2}{2}. \end{cases}$$

Proof: Omitted.

For $r_2 = n$, this network has $(16/3)n^3$ crosspoints versus the $7n^3$ for the standard one-sided rearrangeable network.

V. A c-REARRANGEABILITY THEOREM

We have seen that the c-rearrangeability functions of many networks discussed in previous sections are such that $R_\nu(c)/c \cong \alpha$, a constant, over most of the range of c. Consider $\nu = C(\nu_1, \nu_2, \nu_3, r_1,$

r_2, r_3, n) such that $R_{\nu_i}(c)/c \cong \alpha_i$ for two-sided assignment. What can we say about $R_\nu(c)$ for two-sided assignment? If the blocking in ν_1, ν_2, ν_3 and the blocking due to the structure of ν all act independently, then we should have

$$R_\nu(c) \cong \alpha_1\alpha_2\alpha_3 \min\left\{1, \frac{r_2}{\max\{\alpha_1, \alpha_3\} \times n}\right\}.$$

The reason for the last term is because at most $\max\{\alpha_1, \alpha_3\} \times n$ requests from the same switch can get through to the second stage. However, we show that the blocking in ν_1 and the blocking in ν_3 can be coordinated so that the requests blocked in one stage form a subset of the requests blocked in the other stage. Without loss of generality, suppose $\alpha_1 \leqq \alpha_3$. Then

Theorem 10:

$$R_\nu(c) \geqq \alpha_1\alpha_1 \min\left\{1, \frac{r_2}{\alpha_1 n}\right\}.$$

Proof: For any given assignment, consider its bipartite assignment graph G. Let d_i be the degree of vertex i. We want to find a subgraph G' such that the degree of vertex ν in G' is $d_i' = \alpha_1 d_i$ (treating it as an integer). Then G' is the set of requests that will get through both ν_1 and ν_3.

Let V_1 be the set of vertices corresponding to ν_1 and V_3 the set corresponding to ν_3. Let $d(X)$ denote the sum of degrees over all vertices of X in G, and define $d'(X)$ similarly for X in G'. Finally, let $d_Y(X)$ denote the degree sum of X in G when the set Y is deleted from G. Then a theorem of Gale[11] on network flows has the following interpretation.[12]

Gale's Theorem: *G' exists if and only if there do not exist two sets $S \subseteq V_1$, $T \subseteq V_3$ such that either*

$$d'(S) > d'(T) + d_T(S),$$

or

$$d'(T) > d'(S) + d_S(T).$$

In our case, $d'(S) = \alpha_1 d(S)$ and $d'(T) = \alpha_1 d(T)$. Without loss of generality, suppose $d(S) \geqq d(T)$. Then the second inequality in Gale's theorem certainly cannot hold. To check the first, note that

$$d_T(S) \geqq d(S) - d(T) \geqq \alpha_1 d(S) - \alpha_1 d(T) = d'(S) - d'(T).$$

Hence, the first inequality also does not hold. We conclude that G' exists and Theorem 10 is proved.

When the involved numbers are large, the discrepancy caused by assuming $\alpha_i d_i$ an integer is certainly negligible.

VI. ACKNOWLEDGMENT

The author is grateful to his colleagues for many useful comments on his manuscript.

REFERENCES

1. F. R. K. Chung, "Optimal Rearrangeable Graphs," B.S.T.J., *54* (November 1975), pp. 1647–1661.
2. D. Slepian, unpublished work, March 25, 1952.
3. A. M. Duguid, "Structural Properties of Switching Networks," Brown University Progress Report, BTL-7 (1959).
4. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
5. C. E. Shannon, "A Theorem on Coloring the Lines of a Network," J. Math. Phys., *28* (1949), pp. 148–151.
6. M. C. Paull, "Rearrangeability of Triangular Networks," Proc. Seventh Annual Allerton Conference on Circuit and System Theory, 1969, pp. 487–496.
7. L. A. Bassalygo, I. I. Grushko, and V. I. Neiman, "Some Theorems on Structures of One-Sided Networks for Simultaneous Connection," Problemy peredachi informatsii, *5* (1969), pp. 45–52.
8. A. Zaronni, U. S. Patent 3,041,409, Switching System, June 26, 1962.
9. C. Berge, *Graphs and Hypergraphs*, New York: American Elsevier, 1973.
10. J. Petersen, "Die Theorie der regulären Graphen," Acta. Math., *15* (1891), pp. 193–220.
11. D. Gale, "A Theorem of Flows in Networks," Pac. J. Math., *7* (1957), pp. 1073–1082.
12. F. K. Hwang, "An Inductive Proof of a Theorem of Gale," Bull. Inst. Math., Academia Sinica, *3*, No. 1 (1975), pp. 81–85.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 9, November 1976
Printed in U.S.A.

Optimum Quantizer Design Using a Fixed-Point Algorithm

By A. N. NETRAVALI and R. SAIGAL

(Manuscript received June 10, 1976)

A fixed-point algorithm has been used to obtain the parameters (i.e., decision and representative levels) of an "optimum" quantizer that minimizes a quite general distortion measure, subject to an entropy constraint on its output. Construction of the algorithm starts with a point-to-set mapping whose fixed point satisfies the well-known Karush-Kuhn-Tucker conditions necessary for a local extremum. A computer program is then used to determine a fixed point of this mapping. Several examples are solved, and correspondence with the existing results in the literature is pointed out. Finally, as conjectured, the growth of the computations as a function of dimensionality n (n: number of representative levels) is found to be of the form $a \cdot n^b$ where a is a positive constant and $1.5 \leqq b \leqq 2.0$.

I. INTRODUCTION

Simple quantization[1-3] has been and continues to be a popular method of digitizing analog signals. The relative ease with which quantizers can be implemented in hardware and their near optimum performance has made them withstand the challenge from several new coding schemes.[4-6] Universal use of quantizers has naturally spurred a significant activity in optimizing their performance, some of which is summarized in the next few paragraphs. Our objective in this paper is to show how the problem of obtaining the parameters of an optimum quantizer can be converted to the problem of obtaining fixed points of a suitably constructed mapping and then to use a fixed-point algorithm to solve the problem numerically.

Quantizers have been optimized based on several criteria. In order to discuss these in relation to the problem considered in this paper, we describe the basic quantizer equations. Given a scalar random variable T with probability density $p(t)$, a quantizer Q is a map $Q(t) = y_i$ whenever $x_i \leqq t < x_{i+1}$, where x_i, $i = 1, \cdots, N+1$ and y_i, $i = 1, \cdots, N$ are the decision and representative levels of the quantizer, respectively. The performance of the quantizer is judged generally in terms of two quantities:

the distortion

$$D = \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} g(t - y_i) \times f(t)dt, \tag{1}$$

and the entropy

$$\mathcal{E} = -\sum_{i=1}^{N} (\log_2 p_i) \times p_i, \tag{2}$$

where g is a nonnegative function and f is a nonnegative weighting function that weights the quantization noise and

$$p_i = \int_{x_i}^{x_{i+1}} p(t)dt.$$

Optimum quantizers choose their parameters $\{x_i\}$, $i = 2, \cdots, N$ and $\{y_i\}$, $i = 1, \cdots, N$ (given the end points x_1, x_{N+1}) to optimize a certain combination of D and $\mathcal{E}$.

Most quantization literature uses the weighting function f to be the same as the probability function p, although in some applications[7–9] a different weighting function performs better. Most of the earlier work is concerned with minimizing D for a given number of levels. Panter and Dite[10] have used $g(\cdot) = |(\cdot)|^r (r > 0)$ and obtained an approximate optimum quantizer as one in which each of the quantizing intervals $[x_i, x_{i+1}]$ makes an equal contribution to the integral of $|(t - y_i)|^r$. This allowed them to choose the quantizer parameters for large N. Lloyd[11] and Max[12] have developed an algorithm for $r = 2$, which corresponds to minimizing the mean square error. Bruce[13] has used dynamic programming to solve the same problem in slightly more generality by taking a general function $g(\cdot)$. Simpler suboptimal algorithms and bounds on the performance of the quantizers have been obtained by Roe,[14] Algazi,[15] and Zador.[16]

Representation of the quantizer output by a variable length code allows reduction of the average bit rate of the quantizer when p_i varies with i. Use of Huffman code[17] makes the average bit rate approach the entropy of the quantizer output. Thus, the problem of designing an optimum[18–19] quantizer can be reformulated as that of obtaining the decision and representative levels to minimize D subject to a constraint on the entropy. Goblick and Holsinger have considered this problem for uniform quantizers and have concluded that for gaussian density, for $r = 2$, and for the same distortion, the entropy of the output of the uniform quantizer is higher than the theoretical lower bound based on the rate distortion theory by about $\frac{1}{4}$ bit. Uniform quantizers are also good in an asymptotic sense, since they are optimum for a large number of levels.[21] Moreover, for Laplacian densities, as shown by Berger,[20]

uniform quantizers are optimum for any value of entropy. A different type of distortion measure has been considered by Elias.[22]

The problem we consider is that of obtaining the parameters of quantizers such that D is minimized for a given constraint on the entropy. Although the approach taken here is suitable for a general distortion measure of eq. (1), we consider only the case of $g(\cdot) = (\cdot)^2$, mainly to compare our results to those in the literature. In the next section, we present the necessary conditions that the optimum quantizer must satisfy for a local extremum. Then, in Section III, we construct a point-to-set mapping such that its fixed point satisfies the necessary conditions for a local extremum of our problem. A description of the algorithm is then presented for completeness. In Section IV, we present the results of use of this algorithm for uniform, Laplacian, and gaussian densities. The distortion-entropy curves are presented for each case. We also present a surprising observation on the growth of computations as a function of dimensionality (i.e., the number of quantizer parameters to be optimized).

II. FORMULATION OF THE PROBLEM AND NECESSARY CONDITIONS

Using $g(\cdot) = (\cdot)^2$, the distortion of eq. (1) becomes

$$D = \sum_{j=1}^{N} \int_{x_j}^{x_{j+1}} (t - y_j)^2 f(t) dt. \tag{3}$$

Then the problem is to obtain $\{x_j\}$, $j = 2, \cdots, N$, $\{y_j\}$, $j = 1, \cdots, N$ such that they minimize D subject to $\mathcal{E} \leqq K$, for a given N. The necessary conditions from the Karush-Kuhn-Tucker theory[23] are that there exists a $\lambda \geqq 0$ such that

$$\nabla D(x) + \lambda \nabla \mathcal{E}(x) = 0, \tag{4}$$

where x is a vector of quantizer parameters and ∇ denotes the gradient. For the parameters $\{y_j\}$, since $\mathcal{E}$ is independent of $\{y_j\}$, (4) becomes

$$y_j = \int_{x_j}^{x_{j+1}} t f(t) dt \Big/ \int_{x_j}^{x_{j+1}} f(t) dt, \qquad j = 1, \cdots, N. \tag{5}$$

This implies that the representative levels can be obtained explicitly by knowing the decision levels and therefore they do not add to the dimensionality of the problem. Also, the other necessary conditions are

$$\mathcal{E} \leqq K$$

and

$$\lambda(\mathcal{E} - K) = 0. \tag{6}$$

III. FIXED-POINT APPROACH

In this section, we formulate the quantization problem as a fixed-point problem and give a general description of the algorithm that

solves this problem. This algorithm is based on the theory of complementary pivoting.[24]

Given a point-to-set mapping Γ [i.e., to each point x in R^n it associates a subset $\Gamma(x)$ of R^n], a fixed point of such a mapping is a point x such that $x \in \Gamma(x)$. We show that the problem of finding the parameters of the optimal quantizer can be formulated as a problem of finding a fixed point of a certain point-to-set mapping.

3.1 Fixed-point formulation

Let ∇D and $\nabla \mathcal{E}$ be the gradient vectors of the distortion D and entropy $\mathcal{E}$, respectively. Then, consider the following point-to-set mapping:

$$\Gamma(x) = \begin{cases} x - \{\nabla D(x)\} & \mathcal{E}(x) < K \\ x - \text{hull}\,\{\nabla D(x), \nabla \mathcal{E}(x)\} & \mathcal{E}(x) = K, \\ x - \{\nabla \mathcal{E}(x)\} & \mathcal{E}(x) > K \end{cases} \tag{7}$$

where hull $\{E\}$ is the smallest convex set containing E; i.e., the convex hull of E, and $x - A = \{x - y : y \in A\}$ for a set A in R^n. Note that the mapping as defined is upper semicontinuous* (u.s.c.) and the set $\Gamma(x)$ is convex for each x. As we subsequently see, these properties are needed if the algorithm is to find a fixed point of Γ.

We now show that a fixed point of this mapping satisfies the necessary conditions of Section 2.

Theorem: *Let $x \in \Gamma(x)$. Then, if $\mathcal{E}(x) \leqq K$, x satisfies the necessary conditions of Section 2. Otherwise, x is a local minimizer of $\mathcal{E}(x)$.*

Proof: We construct the required λ and show that (6) is satisfied. Since $x \in \Gamma(x)$ and $\mathcal{E}(x) \leqq K$, we have two cases:

Case (i): $\mathcal{E}(x) < K$. Let $\lambda = 0$ and, since $0 \in \{\nabla D(x)\}$, $\nabla D(x) + \lambda \nabla \mathcal{E}(x) = 0$, satisfying (6). Note that $\lambda[\mathcal{E}(x) - K] = 0$.

Case (ii): $\mathcal{E}(x) = K$. Then, as $0 \in \text{hull}\,\{\nabla D(x), \nabla \mathcal{E}(x)\}$, there exist $\lambda_1 + \lambda_2 = 1$, $\lambda_1 \geqq 0$, $\lambda_2 \geqq 0$ such that

$$\lambda_1 \nabla D(x) + \lambda_2 \nabla \mathcal{E}(x) = 0. \tag{8}$$

Now, in case $\lambda_1 \neq 0$, letting $\lambda = \lambda_2/\lambda_1 \geqq 0$, (4) is satisfied, and $\lambda[\mathcal{E}(x) - K] = 0$. In the contrary case, a constraint qualification would be violated.

In case $x \in \Gamma(x)$ and $\mathcal{E}(x) > K$, then, since $0 \in \{\nabla \mathcal{E}(x)\}$, $\nabla \mathcal{E}(x) = 0$ and we have a local minimizer of $\mathcal{E}(x)$. If $\mathcal{E}(x)$ were a convex function, our problem has no feasible solution [i.e., an x such that

* A mapping Γ is u.s.c. if, for any two sequences $\{x^k\}$, $\{y^k\}$ such that $x^k \to x$, $y^k \in \Gamma(x^k)$, and $y^k \to y$, we have $y \in \Gamma(x)$.

$\mathcal{E}(x) \leqq K]$. In the contrary case, we would conclude the algorithm has failed.

3.2 Description of the algorithm

In this section we give a brief description of the algorithm that computes fixed points of point-to-set mappings. Before going into the details of the algorithm, we introduce some notation.

Given a set C in R^n, and a point-to-set mapping Γ, by $\Gamma(C)$ we represent the set $\bigcup_{x \in C} \Gamma(x)$. Also, given a one-to-one linear mapping r, we say a set C is

(*i*) Γ—complete if $0 \in$ hull $\{\Gamma(C)\}$,

(*ii*) r—complete if $0 \in$ hull $\{r(C)\}$, and

(*iii*) $\Gamma \cup r$—complete if $0 \in$ hull $\{\Gamma(C) \cup r(C)\}$.

The significance of Γ-complete sets is the following: in case Γ is u.s.c. and $\Gamma(x)$ is convex for each x, a sequence C_i, $i = 1, 2, \cdots$ of Γ-complete sets whose diameter approaches 0 as i approaches ∞, converges to a fixed point of Γ (see, for example, Refs. 25–27). The fixed-point algorithms are designed to find such a sequence of Γ-complete sets.

These algorithms work with sets C that are simplexes of appropriate dimension. (An n-dimensional simplex is a convex body obtained by taking the convex hull of $n + 1$ affinely independent points in n-space. A two-dimensional simplex is a triangle; a three-dimensional simplex is a tetrahedron.) They start with a unique r-complete simplex and generate a sequence of $\Gamma \cup r$-complete simplexes that terminate with a Γ-complete simplex. There are essentially two basic algorithms that can be used to generate a sequence of Γ-complete simplexes of decreasing diameters. They are the *restart method* of Merrill[27] and the *continuous deformation method* of Eaves and Saigal.[26] A study of both these methods can be found in Saigal.[28–30]

We now discuss an application of the algorithm. A real number $d > 0$ is chosen. Then the space $R^n \times [0, d]$ is triangulated (i.e., each point in the space lies in an $(n + 1)$-dimensional simplex, and these simplexes overlap only on their boundaries) such that the vertices of the triangulation are only in the set $R^n \times \{d/2^k\}$, $k = 0, 1, \cdots$. In addition, the diameter of each n-dimensional face of each $(n + 1)$-dimensional simplex that lies in $R^n \times [d/2^{k+1}, d/2^k]$ is at most $d/2^k$. Now, an arbitrary starting point x_0 is chosen. We then define

$$r(x) = -x + x_0, \tag{9}$$

which is a one-to-one linear mapping.

The sequence of $\Gamma \cup r$-complete simplexes is then generated as

follows:

Step 1: Start with an r-complete simplex in the triangulation that contains (x_0, d). The triangulation is arranged in such a way that there is a unique such simplex, and that this simplex has exactly one vertex in $R^n \times \{d/2\}$, and $(n+1)$ vertices in $R^n \times \{d\}$. The entering vertex is the one in $R^n \times \{d/2\}$. Design the labeling function L on the vertices of the triangulation with

$$L(x, t) = \begin{cases} z - x & \text{for some } z \in \Gamma(x) \quad \text{if} \quad t < d \\ x^0 - x & \text{if} \quad t = d \end{cases} \tag{10}$$

Step 2: Find the label on the entering vertex.

Step 3: Find a new $\Gamma \cup r$-complete simplex that includes the entering vertex, in place of some vertex of the older simplex. This is equivalent to the basic pivot operation of the simplex method.[31]

Step 4: Find the other $(n+1)$-dimensional simplex that contains the new $\Gamma \cup r$-complete simplex found in Step 3, and determine the entering vertex.

Step 5: If the entering vertex is outside $R^n \times \{d/2^K, d\}$, stop. The earlier $\Gamma \cup r$-complete simplex is actually Γ-complete. Otherwise, go to Step 3.

Having found a Γ-complete simplex τ, say, whose vertices are V^1, V^2, $\cdots$, V^{n+1}, where $V^i = (v^i, d_i)$, $i = 1, \cdots, n+1$, we have determined points $z^i \in \Gamma(v^i)$ and a $\lambda = (\lambda_1, \cdots, \lambda_{n+1}) \geqq 0$ such that

$$\begin{aligned} \sum_{i=1}^{n+1} \lambda_i z^i &= 0 \\ \sum_{i=1}^{n+1} \lambda_i &= 1 \end{aligned} \tag{11}$$

has a solution. In this case, we say that the point x determined by

$$x = \sum_{i=1}^{n+1} \lambda_i v^i \tag{12}$$

is an approximate fixed point (for justification, see Ref. 26).

Since the stopping criterion at Step 5 requires that we generate a vertex in $R^n \times \{d/2^K\}$, we have generated a sequence of Γ-complete sets C_i, the last one of diameter less than $d/2^K$, and have thus found a reasonable solution.

The procedure for triangulation $R^n \times (0, d]$ generally used is called $J3$ in the literature. For a more detailed description of this algorithm, the reader is referred to Ref. 32.

IV. EXAMPLES

In this section, we discuss some examples that we solved using the algorithm described in the previous section. Three of the four examples had $f(\cdot) = p(\cdot)$ corresponding to mean square quantization error as a measure of distortion. The fourth example, on the other hand, uses a different weighting of the quantization noise; it is motivated by the problem of quantizer design for simple element differential coding of picture signals.[9] The examples are:

$$\begin{aligned}
&(i)\quad f(x) = p(x) = \tfrac{1}{32}, && -16 \leqq x \leqq +16\\
&\qquad\quad = 0 && \text{otherwise}\\
&(ii)\quad f(x) = p(x) = \frac{1}{\alpha} e^{-\alpha|x|}, && -\infty < x < +\infty,\ \alpha = 0.1\\
&(iii)\quad f(x) = p(x) = \frac{\exp(-u^2/2\alpha)}{\sqrt{2\pi\alpha}}, && -\infty < x < +\infty,\ \alpha = 1 \qquad (13)\\
&(iv)\quad f(x) = \frac{1}{\beta} e^{-\beta|x|};\quad p(x) = \frac{1}{\alpha} e^{-\alpha|x|}, && -\infty < x < +\infty\\
&\qquad \alpha = 0.18,\ \beta = 0.1; \qquad \text{and} \qquad \alpha = 0.1,\ \beta = 0.065.
\end{aligned}$$

Due to symmetry of functions $f(\cdot)$ and $p(\cdot)$, the optimum quantizers are symmetric and, for simplicity therefore, quantizers were con-

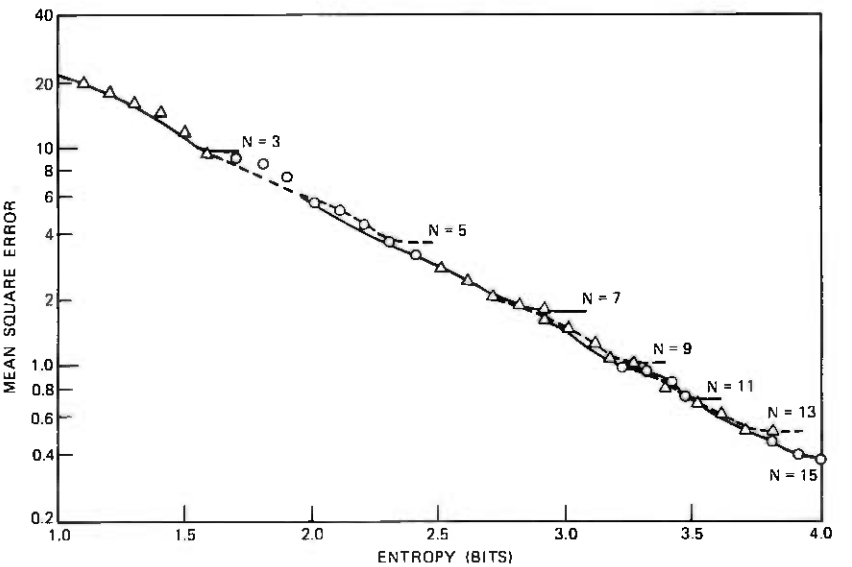


Fig. 1—Quantizer performance for uniform density. Minimum mean square error (MMSE) is plotted against entropy for a fixed number of levels (N). Only odd-level quantizers are considered. For each fixed number of levels, MMSE decreases with entropy up to a certain point, after which there is no further decrease in mean square error.

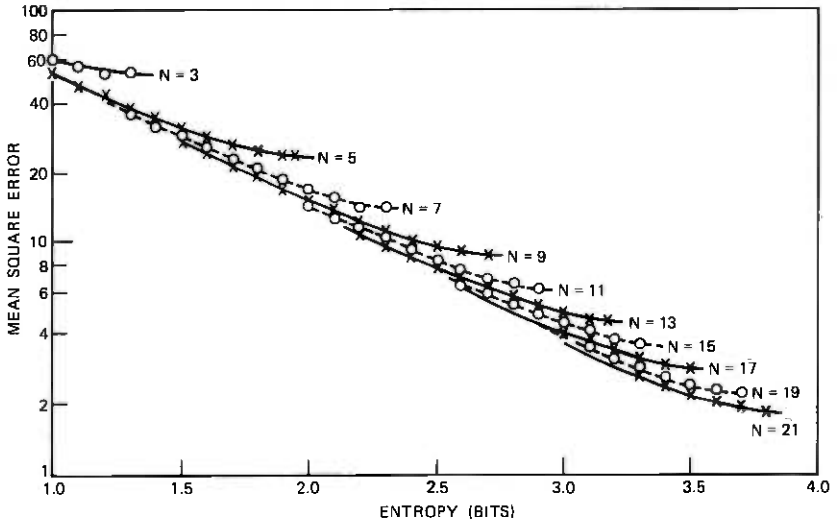


Fig. 2—Quantizer performance for Laplacian density.

strained to be symmetric. Also, without loss of any generality, only quantizers having odd numbers of levels were considered. In each case, several problems were solved by varying the entropy constraint and the number of levels. The number of levels were varied from 3 to 21, and the entropy constraint was varied from 1.0 bit to the largest possible bits using a particular number of levels.

Results of these simulations are given in Figs. 1 through 5. In these

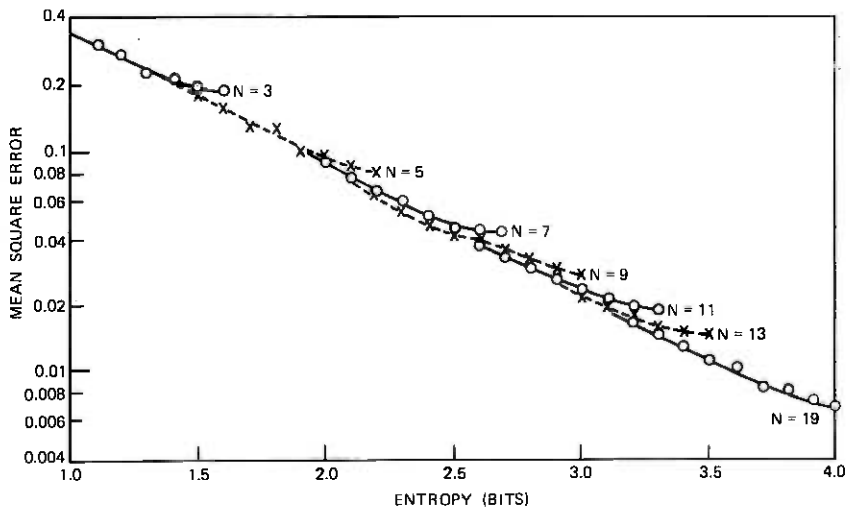


Fig. 3—Quantizer performance for gaussian density.

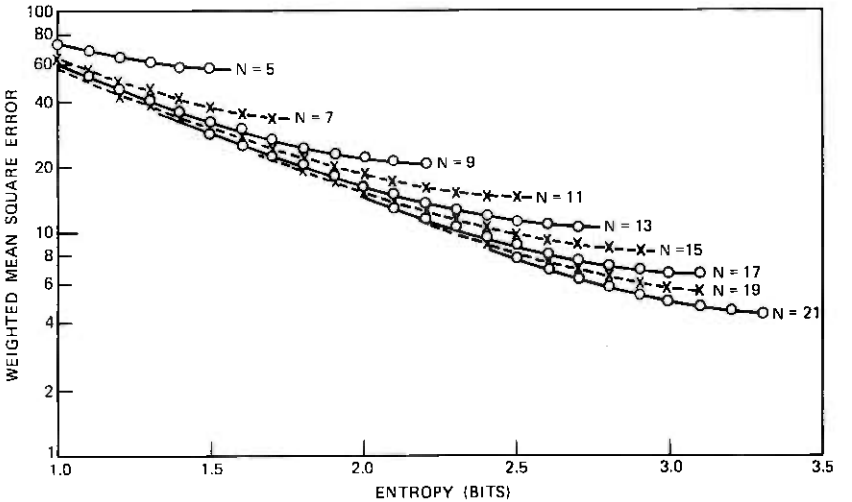


Fig. 4—Quantizer performance for Laplacian density and exponential weighting. Quantization noise is weighted by $1/|\beta|\ \exp(-\beta|x|)$, whereas the probability density is taken to be $1/|\alpha|\ \exp\ (-\alpha|x|)$. Such situations arise in quantization of the prediction errors in predictive coding of the television signals: $\alpha = 0.1, \beta = 0.065$.

figures the distortion is plotted logarithmically on y-axis and the entropy is plotted linearly on x-axis in bits. Alternate solid and broken lines are shown for different values of quantizer levels. For a given number of levels, the minimum distortion decreases approximately exponentially with respect to the entropy up to a certain point and

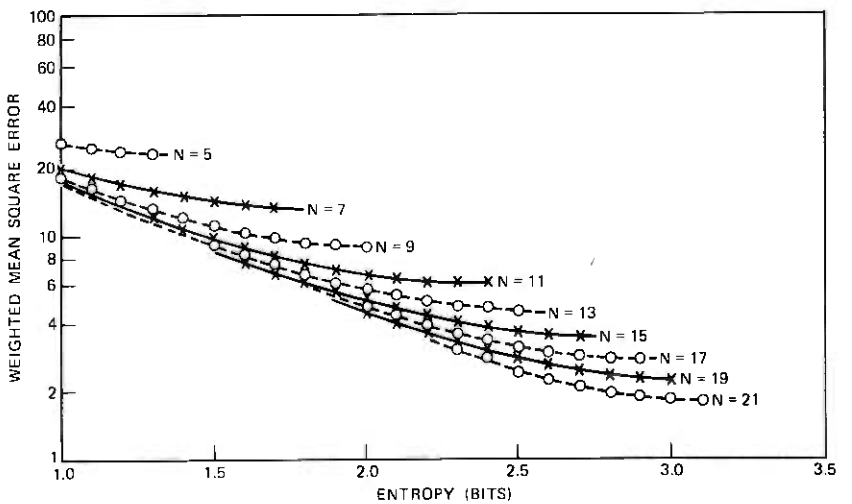


Fig. 5—Quantizer performance for Laplacian density and exponential weighting: $\alpha = 0.18, \beta = 0.1$.

then the entropy constraint is not operative any longer, and consequently the distortion remains a constant. These are indeed Lloyd-Max[11,12] quantizers that minimize the distortion for a given number of levels with no constraint on the entropy. The distortion versus entropy curves are lower bounded by the following functions:

$$\begin{aligned}
\textit{Example}\ (1)\quad & D = 85.3 \exp(-1.39E) \\
\textit{Example}\ (2)\quad & D = 196.18 \exp(-1.32E) \\
\textit{Example}\ (3)\quad & D = 1.40 \exp(-1.39E) \\
\textit{Example}\ (4a)\quad & D = 62.50 \exp(-1.31E) \\
\textit{Example}\ (4b)\quad & D = 176.71 \exp(-1.24E).
\end{aligned} \tag{14}$$

In the case of uniform densities, the optimum quantizer is nonuniform whenever the entropy constraint is operative, but when the entropy constraint is too large and inoperative, the optimum quantizers are uniform. Laplacian densities, on the other hand, always have uniform quantizers as the optimum quantizers. This has been shown by Berger.[20] In the case of gaussian density, the optimum quantizers were not uniform; however, a comparison of our results with those given by Goblick and Holsinger[18] indicates that, although nonuniform quantizers perform better than uniform quantizers, the differences in the performance of the two are somewhat small. This conclusion has also been reached by Wood[19] and Berger.[20] The case of an exponential weighting function falling slower than the probability density function arises in quantization of the prediction error in a simple element differential coding of picture signals. In this case, the density of the prediction error is approximately Laplacian, whereas the perceptual visibility[9,33] of the quantization noise may be approximated by an exponential function decaying somewhat slower than the probability density. The distortion-entropy curves for this case show larger improvement (that is, for a given entropy the distortion decreases much more than in the previous examples) as the number of levels is increased. Also, the optimum quantizers are nonuniform. Improvement in their performance over that of the uniform quantizers is more significant than in the previous examples. It is interesting to note that our algorithm can solve Lloyd-Max problem trivially by setting the entropy constraint to a very high value. This algorithm was also used in other applications related to adaptive quantization[34] of picture signals. The problems in this case were such that they had uniform (constant) weighting functions and two-sided exponentials as the density functions. The resulting quantizers had interesting structure and were used quite successfully.

4.1 Computational effort as a function of n

The increase in computational effort as a function of dimension, n, of the problem is important in the study of algorithms. In the case of fixed-point algorithms, Saigal[28] had speculated that different triangulations would have different effects on this growth and he was the first to propose a measure to describe it. For the triangulation employed in our experimentation, his measure predicted the growth rate of the number of iterations as n^2. Subsequently, Todd[35] refined his measure to predict an "average" growth rate of the iterations as $n^{\frac{1}{2}}$. The measure of Saigal, in some sense, predicts the "worst case" behavior.

The computational experiments in Section IV were ideally suited to test the theoretical predictions of Refs. 28 and 35, since the dimension of the problem was increased in a regular manner, the starting points were chosen in a regular way, and the problems of dimensions varying between 1 and 10 were solved. A number of results for various entropy values were plotted on the log-log paper. A representative plot is given in Fig. 6. It is seen that the experimental points lie on a straight line. The slope of these lines for different cases was a function of the entropy constraint and the probability density used and varied from 1.55 to 1.88, which is between 1.5 predicted by Todd[35] and 2 predicted by Saigal.[28]

Thus, we can conclude, with a high degree of certainty, that the number of iterations of the algorithm to solve a problem of dimension

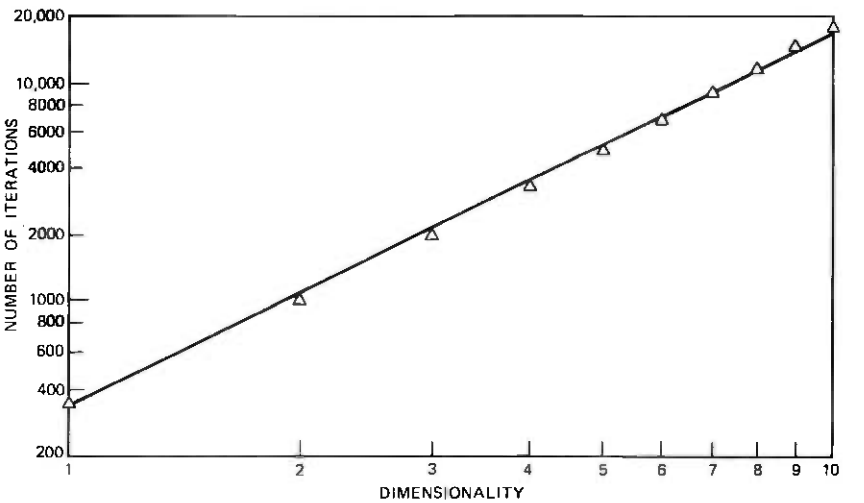


Fig. 6—Growth of computations vs dimensionality. Number of representative levels N is two times the dimensionality n plus 1. Straight line drawn is the minimum mean square error fit to the observations shown by Δ.

n would require an^b for some b between 1.5 and 2.0. Since each iteration requires $0(n^2)$ multiplications and at most one evaluation of the function, the number of function evaluations is bounded by $0(n^2)$ and multiplications by $0(n^4)$.

V. CONCLUSIONS

A fixed-point formulation has been developed to minimize the distortion, using a fairly general distortion measure, with respect to parameters of a quantizer under an entropy constraint on the quantized output. A point-to-set mapping is first developed whose fixed point satisfies the necessary conditions for a local extremum. Then a computer program is developed to compute its fixed points. Several examples are solved to show the usefulness of the algorithm. Finally, the rate of growth of the computations used by the algorithm as a function of the dimensionality of the problem is also discussed.

VI. ACKNOWLEDGMENTS

The authors would like to thank Barry Haskell for many inspiring discussions throughout the course of this work and Ivan Cermak and Ron Olsen for the use of the CDC-6600 computer.

REFERENCES

1. Clavier, Panter, and Greig, "PCM Distortion Analysis," Elec. Eng., *66* (1947), pp. 1110–1122.
2. B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The Philosophy of PCM," Proc. IRE, *36* (November 1948), pp. 1324–1331.
3. B. Widrow, "A Study of Rough Amplitude Quantization by Means of the Nyquist Sampling Theorem," Trans. IRE Prof. Group Circuit Theory, *CT-3* (1956), pp. 266–276.
4. T. Berger, F. Jelinek, and J. K. Wolf, "Permutation Codes for Sources," IEEE Trans. Inform. Theory, *IT-18* (1972), pp. 160–169.
5. J. Anderson, "Algorithms for Tree Source Coding with a Fidelity Criteria," Ph.D. dissertation, Cornell University, Ithaca, N. Y., 1972.
6. R. J. Dick, "Tree Coding for Gaussian Sources," Ph.D. dissertation, Cornell University, Ithaca, N. Y., 1973.
7. J. O. Limb, "Source-Receiver Encoding of Television Signals," Proc. IEEE, *55*, No. 3 (March 1967), pp. 364–380.
8. F. W. Mounts, A. N. Netravali, and B. Prasada, "Design of Quantizers for Real-Time Hadamard Transform Coding of Pictures," unpublished work.
9. A. N. Netravali, "On Quantizers for DPCM Coding of Picture Signals," unpublished work.
10. P. F. Panter and W. Dite, "Quantization Distortion in Pulse-Count Modulation with Nonuniform Spacing of Levels," Proc. IRE, *39* (1951), pp. 44–48.
11. S. P. Lloyd, "Least Squares Quantization in Pulse Count Modulations," unpublished work.
12. J. Max, "Quantizing for Minimum Distortion," IEEE Trans. Inform. Theory, *IT-6* (1960), pp. 7–12.
13. J. D. Bruce, "Optimum Quantization," Research Laboratory of Electronics, M.I.T., Tech. Rept. 429, March 1965.
14. G. M. Roe, "Quantizing for Minimum Distortion," IEEE Trans. Inform. Theory (correspondence), *IT-10* (1964), pp. 384–385.
15. V. Algazi, "Useful Approximations to Optimum Quantization," IEEE Trans. Commun. Technol., *COM-14* (1966), pp. 297–301.

16. P. Zador, "Development and Evaluation of Procedures for Quantizing Multivariate Distribution," Ph.D. dissertation, Stanford University, Stanford, Cal., 1964.
17. D. Huffman, "A Method for the Construction of Minimum Redundancy Codes," Proc. IRE, *40* (September 1952), pp. 1098–1101.
18. T. J. Goblick, Jr. and J. L. Holsinger, "Analog Source Digitization: A Comparison of Theory and Practice," IEEE Trans. Inform. Theory, *IT-13* (April 1967), pp. 323–326.
19. R. C. Wood, "Optimum Quantization," IEEE Trans. Inform. Theory, *IT-15* (March 1969), pp. 248–252.
20. T. Berger, "Optimum Quantizers and Permutation Codes," IEEE Trans. Inform. Theory, *IT-18* (November 1972), pp. 759–765.
21. H. Gish and J. N. Pierce, "Asymptotically Efficient Quantizing," IEEE Trans. Inform. Theory, *IT-14* (September 1968), pp. 676–683.
22. P. Elias, "Bounds on Performance of Optimum Quantizers," IEEE Trans. Inform. Theory, *IT-16* (March 1970), pp. 172–184.
23. W. I. Zangwill, *Nonlinear Programming*, Englewood Cliffs, N. J.: Prentice Hall, 1969.
24. C. E. Lemke, "Bimatrix Equilibrium Points and Mathematical Programming," Management Science, *11* (1965), pp. 681–689.
25. B. C. Eaves, "Homotopies for Computation of Fixed Points," Math. Prog., *3*, No. 1 (1972), pp. 1–22.
26. B. C. Eaves and R. Saigal, "Homotopies for Computation of Fixed Points on Unbounded Regions," Math. Prog., *3*, No. 2 (1972), pp. 225–237.
27. O. H. Merrill, "Applications and Extensions of an Algorithm that Computes Fixed Points of Certain Upper Semi-Continuous Point to Set Mappings," Ph.D. dissertation, University of Michigan, Ann Arbor, 1972.
28. R. Saigal, "Investigations into the Efficiency of Fixed Point Algorithms," to appear in *Fixed Points—Algorithms and Applications*, S. Karamardian, ed, New York: Academic Press.
29. R. Saigal, "Paths Generated by Fixed Point Algorithms," unpublished work.
30. R. Saigal, "On the Convergence Rate of Algorithms for Solving Equations that are Based on Complementary Pivoting," unpublished work.
31. G. B. Dantzig, "Linear Programming and Extensions," Princeton: Princeton University Press, 1963.
32. R. Saigal, "Fixed Point Computing Methods," *Encyclopedia of Computer Science and Technology*, M. Dekker, Inc., 1976.
33. J. C. Candy and R. H. Bosworth, "Methods for Designing Differential Quantizers Based on Subjective Evaluations of Edge Busyness," B.S.T.J., *51*, No. 7 (September 1972), pp. 1495–1516.
34. A. N. Netravali and B. Prasada, "Adaptive Quantization of Picture Signals Based on Spatial Masking," unpublished work.
35. M. J. Todd, "On Triangulations for Computing Fixed Points," Tech. Report No. 249, Department of Operations Research, Cornell University, Ithaca, New York, March 1975.

Contributors to This Issue

Robert R. Boorstyn, B.E.E., 1958, City College of New York; M.S., 1963, Ph.D., 1966, Polytechnic Institute of Brooklyn; Sperry Gyroscope Company, 1958–1961; Polytechnic Institute of Brooklyn, 1961—. Presently an associate professor, Mr. Boorstyn's current research interest is in computer communication networks. Chairman, New York Metropolitan chapter of IEEE Information Theory Group; Associate Editor, *Networks* journal; member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

John R. Cavanaugh, B.S. (Electrical Engineering), 1961, Ohio University; M.S. (Electrical Engineering), 1963, New York University; Bell Laboratories 1961—. Mr. Cavanaugh has worked on determining acceptable picture quality standards for broadcast television transmission. He is currently a member of the Network Objectives Department working on subjective evaluations of digital telephone systems. Member, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu.

Richard W. Hatch, B.S. (Electrical Engineering), 1952, Northeastern University; M.S. (Mathematics) 1958, Stevens Institute of Technology; Bell Laboratories, 1952—. Mr. Hatch worked for several years on the design of microwave radio relay systems and in 1961 and 1962 supervised groups working on the ground transmitter, systems analysis, and communications tests for the TELSTAR® satellite. In 1962 he became head of a department engaged in transmission systems engineering studies of network performance and maintenance. He currently heads a department responsible for studies of customer opinion and network performance leading to recommendations concerning network performance objectives. Member, Tau Beta Pi, Eta Kappa Nu.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., City University of New York; Ph.D. (Statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang spent the fall of 1970 visiting the Department of Mathematics of National Tsing-Hua University. He has been engaged in research in statistics, computing science, discrete mathematics, and switching networks.

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.

J. B. Lastovka, B.S. (Physics), 1962, John Carroll University; Ph.D. (Physics), 1967, Massachusetts Institute of Technology; Staff Member in the Division of Sponsored Research, Massachusetts Institute of Technology, 1967–1970; Bell Laboratories, 1970—. Since joining Bell Laboratories Mr. Lastovka has been engaged in research related to various aspects of hydrodynamic instabilities.

Dietrich Marcuse, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–1957; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research and studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of three books. Fellow, IEEE; member, Optical Society of America.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, Ph.D., 1970 (Electrical Engineering), Rice University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on various aspects of signal processing. Member, Tau Beta Pi, Sigma Xi.

David C. Rife, B.S.E.E., 1960, University of Washington; M.E.E., 1962, New York University; Ph.D. (E.E.), 1973, Polytechnic Institute of Brooklyn; Bell Laboratories, 1960—. Mr. Rife has worked on data carrier systems, automatic calling units, and data test equipment. He is currently supervisor of a group engaged in the development of an automated data station testing system. Senior member, IEEE; member, Tau Beta Pi, Phi Beta Kappa, Sigma Xi.

Romesh Saigal, B. Tech. (Honors) (Mechanical Engineering), 1961, M. Tech (Industrial Engineering), 1963, Indian Institute of Technology, Kharagpur, India; Ph.D. (Operations Research), 1968, University of California (Berkeley); Assistant Professor, University of California, Berkeley, 1967–1973; Bell Laboratories, 1973–1976; North-

western University, 1976—. Mr. Saigal has worked primarily on mathematical programming. Member, ORSA, Mathematical Programming Society.

Marvin R. Sambur, B.E.E. (1968), City College of New York; S.M. (1969) and Ph.D. (1972), Massachusetts Institute of Technology; Bell Laboratories, 1972—. Mr. Sambur is a member of the Acoustics Research Department and has worked in the areas of speech recognition, speaker recognition, low-bit-rate vocoder systems, encryption techniques, and digital waveform coders. Member, MPA-TC subcommittee on Speech Recognition and Understanding, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

John L. Sullivan, B.S.E.E., 1953, Iowa State University; M.S.E.E 1959, Newark College of Engineering; Bell Laboratories, 1953—. Mr. Sullivan has worked on television test equipment and on establishing objectives for exchange area trunks, wide-area data service, electronic telephone sets and TOUCH-TONE® calling. He presently supervises a group responsible for recommending local transmission objectives. Group activities include subjective testing to determine observer reaction to telephone message impairments, and using these test results in studies of present and future network transmission performance for speech. In addition, he is involved in Bell System, national, and international standards activities. Member, Acoustical Society of America.

Abstracts of Papers by Bell System Authors Published in Other Journals

CHEMISTRY

Investigations of an Electrodeposited Tin-Nickel Alloy: II. Surface Passivity Studied with Auger Electron Spectroscopy. H. G. Tompkins and J. E. Bennett, J. Electrochem. Soc., *123* (July 1976), pp. 1003–1006. This work investigates the surface film which is formed in air, using Auger electron spectroscopy and depth profiling. The surface film appears to be primarily a tin-rich oxide which is several atoms thick. The passive film which forms may not depend on the metastable nature of the alloy.

Time-Resolved Spectroscopy of Hemoglobin and Its Complexes with Subpicosecond Optical Pulses. C. V. Shank, E. P. Ippen, and R. Bersohn,* Science, *193* (July 2, 1976), pp. 50–51. Subpicosecond optical pulses have been used to study the photolysis of hemoglobin complexes. Photodissociation of carboxyhemoglobin is found to occur in less than 0.5 picosecond. In hemoglobin and oxyhemoglobin, a nondissociative excited state recovery in 2.5 picoseconds is observed. * Dept. of Chemistry, Columbia University.

ELECTRICAL AND ELECTRONIC ENGINEERING

Microwave Switching by Picosecond Photoconductivity. A. M. Johnson and D. H. Auston, IEEE J. Quantum Electron., *QE-11* (June 1975), pp. 283–287. Bulk photoconductivity produced by the absorption of picosecond optical pulses in silicon transmission-line structures has been used to switch and gate microwave signals. The technique permits the generation of microwave and millimeter-wave pulses as short as a single cycle, and requires only a few microjoules of optical energy.

Transistor Design Considerations for Low-Noise Preamplifiers. R. B. Fair, IEEE Trans. Nuc. Sci., *NS-23* (February 1976), pp. 218–225. A review is presented of design considerations for GaAs Schottky-barrier FETs and other types of transistors in low-noise amplifiers for capacitive sources. Ultimate limits of g_m/C and gate leakage currents are presented.

Wide-Band Matched Amplifier Design Using Dual Loop Feedback and Two Common Emitter Transistor Stages. T. J. Aprille, Jr., IEEE Trans. Circuits Syst. *CAS-23*, No. 7 (July 1976), pp. 434–442. Practical synthesis procedures for the design of shaped gain, wide-band, matched, dual loop feedback amplifiers that use a cascade of two common emitter transistor stages are treated. Circuits realized from these procedures are shown to be more than viable alternatives to existing single loop designs.

GENERAL MATHEMATICS AND STATISTICS

The Structure of a Utility Function Under Strong Risk Invariance. J. A. Morrison, SIAM J. Appl. Math., *31* (July 1976), pp. 93–98. The utility function of a decision-maker who faces alternatives with multidimensional consequences, and acts as if to maximize his expected utility, is considered. The functional form of the utility function, under strong risk invariance, is determined in the case of nonconstant risk aversion, verifying a conjecture of R. Willig.

MATERIALS SCIENCE

Anodic Passivation and Coating of AlAs in Aqueous Solutions. W. D. Johnston, Jr., J. Electrochem. Soc., *123* (March 1976), pp. 442–443. Uniform, stable oxide films have been grown on VPE deposited AlAs layers by anodization at constant current in pH 2.0 water. The oxide has refractive index ~1.8 and is suitable as a protective and anti-reflective coating for solar cells, LEDs, and other devices made from AlAs.

Enhanced Solubility and Ion Pairing of Cu and Au in Heavily Doped Silicon at High Temperatures. R. L. Meek and T. E. Seidel, J. Phys. Chem. Solids, *36* (1975), pp. 731–740. The equilibrium solubilities of Cu and Au in silicon have been calculated for high temperatures (900–1100°C) and heavy dopings (10^{19}–10^{21} cm^{-3}) and are

compared with experimental results for uniformly bulk doped and diffusion doped material. For strongly extrinsic n type material, a large solubility enhancement (about 10^3 times the intrinsic solubility) is calculated, due to ion pairing of the substitutional metal acceptor with donors. The saturation metal solubilities observed in bulk samples and diffused layers are in substantial agreement (within a factor of ~2) with calculations for all temperatures and doping levels.

Melting Point Depression and Kinetic Effects of Cooling on Crystallization in Poly(Vinylidene Fluoride)—Poly(Methyl Methacrylate) Mixtures. T. Nishi* and T. T. Wang, Macromolecules, *8* (November 1975), pp. 909–915. Melting point depression has been observed in mixtures of poly(vinylidene fluoride) and poly(methyl methacrylate). The phenomenon is ascribed to thermodynamic mixing of two compatible polymers. An appropriate expression is derived from which the interaction parameter for the polymer pair was found to be −0.295 at 160°C.
* Bridgestone Tire Co., Ltd., Tokyo, Japan.

Probability of Static Fatigue Failure In Optical Fibers. D. Kalish and B. K. Tariyal, Appl. Phys. Lett., *28* (June 15, 1976), pp. 721–723. An expression for the probability of static fatigue failure in glass is developed based upon a Weibull-type cumulative strength distribution and a fracture mechanics slow crack growth law. Good agreement between the model and experimental results is demonstrated. Examples for using this model as a predictive tool are presented.

Sputtering Techniques and Applications. A. K. Sinha, Electron. Packag. Prod., *15*, No. 8 (October 1975), pp. V10–V14. Important concepts are stated that should help understand and utilize sputtering technique for thin-film deposition. A state-of-the-art description is given of commercially available sputtering variants; namely, rf-diode sputtering, dc-diode sputtering, triode sputtering, and the magnetron sputter source. Effect of deposition variables is described on film composition, microstructure, and properties.

PHYSICS

Concentration-Dependent Absorption and Spontaneous Emission of Heavily Doped GaAs. H. C. Casey, Jr., and F. Stern, J. Appl. Phys., *47* (February 1976), pp. 631–643. A model for the calculation of the absorption and emission spectra for GaAs at carrier concentrations in excess of 1×10^{18} cm^{-3} is described. Calculated absorption and emission spectra are compared to previous experimental results, which permits assignment of the concentration dependence of the energy gap. The concentration-dependent thermal equilibrium electron-hole density product and radiative lifetime are calculated for p-type GaAs.

Determination of the Stress in Optical Fibers by Means of a Polariscope. M. J. Saunders, Rev. Sci. Instrum., *47* (April 1976), pp. 496–500. A polariscope is used in conjunction with a diameter-measuring instrument to determine the relationship between the tension of a Vycor-clad quartz fiber, as it is being drawn, and the diameter of the fiber. The polariscope is also used to determine the stress optical coefficient of optical fibers and preforms.

The Effects of Soft Modes on the Structure and Properties of Materials. P. A. Fleury, Ann. Rev. Mat. Sci., *6* (1976), pp. 157–180. This paper reviews recent developments regarding the role of crystal lattice instabilities (soft modes) in structural phase transitions and the associated enhanced or anomalous physical properties of materials. Material systems considered include ferroelectrics, superconductors, and metastable alloys. Phenomena considered include light scattering, critical fluctuations, normal mode interactions, etc. Device applications of soft mode effects are discussed.

Metal-Induced Extrinsic Surface States on Si, Ge, and GaAs. J. E. Rowe, J. Vacuum Sci. Technol., *13* (January/February 1976), pp. 248–250. Evidence for extrinsic metal-induced empty surface states during the Schottky-barrier formation on Si(111), GaAs($\bar{1}\bar{1}\bar{1}$), Ge(111), and Ge(100) is obtained with electron-energy loss spectroscopy. UV photoemission spectroscopy provides similar evidence for occupied extrinsic states on Si(111). The anomalous results on (110) surfaces are discussed in terms of a simple structural model.

Optically-Induced Energy Level Shifts for Intermediate Intensities. P. F. Liao and J. E. Bjorkholm, Opt. Commun., *16*, No. 3 (March 1976), pp. 392–395. We report measurements of optically-induced energy level shifts by nonresonant light at intermediate intensities in atomic sodium vapor. The intensity-dependence of the shifts departs substantially from the linear behavior predicted by second order perturbation theory, but is in good agreement with more exact calculations and yields 0.10 for the $3P_1$-4D oscillator strength.

Plasmon-Interband Coupling in Gallium Compounds. J. E. Rowe, J. C. Tracy, and S. B. Christman, Surface Sci., *52* (1976), pp. 277–284. Electron energy loss spectra have been measured for metallic Ga and the semiconductor compounds GaSb, GaAs, GaP, and GaN. The intensity at the measured plasmon energy decreases with increasing ionicity of the compound semiconductor. This is explained by a simple model involving a coupling of plasmon oscillations and interband transitions.

Production of Stabilized Coloration in Alkali Halides by a Two-Photon Absorption Process. L. F. Mollenauer, G. C. Bjorklund, and W. J. Tomlinson, Phys. Rev. Lett., *35*, No. 24 (December 15, 1975), pp. 1662–1665. We have measured the stabilized coloration produced from the conversion of *U* centers in KCl by means of a two-photon process. The measurements show that the coloration involves energy transfer via electron-hole pairs, rather than direct photoexcitation of the *U* centers. The absolute two-photon absorption cross section of KCl at $\lambda = 266$ nm was determined.

Pyroelectric Sign of $LiIO_3$. E. H. Turner, J. Appl. Crystallog., *9* (February 1976), p. 52. The sign of the total pyroelectric coefficient, p_3 of $LiIO_3$ is related to the absolute configuration. Using the piezoelectric sign convention, p_3 is found to be positive.

SYSTEMS ENGINEERING AND OPERATIONAL RESEARCH

On the Output of a GI/M/N Queuing System with Interrupted Poisson Input. H. Heffes, Oper. Res., *24* (May–June 1976), pp. 530–542. We characterize the departure process as a semi-Markov process and give results for the joint distribution of the number of customers in the system and the state of the input process at service completions. We also present results relating to the interdeparture time distribution and the distribution of the nonbusy period and compare the results with some known results for single-server systems.