THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV FEBRUARY 1965 NUMBER 2



# Numerical Evaluation of Electron Image Phase Contrast

By R. D. HEIDENREICH and R. W. HAMMING

(Manuscript received April 17, 1964)

*This paper is concerned with phase contrast in electron images with emphasis on a periodic scattering object. The Kirchoff diffraction or imaging integral over the back focal plane is formulated in terms of the amplitude and phases of the scattered wave, using coordinates adapted to numerical methods. The integral was programmed for evaluation on the IBM 7090 and the image plane amplitude displayed on a microfilm plotter. The effects of spherical aberration, aperture size, defocus and thermal motion of the scattering atoms on image contrast of atom positions for chains of nickel and of gold atoms were investigated for separations of 2 Å and 8 Å.*

*In the range of atomic separations, spherical aberration is the most destructive single factor in the loss of phase contrast. It appears that an objective lens with a spherical aberration coefficient less than 0.2 mm will be necessary if phase contrast images of atom locations are to be attained. In addition, a practical quarter-wave phase plate is essential for the objective lens system and will be much more effective than defocus contrast. Even so the contrast is marginal for photographic recording except for the case of a thin perfect crystal. Amplitude contrast is very small for atom positions and should be minimized by the use of a large objective aperture.*

*Phase contrast should improve with increasing accelerating potentials due to reduced inelastic cross sections compensating the loss in elastic contrast near* $2 \times 10^5$ *volts.*

## I. INTRODUCTION

The resolving power and general quality of electron microscope images are determined basically by image plane contrast. There are two distinct contrast mechanisms:

(1) *amplitude contrast* — produced by removal from the image plane of electrons scattered *outside* the objective half-angle, $\beta_{obj}$

(2) *phase contrast* — produced by suitable recombination at the image plane of waves scattered within the objective aperture.

The first mechanism, amplitude contrast, is the one commonly operating in images of objects which are greater than about 10 Å in size. The contrast between two image points for this type of object is either *mass thickness* or *diffraction* and is given approximately by

$$\Delta I/I = \Delta(Qt) = Q\Delta t \quad \text{or} \quad t\Delta Q. \tag{1a}$$

for amorphous materials. (See Ref. 1, Ch. IX, for diffraction contrast.) Here $Q$ is the cross section for scattering outside the objective aperture and $t$ is the object thickness. In turn, the cross section for scattering outside the aperture is

$$Q = \frac{N_0\sigma_{atom}}{A}\rho \tag{1b}$$

with $N_0 = 6.02 \times 10^{23}$ being Avogadro's number. Here, $A$ is the atomic weight and $\rho$ the density. The cross section per atom for scattering outside the aperture is $\sigma_{atom}$ and consists of an elastic and an inelastic part

$$\sigma_{atom} = \sigma_{el} + \sigma_{inel}\,. \tag{1c}$$

The cross section $\sigma_{atom}$ is the fraction of incident electrons scattered beyond the objective half-angle, $\beta_{obj}$, and is the integral of the differential cross section, $D(\beta)$, over the scattering angle $\beta$

$$\sigma_{atom} = 2\pi \int_{\beta_{obj}}^{\pi} D(\beta) \sin\beta \, d\beta. \tag{1d}$$

The differential cross section, $D(\beta)$, determines the intensity distribution at the back focal plane of the objective lens. For elastic scattering by an isolated atom $D(\beta)$ is simply the square of the atomic scattering amplitude or

$$D(\beta) = |f(s)|^2 \tag{2a}$$

where $f(s)$ is the electron scattering amplitude per atom as a function of the scattering parameter

$$s \equiv 4\pi \frac{\sin(\beta/2)}{\lambda} \tag{2b}$$

with $\beta$ the scattering angle and $\lambda$ the incident electron wavelength. For the small scattering angles encountered with fast electrons, $\sin(\beta/2) \approx \beta/2$ and

$$s \approx 4\pi \frac{\beta}{2\lambda}. \tag{2c}$$

Tables of scattering amplitudes [Ref. 2, Tables 3.3.3 A(1) and A(2)] list values of $f(s)$ in Å or cm for chosen values of the parameter $\beta/2\lambda$.

The differential cross section for a solid object is determined by interference among the waves scattered by the individual atoms composing the object. This, in turn, depends upon the spatial arrangement of the atoms. The differential cross section for a liquid or glass is diffuse, with broad maxima corresponding to those interatomic distances occurring most frequently. On the other hand, that for a single crystal is a discrete set of spots.

The portion of the scattered amplitude distribution falling outside the objective aperture determines the amplitude contrast of an image point as sketched in Fig. 1. That part of the distribution falling within the aperture is available for phase contrast at the image plane. The extent to which the phase information can be used in the formation of an image is presently limited by spherical aberration which essentially "scrambles" the phases to a degree increasing as the fourth power of the scattering angle. Phase information concerning the distance between two object points is thus increasingly garbled as the distance between the points decreases. This follows since the scattering angle for a distance $a$ in the object is

$$\beta \approx \lambda/a. \tag{2d}$$

Only values of $a$ such that $\beta < \beta_{obj}$ have any possibility of contributing to phase contrast.

Although numerical calculation of phase contrast is the purpose of this paper, it is instructive to consider the limitations on amplitude contrast in the region of interatomic separations before discarding it.

## II. AMPLITUDE CONTRAST

The total amplitude contrast for a single atom in the object plane of a perfect lens would be

$$\frac{\Delta I}{I_0} \approx -\frac{\sigma_{el}}{\pi\langle R_{el}^2\rangle} - \frac{\sigma_{inel}}{\pi\langle R_{ex}^2\rangle} \tag{3a}$$

where $R_{el}$ is the scattering radius for elastic collisions and $R_{ex}$ that for

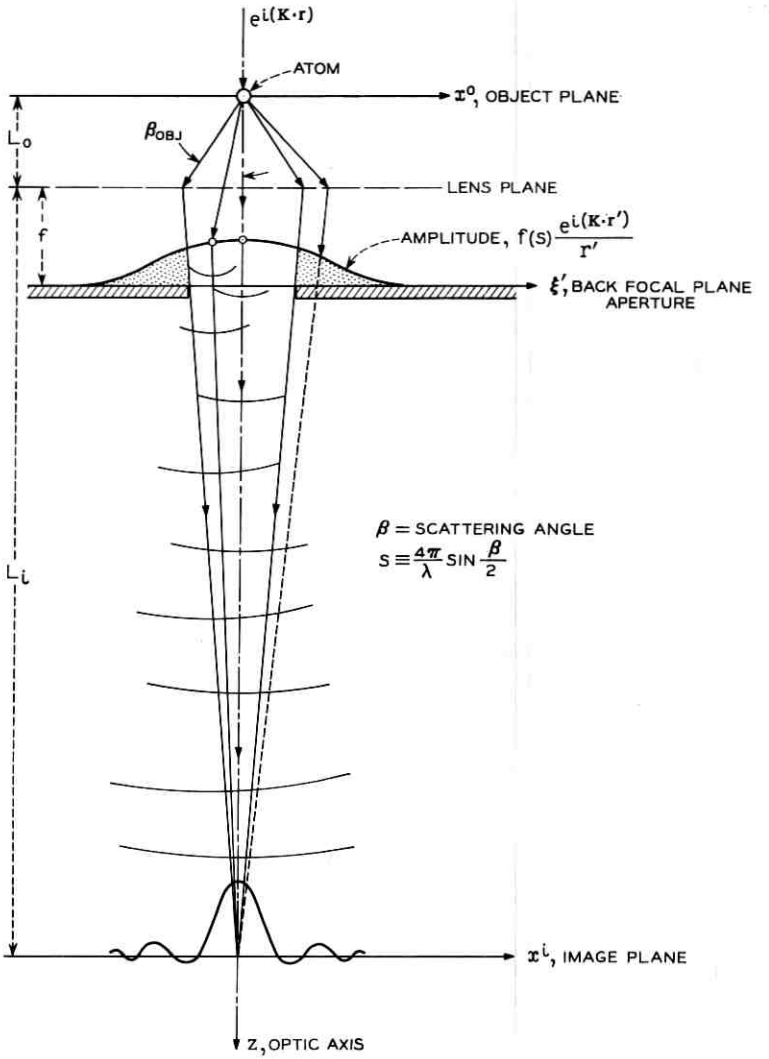


Fig. 1 — Schematic diagram showing amplitude scattered by single atom in object plane. The diffracted amplitude at the back focal plane produces image plane contrast depending upon the size of the aperture. The portion (shaded) falling outside the aperture results in deficiency amplitude contrast. The portion within the aperture can produce phase contrast.

inelastic. If the resolving power of an imperfect objective lens is $\delta$, then $R_{el} \approx \delta$ in (3a). Both terms of (3a) rapidly diminish with increasing incident electron velocity and fixed aperture. The negative sign denotes deficiency contrast for an image point.

Since the scattering contracts into the forward direction as the electron velocity increases, it is advisable to consider the scattering amplitude at constant values of $\beta_{obj}/2\lambda$ (the scattering parameter). This amounts to shrinking the objective aperture as the wavelength decreases so that the Airy diffraction disc from the aperture maintains a constant radius. This is accomplished by changing the variable of integration in (1d) from $\beta$ to the scattering parameter $s$ of (2b). The elastic cross section for scattering outside the objective aperture (per atom) becomes

$$\sigma_{el} = \frac{\lambda^2(1 - \gamma^2)^{-1}}{2\pi} \int_{(2\pi/\lambda)\beta_{obj}}^{4\pi/\lambda} |f^0(s)|^2 s \, ds \tag{3b}$$

with $\gamma \equiv v/c$, the ratio of electron velocity to that of light. Here $f^0(s)$ is the scattering amplitude based on the electron rest mass. It is noted that the velocity dependence of (3b) results in a rapidly decreasing value of $\sigma_{el}$ as the accelerating voltage increases, with consequent loss of elastic amplitude contrast in (3a). The elastic cross sections as a function of atomic number have recently been calculated by Burge[3] and Smith using numerical methods with (3b).

The second term of (3a) depends upon the inelastic cross section $\sigma_{inel}$ and the inelastic excitation radius, $R_{ex}$. Williams[4] arrived at an excitation distance $R_{ex}$ for an atom of average excitation energy $\langle \Delta E \rangle$ having a velocity dependence

$$R_{ex} \approx \frac{0.2vh}{\langle \Delta E \rangle} (1 - \gamma^2)^{-\frac{1}{2}}. \tag{4}$$

For carbon, $R_{ex}$ is about 40 Å at $V_a = 10^5$ volts taking $\langle \Delta E \rangle \approx 44$ volts, so that the location of the energy loss event is very diffuse. For heavier atoms, $R_{ex}$ is generally less, being about 4 Å for Al at $10^5$ volts with $\langle \Delta E \rangle \approx 300$ volts. The inelastically scattered electrons thus do not carry localized information on atom locations (unless they are subsequently diffracted). The relatively large values of $R_{ex}$ make the amplitude contrast due to inelastic scattering trivial except for relatively large objects (50–100 Å).

The amplitude contrast resulting from elastic scattering is so small as to offer little hope of imaging single atoms. For example, the elastic cross section for a nickel atom is about $7 \times 10^{18}$ cm$^2$ at $V_a = 5 \times 10^4$ volts

for an objective aperture $\beta_{obj} \approx 5 \times 10^{-3}$ rad. The contrast is only about 0.8 per cent for $\delta \approx 2$ Å, or so low as to be of little importance. The chief hope for imaging atoms with the electron microscope lies in phase contrast, wherein intensities are very sensitive to the relative phases of the scattered waves reconstituted at the image plane.

## III. PHASE CONTRAST[1]

The dominant contrast mechanism for scattering objects exhibiting a periodic structure or with a detail size approaching atomic dimensions is that of phase contrast. This requires that the image plane amplitude be evaluated by superposition of all the waves which are scattered by the object and enter the objective aperture of the lens. In its most general form this superposition is the imaging integral.[5] For a perfect, aberration-free objective lens the imaging integral is the Fourier transform of the diffracted amplitude distribution at the back focal plane, which is itself the Fourier transform of the distribution of scattering potential in the object. The imaging integral is thus a magnified representation of the Fourier projection in the object plane of the electrostatic potential distribution[6] in the object itself. In the case of a real objective lens, the fidelity of the image plane amplitude distribution is determined by the lens aberrations and by the size of the objective aperture. The objective aperture limits the distance information available to the image plane, removing the higher "spatial frequencies" in the diffraction pattern at the back focal plane. This is in addition to its own diffraction pattern.

A description of the intensity distribution in the image plane involves a consideration of three planes in the objective lens. These are: (1) the object plane with rectangular coordinates, the $x^o$-$y^o$ plane, (2) the back focal plane or $\xi'$-$\eta'$ plane containing the objective aperture and (3) the image or $x^i$-$y^i$ plane. The three planes are normal to the optic axis $z$ as shown in Fig. 2. The object and image distances measured from the lens plane are $L_0$ and $L_i$ respectively. The distance from the lens plane to the back focal plane is the focal length of the lens $f$. The magnification of the image is $M = L_i/L_0$, and $f \approx L_0$ when $L_i \gg L_0$.

Assume a plane parallel electron beam of wave vector $\mathbf{K}$ along the optic axis incident on the object plane from above. Let its amplitude be unity. Electron waves scattered by the object in the object plane produce a diffraction or scattering pattern at the back focal plane. This pattern is nearly a plane section through the reciprocal lattice of the object with a scale factor $L_0\lambda$, where $\lambda$ is the electron wavelength and $|\mathbf{K}| = 2\pi/\lambda$. The diffracted amplitude distribution $\psi$ in the back focal plane is in turn a source for Huygens wavelets which propagate to the

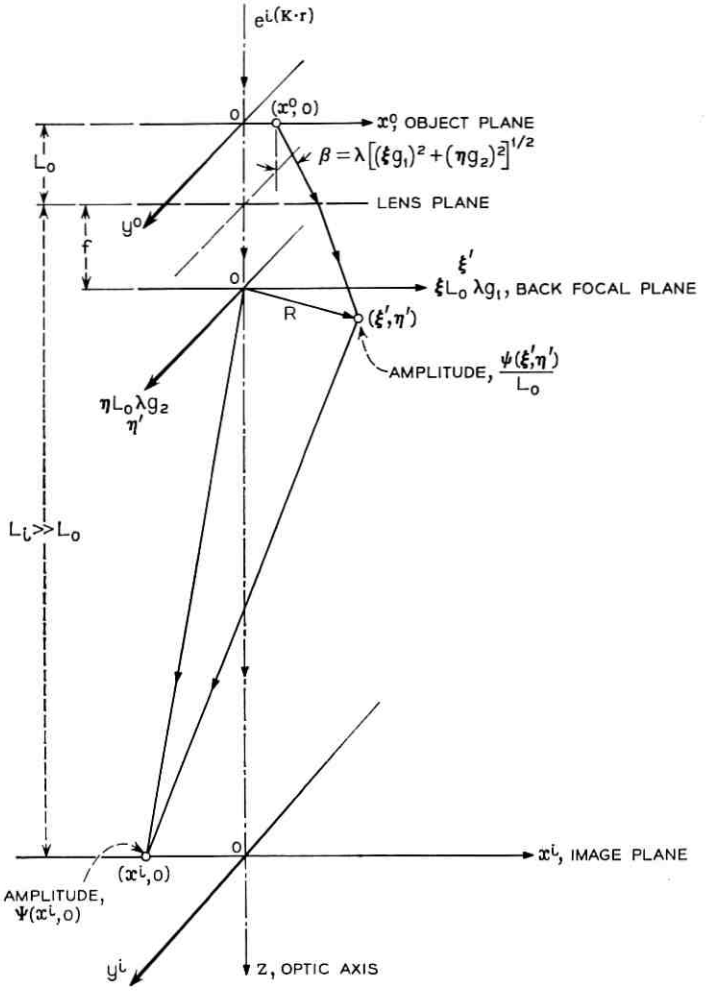


Fig. 2 — The three planes — object, back focal and image — that are of concern in contrast calculations. The scale in the back focal plane is for a periodic scattering object with reciprocal lattice vectors $g_1$ and $g_2$ .

image plane and recombine according to their phases and so produce the image amplitude distribution. The entire process can be formally described in terms of Fourier transforms, as already mentioned. However, when lens aberrations are present the imaging integral describing the image plane amplitude distribution for the imperfect image is no longer identical with the object plane Fourier projection of potential in the object. It is usually necessary to employ numerical methods in evaluating

the image plane amplitude, since the integration cannot in general be carried out in analytic form. The integral to be considered[7] is

$$\Psi(x^i, y^i) = \frac{1}{L_0 L_i \lambda} \int_{\xi'} \int_{\eta'} \psi(\xi', \eta') \exp [i\chi(\xi', \eta')] \cdot \exp [-iK(\xi' x^i + \eta' y^i)/L_i] \, d\xi' \, d\eta'. \tag{5}$$

$\chi(\xi', \eta')$ is the phase of the amplitude at $(\xi', \eta')$ in the back focal plane.

Since the primary concern is with crystalline or periodic scattering objects, the coordinate system chosen in the back focal plane is adapted both to this situation and to a formulation lending itself to numerical evaluation. Distances $R$ in the back focal plane are related to distances $d$ in the object by the relation $| R | = L_0\lambda/d$. $L_0\lambda$ is the camera constant or scale factor. A single-crystal object produces an array of diffraction spots in the back focal plane located at $\mathbf{R} = L_0\lambda\mathbf{g}_1 + L_0\lambda\mathbf{g}_2$ where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the operating reciprocal vectors in the object plane. The generating vectors in the object are $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ and the generating vectors in reciprocal space are $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$, subject to the condition

$$\begin{aligned} (\mathbf{a}_i \cdot \mathbf{b}_j) &= \delta_{ij} \\ &= 0 \quad \text{if} \quad i \neq j \\ &= 1 \quad \text{if} \quad i = \text{j} \end{aligned} \tag{6a}$$

so that $| \mathbf{b}_1 | = | \mathbf{a}_1 |^{-1}$ etc. and

$$\mathbf{g} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3 \tag{6b}$$

where $h,k,l$ are the Miller indices denoting the reciprocal lattice vector $\mathbf{g}$.

Let $\xi$ and $\eta$ be numbers such that an arbitrary point in the back focal plane is the terminus of the vector $\mathbf{R}$

$$\mathbf{R}(\xi', \eta') = \xi L_0\lambda\mathbf{g}_1 + \eta L_0\lambda\mathbf{g}_2 . \tag{6c}$$

If $\xi$ and $\eta$ are integers, $R$ locates a diffraction spot at a distance

$$| R | = [(\xi L_0\lambda g_1)^2 + (\eta L_0\lambda g_2)^2]^{\frac{1}{2}} \tag{6d}$$

since the $\xi$ and $\eta$ axes are rectangular. The scattering angle $\beta$ for electrons to the point $\mathbf{R}$ is, for small angles ($\sin\beta \approx \beta$)

$$\beta = \frac{| R |}{L_0} = \lambda[(\xi g_1)^2 + (\eta g_2)^2]^{\frac{1}{2}} \tag{6e}$$

and the scattering parameter $s$ is

$$s \equiv \frac{4\pi}{\lambda} \sin\frac{\beta}{2} \approx 2\pi[(\xi g_1)^2 + (\eta g_2)^2]^{\frac{1}{2}}. \tag{6f}$$

The amplitude at a point $(x^i, y^i)$ in the image plane is the diffraction integral[8] over the aperture

$$\begin{aligned}\Psi(x^i, y^i) &= \frac{1}{L_0 L_i \lambda} \int_{-\max}^{+\max} \int \psi(\mathbf{R}) e^{i\chi(\mathbf{R})} \\ &\quad \cdot \exp\left[-2\pi i(\xi x^0 g_1 + \eta y^0 g_2)\right] d(L_0\lambda\xi g_1) d(L_0\lambda\eta g_2) \\ &= \frac{(L_0\lambda)^2 g_1 g_2}{L_0 L_i \lambda} \int_{-\xi_{\max}}^{\xi_{\max}} \int_{-\eta_{\max}}^{\eta_{\max}} \psi(\xi,\eta) e^{i\chi(\xi,\eta)} \\ &\quad \cdot \exp\left[-2\pi i(\xi x^0 g_1 + \eta y^0 g_2)\right] d\xi\, d\eta. \end{aligned} \tag{7a}$$

In order to evaluate (7a) it is necessary to know the amplitude distribution $\psi(\xi, \eta)$ and the phase $\chi(\xi, \eta)$ in the back focal plane.

If the diffraction pattern has a center of symmetry at the origin of the back focal plane, the periodic object likewise has a center of symmetry at the origin of the object plane. In this case the imaging integral (7a) can be simplified for the purposes of calculation to

$$\begin{aligned}\Psi(x^i, y^i) &= \frac{2\lambda g_1 g_2}{M} \int_0^{\xi_{\max}} \int_0^{\eta_{\max}} \psi(\xi,\eta)[\cos\chi(\xi,\eta) + i \sin\chi(\xi,\eta)] \\ &\quad \times \cos 2\pi\xi x^0 g_1 \cos 2\pi\eta y^0 g_2 d\xi\, d\eta. \\ &= \frac{2\lambda g_1 g_2}{M} S^i \end{aligned} \tag{7b}$$

where $S^i$ denotes the phase integral in (7b).

It is now necessary to choose expressions for the amplitude $\psi(\xi, \eta)$ at the back focal plane occurring in (7b). This amplitude depends upon the detailed spatial arrangement of atoms in the object plane and upon the atomic scattering amplitude $f(s)$. There are two approaches to $\psi(\xi, \eta)$ — the kinematic and the dynamical theories of electron diffraction.[9]

The kinematic approach is much the simpler of the two and will illustrate the pertinent features in phase contrast. Since the main interest here is with scattering objects which are basically periodic, the kinematic amplitudes at the back focal plane can be immediately written down. Let the object be a crystal sheet with an atom at the origin of coordinates in the object plane. If the lateral extent of the crystal is $n_1$ atoms along $x^0$ and $n_2$ atoms along $y^0$ and if the extent in the $z$ direction is $n_3$ atoms, the kinematic amplitude is the familiar expression[10]

$$\psi(\xi,\eta) = (1 - \gamma^2)^{-\frac{1}{2}} f^0(s) n_3 \frac{\sin \pi n_1 \xi}{\sin \pi\xi} \cdot \frac{\sin \pi n_2 \eta}{\sin \pi\eta} \tag{8}$$

where $n_1$, $n_2$, and $n_3$ are odd integers. If even integers are employed, phase factors must be included in (8). The discussion is therefore confined to odd values. The atomic scattering amplitude $f^0(s)$ is based on the rest mass of the electron and $(1 - \gamma^2)^{-\frac{1}{4}}$ is the relativistic correction (with $\gamma$ the ratio of electron velocity to that of light). It is noted that

$$\pi n_1 \xi = \frac{|K|}{2} n_1 \frac{\beta_1}{g_1}$$

and

$$\pi n_2 \eta = \frac{|K|}{2} n_2 \frac{\beta_2}{g_2}.$$

The phase $\chi$ in the back focal plane is made up of several terms[11]

$$\chi(\xi,\eta) \equiv \frac{\pi}{2} - \frac{2\pi}{\lambda} C_{\bar{o}} \beta^4 + \frac{\pi}{\lambda} \pi \Delta f \beta^2. \tag{9a}$$

The phase change upon diffraction by the object is $\pi/2$, while the second term is the phase due to spherical aberration. The third term is the phase introduced by defocusing the lens an amount $\Delta f$. There is a fourth term which is the phase associated with the electron scattering amplitude $f(s)$ for an atom, but it is small except for the heaviest atoms and at large scattering angles and will be neglected here.

It should be mentioned that the spherical aberration phase term in (9a) is four times larger than that commonly used in discussions[12,13] of the effect of spherical aberration. The term in (9a) applies to a *single* ray at angle $\beta$ with a corresponding circle of aberration of radius $C_{\bar{o}}\beta^3$. The unweighted *average* radius for a bundle of rays filling the range $0 < \beta < \beta_{\max}$ is

$$\langle \Delta x^0 \rangle_{sp} = \frac{C_{\bar{o}}}{\beta_{\max}} \int_0^{\beta_{\max}} \beta^3 d\beta = \frac{1}{4} C_{\bar{o}} \beta_{\max}{}^3$$

with a resultant average phase

$$\langle \chi \rangle_{sp} = \frac{2\pi}{\lambda} \frac{C_{\bar{o}} \beta_{\max}{}^3}{4}. \tag{9b}$$

The use of (9b) in (9a) in evaluating the imaging integral is not correct, since it suppresses the destructive effect of spherical aberration.

The total amplitude at an image point is the sum of the scattered amplitude (7b) and the unscattered axial wave. If the incident amplitude is unity, then in the kinematic approximation of weak scattering the unscattered wave leaving the object plane is very nearly of unit ampli-

tude. Its amplitude at the origin of the back focal plane is large, but at the image plane it is $(M)^{-1}$. Adding the unscattered amplitude to (7b) yields the total image point amplitude

$$\Psi(x^i, y^i)_{\text{total}} = \frac{1}{M}\left[1 + 2\lambda g_1 g_2 (1 - \gamma^2)^{-\frac{1}{2}} S^i\right]. \tag{10a}$$

The mass correction factor $(1 - \gamma^2)^{-\frac{1}{2}}$ has been removed from $S^i$ so that the velocity dependence of the amplitude can be easily seen. The expression (10a) does not apply when dynamical conditions[14] are realized in the object, since then the axial wave may be weaker than the scattered amplitude and the approximation is no longer valid.

The image point intensity obtained by multiplying (10a) by its complex conjugate is

$$|\Psi(x^i, g^i)|^2 \approx \frac{1}{M^2}\left[1 + 4\lambda g_1 g_2 (1 - \gamma^2)^{-\frac{1}{2}} S^i{}_{\text{real}}\right] \tag{10b}$$

since the term in $|S^i|^2$ will generally be trivial compared to the cross product term until the mass correction term becomes large at very high accelerating voltages ($\approx 10^6$ volts). The imaginary part of the scattered amplitude (7b) is here neglected.

The kinematic contrast $G$ between two image points is defined to be the intensity difference between the respective points divided by the background intensity. Using (10b) the result is

$$G \approx 4\lambda g_1 g_2 (1 - \gamma^2)^{-\frac{1}{2}} \Delta S^i \tag{10c}$$

with all lengths in angstroms. $\Delta S^i$ is the amplitude differential between the image points in question obtained by numerical evaluation of

$$S^i = n_3 \int_0^{\xi_{\max}} \int_0^{\eta_{\max}} f^0(s) \frac{\sin n_1 \pi \xi}{\sin \pi \xi} \frac{\sin n_2 \pi \eta}{\sin \pi \eta} \times \cos \chi(\xi,\eta) \cos 2\pi \xi x^0 g_1 \cos 2\pi \eta y^0 g_2 \, d\xi \, d\eta \tag{10d}$$

at each point.

For the purpose of phase contrast calculations it is sufficient to consider a single chain of atoms of spacing $a$ lying along the $x^0$ axis. The number of atoms in the chain is $n_1$, with the middle atom on the optic axis. Setting $n_2 = n_3 = 1$ in (10d) gives the integral to be evaluated at a point on the $x^i$ axis as

$$S^i = \int_0^{\xi_{\max}} \int_0^{\eta_{\max}} f^0(s) \frac{\sin \pi n_1 \xi}{\sin \pi \xi} \cos \chi(\xi,\eta) \cos 2\pi \left(\frac{x^0}{a}\right) \xi \, d\xi \, d\eta \tag{10e}$$

having maxima at the image points corresponding to atoms or where $x^0 = 0, \pm a, \pm 2a, \cdots$ etc. The amplitude maxima for atomic positions approach

$$\int_0^{\xi_{\max}} \int_0^{\eta_{\max}} f^0(s) \frac{\sin \pi n_1 \xi}{\sin \pi \xi} \cos \chi(\xi,\eta) d\xi \, d\eta \to \int_0^{\infty} \int_0^{\infty} f^0(s) \cos \chi(\xi,\eta) d\xi \, d\eta$$

as the size of the objective aperture, $\xi_{\max}$, $\eta_{\max}$, increases. The value of the maxima is seen to be sensitive to the phase $\chi(\xi, \eta)$. For a perfect lens, the scattered amplitude in the Gaussian image plane vanishes, since $\chi \approx \pi/2$, so there will be only background intensity. On the other hand, the introduction of a quarter-wave phase plate[15] into the scattered beams can make $\chi = 0$ or $\pi$ and the amplitude will be a maximum. The sign of the maxima will be that of $\cos \chi$, so that atom positions can be either bright or dark against the background, depending upon the phase.

## IV. NUMERICAL RESULTS

### 4.1 *Spherical Aberration at the Gaussian Image Plane*

Numerical evaluation of the phase integral $S^i$ in (7b) requires that the phase (9a) be expressed in terms of the dimensionless coordinates $(\xi, \eta)$ using (6e) to give

$$\chi(\xi, \eta) = (\pi/2) - 2\pi C_5 \lambda^3 [(\xi g_1)^2 + (\eta g_2)^2]^2 + \pi \Delta f \lambda [(\xi g_1)^2 + (\eta g_2)^2].$$

For the single-atom chain of spacing $a$ this reduces to

$$\chi(\xi, \eta) \approx (\pi/2) - \pi C_5' [\xi^2 + \eta^2]^2 + \pi \Delta f' (\xi^2 + \eta^2) \qquad (11a)$$

with $C_5' = 2C_5\lambda^3/a^4$ and $\Delta f' = \Delta f \lambda / a^2$ being dimensionless aberration and defocus parameters. It is noted that $\frac{1}{2}C_5'$ is the number of wavelengths of spherical aberration at the first diffraction maximum ($\xi = 1$, $\eta = 0$) while $\frac{1}{2}\Delta f'$ is the number of wavelengths of defocus.

In the Gaussian image plane $\Delta f = 0$ and (11a) becomes

$$\chi(\xi, \eta) = (\pi/2) - \pi C_5' [\xi^2 + \eta^2]^2. \qquad (11b)$$

The real phase integral $S^i$ along the chain, $y^0 = 0$, is now

$$S^i = \int_0^{\xi_{\max}} \int_0^{\eta_{\max}} \psi(\xi,\eta) \cos \chi(\xi,\eta) \cos 2\pi \xi x^0/a \, d\xi \, d\eta \qquad (12)$$

and the absolute amplitude is $2\lambda a^{-2} S^i$. If the phase (11b) is used in (12) it is evident that $\cos \chi = \sin \pi C_5' [\xi^2 + \eta^2]^2$. The trivial contrast observed is produced by spherical aberration. Fidelity contrast at the

Gaussian image plane requires the use of a $\lambda/4$ wave plate in the back focal plane. If the plate advances only the diffraction spectra but not the unscattered or zero-order beam, the phase term in (12) is

$$\cos \chi = -\cos \pi C_5' [\xi^2 + \eta^2]^2$$

and the image will display atom positions dark relative to the background.

Since analytical expressions for the atom scattering factors are not available, the amplitude in the back focal plane for this purpose is obtained numerically from tabular values as described in the Appendix. A discussion of the choice of sampling intervals along the $\xi$, $\eta$ and $x^0/a$ axes is also found in the Appendix and some of the artifacts that may occur are pointed out.

If $C_5' = 0$ and $n_1 = 1$ in (12), it will be noted that $S^i$ is the profile of the Airy disk[16] for a single atom in the object plane due to diffraction by the objective aperture. Fig. 9(b) of the Appendix shows a computer plot (IBM-7090 microfilm plotter) of (12) for this case of a single nickel atom. The absolute maximum amplitude is 7.7 $\lambda$ angstrom, or a contrast of about 60 per cent relative to background for 100-kv electrons. The value $a = 2$ Å was employed in (10c), since the scaled $f^0(s)$ curve was at this value of $a$. The contrast is independent of the number of atoms, $n_1$, in the chain but is proportional to $n_3$, the number of atoms along the optic axis if chains are stacked one directly above the other.

Since the electron microscope objective aperture imposes a finite upper bound on the imaging integral the result is equivalent to terminating or truncating a Fourier series to a finite number of terms. As a consequence, the imaging integral does not converge to the value expected for an infinite upper bound but oscillates about this value. The Gibbs phenomenon[17] in the particular case of the objective aperture manifests itself as the experimentally observed diffraction pattern of the aperture. The same phenomenon gives rise to artifacts between atom positions in the image when a finite number of diffraction spectra are admitted by the aperture.[18]

The devastating effect of spherical aberration upon phase contrast in the image of an atom chain is illustrated in the series of amplitude profiles of Fig. 3. The numerical values are for nickel with a spacing of 2Å. Profiles were obtained for chain lengths of 5 and 17 atoms which are within the transverse coherence[19] length of double condenser illumination in present microscopes. Inelastic scattering and thermal motion are neglected. The former is reduced by increasing the accelerating potential and the latter by reducing the temperature. The profiles of Fig. 3 speak

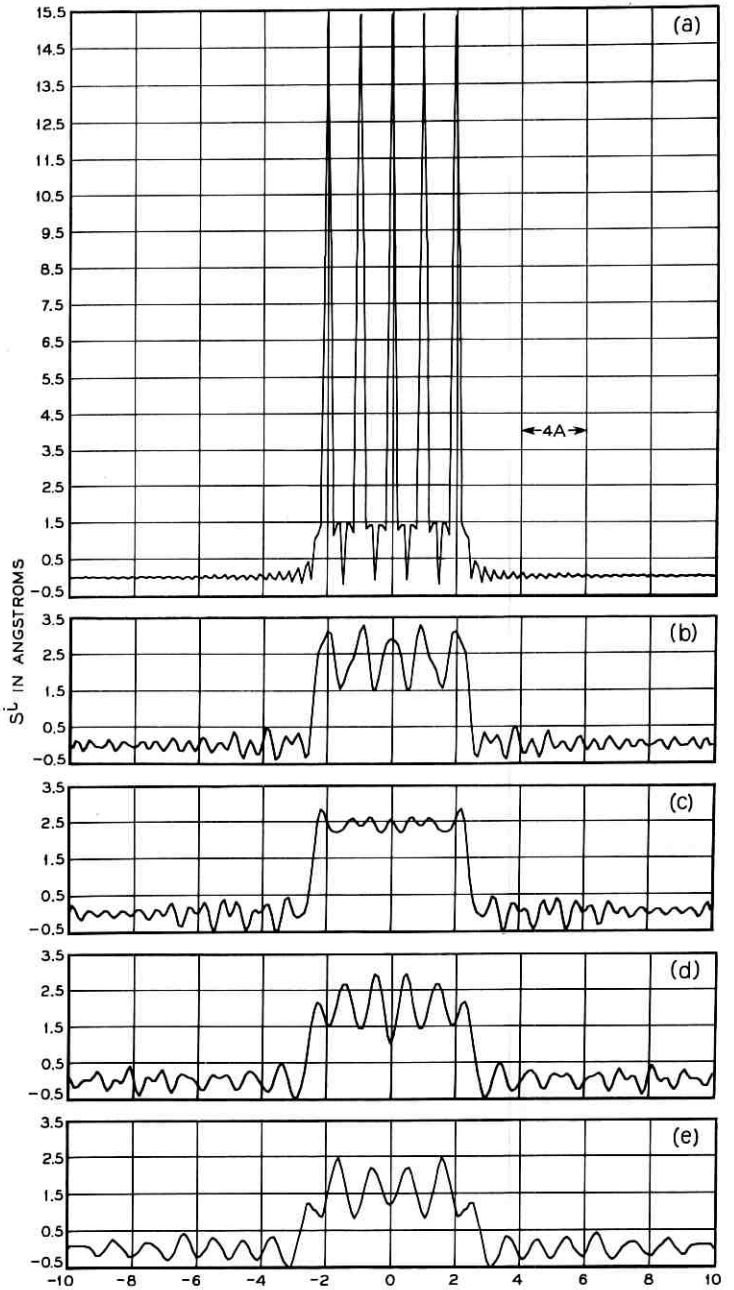


Fig. 3 — Series of numerical image amplitude profiles for a 5-atom chain of nickel atoms with spacing $a = 2$ Å. The effect of increasing spherical aberration parameter $C_{\bar{0}}'$ ($\frac{1}{2} C_{\bar{0}}'$ is the number of wavelengths of aberration at the first diffraction maximum) is illustrated in the series (a)–(e): (a) $C_{\bar{0}}' = 0$, (b) 0.2, (c) 0.3, (d) 0.5, and (e) 1.0. A quarter-wave phase plate is assumed present in the back focal plane. The contrast of about 60 per cent in (a) falls rapidly to about 9 per cent in (b) for 100-kv electrons with $f = 3$ mm.

for themselves, showing a loss in atom position amplitude from 15.5 Å with $C_{\bar{o}}' = 0$ to 2 Å with the introduction of 0.1 wavelength of spherical aberration at the first diffraction maximum where $C_{\bar{o}}' = 0.2$.

If the diffraction spectra in the back focal plane were points, as for an infinite, perfect crystal, the phase contrast would be periodic with $C_{\bar{o}}'$ and would have maxima where $(\Delta x^0)_{sph} = C_{\bar{o}}\beta^3$ is an integral number of half-lattice spacings. For the finite chains of 5 and 17 atoms, the damped periodic nature of the contrast is illustrated in Figs. 4(a) and (b) compared to that for a single atom. The aperture has been decreased from 3.5 diffraction maxima in Fig. 4(a) to 1.5 in Fig. 4(b) with a sizable reduction in contrast. The reduction in aperture size further limits the amount of spherical aberration but this is more than compensated for by the loss of distance information in the diffraction pattern and the increased diffraction by the aperture. It must be concluded that if useful phase contrast of atom positions is to be obtained spherical aberration must be minimized. Other contrast enhancing devices cannot overcome this defect of disturbed phase information.

### 4.2 *Contrast by Defocus with Spherical Aberration*

A case approximating the present state of the electron microscope objective lens is that for no phase plate with $C_{\bar{o}} \approx 3$ mm and contrast enhancement by defocus. This example is a chain of 5 gold atoms with a spacing of 8 Å and was chosen to approximate the situation of gold-stained sites on a DNA molecule.[20] No allowance is made for a substrate or inelastic scattering.

The phase (11a) for this particular case retains $\pi/2$ (with no phase plate) as well as the defocus term. The scaled aberration parameter is now $C_{\bar{o}}' = 0.74$ or 0.37 wavelengths at the first diffraction maximum using 100-kv electrons. A series of amplitude profiles was computed on the IBM 7090 for a range of values $0 \leqq \Delta f' \leqq 4$. From these results the relative amplitude of atom positions $\Delta S^i$ was plotted against $\Delta f'$ as shown in Fig. 5. The oscillation of contrast with changing defocus is typical with near zero contrast at exact focus. Maximum contrast is obtained by weakening the objective lens ($\Delta f$ positive) as is well known. Even so the contrast does not rise above about 6 per cent, which is submarginal for seeing the gold atom positions in the image.

If the spherical aberration coefficient is reduced to 2 mm or $C_{\bar{o}}' \approx 0.5$ with 0.25 wavelengths at the first diffraction maxima, the maximum defocus contrast rises to around 9 per cent for a gold atom with 100-kv electrons. The neglect of substrate, thermal motion and inelastic scattering again renders the visibility of single gold atoms marginal at best.

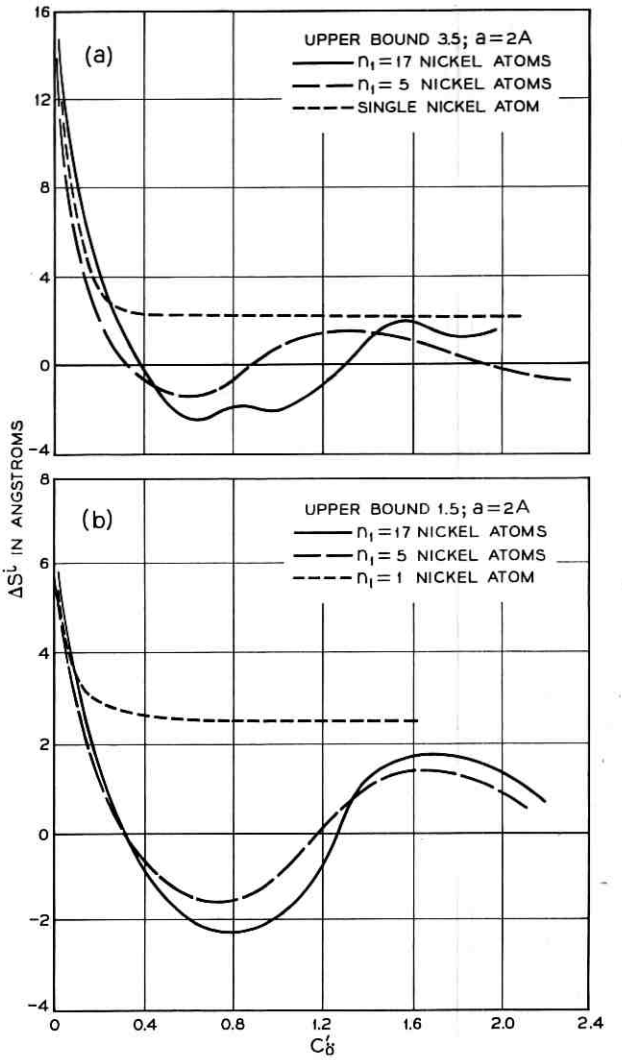


Fig. 4 — Plot of $\Delta S^i$ vs $C_\delta' = 2C_\delta \lambda^3/a^4$ for nickel atom chains of 5 and 17 atoms. The curves approximate damped, periodic functions. The kinematic contrast is $G = 4.4\ \Delta S^i$ per cent for $f = 3$ mm using 100-kv electrons. The upper bound is 3.5 diffraction maxima in (a) and 1.5 in (b).

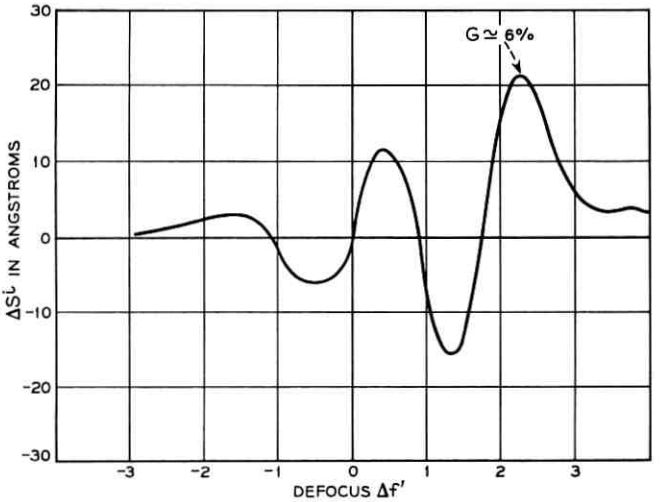


Fig. 5 — Defocus phase contrast for a chain of 5 gold atoms with 8 Å spacing and a spherical aberration parameter $C_5' = 0.74$ or 0.37 wavelengths at the first diffraction maximum ($C_5 = 3$ mm). The defocus is $\frac{1}{2}\Delta f'$ wavelengths or $\Delta f = 1730\ \Delta f'$ angstrom. The contrast is $G \approx 0.28\ \Delta S'$ per cent at 100 kv.

However, a clump of three gold atoms at the staining sites might be visible in the image, since this would multiply the contrast by a factor of roughly two. There might be also some benefit from amplitude contrast providing the phase and amplitude contrast are the *same* sign and that the image points for the two coincide. A small shift in image points by unsymmetric phase would separate the phase and amplitude image points, resulting in confusion.

If the diffraction spectra at the back focal plane are discrete points, the condition for maximum phase contrast by defocus is that the phase (11a) be $n\pi$ where $n$ is an integer. The optimum defocus parameter $\Delta f'$ is then

$$\Delta f' = (n - \tfrac{1}{2}) + C_5'$$

a relation useful in estimating the amount of defocus for best contrast. Because defocus can optimize only one object plane spacing at a time, a phase plate is much to be preferred wherein all spacings are maximized in the same image plane.

### 4.3 *Effect of Thermal Motion*

The scattering amplitude $\psi(\xi, \eta)$ in (12) assumes that the atoms in the object plane are stationary. This is not true, since they possess ther-

mal motion which is temperature-dependent and zero-point motion at the absolute zero of temperature. Detailed analysis of thermal vibration amplitudes in a solid is a complex problem[21] which need not be discussed here. To a first approximation the effect of an isotropic thermal motion is described by the Debye-Waller factor $e^{-M}$ to produce an effective atom scattering amplitude

$$(1 - \gamma^2)^{-\frac{1}{2}} f^0(s) e^{-M}.$$

This dependence has been recently experimentally verified by Horstmann and Meyer.[22] The uncertainty lies in the evaluation of $M$. For the isotropic averaged vibration case, $M$ is given by

$$\frac{8\pi^2}{\lambda^2} \langle u^2 \rangle \sin^2 \beta/2 \approx 2\pi^2 \langle u_g^{\,2} \rangle [(\xi g_1)^2 + (\eta g_2)^2]$$

for small angles. For the simple one-dimensional grating $\langle u_a \rangle^2$ is the mean square atom displacement along the chain. The Debye factor is then

$$M \approx 2\pi^2 \frac{\langle u_a^{\,2} \rangle}{a^2} (\xi^2 + \eta^2) \tag{13}$$

and the effect of thermal vibrations on contrast can be approximated by introducing $e^{-M}$ into the phase integral (12). The thermal diffuse amplitude is neglected, as is the inelastic scattering, so that again the computed contrast will be higher than could actually be expected. The effect on contrast of thermal motion alone arising through diminution of the diffracted amplitudes is illustrated using (12) and setting $C_5' = \Delta f' = 0$ and inserting a quarter-wave plate in the back focal plane. Amplitude profiles for a range of values of the relative mean square displacement were computed. The results are summarized in Fig. 6, showing $\Delta S^i$ as a function of $\langle u_a^{\,2} \rangle^{\frac{1}{2}}/a$. The contrast for stationary atoms is 60 per cent, as before, but falls to 30 for a relative mean square thermal displacement of 0.1. If spherical aberration were introduced, the contrast of atom positions would rapidly fall below that necessary for visibility in the usual microscope viewing system.

## V. DISCUSSION

The foregoing numerical results serve to point up the rather severe requirements for an objective lens system capable of yielding phase contrast images of atoms. The highly destructive influence of spherical aberration is a major hurdle that must be reduced to a minimum. If an objective lens with $C_5 \approx 1$ mm can be realized, the best contrast obtain-

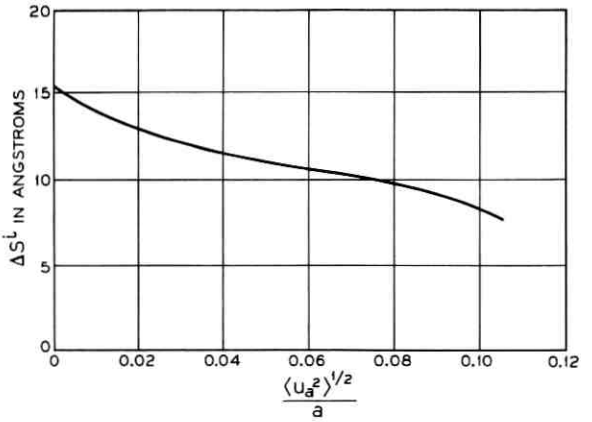


Fig. 6 — Effect of thermal motion on image profile maximum amplitude; $\langle u_a^2 \rangle^{\frac{1}{2}}$ is the root mean square amplitude along the chain. The values are for nickel atoms with $a = 2$ A and $C_5' = \Delta f' = 0$.

able with an ideal phase plate will be around 10 per cent for an individual nickel atom in a chain. The situation is much more favorable for a thin crystal of the order of 50 Å thick, wherein the diffraction spectra are greatly reinforced. For the single layer of atoms it may be necessary to turn to image intensifiers in combination with communication techniques of extracting a useful signal from background noise.

Practical considerations demand ultra-high vacuum to eliminate contamination by electron bombardment. A cryogenic stage will be useful to reduce loss of contrast due thermal motion. The problem of background noise in the image from a substrate can be circumvented either by a self-supporting specimen over a small hole or by using thin single-crystal substrates. The noise level from carbon substrates is intolerably high for this purpose.

The rapid decline of inelastic scattering cross sections with increasing accelerating voltage should offset the reduced elastic contrast predicted by (10c), suggesting that potentials in the range 150–200 kv should be appropriate. This is just opposite to the use of lower electron velocities for amplitude contrast.

It appears necessary to develop a practical quarter-wave contrast plate for the back focal plane. The thickness $t$ of a region of potential $V_0$ required to introduce a phase advance of $\pi/2$ is[23]

$$\frac{\pi V_0 t}{\lambda V_a} = \frac{\pi}{2} \quad \text{or} \quad t = \frac{\lambda V_a}{2V_0}$$

where $V_a$ is the accelerating voltage. A material film of inner potential

$V_0$ is a possibility, but the elastic and inelastic scattering by the film itself must be reckoned with.

The authors would like to acknowledge the invaluable aid of Miss Barbara Dale, who programmed the imaging integral and carried out the computational procedures and checks.

## APPENDIX

The appropriate numerical values of the amplitude in the back focal plane are obtained from tabular values of the atom scattering factors. The scale factor for $f^0(s)$ in the back focal plane requires that

$$\sqrt{\xi^2 + \eta^2} = 2a \frac{\sin \beta/2}{\lambda}.$$

For a nickel atom chain with $a = 2$Å, the first diffraction maximum is at $(\sin \beta/2)/\lambda = 0.25$ and $\sqrt{\xi^2 + \eta^2} = 1$, for which $f^0 = 3.22$ Å, corresponding to a scattering angle $\beta = \lambda/a \sqrt{\xi^2 + \eta^2} = 1.8 \times 10^{-2}$ rad. The empirical curve $f^0(\sqrt{\xi^2 + \eta^2})$ is obtained from the tabular values using a third-order Lagrange interpolation[24] stored on tape for use in evaluating the integral (12) on the IBM 7090.

The phase integral $S^i$ is put into suitable form for machine computation by dividing each of three axes, $\xi$, $\eta$ and $x/a$ into equally spaced intervals. The intervals along the three axes are $\overline{\Delta\xi}$, $\overline{\Delta\eta}$, and $\Delta(x/a)$ respectively. A point in the $\xi$-$\eta$ plane now becomes a point in a grid with coordinate $(m_1\overline{\Delta\xi}, m_2\overline{\Delta\eta})$, where $m_1$ and $m_2$ are integers. The value of the integral (12) is now approximated by a summation over the integers $m_1$ and $m_2$

$$\sum_{m_1=0} \sum_{m_2=0} f^0 \left(\sqrt{(m_1\overline{\Delta\xi})^2 + (m_2\overline{\Delta\eta})^2}\right) \frac{\sin \pi n_1 m_1 \overline{\Delta\xi}}{\sin \pi m_1 \overline{\Delta\xi}} \cos 2\pi \left(\frac{x}{a}\right) m_1 \overline{\Delta\xi}. \quad (14)$$

The Fourier integral (12) is thus approximated by a Fourier series (14). This raises the question of how well the series converges to the integral, which in turn is determined by the size of the intervals $\overline{\Delta\xi}$ and $\overline{\Delta\eta}$ and by $m_1$ and $m_2$. Since a Fourier series is periodic, it not only approximates the integral in the range of the function $S^i$ but produces repetitive images outside the range.

The importance of the intervals $\overline{\Delta\xi}$ and $\overline{\Delta\eta}$ in the back focal plane lies in the fact that they set a limit to the information available at the image plane. High fidelity of the image point amplitudes requires that $\overline{\Delta\xi}$ and $\overline{\Delta\eta}$ be as small as possible relative to the amplitude detail in the back focal plane. Along the $\eta$ axis, the amplitude is a monotonic decreasing

function with gradual changes in slope. Along the $\xi$ axis the situation is quite different, as seen in Fig. 7, where IBM 7090 microfilm plots of the diffraction amplitude or Fourier coefficient in (10d) are shown for $n_1 = 5$ and $n_1 = 17$ atoms. The interval along $\xi$ is $\overline{\Delta\xi} = 0.01$. The half width of the primary maxima is $(\Delta\xi)_{\frac{1}{2}} = 1/n_1$, so that a good approximation first requires that

$$\overline{\Delta\xi} \ll (\Delta\xi)_{\frac{1}{2}} = \frac{1}{n_1}$$

or

$$\overline{\Delta\xi} n_1 \ll 1. \tag{15}$$

Between the principle maxima of spacing unity there are $(n_1 - 2)$ subsidiary maxima and minima. The actual spacing of these subsidiaries is

$$\frac{1 - 2(\Delta\xi)_{\frac{1}{2}}}{n_1 - 2} = \frac{1}{n_1}$$

and at least three points are required to locate a maximum, zero, and then the minimum. If it is arbitrarily assumed that six points are a reasonable sampling density between adjacent maxima and minima, the sampling interval must be

$$\overline{\Delta\xi} \leqq \frac{1}{6n_1}$$

or

$$\overline{\Delta\xi} n_1 \leqq \frac{1}{6}.$$

On this basis, then, $\overline{\Delta\xi} = 0.01$ is adequate for $n_1 = 5$. If $n_1 = 17$, however, then $\overline{\Delta\xi} \leqq 1/102$ or the interval $\overline{\Delta\xi}$ is just adequate. As the number of atoms in the chain increases further the detail in the amplitude distribution soon becomes smaller than $\overline{\Delta\xi} = 0.01$. The interval $\overline{\Delta\xi} = 0.01$ is thus considered too crude for $n_1 > 25$ atoms. This is admittedly rather arbitrary, and a more detailed analysis might discover a better criterion.

Along the $\eta$ axis, on the other hand, an interval $\overline{\Delta\eta} = 0.1$ is quite adequate and reduces the number of sampling points. Under these conditions the number of sample points in the back focal plane for each image point is

$$\frac{\eta_{\max}}{\Delta\eta}\frac{\xi_{\max}}{\eta\xi} = 12.7 \times 10^3; \qquad \xi_{\max} = \eta_{\max} = 3.5. \tag{16}$$

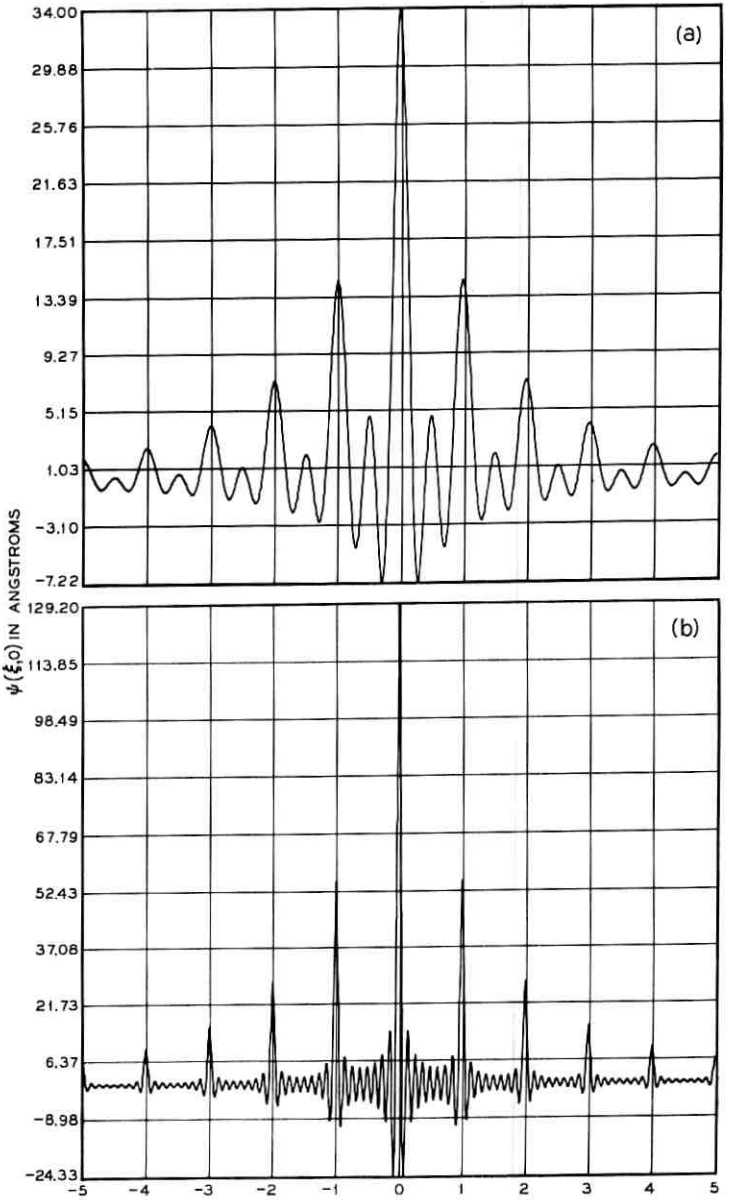


Fig. 7 — Diffracted amplitude distributions at the back focal plane as obtained from the microfilm tracer on the IBM 7090 for a sampling interval $\overline{\Delta\xi} = 0.01$; the grating spacing is $a = 2$ Å: (a) for a chain of 5 nickel atoms; (b) for a chain of 17 nickel atoms.

The remaining consideration relating to the sample intervals is *aliasing*. This behavior can be illustrated by returning to (14). Writing the coefficient as $\phi(\xi, \eta)$, the series sums at $(x/a) = 0$ to

$$S_{\text{real}}{}^{i}(0) = \sum_{m_1=0} \sum_{m_2=0} \phi(\xi,\eta). \tag{17a}$$

If the sampling interval along $x$ is $\Delta(x/a)$, then the sum at an image point $x = j\Delta(x/a)$ with $j$ an integer is

$$S_{\text{real}}{}^{i}\left(j\Delta\left(\frac{x}{a}\right)\right) = \sum_{m_1=0} \sum_{m_2=0} \phi(\xi,\eta) \cos 2\pi j\Delta\left(\frac{x}{a}\right) m_1\overline{\Delta\xi}. \tag{17b}$$

If $2\pi j\Delta(x/a)m_1\overline{\Delta\xi}$ = multiple of $2\pi$ then the cosine terms are unity and (17a) and (17b) are equal. Thus

$$j\Delta\left(\frac{x}{a}\right)\overline{\Delta\xi} = \text{integer}. \tag{17c}$$

The *aliasing period*[25] is then

$$\Delta j = \frac{1}{\Delta\left(\frac{x}{a}\right)\overline{\Delta\xi}}$$

or

$$\text{alias period} = \frac{1}{\overline{\Delta\xi}}. \tag{17d}$$

Thus the critical sampling interval for these computations is $\overline{\Delta\xi}$, since it determines both the resolution (15) and the aliasing period. The effect is well illustrated in Fig. 8, comparing the profiles for $n_1 = 5$ with $\overline{\Delta\xi} = 0.01$ and $\overline{\Delta\xi} = 0.1$. The repeating nature of the image or aliasing is evident for $\overline{\Delta\xi} = 0.1$ when the aliasing period is only ten. With $\overline{\Delta\xi} = 0.01$, the period is 100 and not seen on the plot. If the number of atoms $n_1$ were now increased to nine the image profile with $\overline{\Delta\xi} = 0.1$ would show no break at the end of an atom chain and so would appear to be an infinite chain.

The sampling interval along the $x^i$ axis in the image plane has an effect on the representation of the finer details in the profile. When the detail approaches the interval length $\Delta(x/a)$ the representation becomes inaccurate with "bumps" and "angles" rather than a smooth curve. This is illustrated in Fig. 9 for the case of a single atom, $n_1 = 1$, for which

$$S^i\left(j\Delta\left(\frac{x}{a}\right)\right) = \sum_{m_1=0} \sum_{m_2=0} f(\sqrt{(m_1\overline{\Delta\xi})^2 + (m_2\overline{\Delta\eta})^2}) \cos 2\pi j\Delta\left(\frac{x}{a}\right) m_1\overline{\Delta\xi}$$

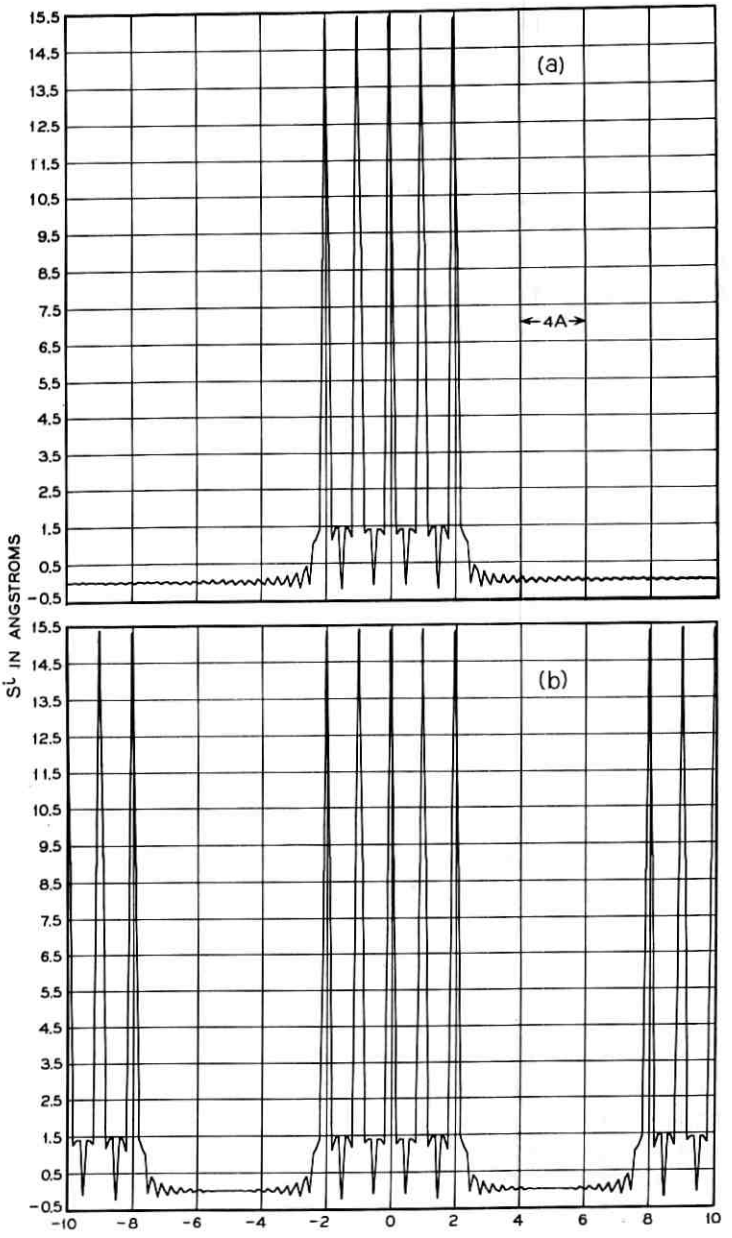


Fig. 8 — Computer-drawn image plane amplitude profiles along the axis $x^\circ$ of the nickel atom chain. For this case $a = 2$ Å, $n_1 = 5$ and $C_{\ddot{0}}' = \langle u_a^2 \rangle = 0$ in equation (12a). The sampling intervals in (a) are $\overline{\Delta\xi} = 0.01$ and $\Delta(x^\circ/a) = \overline{\Delta\eta} = 0.1$. In (b) the sampling interval $\overline{\Delta\xi} = 0.1$ and the profile shows an aliasing period of $(\overline{\Delta\xi})^{-1} = 10$. The aliasing period in (a) is 100.

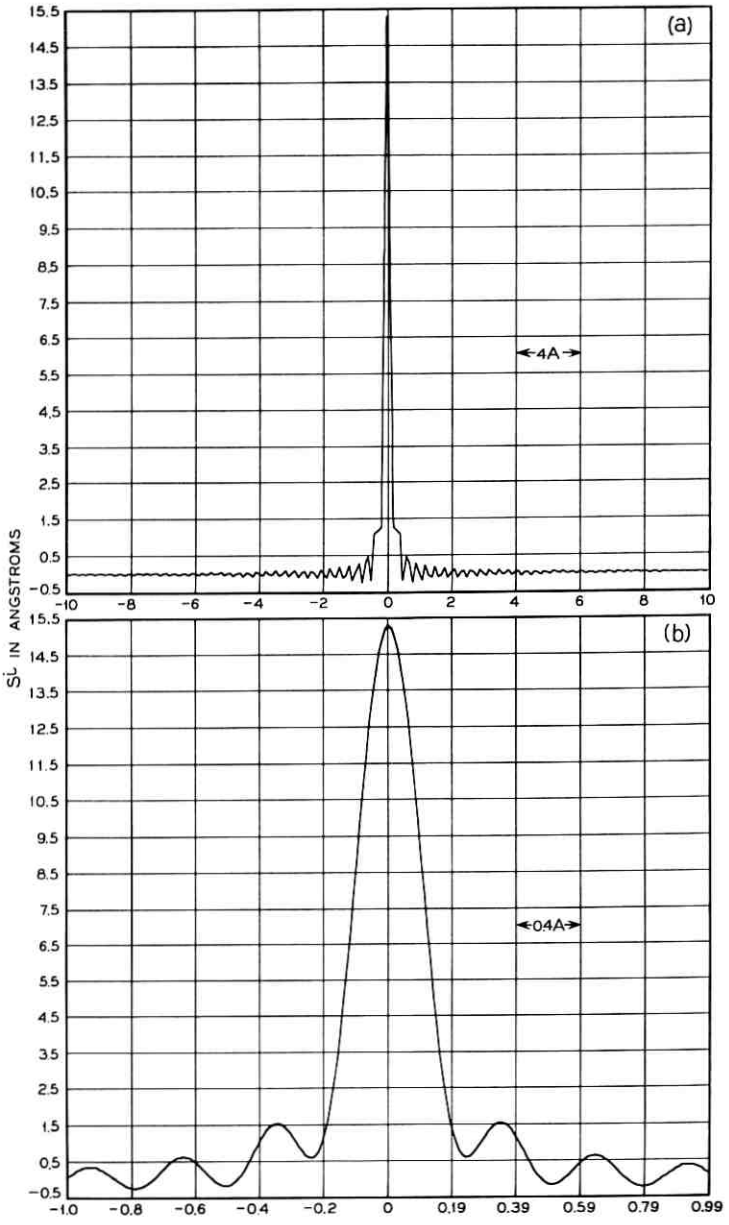


Fig. 9 — Profiles for equation (12) with $n_1 = \underline{1}$ nickel atom, $C_5' = 0$ and upper bounds $\xi_{max} = \eta_{max} = 2.5$ diffraction maxima; $\overline{\Delta\xi} = 0.01$ and $\Delta\eta = 0.1$. (a) Interval $\Delta(x/a) = 0.1$ resulting in unresolved structure near origin. (b) Interval $\Delta(x/a) = 0.01$ with expanded scale showing resolution of structure.

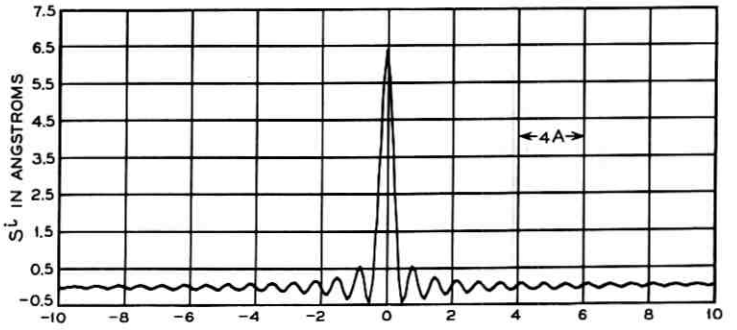


Fig. 10 — Effect of upper bound on $\Delta S^i$: $n_1 = 1$ nickel atom and $C_\delta' = 0$. The upper bound is 1.5 diffraction maxima and $\Delta(x/a) = \overline{\Delta\eta} = 0.1$ and $\overline{\Delta\xi} = 0.01$. Compare with Fig. 9.

with $\overline{\Delta\xi} = 0.01$, $\overline{\Delta\eta} = 0.1$ and $\Delta(x/a) = 0.1$. The profile is shown in Fig. 9(a). The number of points involved in this profile is

$$\frac{\eta_{\max}}{\overline{\Delta\eta}} \cdot \frac{\epsilon_{\max}}{\overline{\Delta\xi}} \cdot \frac{20}{\Delta\left(\frac{x}{a}\right)} = 25.4 \times 10^5.$$

The neighborhood of the central maximum shows irregularities that are smoothed out to a more faithful representation when $\Delta(x/a) = 0.01$ as seen in Fig. 9(b). The number of points involved in the profile is the same since the range has been reduced by a factor of ten.

The subsidiary maxima and minima near the principal maximum of Fig. 9(b) are due to aperture diffraction and would smooth out if the upper limit of the integral were extended to infinity. The radius of the atom in the profile is not distinct, since the shoulder of the maximum will approach the axis asymptotically.

The size of the aperture given by the upper bounds $\xi_{\max}$ and $\eta_{\max}$ is reduced from 3.5 in Fig. 9(b) to 1.5 in Fig. 10. The reduction in aperture results in a loss of contrast to about 8 per cent. The upper bound 1.5 corresponds to an objective half-angle of $2.8 \times 10^{-2}$ rad.

REFERENCES

1. Heidenreich, R. D., *Fundamentals of Transmission Electron Microscopy*. Interscience, New York, 1964; see Ch. V for elementary discussion.
2. *International Crystallographic Tables*, Vol. III, Kynoch Press, Birmingham, England, 1962.
3. Burge, R. E., and Smith, G. H., Proc. Phys. Soc., *79*, 1962, p. 673.
4. Williams, E. J., Proc. Roy. Soc., *A139*, 1933, p. 163.
5. Ref. 1, p. 308.

6. Ref. 1, Ch. X.
7. Ref. 1, p. 308.
8. Ref. 1, p. 107.
9. See Ref. 1, Ch. VIII for a detailed discussion of both the kinematic and dynamical theories.
10. Ref. 1 and also James, R. W., *Optical Principles of the Diffraction of X-Rays*, G. Bell and Sons, Ltd., London, 1954.
11. See Ref. 1, Appendix B.
12. Scherzer, O., J. Appl. Phys., *20*, 1949, p. 20.
13. Haine, M. E., *Electron Microscope*, Interscience, New York, 1961, p. 54.
14. See Ref. 1, p. 230.
15. Ref. 1, p. 140.
16. Ref. 1, p. 109.
17. For a discussion, see Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1962, p. 295.
18. See Ref. 1, p. 330.
19. See Ref. 1, p. 312.
20. Beer, M., J. Molecular Biology, *7*, 1963, p. 70.
21. Maradudin, A. A., Montroll, E. W., and Weiss, G. H., Solid-State Physics, *Advances in Research and Application*, Supp. 3, Academic Press, New York, 1963.
22. Horstmann, M., and Meyer, G., Phys. Kondens. Materie, *1*, 1963, p. 208.
23. Ref. 1, pp. 127 and 140.
24. Ref. 17, Ch. 8.
25. Ref. 17, p. 276.

# Design of Bandlimited Signals for Binary Communication Using Simple Correlation Detection*

By B. R. SALTZBERG and L. KURZ

(Manuscript received July 8, 1964)

*This paper considers the design of binary bandlimited signals for transmission over a channel with additive white Gaussian noise, the signals to be received by a memoryless correlation detector. A signal waveform is found which allows communication at the Nyquist rate without intersymbol interference and with 1.3 db degradation compared to an optimum communication system. Other waveforms, consisting of the sum of a few prolate spheroidal functions, are also investigated.*

## I. INTRODUCTION

In the reception of serial binary data transmitted over a noisy bandlimited channel, errors result from the combined effects of intersymbol interference and noise. Minimization of the error rate involves appropriate design of both the transmitted signal and the method of detection, taking into account the effects of both causes of degradation.

Nyquist has shown how bandlimited signals may be designed so as to eliminate intersymbol interference when detection is accomplished by periodic instantaneous sampling.[1] Sunde has shown that optimum performance over a channel with white Gaussian noise is achieved when the shaping is divided equally between the transmitter and receiver.[2] Tufts has developed a technique of long memory detection which eliminates intersymbol interference and optimizes noise performance subject to that constraint, for an arbitrary transmitted signal.[3] Kurz and Trabka have studied the design of signals for transmission in the presence of nonwhite noise without the problem of intersymbol interference.[4,5]

* This paper is based on parts of a thesis accepted by the faculty of the Graduate Division of the School of Engineering and Science of New York University in partial fulfillment of the requirements for the degree of Doctor of Engineering Science.

This paper discusses the design of bandlimited signals for communication in the presence of white Gaussian noise, when the detector is a memoryless correlator. Memoryless correlation is a widely used suboptimum means of detection. It will be shown in Section III that we can communicate without intersymbol interference at the Nyquist rate using memoryless correlation. In Section IV we investigate another form of signaling for communication with memoryless correlation. Here signals are chosen which do not eliminate intersymbol interference, but lead to low error probability for the most adverse message sequence.

## II. PRELIMINARIES

In serial binary transmission, the $n$th binary digit of the message is transmitted by sending either $s_0(t - nT)$ or $s_1(t - nT)$. We will assume that the a priori probabilities of $s_0$ and $s_1$ are $1/2$ and that all digits are independent. The transmitted information rate is therefore $1/T$ bits per second.

If the signal is perturbed by additive white Gaussian noise, the optimum detector is well known to be a simple correlator if $s_0(t)$ and $s_1(t)$ are time limited to an interval of length $T$.[6] Such a detector chooses $s_0$ if and only if

$$\int v(t)s_0(t - nT)dt - \frac{1}{2}\int s_0^2(t - nT)dt > \int v(t)s_1(t - nT)dt - \frac{1}{2}\int s_1^2(t - nT)dt$$

where $v(t)$ is the received signal and the integration is taken over the interval of length $T$.

A polar signal leads to minimum error probability:[7]

$$s_0(t) = -s_1(t) = f(t).$$

The correlation detector then chooses $s_0$ if and only if

$$\int v(t)f(t - nT)dt > 0.$$

If $f(t)$ is not time limited to an interval of length $T$, as is inevitable if it is bandlimited, then the memoryless correlator is a suboptimum detector because it does not make use of the signal energy outside the interval. An infinite memory correlator or, equivalently, a matched filter and sampler, is the optimum detector, provided that intersymbol interference can be eliminated. The memoryless correlator, however,

has found extensive practical application. With proper choice of $f(t)$, the degradation as compared with optimum detection need not be too large.

Aein and Hancock have shown that some improvement of the memoryless correlator can be obtained in the presence of intersymbol interference by modifying the correlating function.[8] However, this procedure is sensitive to amplitude variations of both the signal and the noise, whereas the simple correlation detector is not. We will therefore use the simple correlator and seek to minimize error probability through the choice of $f(t)$.

We will shift the time axis so that the origin is in the center of the bit to be detected, and assume that an infinite number of bits has been transmitted both before and after the bit currently being detected.

$$v(t) = \sum_{k=-\infty}^{\infty} a_k f(t + kT) + n(t)$$

where $a_k = \pm 1$ and $n(t)$ is a member function of a stationary Gaussian random process with autocorrelation $N[\delta(t)]/2$. The one-sided spectral density of the noise is therefore $N$.

Since we are using a simple correlation detector, $a_0$ will be chosen as

$$\{a_0\} = \operatorname{sgn} \int_{-T/2}^{T/2} v(t) f(t) dt$$

where

$$\operatorname{sgn} x = \frac{x}{|x|}.$$

The choice of $a_0$ when

$$\int_{-T/2}^{T/2} v(t) f(t) dt = 0$$

is not important, since this event occurs with zero probability.

$$Q = \int_{-T/2}^{T/2} v(t) f(t) dt$$

is a linear functional of a Gaussian process and is therefore itself normally distributed for a given sequence $[a_k]$. Its expected value is

$$E(Q) = \sum_{k=-\infty}^{\infty} a_k \int_{-T/2}^{T/2} f(t + kT) f(t) dt,$$

which may be written as

$$E(Q) = a_0 d^2 + \sum_{k \neq 0} a_k \rho_k d^2$$

where

$$d^2 = \int_{-T/2}^{T/2} f^2(t)dt \tag{1}$$

and

$$\rho_k = \frac{1}{d^2} \int_{-T/2}^{T/2} f(t + kT) f(t) dt. \tag{2}$$

The variance of $Q$ is

$$\text{Var } (Q) = \frac{N}{2} d^2.$$

The probability density of $Q$ is therefore

$$p(Q) = \frac{1}{\sqrt{\pi N}\, d} \exp\left[ -\frac{1}{Nd^2} \left(Q - a_0 d^2 - \sum_{k \neq 0} a_k \rho_k d^2\right)^2 \right].$$

We may now calculate the probability of error as

$$p(e \mid a_0 = +1) = \frac{1}{\sqrt{\pi N}\, d} \int_{-\infty}^{0} \exp\left[ -\frac{1}{Nd^2} \left(Q - d^2 - \sum_{k \neq 0} a_k \rho_k d^2\right)^2 \right] dQ$$

$$p(e \mid a_0 = -1) = \frac{1}{\sqrt{\pi N}\, d} \int_{0}^{\infty} \exp\left[ -\frac{1}{Nd^2} \left(Q + d^2 - \sum_{k \neq 0} a_k \rho_k d^2\right)^2 \right] dQ.$$

These expressions reduce to

$$p(e \mid a_0 = +1) = \frac{1}{2} \text{erfc}\left[ \frac{d}{\sqrt{N}} \left(1 + \sum_{k \neq 0} a_k \rho_k\right) \right]$$

$$p(e \mid a_0 = -1) = \frac{1}{2} \text{erfc}\left[ \frac{d}{\sqrt{N}} \left(1 - \sum_{k \neq 0} a_k \rho_k\right) \right]$$

where

$$\text{erfc } (x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt.$$

The maximum probability of error occurs when

$$a_k = -a_0 \text{ sgn } \rho_k \quad \text{for all} \quad k \neq 0 \tag{3}$$

$$p_{\max} = \frac{1}{2} \text{erfc}\left[ \frac{d}{\sqrt{N}} \left(1 - \sum_{k \neq 0} |\, \rho_k \,|\right) \right]. \tag{4}$$

It may be noted that if $\rho_k = 0$ for all $k \neq 0$, then the probability of

error is independent of the message sequence and is equal to

$$\tfrac{1}{2} \operatorname{erfc} (d/\sqrt{N}).$$

This is the case of no intersymbol interference, and the error probability is a monotone decreasing function of $d$. If intersymbol interference does exist, the error probability is greatest for the sequence (3) and is given by (4). The average error probability is tedious to calculate, but may be readily approximated.[9]

Equation (4) may be compared with the error probability for optimum detection[10]

$$p_e = \tfrac{1}{2} \operatorname{erfc} (A/\sqrt{N})$$

where

$$A^2 = \int_{-\infty}^{\infty} f^2(t)dt).$$

It is extremely desirable that the system perform error-free in the absence of noise, $N = 0$. Since

$$\begin{aligned} \lim_{N \to 0} p_{\max} &= 0, \quad \text{if} \quad \sum_{k \neq 0} |\rho_k| < 1 \\ &= \tfrac{1}{2}, \quad \text{if} \quad \sum_{k \neq 0} |\rho_k| = 1 \\ &= 1, \quad \text{if} \quad \sum_{k \neq 0} |\rho_k| > 1, \end{aligned}$$

we will reject any system for which $\sum_{k \neq 0} |\rho_k| \geqq 1$, since in this case there will be some sequence of binary digits that cannot be received without error.

## III. SIGNALS WITHOUT INTERSYMBOL INTERFERENCE

In order to avoid intersymbol interference with memoryless correlation detection, it is necessary that

$$\rho_k = \int_{-T/2}^{T/2} f(t)f(t + kT)dt = d^2\delta_{0k}. \tag{5}$$

We will seek bandlimited functions $f(t)$ which satisfy (5) by using an unpublished method of H. O. Pollak.

Let $F(\omega)$ be the Fourier transform of $f(t)$. If $f(t)$ is bandlimited to $|\omega| < \omega_c$, then

$$f(t) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} F(\omega)e^{j\omega t}d\omega$$

and

$$\rho_k = \frac{1}{2\pi}\int_{-\omega_c}^{\omega_c}\int_{-\omega_c}^{\omega_c} F(\omega)F^*(x)\,\frac{\sin\,(\omega - x)\,\frac{T}{2}}{\pi(\omega - x)}\,e^{j\omega kT}d\omega\,dx.$$

Let

$$G(\omega) = \int_{-\omega_c}^{\omega_c} F^*(x)\,\frac{\sin\,(\omega - x)\,\frac{T}{2}}{\pi(\omega - x)}\,dx. \tag{6}$$

Then

$$\rho_k = \frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} F(\omega)G(\omega)e^{j\omega kT}d\omega.$$

We now divide the interval $(-\omega_c, \omega_c)$ into subintervals of length $2\pi/T$

$$\rho_k = \frac{1}{2\pi}\sum_{n=-N}^{N}\int_{(2n-1)\pi/T}^{(2n+1)\pi/T} F(\omega)G(\omega)e^{j\omega kT}d\omega$$

where

$$N \geqq \frac{1}{2}\left(\frac{\omega_c T}{\pi} - 1\right)$$

$$\rho_k = \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T}\sum_{n=-N}^{N} F\left(\omega + \frac{2n\pi}{T}\right)G\left(\omega + \frac{2n\pi}{T}\right)e^{j\omega kT}d\omega. \tag{7}$$

Equation (7) indicates that the $\rho_k$'s are the Fourier coefficients of the function

$$H(\omega) = \sum_{n=-N}^{N} F\left(\omega + \frac{2n\pi}{T}\right)G\left(\omega + \frac{2n\pi}{T}\right). \tag{8}$$

Since $\rho_k = 0$ for all $k \neq 0$, $H(\omega)$ must be a constant independent of $\omega$. Using (5) and (7), we find that

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

If $\omega_c T \leqq \pi$, then we may choose $N = 0$, and (9) reduces to

$$F(\omega)G(\omega) = Td^2, \qquad -\frac{\pi}{T} \leqq \omega \leqq \frac{\pi}{T}. \tag{10}$$

Equation (10) cannot be satisfied if $F(\omega) = 0$ for any $\omega$ in the in-

terval $(-\pi/T,\pi/T)$. Therefore intersymbol interference cannot be avoided if $\omega_c < \pi/T$.

Let us now investigate the case $\omega_c = \pi/T$. Substituting (6) into (10)

$$F(\omega) \int_{-\pi/T}^{\pi/T} F^*(x) \frac{\sin (\omega - x) \frac{T}{2}}{\pi(\omega - x)} dx = Td^2. \tag{11}$$

In an unpublished work, Pedro Nowosad has proved that the quadratic integral equation (11) has a continuous, real, positive solution $F(\omega)$.

Equation (11) has been solved numerically by assuming an arbitrary $F_0(\omega)$ and iteratively finding

$$\frac{1}{F_n(\omega)} = \int_{-\pi/T}^{\pi/T} F_{n-1}(x) \frac{\sin (\omega - x) \frac{T}{2}}{\pi(\omega - x)} dx.$$

The resultant amplitude spectrum $F(\omega)$ is plotted in Fig. 1. The corresponding time function $f(t)$ is plotted in Fig. 2. Since both $F(\omega)$ and $f(t)$ are even functions, only the positive abscissas are shown. The

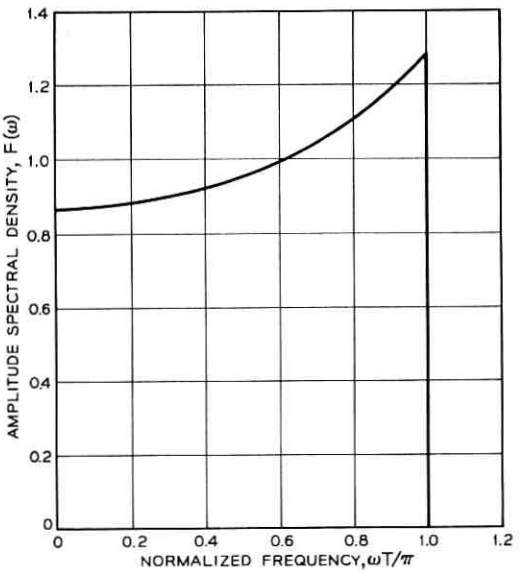


Fig. 1 — Spectrum of the signal which permits transmission without intersymbol interference.

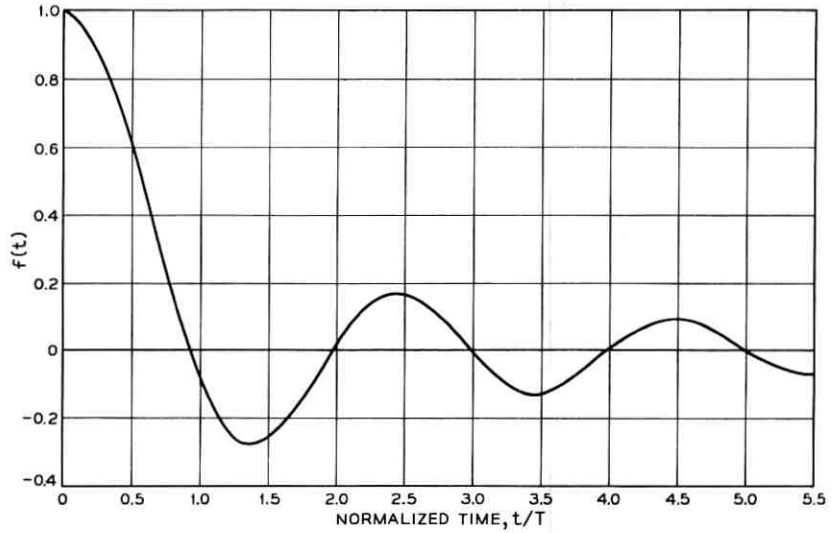


Fig. 2 — Bandlimited signal which permits transmission without intersymbol interference.

signal has been normalized for unit total energy. This time function does indeed satisfy (5) with $d^2 = 0.744$.

Digital communication using bandlimited signals and memoryless correlation detection can therefore be achieved without intersymbol interference at the Nyquist rate, $1/T = \omega_c/\pi$. The resultant degradation in the presence of white Gaussian noise, when compared with optimum detection of the signal $\sin \omega_c t/t$, is $-10 \log_{10} 0.744 = 1.3$ db.

A disadvantage which $f(t)$ as shown in Fig. 2 shares with $\sin \omega_c t/t$ is that $\Sigma f(t + nT)$ does not converge absolutely. Very large amplitudes may therefore be caused by certain sequences. If $\omega_c T > \pi$, then signals which converge more rapidly can easily be designed for detection by sampling. It is expected that solutions of (9) exist which also make such use of the additional available bandwidth. However, no such signals have as yet been found.

## IV. OTHER SIGNALS

It is not at all necessary that intersymbol interference be eliminated in order to achieve reliable digital communication. For use with memoryless correlation, a signal with some intersymbol interference may very likely lead to a lower error probability than a signal with no intersymbol interference but with less of its energy in the principal time interval.

In this section we will drop the constraint of no intersymbol interference. The probability of error is therefore dependent on the message sequence. We will use a minimax type of criterion for designing the bandlimited signal. That is, we will attempt to minimize the probability of error for the worst sequence. It is believed that the minimax criterion may be more realistic than an average error rate criterion, since the latter approach does not prevent the possibility of having some extremely sensitive message sequences. It is possible that such sensitive messages cannot be transmitted without error even over a noiseless channel. A further advantage of the minimax criterion is that it leads to a solution which does not require knowledge of the noise level.

An additional constraint that will be imposed is that the signal amplitude remain bounded for any message sequence.

We will attempt to minimize $p_{\max}$ as given by (4). From (4), $p_{\max}$ is a monotone decreasing function of

$$D = d(1 - \sum_{k \neq 0} |\rho_k|). \tag{12}$$

We can therefore satisfy the minimax criterion by maximizing the separation function, $D$. It is convenient to scale the amplitude of $f(t)$ so that $d = 1$. Such scaling, of course, affects the total energy of the signal. However, the quantity $D/\sqrt{E}$ remains invariant under such scaling, and we may accordingly maximize the quantity

$$D' = \frac{1}{\sqrt{E}}\left[1 - \sum_{k \neq 0} \left| \int_{-T/2}^{T/2} f(t)f(t + kT)dt \right| \right] \tag{13}$$

and the resultant $f(t)$ may later be multiplied by the factor $\sqrt{A^2/E}$ in order to satisfy the fixed energy requirement. Here, $A^2$ is the required energy, while $E$ is the energy of the scaled signal.

It is also convenient to scale the time axis so that $T = 2$ and $\omega_c = c$, where the normalized bandwidth, $c = \frac{1}{2}\omega_c T$. Note that $c = \pi/2$ corresponds to transmission at the Nyquist rate.

We will make use of the properties of the prolate spheroidal functions, $\psi_i(t)$, which are extensively discussed and plotted by Slepian, Landau and Pollak.[11,12] Some of these properties are

$$\int_{-\infty}^{\infty} \psi_i(t)\psi_j(t)dt = \delta_{ij}$$

$$\int_{-1}^{1} \psi_i(t)\psi_j(t)dt = \lambda_i\delta_{ij},$$

where $\lambda_i$ is the $(i + 1)$th largest eigenvalue of

$$\lambda\psi(t) = \int_{-1}^{1} \psi(v) \frac{\sin c(t - v)}{\pi(t - v)} dv.$$

$\psi_i(t)$ is the eigenfunction corresponding to $\lambda_i$. Both $\psi_i$ and $\lambda_i$ depend on the parameter $c$.

Since $f(t)$ is a bandlimited function, it may be expressed as a series of prolate spheroidal functions[11]

$$f(t) = \sum_{n=0}^{\infty} \gamma_n \psi_n(t). \tag{14}$$

If we set

$$\gamma_n = \frac{\beta_n}{\sqrt{\lambda_n}},$$

then

$$f(t) = \sum_{n=0}^{\infty} \beta_n \frac{\psi_n(t)}{\sqrt{\lambda_n}}. \tag{15}$$

The functions $\psi_n(t)/\sqrt{\lambda_n}$ are orthonormal over the interval $(-1,1)$. $f(t)$ can be expressed as a vector $F = [\beta_0, \beta_1, \cdots]$, with orthonormal basis

$$\left[\frac{\psi_0(t)}{\sqrt{\lambda_0}}, \frac{\psi_1(t)}{\sqrt{\lambda_1}}, \cdots\right].$$

The energy in the interval $(-1,1)$ is equal to

$$\int_{-1}^{1} f^2(t)dt = FF^t = \beta_0^2 + \beta_1^2 + \cdots = 1,$$

where $F^t$ is the transpose of $F$. The total energy is equal to

$$E = \int_{-\infty}^{\infty} f^2(t)dt = \frac{\beta_0^2}{\lambda_0} + \frac{\beta_1^2}{\lambda_1} + \cdots = F\Lambda F^t$$

where $\Lambda$ is a diagonal matrix with elements $\Lambda_{ij} = \delta_{ij}/\lambda_i$.

Since $f(t)$ is bandlimited, $f(t)$ in the interval $(-1,1)$ determines $f(t)$ for all time.

$$f(t + 2k) = \sum_{n=0}^{\infty} \beta_n \frac{\psi_n(t + 2k)}{\sqrt{\lambda_n}}.$$

$\psi_n(t + 2k)/\sqrt{\lambda_n}$ can itself be expanded as

$$\frac{\psi_n(t+2k)}{\sqrt{\lambda_n}} = \sum_{m=0}^{\infty} t_{mnk} \frac{\psi_m(t)}{\sqrt{\lambda_m}}$$

where

$$t_{mnk} = \frac{1}{\sqrt{\lambda_m \lambda_n}} \int_{-1}^{1} \psi_m(t)\psi_n(t+2k)dt \tag{16}$$

so that

$$f(t+2k) = \sum_n \beta_n \sum_m t_{mnk} \frac{\psi_m(t)}{\sqrt{\lambda_m}}$$

or, in matrix form,

$$F_k = FT_k{}^t \tag{17}$$

where the elements of $T_k$ are $t_{ijk}$ as given by (16).

We can now express the intersymbol interference terms as

$$\int_{-1}^{1} f(t)f(t+2k)dt = FF_k{}^t = FT_kF^t. \tag{18}$$

Then

$$D' = \frac{1 - \sum_{k\neq 0} b_k F T_k F^t}{\sqrt{F\Lambda F^t}} \tag{19}$$

where

$$b_k = \text{sgn } (FT_kF^t).$$

We seek suboptimum solutions by confining $F$ to $M$ dimensions. That is, we will seek an optimum $f_M(t)$ of the form

$$f_M(t) = \sum_{n=0}^{M-1} \frac{\beta_n \psi_n(t)}{\sqrt{\lambda_n}}. \tag{20}$$

Such an approach is justified if

$$\lim_{M\to\infty} f_M(t) = f(t),$$

the true optimum solution, and this convergence is sufficiently rapid. All vectors in the previous development are now $M$-dimensional and all square matrices are $M \times M$. Note that (17) is no longer strictly correct, but instead gives the projection of $F_k$ in the $M$-dimensional space. Equations (18) and (19), however, remain valid.

At this point we will introduce the constraint which requires that the

total signal amplitude remain bounded for any message sequence. This is highly desirable physically, because of the effects of inexact timing and the technical impossibility of handling unbounded signals.

For the worst sequence,

$$s_{\max}(t) = \sum_{k=-\infty}^{\infty} |f(t + 2k)|$$

and we wish to constrain $f(t)$ so that $s_{\max}(t)$ remain bounded. We first express $\psi_n(t)$ as a multiple of the radial prolate spheroidal function:[13]

$$\psi_n(t) = \frac{R_n(t)}{K_n}$$

where

$$K_n^2 = \int_{-\infty}^{\infty} R_n^2(t)dt.$$

Since[11]

$$\lambda_n = \frac{2c}{\pi} R_n^2(1)$$

we can also express $K_n^2$ as

$$K_n^2 = \frac{\pi \int_{-1}^{1} R_n^2(t)dt}{2cR_n^2(1)}.$$

Then

$$f(t) = \sum_{n=0}^{M-1} \frac{\beta_n}{K_n\sqrt{\lambda_n}} R_n(t)$$

$$s_{\max}(t) = \sum_{k=-\infty}^{\infty} \left| \sum_{n=0}^{M-1} \frac{\beta_n}{K_n\sqrt{\lambda_n}} R_n(t + 2k) \right|.$$

For large $|t|$, $R_n(t)$ can be expressed asymptotically by[13]

$$R_n(t) = (-1)^{n/2} \frac{\sin ct}{ct} + O(t^{-2}), \qquad n \text{ even}$$

$$R_n(t) = (-1)^{(n+1)/2} \frac{\cos ct}{ct} + O(t^{-2}), \qquad n \text{ odd}.$$

Let us examine

$$s_N(t) = \sum_{|k|=N}^{\infty} \left| \sum_{n=0}^{M-1} \frac{\beta_n}{K_n\sqrt{\lambda_n}} R_n(t + 2k) \right|$$

$$s_N(t) \leqq \sum_{|k|=N}^{\infty} \left| \sum_{n=0}^{M-1} \frac{\beta_n}{K_n\sqrt{\lambda_n}} O(k^{-2}) \right| + \sum_{|k|=N}^{\infty} \left| \sum_{n \text{ even}} (-1)^{n/2} \frac{\beta_n}{K_n\sqrt{\lambda_n}} \frac{\sin c(t+2k)}{c(t+2k)} \right| + \sum_{|k|=N}^{\infty} \left| \sum_{n \text{ odd}} (-1)^{(n+1)/2} \frac{\beta_n}{K_n\sqrt{\lambda_n}} \frac{\cos c(t+2k)}{c(t+2k)} \right|.$$

The last two series diverge, except for isolated values of $c$ and $t$. Sufficient conditions for $s_{\max}(t)$ to be bounded are therefore

$$\sum_{n \text{ even}} (-1)^{n/2} \frac{\beta_n}{K_n\sqrt{\lambda_n}} = 0 \tag{21}$$

and

$$\sum_{n \text{ odd}} (-1)^{(n+1)/2} \frac{\beta_n}{K_n\sqrt{\lambda_n}} = 0. \tag{22}$$

These equations confine $F$ to an $(M - 2)$-dimensional subspace orthogonal to the two vectors

$$V_0 = \left[ \frac{1}{K_0\sqrt{\lambda_0}}, 0, -\frac{1}{K_2\sqrt{\lambda_2}}, 0, \frac{1}{K_4\sqrt{\lambda_4}}, \cdots \right]$$

$$V_1 = \left[ 0, \frac{1}{K_1\sqrt{\lambda_1}}, 0, -\frac{1}{K_3\sqrt{\lambda_3}}, 0, \cdots \right]$$

$$FV_0{}^t = FV_1{}^t = 0.$$

We can form an orthogonal matrix $V$ in which the first two rows are $CV_0$ and $KV_1$, and the remaining $M - 2$ rows are any vectors such that the $M$ rows form an orthonormal set. The last $M - 2$ rows may, for example, be chosen by the Gram-Schmidt orthogonalization process. We may then form

$$G = FV^t = FV^{-1}$$

since $V^t = V^{-1}$ for an orthogonal matrix. Due to the above constraints, the first two components of $G$, $g_1$ and $g_2 = 0$. Since $V$ is an orthogonal transformation, $GG^t = FF^t = 1$.

We may also form matrices $U_k$ from $T_k$. Since $T_k$ is used only in the quadratic form (18), we need only consider its symmetric component, $T_k'$, in which $t_{ijk}{}' = t_{jik}{}' = \frac{1}{2}(t_{ijk} + t_{jik})$. Then

$$FT_kF^t = FT_k'F^t = GVT_k'V^tG^t.$$

Let $U_k = VT_k'V^t$. $U_k$ is a symmetric matrix since it is congruent to $T_k'$, a symmetric matrix.

$$FT_kF^t = GU_kG^t.$$

If we also let $\Theta = V\Lambda V^t$

$$D' = \frac{1 - \sum_{k\neq 0} b_k GU_kG^t}{\sqrt{G\Theta G^t}}. \tag{23}$$

We find the optimum $M$-dimensional signal $f(t)$ by varying the unit-length, $(M-2)$-dimensional vector $G$ so as to maximize $D'$ given by (23), and then perform the inverse transformation and scaling. Note that if $f(t)$ is constrained to be either an even or an odd function, only terms of even or odd $n$ appear in (14), and only one of the constraints (21) or (22) is needed.

The resultant $f(t)$ is of the form (20). Landgrebe and Cooper have shown that the Fourier transform of $\psi_n(t)$ is[14]

$$\begin{aligned} \mathfrak{F}[\psi_n(t)] &= j^{-n}\sqrt{\frac{2\pi}{\lambda_n c}}\,\psi_n\left(\frac{\omega}{c}\right), && |\omega| < c \\ &= 0, && |\omega| > c. \end{aligned}$$

Therefore the optimum $f(t)$ may be generated by passing an impulse through a filter whose frequency response is

$$\begin{aligned} H(\omega) &= K\sum_{n=0}^{M-1} j^{-n}\beta_n\psi_n\left(\frac{\omega}{c}\right), && |\omega| < c \\ &= 0, && |\omega| > c. \end{aligned}$$

If $M$ is reasonably small, $H(\omega)$ is well behaved, except at $\omega = c$, and may be readily approximated by a physically realizable filter.

The optimum signals and their separation functions have been computed for several low-dimensional cases, each for several values of $c$. The total energy of the signals was set to unity in all cases.

The simplest signal is a two-dimensional even or odd function. It is completely determined by its energy and constraint (21) or (22). Three such signals have been examined. The components of these three signals are $\psi_0$ and $\psi_2$, $\psi_1$ and $\psi_3$, and $\psi_2$ and $\psi_4$, respectively. For all values of $c$, it was found that the first signal led to the highest value of the separation function $D$, while the third signal gave the lowest value of $D$. This result would be anticipated by energy considerations alone.

The $\gamma_0$ and $\gamma_2$ components of the optimum two-dimensional signals are plotted in Fig. 3 as a function of the normalized bandwidth, $c$. The values of the separation function for these signals are shown in Fig. 5.

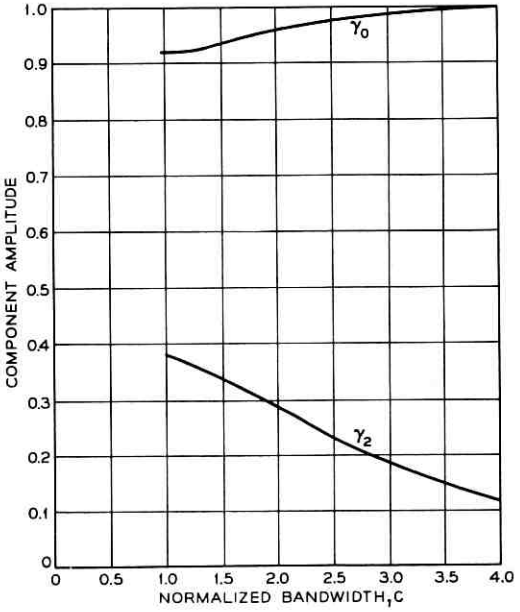


Fig. 3 — Components of optimum two-dimensional bandlimited signals.

A three-dimensional signal may be formed with $\psi_0$, $\psi_2$ and $\psi_4$ components. One degree of freedom is available for adjusting the coefficients of these components so as to maximize $D$. The optimum coefficients for signals of this form are shown in Fig. 4. The resultant values of $D$ are plotted in Fig. 5. It is seen that substantial improvement over the two-dimensional signal is obtained over a large range of $c$.

A four-dimensional signal consisting of $\psi_0$, $\psi_1$, $\psi_2$ and $\psi_3$ components was also investigated. Constraints (21) and (22) and the energy requirement permit one degree of freedom in the signal design. It was found that no significant improvement over the two-dimensional signal could be obtained by using this form of signal.

For an ideal signal which has all of its unit energy in the interval $(-1,1)$, $D = 1$. Fig. 5 may be considered to be a comparison of the worst error probabilities of bandlimited signals to the error probability of an ideal signal. If $D > 0$, then the power of the bandlimited signal must be increased by $-20 \log_{10} D$ db in order for its error probability for the worst sequence to be equal to the error probability of an ideal signal.

It should be noted that for these signals, $D < 0$ when $c < \pi/2$.

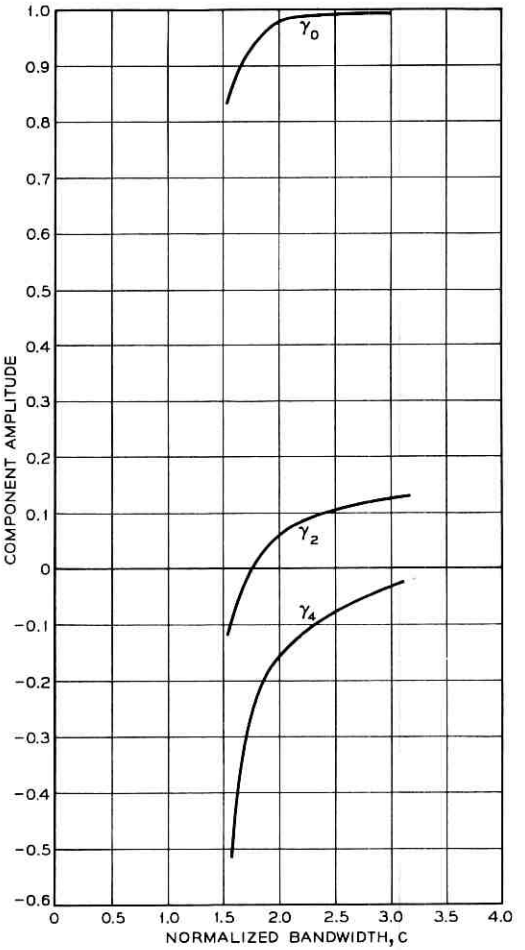


Fig. 4 — Components of optimum three-dimensional bandlimited signals.

We must therefore transmit slower than the Nyquist rate in order to achieve error-free performance in the absence of noise.

## V. CONCLUSIONS

Memoryless correlation is a suboptimum but useful method of detecting binary signals. With proper choice of the transmitted signal, the performance of a communication system using memoryless correlation can be made to be almost as good as that of an optimum system.

Communication at the Nyquist rate without intersymbol interference

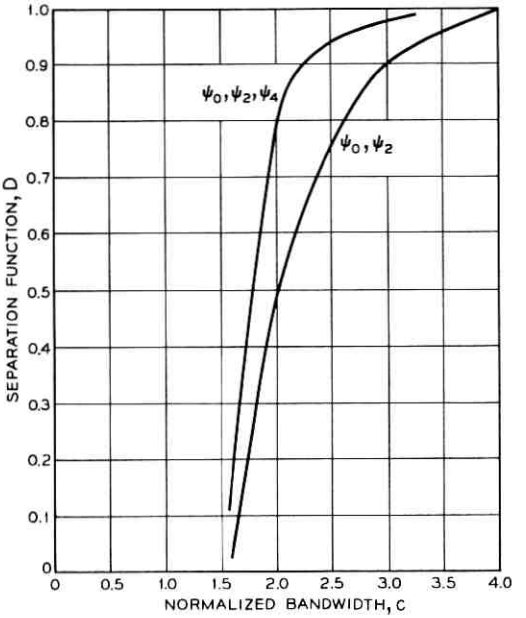


Fig. 5 — Separation functions for optimum two- and three-dimensional bandlimited signals.

using memoryless correlation detection is possible when the function shown in Fig. 2 is used as the transmitted signal. The resultant performance in the presence of noise is 1.3 db worse than that of an optimum system.

Bandlimited signals may also be designed so as to lead to low error probabilities in spite of intersymbol interference. Signals consisting of linear combinations of a finite number of prolate spheroidal functions accomplish this purpose. These signals may be designed so as to remain bounded for all message sequences.

## VI. ACKNOWLEDGMENTS

The authors wish to thank Professor Peter Lax and Mr. Pedro Nowosad of New York University and Dr. Henry J. Landau of Bell Telephone Laboratories for their considerable mathematical assistance.

## REFERENCES

1. Nyquist, H., Certain Topics in Telegraph Transmission Theory, Trans. A.I.E.E., *47*, April, 1928, pp. 617–644.
2. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., *38*, Nov., 1959, pp. 1357–1426.

3. Tufts, D. W., *Matched Filters and Intersymbol Interference*, Tech. Rpt. No. 345, Cruft Laboratory, Harvard University, Cambridge, Mass., 1961.
4. Kurz, L., A Method of Digital Signaling in the Presence of Additive Gaussian Noise, Trans. I.R.E., *IT-7*, Oct., 1961, pp. 215–223.
5. Trabka, E. A., *Signal Waveforms for Transmitting Binary Data over a Dispersive Channel with Independent Noise Sources at Input and Output*, DETECT Memo, No. 20, Cornell Aeronautical Lab., Inc., Buffalo, N. Y., 1962.
6. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, 1958, pp. 338–345.
7. Lerner, R. M., Modulation and Signal Selection for Digital Data Systems, Proc. N.E.C., *16*, 1960, pp. 2–15.
8. Aein, J. M., and Hancock, J. C., Reducing the Effects of Intersymbol Interference with Correlation Receivers, Trans. IEEE, *IT-9*, July, 1963, pp. 167–175.
9. Saltzberg, B. R., Error Probabilities for a Binary Signal Perturbed by Intersymbol Interference and Gaussian Noise, Trans. IEEE, *CS-12*, March, 1964, pp. 117–120.
10. Oliver, B. M., Pierce, J. R., and Shannon, C. E., The Philosophy of PCM, Proc. I.R.E., *36*, Nov., 1948, pp. 1324–1331.
11. Slepian, D., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — I, B.S.T.J., *40*, Jan., 1961, pp. 43–63.
12. Landau, H. J., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — II, B.S.T.J., *40*, Jan., 1961, pp. 65–84.
13. Flammer, C., *Spheroidal Wave Functions*, Stanford University Press, Stanford, California, 1957.
14. Landgrebe, D. A., and Cooper, G. R., Two-Dimensional Signal Representation Using Prolate Spheroidal Functions, Trans. IEEE, Comm. and Elect. *82*, March, 1963, pp. 30–40.

# Four-Phase Data Systems in Combined Delay Distortion, Gaussian Noise, and Impulse Noise

By M. A. RAPPEPORT

(Manuscript received August 26, 1964)

*The performance of a four-phase data system over delay- and attenuation-distorted transmission lines in the presence of impulse noise has been simulated on a digital computer previously. This paper describes an extension of that simulation to performance with the additional degradation of additive Gaussian noise. Results are presented for Gaussian noise alone in terms of absolute error rate, and for Gaussian plus impulse noise in terms of conditional error rate given the occurrence of an impulse. Some conclusions on the limiting effects in various situations are given.*

## I. INTRODUCTION

A previous paper on digital computer simulation of a four-phase data system[1] was concerned with performance in the presence of delay and attenuation distortion, and impulse noise. This paper extends those results to include Gaussian noise. Combinations of Gaussian noise, impulse noise, and delay and attenuation distortion are presented and discussed.

The prime effect of delay and attenuation distortion is to reduce system margin against error. Adding Gaussian noise to the distorted signal can have one or both of two effects. Either errors occur directly due to the Gaussian noise alone, or the margin against other disturbances, in particular impulse noise, is decreased. On most telephone lines it is generally feasible to keep the direct errors produced at a very low level. The generally steep slope of a curve of signal to Gaussian noise ($S/N$) versus probability of error, (as much as a factor of 10 : 1 in error rate for one db change in $S/N$) makes the system particularly sensitive to changes in the amount of Gaussian noise. Thus, at least on land facilities, Gaussian noise is usually kept to a level where it enters basically as a

degradation in system margin, i.e., an increase in conditional probability of error given a noise impulse.

## II. METHOD

The main problem in straightforward introduction of Gaussian noise into digital simulations is the computation time. The accuracy of the results basically depends upon the number of errors obtained. A sufficiently large number of noise samples to give reliable information about absolute error rates of the order of $10^{-5}$ or less is extremely time consuming. A rough estimate of the accuracy obtainable can be obtained by considering a test as consisting of $N$ independent trials each with common probability of failure (i.e., of making an error). Call this probability $p$.

The estimate

$$\hat{p} = \frac{\text{number of errors}}{\text{number of samples}} \tag{1}$$

has expected value $p$.

The standard deviation of $\hat{p}$ is

$$\sigma(\hat{p}) = \frac{\sqrt{pq}}{\sqrt{n}} \approx \sqrt{\frac{p}{n}} \tag{2}$$

since $q \approx 1$. Thus for $p = 0.001$, $\sigma(\hat{p})$ would be 0.00014 for $n = 50{,}000$.

We emphasize that these results are very rough, since the assumption of independent trials (i.e., independence of successive bit periods) is clearly not too accurate. However, they do give an idea of the number of samples required to obtain accurate results.

When impulse noise is present, this problem is solved by computing the conditional probability of error given the occurrence of an impulse. The long quiet intervals which characterize impulse noise in the telephone plant are in effect taken as having an error probability of zero. For Gaussian noise alone this procedure is not realistic, since the noise amplitudes are not segregated into very large impulses and very low quiet periods.

There are two problems: Gaussian noise alone and Gaussian noise as a degrading factor in the performance of a system with impulse noise. The solution to both of these revolves around a program which uses a tape of approximately 25,000 one-dibit intervals of Gaussian noise bandlimited by the input receiving filter of the data set. A train of

512 dibits is used.* Fifty one-dibit samples of noise are introduced into each signal dibit. To find the effect of Gaussian noise alone, the demodulation is then performed and the errors simply counted. The result is the error rate due to Gaussian noise.

For the second problem, Gaussian noise is introduced and the effect of an impulse is systematically examined in a pseudo-random way. The details of this process are essentially the same as those outlined in Ref. 1 for finding conditional probability of error.

## III. SIMULATION ACCURACY

For those facilities in which the effect of Gaussian noise alone was desired, two tests were run for each pattern of delay distortion. Thus, noise was introduced into 50,000 separate dibit intervals and the resulting errors counted. As mentioned above, it is reasonable to trust these results without further evidence only to about an error level of $10^{-3}$. At that level about 50 errors are obtained and the accuracy is still quite good.

The curves can be directly extrapolated to perhaps $2 \times 10^{-4}$, because determining $P(e)$ to an accuracy of perhaps 3:1 (e.g., from $10^{-4}$ to $3 \times 10^{-4}$) is quite suitable for most applications.

However, we often desire results to error levels of $10^{-6}$ or lower. To obtain such results we make use of a conjectured property of the $P(e)$ curves which we will call "convexity." If a $P(e)$ versus $S/N$ curve is plotted on log vs db paper then the curve is convex down.

The justification for this conjecture is primarily experience. All experience on laboratory, field test, particular extended computer runs shows this to be so. In addition, a preliminary proof by the author has indicated that the second and third derivatives of the curve with respect to the noise power are both $\leqq 0$. The details of this proof are being clarified, and should be presented in a future paper.

Therefore, an upper bound for any $P(e)$ versus $S/N$ curve can be obtained by extending the curve with a straight line (i.e., second derivative = 0; see Fig. 1). Similarly, a lower bound can be obtained by setting the third derivative = 0, that is, holding the second derivative constant. The results given sharp bounds, since the curves at $P(e) = 5 \times 10^{-4}$ in general have a severe slope (i.e., 5 or 10 to 1 change in $P(e)$ for 1 db change in $S/N$). Taking into account that the allowable measured error rate range generally increases, in practice, as the $P(e)$ drops [e.g.,

* This length pattern gives all possible combinations of four dibits, where the first dibit can take any one of the eight phases used in the modem.

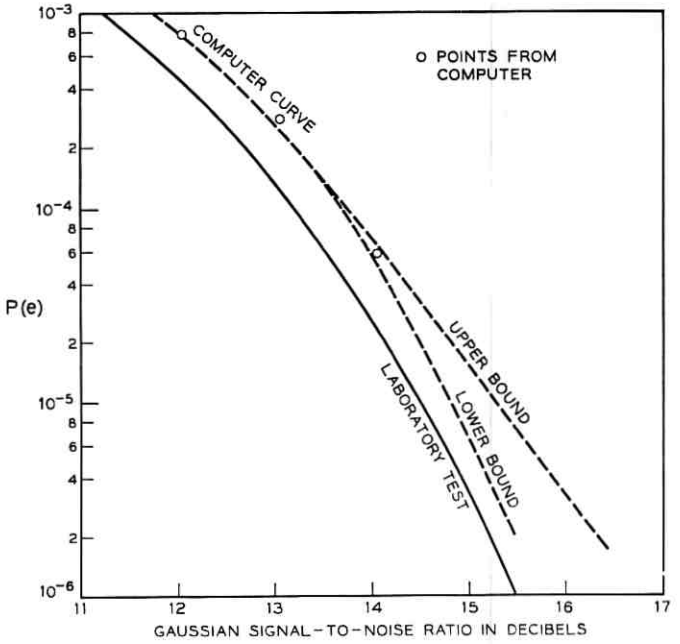


Fig. 1 — Undistorted performance with Gaussian noise, showing upper- and lower-bound convexity projections.

perhaps a 5:1 range in $P(e)$ generally allowable at $P(e) = 10^{-6}$], useful estimates of $P(e)$ can be obtained for $P(e)$ as small as $10^{-6}$ in most cases. Fig. 1 also shows a laboratory test curve.

In an impulse noise environment, the simulation has inherently much greater accuracy. This is because the aim is to produce conditional probability of error $P_N(e)$ — that is, to count the errors per impulse — and the desired performance range is more in the order of $10^{-2}$ or perhaps $10^{-3}$. Thus, the number of trials, i.e., the number of introductions of Gaussian noise, is sufficiently large to give very good convergence to real conditional probabilities of error.

## IV. RESULTS

The four-phase system considered was one using a cycle and a half of carrier per dibit, (e.g., an 1800-cycle carrier with a 1200-dibit or a 2400-bit per second system). In keeping with previous results on the bandwidth in which delay distortion degrades the signal,[1] the delay was specified from the carrier to plus or minus a number of cycles equal to

$\frac{2}{3}$ the dibit speed. For example, for the system considered, delay was introduced between $(f_c - 800)$ and $(f_c + 800)$ cycles. A sequence of peak delays in this range was considered. Because the shape of delay (as well as its peak magnitude) is significant, several delay shapes were considered. These were chosen, based on previous work, to give results which are typical of a wide class of transmission facilities used for data transmission. The delays used were respectively parabolic, centered at the carrier frequency, and a sinusoidal shape with three cycles of sinusoid across the transmission band.

The results are shown in Figs. 2 through 5. Fig. 2 shows the error rates obtained for Gaussian noise alone for the parabolic and sinusoidal lines respectively, with extrapolation obtained by using a compromise between computed points and the derived upper bounds.

Figs. 3, 4 and 5 give the effect of introducing impulse noise and Gaussian noise simultaneously with delay distortion present. Each set

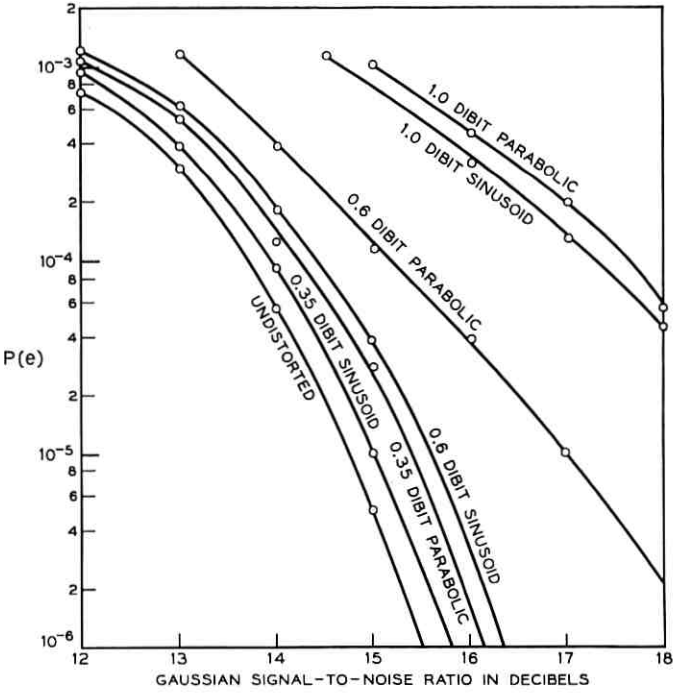


Fig. 2 — Probability of error vs Gaussian noise level for various delays. Parabolic delay specified by delay at $\omega_c \pm 0.7\ \omega_{bit}$ ; sinusoidal delay specified by delay in passband of peak-to-peak 3-cycle sinusoid.

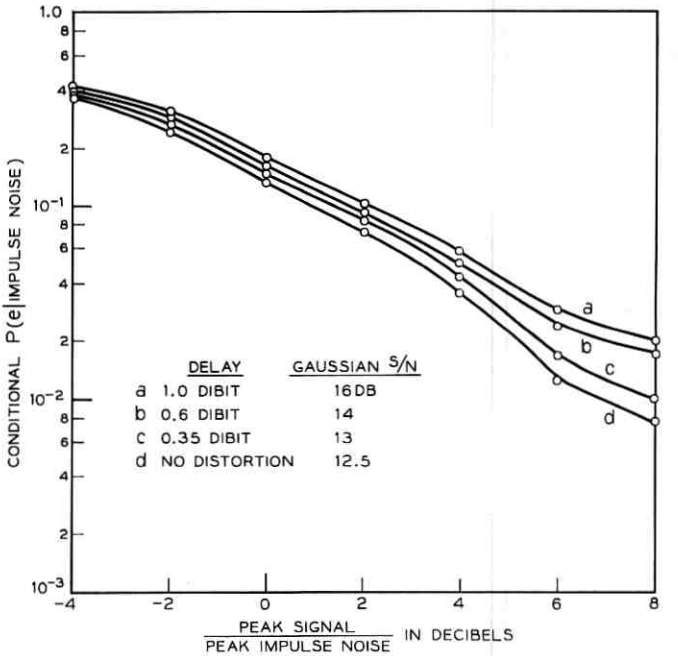


Fig. 3 — Conditional probability of error given an impulse noise for Gaussian noise such that the error rate due to Gaussian noise alone equals $5 \times 10^{-4}$. Curves averaged over various impulse shapes and two delay shapes (see Fig. 2).

of curves gives conditional probability of error for a Gaussian noise level chosen to yield some specific error rate. The Gaussian noise was chosen to give error rates of approximately $5 \times 10^{-4}$, $10^{-5}$, and $5 \times 10^{-7}$. Each curve was averaged over a variety of impulse noise shapes. More exact information would have to be known on the noise expected on a given transmission line before precise absolute error rates could be obtained. However, reasonable results should be achieved in most practical situations by specifying an allowable number of counts from the averaged curves.

## V. CONCLUSIONS

The curves as given, i.e., the results of the simulation, are primarily useful for design purposes. They can be used to find the trade-off obtainable between Gaussian noise, delay distortion, and number of impulse noise counts. Thus the curves can be used for over-all design of transmission facilities at given error rates with a four-phase data trans-

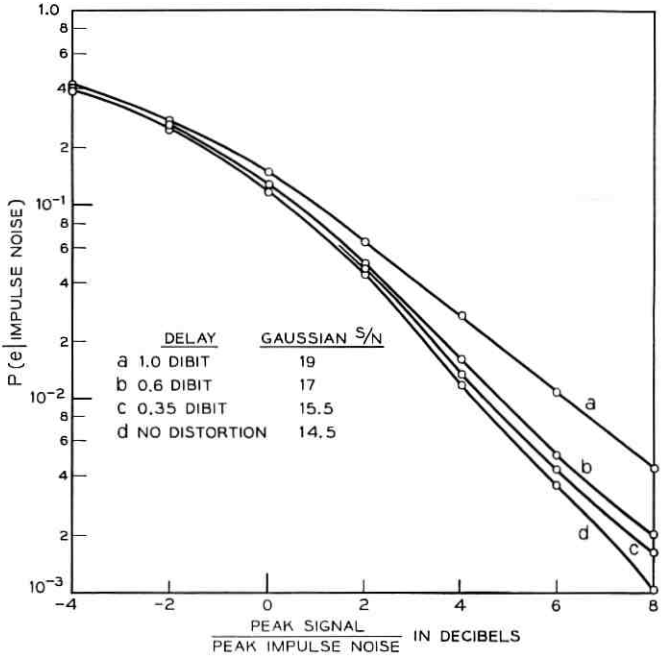


Fig. 4 — Conditional probability of error given an impulse noise for Gaussian noise such that error rate due to Gaussian noise alone equals $10^{-5}$. Curves averaged over various impulse shapes and two delay shapes (see Fig. 2).

mission system. Here it is worth noting that the effects of frequency shift on such a system were also investigated as part of the program, and frequency shifts up to $\pm 5$ cycles were found to have essentially no effect.

One factor that stands out in the results is the rapid deterioration in performance of the system as delay distortion is increased beyond a certain amount. This is true either for Gaussian noise alone or for the combined effects of Gaussian and impulse noise. It seems that, for the data modem and conditions assumed, signal-to-noise ratios worse than those actually occurring in most of the plant are not harmful for reasonable delay ranges. Thus it seems reasonable that the prime consideration in designing transmission facilities for four-phase data systems must be the minimization of the effect of delay distortion, while the present level of Gaussian noise, at least for land-line systems, does not seem too critical. However, recent data on certain carrier systems show that for data systems using a larger number of phases (e.g., eight- or sixteen-phase systems), but maintaining constant signal element rates, the noise levels might be high enough that the systems would be limited by Gaus-

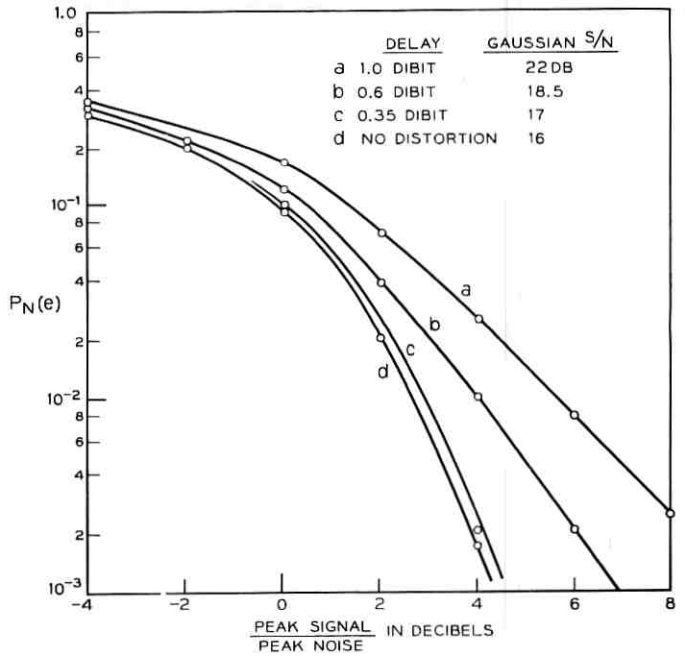


Fig. 5 — Conditional probability of error given an impulse noise for Gaussian noise such that error rate due to Gaussian noise alone equals approximately $5 \times 10^{-7}$. Curves averaged over various impulse shapes and two delay shapes (see Fig. 2).

sian noise rather than delay. Thus caution should be used in systems employing more than four phases or levels.

Indeed, this emphasizes that any comparison of multiphase and multilevel systems which is based on their performance under certain degradation (e.g., Gaussian noise) may be misleading. Instead, comparison must be based on the critical factors, which for many systems may well be how they are affected by delay distortion.

REFERENCE

1. Rappeport, M. A., Digital Computer Simulation of a Four-Phase Data Transmission System, B.S.T.J., *43*, May, 1964, p. 927.

# Atomic Hydrogen as a Reducing Agent

By A. A. BERGH

(Manuscript received September 4, 1964)

*Atomic hydrogen is a powerful chemical reducing agent. It may be used to perform reductions at lower temperatures than those achieved in molecular hydrogen, and since it is electrically neutral it is not subject to shielding effects. Because of the contradictory results reported in the literature, an investigation was undertaken to determine optimum ways to produce and transfer atomic species. It is concluded that electrodeless discharge is the most reliable way of production, and that, although Teflon and phosphoric acid coated glass have very low catalytic activities toward recombination, Pyrex and quartz are satisfactory container materials. Reduction temperatures in both molecular and atomic hydrogen were established for a variety of oxides, and the latter were found to be substantially lower. Finally, the advantages and limitations of atomic hydrogen as a reducing agent are considered.*

## I. INTRODUCTION

The removal of surface oxides is an important step in the manufacture of electron devices. The objective of this process is usually to facilitate the wetting of solid surfaces by molten metals or to improve some electrical properties of the device by changing the surface properties of certain components. Oxide-free surfaces can be obtained by wet chemical treatments, by bombardment with accelerated particles or by reduction in a gaseous ambient. Each of these processes exhibits certain limitations. Liquid reagents leave residues and require additional cleaning, electron or ion bombardment is vulnerable to shielding effects, and the high temperatures necessary for gaseous reduction restrict its application on most assembled electron devices. A powerful gaseous reducing agent capable of removing oxygen at relatively low temperatures combines the virtues of all these methods.[1] A study was undertaken, therefore, with such a reagent, atomic hydrogen, to establish the feasibility of its use. The following discussion gives our findings on the following three questions: (1) how to produce and transfer atomic species, (2) how much decrease in reduction temperature can be expected compared

with those observed in molecular hydrogen, and (3) what are the limitations and advantages of using atomic hydrogen in device processing?

## II. PRODUCTION AND TRANSFER OF ATOMIC HYDROGEN

To study reduction in atomic hydrogen, a source of atoms and a way to transfer them to the oxide surface must be established. There are several methods known to dissociate molecular hydrogen (ultraviolet radiation, thermal or electrical energies, etc.). The data in the literature, however, are contradictory to the extent that we found it necessary to evaluate them for our purposes.

Dissociation of molecular hydrogen on hot tungsten filaments was reported in 1911;[2] experimental conditions have been well established,[3] and the method has recently been employed in absorption studies.[4,5] Our experiments, however, showed it impractical for reduction purposes. The water vapor evolved during reduction oxidizes the tungsten filament, deteriorating the wire and contaminating the system with evaporated tungsten oxide.

Contamination is the major disadvantage of the electrical discharge method first described by Wood.[6] The sputtering of the electrodes gradually contaminates the walls of the apparatus and, by increasing their catalytic activity, reduces the transfer of atoms to the probe. Cooling the electrodes by compressed air delays wall poisoning only by several hours. Similar observations were reported by Linnett and Marsden[7] for an oxygen discharge. Many previous investigations using Wood discharges neglected this effect, suggesting that their results should be treated with caution.

The theory of the electrodeless discharge was outlined by Thomson[8] in 1928, and an efficient way to produce high concentrations of atomic hydrogen was described by Jennings and Linnett.[9] This is the method we have essentially adopted. A schematic diagram of the apparatus is shown in Fig. 1. The quartz or Pyrex discharge tube, 50 mm in diameter, was surrounded by a gold-plated copper electrostatic screen to restrain the plasma from spreading beyond the coil. The screen was connected to ground and a distance of 3–4 mm was maintained between the screen and the discharge coil with the help of a quartz-tube spacer. The discharge was established with a variable-frequency (5–16 mc), 5-kw Lepel high-frequency generator, and the exciting coil consisted of eight turns of 6-mm copper tubing. The temperature of the tube was kept below 300°C by a jet of compressed air. The apparatus was connected to a mechanical vacuum pump, and a steady-state pressure of 200 $\mu$ was maintained by means of an adjustable leak. Pressure was measured

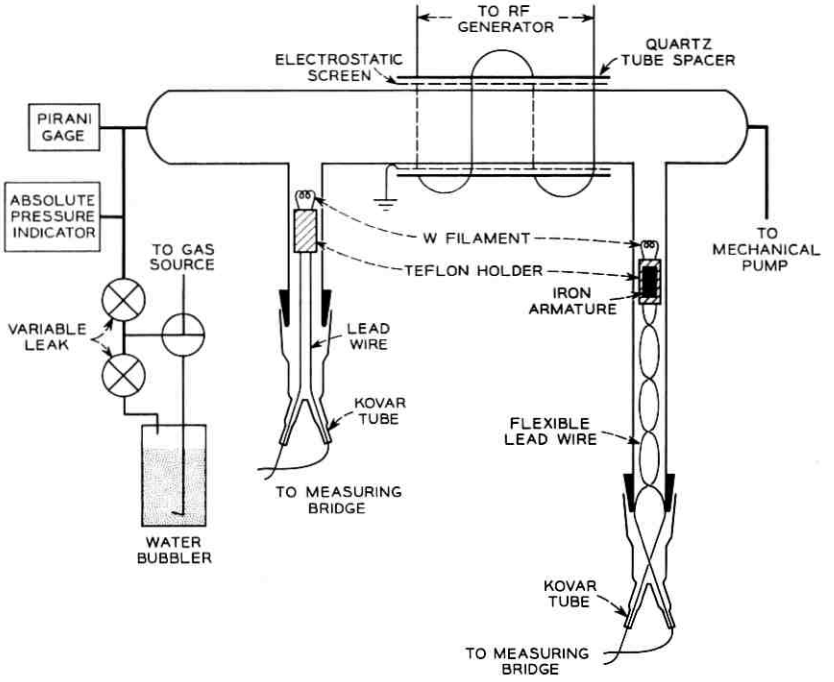


Fig. 1 — Schematic diagram of the electrodeless discharge apparatus.

with a conventional Pirani gauge and a Wallace and Tiernan absolute pressure indicator. Since all the experiments were carried out in the side tubes, the flow of atoms in the reduction zone was purely diffusional.

The concentration of the atomic hydrogen was measured calorimetrically by a method similar to that described by Wood and Wise.[10] Since the recombination on tungsten has been reported to change at lower temperatures due to surface poisoning,[11] the probe temperature was always kept above 150°C by a constant-voltage dc source (energy $w_1$). The heat evolved by recombination of hydrogen atoms increased the filament temperature; this was measured by supplying equivalent energy, $w_2$, from the dc source after the discharge was terminated. The energy difference, $w_2 - w_1 = \Delta w$, is directly proportional to the number of atoms recombining on the probe. The side arms contained identical filaments, one of them serving only to check the stability of the atomic hydrogen concentration in the discharge tube. Considering the high concentration of atoms obtained, the reproducibility of the atomic level from one experiment to another and the stability of the

atomic hydrogen concentration during individual runs, the electrodeless discharge proved to be superior among the three methods tested.

## III. THE TRANSFER OF ATOMIC HYDROGEN

The high reactivity of atomic hydrogen imposes serious difficulties in transportation and optimum utilization. Walls of low catalytic activity are needed to prevent recombination on the wall of the apparatus before reaction with the oxide. Many investigators have attempted to prepare surfaces of low catalytic activities and have offered explanations for the cause of this effect. Almost all have used, however, different experimental techniques and arrived at contradictory conclusions. The discrepancies are closely related to the question of whether or not hydrogen atoms can be produced in an electric discharge from dry gases.

Several investigators observed that the dissociation of hydrogen is greatly reduced[6] or eliminated[12,13] in the absence of water vapor. Coffin,[14] on the other hand, found no appreciable decrease in the intensity of the Balmer lines in dry hydrogen. One explanation for the effect of water is that it poisons the catalytic activity of glass;[3,6,15,16,18] however, two independent measurements using wet[15] and dry[19] hydrogen show practically identical catalytic activities. Smith[15] obtained similar results after cleaning his apparatus with KOH, $KHCrO_4$, $H_2SO_4$, or HF, while others reported recombination coefficients different by three orders of magnitude for Pyrex rinsed in $HNO_3$[19] and Pyrex rinsed in HF.[20] It can be concluded, therefore, that the role of water vapor in the production and recombination of atomic hydrogen is open for further clarification. Several attempts were made to poison glass and quartz surfaces by coating with phosphoric acid,[3,15-17] with a mixture of dimethylchlorosilane and methyldichlorosilane ("Dry-film"),[21,22] and more recently with Teflon.[23] No attempt was made, however, to evaluate the relative catalytic activities of the different surfaces.

In order to establish high concentration of atoms and walls of low catalytic activity, a series of experiments was carried out using both dry (water content $<20$ ppm) and wet hydrogen (by passing the gas through a water bubbler at 25°C) while changing the wall surface of the side arm. The inside diameter of the side arm was about 20 mm and the probe was positioned at a distance of $L/R = 10$, where $L$ is the distance from the atom source and $R$ is the radius of the tube. The stability of the atomic concentration of the source was checked continually by the probe in the second side arm. The results of all experiments are summarized in Table I. It is important to note that the heat evolved at the

TABLE I — THE ENERGY, $\Delta w$, RELEASED BY THE RECOMBINATION OF ATOMIC HYDROGEN ON THE TUNGSTEN FILAMENT AFTER DIFFUSING IN CYLINDERS OF DIFFERENT CATALYTIC ACTIVITIES. ($L/R$ = 10, $\Delta w_0$ = 150 mw.)

| Cylinder Wall | $\Delta w$ (mw) in Dry Hydrogen | $\Delta w$ (mw) in Wet Hydrogen |
|---|---|---|
| Teflon | 86 | 85 |
| Phosphorous coated Pyrex | 80 | 81 |
| Quartz rinsed with $HNO_3$ | 77 | 78 |
| Pyrex rinsed with $HNO_3$ | 74 | 76 |
| Pyrex rinsed with HF | 75 | 74 |
| Pyrex coated with dry-film | 10–45 | 10–48 |

probe, $\Delta w$, is not a linear function of the catalytic activity of the wall[24,25] and can be used only to establish a series of activities. Due to the limitations of Teflon toward heat treatment and the tedious cleaning procedures required with phosphoric acid coating,[3] all the reduction studies were carried out in $HNO_3$-rinsed Pyrex or quartz tubes with dry hydrogen.

## IV. COMPARISON OF THE REDUCTION TEMPERATURES IN ATOMIC AND MOLECULAR HYDROGEN

The most important advantage of using atomic instead of molecular hydrogen is that a substantial decrease in the reduction temperature can be expected. An attempt was made, therefore, to compare the corresponding reduction temperatures in the two reducing atmospheres. Although some kinetic data for the reduction of metal oxides in hydrogen have been published before,[26] we have carried out reduction in both atmospheres to assure identical properties (impurity level, particle size, etc.) of the oxides. Since the reduction temperatures in atomic hydrogen were determined by observing the color changes of the oxides, it was also necessary to establish corresponding color-change temperatures in molecular hydrogen.

## V. REDUCTION TEMPERATURES IN MOLECULAR HYDROGEN

Reduction in molecular hydrogen was carried out in the apparatus shown in Fig. 2. Reagent grade oxides in quantities producing about $3 \times 10^{-3}$ moles of water upon reduction were placed into a quartz tube of 25 mm ID. The apparatus was flushed with dry nitrogen for several hours at temperatures ranging from 100° to 250°C. After cooling to room temperature the nitrogen flow was replaced by hydrogen (560

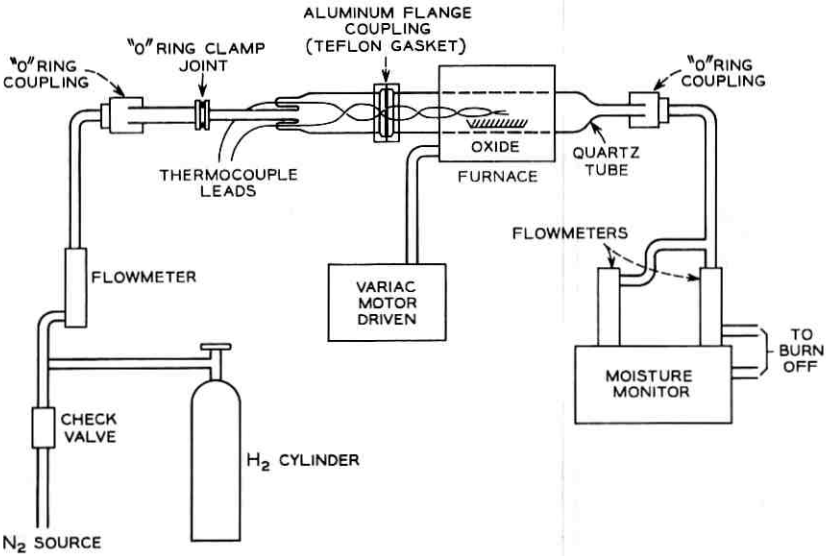


Fig. 2 — Hydrogen reduction apparatus.

cc/min) and the reduction cycle started, provided that the moisture content of the emerging gas was less than 10 ppm. The temperature of the tube furnace was raised at 5 ± 1°C/min, and the moisture content of the gas was monitored continuously. The change in the moisture content as a function of temperature for the different oxides is shown in Figs. 3, 4, and 5. The temperatures at which color changes occurred were also noted. Reduction ranges and color change temperatures are summarized in Table II.

## VI. REDUCTION TEMPERATURES IN ATOMIC HYDROGEN

Several investigators[3,27] reduced metal oxides in atomic hydrogen; no attempt was made, however, to determine minimum reduction temperatures. The role of the heat of recombination on the reduction temperature was noted by Kroepelin and Vogel.[3] They could reduce $Cr_2O_3$ only if the oxide particle contained catalytically active impurities and the heat of recombination raised the temperature above a critical point. The control of oxide temperatures imposes difficulties in atomic hydrogen. Since the recombination coefficient of Pyrex increases with increasing temperature,[19] the apparatus wall cannot be heated along with the oxide without reducing atomic concentrations. Due to the

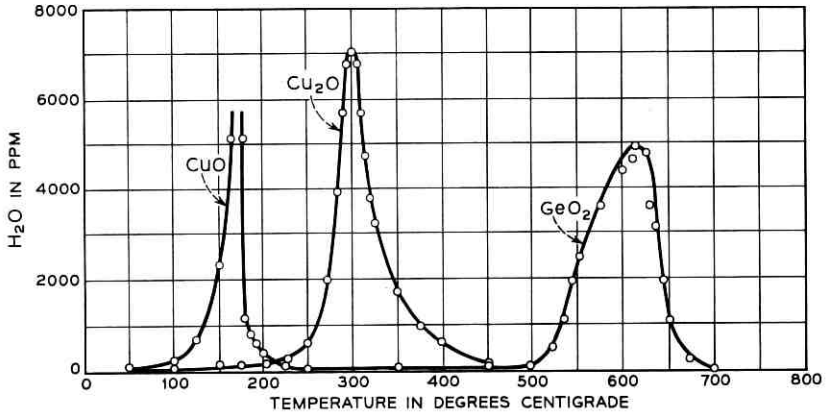


Fig. 3 — Reduction of CuO, $Cu_2O$, and $GeO_2$ in hydrogen: $[O_2]$ in oxides = $1.5 \times 10^{-3}$ M; $H_2$ flow rate = 560 cc/min; temperature rise $\approx 5°$ C/min.

various catalytic activities and accommodation coefficients (the fraction of the heat of recombination transferred to the surface on which the recombination occurs) of the different oxides the heat of recombination yields different temperatures on each sample. Finally, the much higher catalytic activities of pure metals than those of the corresponding oxides result in a sudden temperature rise after reduction occurs. To overcome these difficulties, all oxide samples, 0.1 $cm^2$ in area and several

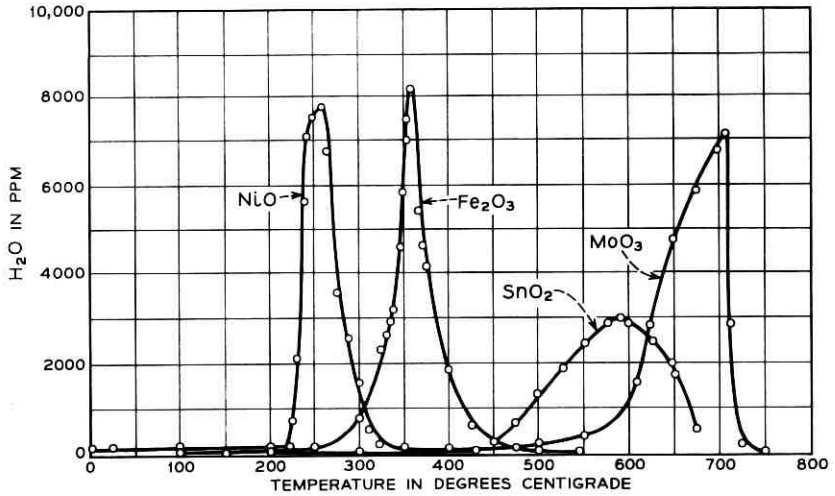


Fig. 4 — Reduction of NiO, $Fe_2O_3$ $SnO_2$, and $MoO_3$ in hydrogen: $|O_2|$ in oxides = $1.5 \times 10^{-3}$ M; $H_2$ flow rate = 560 cc/min; temperature rise $\approx 5°$ C/min.

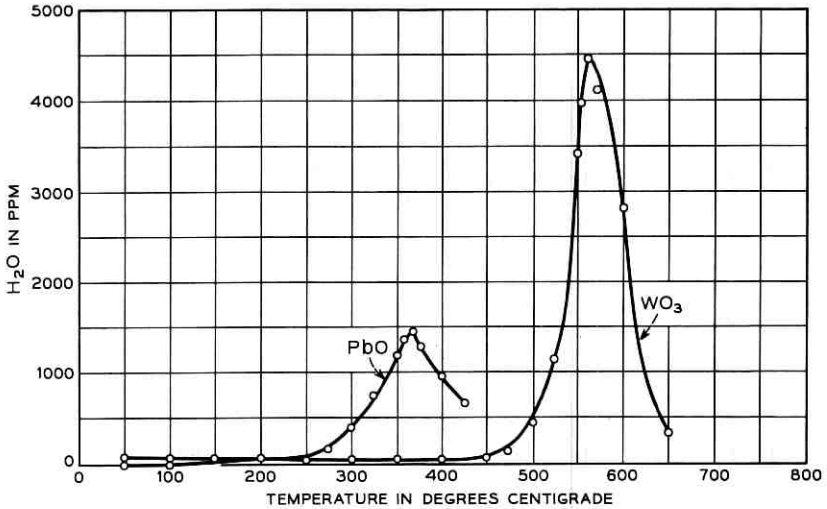


Fig. 5 — Reduction of PbO and $WO_3$ in hydrogen: $|O_2|$ in oxides = $1.5 \times 10^{-3}$ M; $H_2$ flow rate = 560 cc/min; temperature rise $\approx 5°$ cc/min.

microns thick, were placed on identical ceramic wafers. The wafers were 0.7 $cm^2$ in area and 2.0 mm thick, containing a hole for a thermocouple probe. The oxides were taken from the same lots supplying the material for the experiments in molecular hydrogen, and the reduction was continued until color changes — similar to those observed in $H_2$ — could be detected. The ceramic blocks were placed near to the entrance of the side arm and heated by the recombination of atoms on the block. The temperature rise was about 20°C/min and, due to the small areas involved, the presence of oxides did not influence this rate. The color

TABLE II — REDUCTION TEMPERATURES FOR DIFFERENT OXIDES IN MOLECULAR AND ATOMIC HYDROGEN (Temperature rise 5 ± 1°C/min in $H_2$)

| Oxide | Reduction Temp. in $H_2$ (°C) | Color Change in $H_2$ (°C) | Color Change in H (°C) |
|---|---|---|---|
| $MoO_3$ | 500–725 | 610 | 43 |
| $GeO_2$ | 500–700 | 560 | 35 |
| $WO_3$ | 475–650 | 535 | 25 |
| $SnO_2$ | 400–675 | 490 | 100 |
| $Fe_2O_3$ | 250–500 | 310 | 40 |
| PbO | 225–475 | 300 | 25 |
| $Cu_2O$ | 225–450 | 265 | 25 |
| NiO | 225–325 | 250 | 62 |
| CuO | 100–225 | 140 | 25 |

change temperatures with those obtained in molecular hydrogen are listed in Table II. Although all reduction temperatures are substantially lower in atomic hydrogen, no quantitative correlation could be established between the two sets of data.

## VII. DISCUSSION

The most important advantage of using atomic hydrogen is lower reduction temperatures than those obtained in molecular hydrogen. This can be especially significant for stable oxides, not reducible in hydrogen below the melting point of the metal, such as $Al_2O_3$, $SiO_2$, etc. Removal of the oxide from aluminum inside an electrodeless discharge tube has been reported,[10] and we have succeeded in removing $SiO_2$ films more than a micron thick from silicon slices.

The high reactivity of hydrogen atoms toward organic compounds makes it a potential remover of organic residues. Propane can be easily converted to methane at room temperatures[28] by the use of atomic hydrogen, and higher paraffins probably react in a similar manner. The formation of gaseous hydrocarbons was also reported on exposure of graphite to atomic hydrogen.[29]

There are, however, several limitations to the use of atomic hydrogen. One is the difference in catalytic activities of different oxides, and that between metals and the corresponding oxides, those of the metals being much higher. If a complex system containing different materials were exposed to atomic hydrogen, it is likely that first the oxides with the largest catalytic activities would be reduced, and that the heat of recombination on the metals could melt some components before more stable oxides could be reduced. To determine whether a system can be exposed to certain atomic concentrations, a knowledge of the respective recombination and accommodation coefficients is required. While there is very little known about the latter quantity, the available data on the former are highly contradictory, as shown in Table III, and there are almost as many theories as authors for the property of the material determining its magnitude.

Another potential difficulty arises from the embrittlement of metals exposed to atomic hydrogen. Although no data are known about the solubility of atomic hydrogen in metals, it is reasonable to assume that it is greater than that of $H_2$, especially for metals of group A (see Ref. 26, p. 518).

Finally, chemical interactions between atomic hydrogen and metals and metal oxides other than reduction should be considered. Partial reduction or hydride formation may occur with the formation of vola-

TABLE III — RELATIVE VALUES OF THE RECOMBINATION COEFFICIENT OF ATOMIC HYDROGEN ON METALS
($\gamma_{Pt} = 1.00$)

| Metal | Reference | | | | | |
|---|---|---|---|---|---|---|
| | 27* | 32 | 33 | 34 | 35 | 10 |
| Pt | A | 1.00 | 1.00 | | 1.00 | 1.00 |
| Co | | 0.98 | | 1.00 | 0.72 | |
| Pd | B | | 0.87 | | 0.80 | 3.5 |
| Ni | | 0.91 | | 0.83 | 0.72 | 9.0 |
| W | C | | | 0.70 | | 3.0 |
| Fe | D | 0.81 | | 0.83 | 0.68 | |
| Cr | E | 0.71 | | 0.60 | 0.64 | |
| Ag | F | | 0.71 | | 0.52 | |
| Cu | G | 0.74 | 0.66 | | 0.76 | 5.5 |
| Ti | | | | 0.68 | 0.40 | 19.0 |
| Au | | | | | 0.40 | 4.0 |
| Al | | | 0.47 | | $2.5 \times 10^{-2}$ | 13.5 |

* A, B, C, ··· : decreasing order of catalytic activities

tile products. Pietsch[30] formed compounds similar to lithium hydride by reacting atomic hydrogen with silver, beryllium, gallium, indium, and tantalum, while others[31] formed volatile metal hydrides with germanium, tin, arsenic, antimony, and tellurium. In removing $SiO_2$ films, sometimes a black deposit on the discharge tube could be detected, the composition of which is not yet determined, and prolonged exposure of silicon to atomic hydrogen produced a pitted rough surface characteristic of gas phase etchings. It can hence be seen that the high reactivity of hydrogen can yield undesired effects such as the removal or damage of important surface areas and the contamination of certain parts of the system.

REFERENCES

1. Thomas, C. O., and Koontz, D. E., Electrochem. Tech., *2*, 1964, p. 115.
2. Langmuir, I., Trans. Amer. Electrochem. Soc., *20*, 1911, p. 225; J. Am. Chem. Soc., *34*, 1912, p. 860 and p. 1310; *36*, 1914, p. 1708; *37*, 1915, p. 417.
3. Kroepelin, H., and Vogel, E., Naturwiss., *20*, 1932, p. 821; Abhandl. braunschweig wiss. Ges., *6*, 1954, p. 73.
4. Law, J. T., J. Phys. Chem., *59*, 1955, p. 543.
5. Bennett, M. J., and Tompkins, F. C., Proc. Roy. Soc., *259A*, 1960, p. 28; Trans. Faraday Soc., *58*, 1962, p. 816.
6. Wood, R. W., Proc. Roy. Soc., *A97*, 1920, p. 455; Phil. Mag., *42*, 1921, p. 729; *44*, 1922, p. 538; Proc. Roy. Soc., *A102*, 1922, p. 1.
7. Linnett, J. W., and Marsden, D. G. H., Proc. Roy. Soc., *A234*, 1956, p. 489.
8. Thomson, J. J., Proc. Phys. Soc., *40*, 1928, p. 79; Thomson, J. J., and Thomson, G. P., *Conduction of Electricity Through Gases*, Cambridge University Press, 3rd ed., 1928.
9. Jennings, K. R., and Linnett, J. W., Nature, *182*, 1958, p. 598.
10. Wood, B. J., and Wise, H., J. Phys. Chem., *65*, 1961, p. 1976.

11. Fox, J. W., Smith, A. C. H., and Smith, E. J., Proc. Phys. Soc., London, *73*, 1959, p. 533.
12. Finch, G. I., Proc. Phys. Soc., *62*, 1949, p. 464.
13. Deleplace, R., Compt. Rend., *202*, 1936, p. 1986.
14. Coffin, F. D., J. Chem. Phys., *30*, 1959, p. 593.
15. Smith, W. V., J. Chem. Phys., *11*, 1943, p. 110.
16. Warttenberg, H. V. and Schultz, G., Z. Phys. Chem., *B6*, 1930, p. 261.
17. Poole, H. G., Proc. Roy. Soc., *A163*, 1937, p. 404.
18. Nowak, E. J., Nurzuis, S., Deckers, J., and Boudart, M., Surface Recombination of Hydrogen Atoms in the Presence of Water Vapor, Armed Services Technical Information Agency, Report No. AD247-517, 1960.
19. Wood, B. J., and Wise, H., J. Phys. Chem., *66*, 1962, p. 1049.
20. Collins, R. L., and Hutchins, J. W., Bull. Am. Phys. Soc., *7*, 1962, p. 114, Abstract C1.
21. Wittke, J. P., and Dicke, R. H., Phys. Rev., *103*, 1956, p. 620.
22. Shaw, T. M., J. Chem. Phys., *30*, 1959, p. 1366.
23. Berg, H. C., and Kleppner, D., Rev. Sci. Inst., *33*, 1962, p. 248.
24. Graves, J. C., and Linnett, J. W., Trans. Faraday Soc., *55*, 1959, p. 1338.
25. Motz, H., and Wise, H., J. Chem. Phys., *32*, 1960, p. 1893.
26. Dushman, S., *Scientific Foundations of Vacuum Technique*, 2nd Ed., John Wiley & Sons, 1962, p. 750.
27. Bonhoeffer, K. F., Z. Physik. Chem. *113*, 1934, p. 199.
28. Steacie, E. W. R., and Parlee, N. A. D., Trans. Faraday Soc., *35*, 1939, p. 854.
29. Blackwood, J. D., and McTaggart, F. K., Austral. J. Chem., *12*, 1959, p. 114 and p. 533.
30. Pietsch, E., Z. Electrochem., *39*, 1933, p. 577.
31. Pearson, T. G., Robinson, P. L., and Stoddart, E. M., Proc. Roy. Soc., *A142*, 1933, p. 275.
32. Katz, S., Kistiakowski, G. B., and Steiner, R. F., J. Am. Chem. Soc., *71*, 1949, p. 2258.
33. Wood, B. J., and Wise, H., J. Chem. Phys., *29*, 1958, p. 114.
34. Suhrmann, R., and Csesech, H., Z. Physik. Chem., *B28*, 1935, p. 215.
35. Nakada, K., Bull. Chem. Soc. Japan, *32*, 1959, p. 809.

# Measured $TE_{01}$ Attenuation in Helix Waveguide with Controlled Straightness Deviations

By D. T. YOUNG

(Manuscript received September 1, 1964)

*A helix circular waveguide 380 feet long has been deformed into three different curves. The added mode conversion loss for a $TE_{01}$ signal mode has been measured and calculated theoretically, with good agreement. For a curve with 30-inch deflection over a distance of 100 feet, peak-to-valley (200-ft period), the measured added loss at 55.5 kmc was 0.23 db/mile, the calculated 0.20 db/mile.*

## I. INTRODUCTION

The added mode conversion loss in a circular waveguide due to random straightness deviations of the guide axis has been calculated by Rowe and Warters.[1] These calculations show that if the mechanical spectrum of the axis wiggles has high density near the beat wavelengths of the coupling modes, then the transmission loss for a $TE_{01}$ signal mode can be excessive for axis deviations of fractions of a mil. However, as emphasized in this paper and in other calculations,[2,3] the requirement on the straightness depends strongly on the mechanical spectrum of the deviations. If the significant part of the mechanical spectrum is far from the beat wavelength of the coupling modes, then extremely large deviations (several feet) do not seriously affect the $TE_{01}$ transmission. Here we offer experimental evidence of these facts for a 2-inch diameter helix waveguide, and associated theoretical calculations which agree favorably with the experimental results.

In Section II we describe the experiment and discuss briefly the measured results of the experiment. Section III contains some theoretical calculations and, finally, in Section IV we discuss the comparison of experimental and theoretical results and some conclusions of the experiment.

## II. DESCRIPTION OF EXPERIMENT AND EXPERIMENTAL RESULTS

A steel-jacketed helix waveguide 380 feet long which was constructed at the Holmdel, N.J., Bell Laboratories[4] was used in this experiment. The guide was bent in the vertical plane to conform to three different curves. A photograph of the guide in one case appears in Fig. 1. The first two curves were roughly sinusoidal and so had a very limited spectral content. The first curve had a period of 40 feet, and the peak-to-peak deflection was 2.4 inches. The second curve had a period of 200 feet and a peak-to-peak deflection of 30 inches. The third curve was a simple model of a deviation which might be caused by random laying errors; basically, the guide was supported every 10 feet by a support which was either 1.2 inches higher or lower than the previous support (mechanical nonuniformities such as the brass couplings prevented rigid adherence to this format at four supports). The actual displacements are given in Table I.

The electrical measurements were performed with the above-ground waveguide test set at Holmdel, using the shuttle-pulse technique.[5] The measured $TE_{01}$ loss over the 50–60 kmc band is shown on an expanded scale in Fig. 2. The measured added losses were 0.65 db/mile

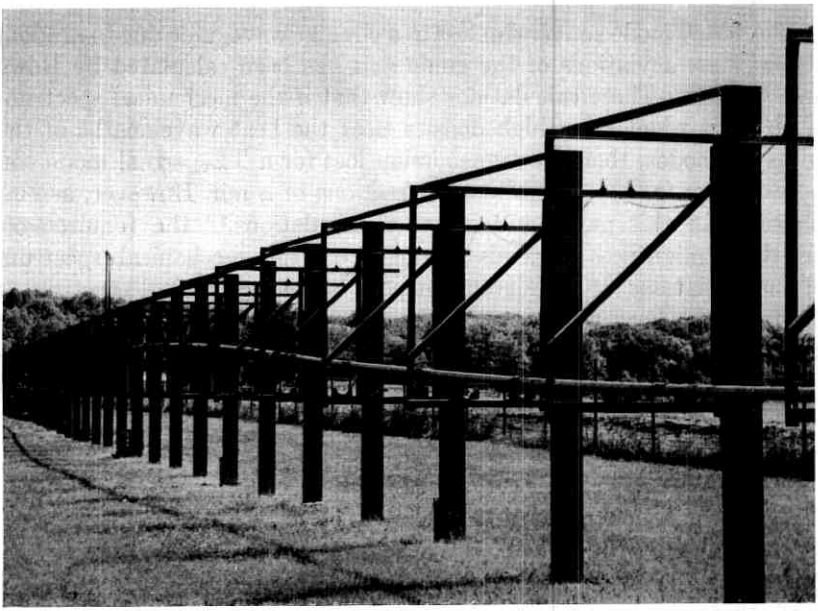

Fig. 1 — 30-inch peak, 200-foot period bend.

TABLE I — ACTUAL DISPLACEMENTS IN EXPERIMENTAL 380-FOOT HELIX WAVEGUIDE

| y (inches) | x (feet) | y (inches) | x (feet) | y (inches) | x (feet) |
|---|---|---|---|---|---|
| 0.0 | 0 | 1.2 | 130 | 2.2 | 260 |
| 0.6 | 10 | 2.4 | 140 | 2.4 | 270 |
| 1.2 | 20 | 1.2 | 150 | 3.6 | 280 |
| 1.2 | 30 | 0.0 | 160 | 4.8 | 290 |
| 0.0 | 40 | 0.0 | 170 | 3.6 | 300 |
| 1.2 | 50 | 1.2 | 180 | 2.4 | 310 |
| 2.4 | 60 | 2.4 | 190 | 1.2 | 320 |
| 3.6 | 70 | 1.2 | 200 | 0.0 | 330 |
| 2.4 | 80 | 0.0 | 210 | 0.0 | 340 |
| 1.2 | 90 | 1.2 | 220 | 1.2 | 350 |
| 1.2 | 100 | 2.4 | 230 | 0.6 | 360 |
| 0.0 | 110 | 1.8 | 240 | 0.0 | 370 |
| 0.0 | 120 | 2.4 | 250 | 1.2 | 380 |

0.23 db/mile, and 0.30 db/mile over the 2.55 db/mile measured at 55.5 kmc with the waveguide straight. The smallest added loss was associated with the bend with the largest amplitude and longest period, thus emphasizing that slow bends cause little added loss even for quite large deviations of the guide axis.

## III. THEORY

Calculation of the added mode conversion loss requires three steps: (1) calculation of the curvature of the guide; (2) calculation of the wall

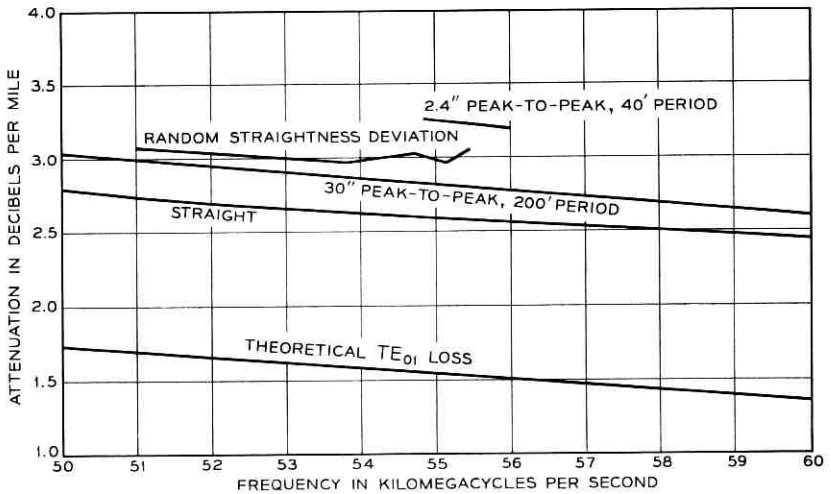


Fig. 2 — Measured $TE_{01}$ attenuation.

impedance in order to determine the propagation constants and coupling coefficients; (3) application of the Rowe-Warters perturbation calculation.

For the curvature calculation we assume the following mechanical model: a uniform homogeneous circular tube of outer radius $b$ and inner radius $a$ with a modulus of elasticity $E$ and a distributed weight $\omega_0$ pounds/foot. The moment of inertia about the axis is $I = (\pi/4)(b^4 - a^4)$. The tube is supported by $(N + 1)$ point supports a distance $l$ apart and the ends are free, so that $y''(0) = y''(Nl) = 0$, where $y(x)$ is the vertical displacement from a preselected reference. Now the simple bending equation for beams is

$$1/R = M/EI \tag{1}$$

where $R$ is the radius of curvature, $M$ is the bending moment and $1/R = y''/[1 + (y')^2]^{\frac{3}{2}}$. For the displacements of interest here $|y'| \ll 1$ and $(1/R) \approx y''$. Now we assume further that the external forces at the supports are concentrated at discrete points. Then, if we differentiate (1) twice, we have

$$EIy^{IV} \approx -\omega_0 + \sum_{i=1}^{N+1} F_i \delta(x - x_i) \tag{2}$$

where $F_i$ is the force at the $i$th support, $\delta(x - x_i)$ is the Dirac delta function and $x_i$ is the location of the $i$th support.

The $F_i$ are difficult to measure and (2) cannot be solved in terms of a simple analytic function. However, if we consider the equation between supports the analytic form of the solution is immediate, and we need to use our end and continuity conditions to evaluate the necessary constants.

Let $z_i = x - x_i$ and $0 \leqq z_i \leqq l$ for all $i$; then for the $i$th section (2) becomes

$$y_i^{IV}(z_i) = -\omega_0/EI \tag{3}$$

which has the solution

$$y_i(z_i) = -\frac{\omega_0}{24EI} z_i^4 + A_i z_i^3 + B_i z_i^2 + C_i z_i + D_i . \tag{4}$$

The conditions necessary to obtain the constants are: $y_1''(0) = 0$, $y_N''(l) = 0$ and $y_i$, $y_i'$, $y_i''$ continuous at the supports. With $N + 1$ supports we have to evaluate $4N$ constants. The continuity conditions are $3(N - 1)$ in number. These, plus the $N + 1$ known displacements at the supports and the two end conditions, yield a total of $4N$ conditions. The actual calculation is long, but straightforward.

H. -G. Unger[6,7] has solved the helix waveguide problem of calculating propagation constants and curvature coupling coefficients assuming a model with the following boundary conditions at the helix:

$$\begin{aligned} E_\theta &= 0 \\ E_z &= -ZH_\theta . \end{aligned} \tag{5}$$

Unger's analysis gives a very general solution if the wall impedance $Z$ due to a complex jacket outside the helix can be calculated. Strictly speaking, the wall impedance will depend on the mode through the propagation constant, so that an exact solution is not possible, but in the oversize waveguide of interest here the wall impedance may be calculated by assuming the propagation constant of each mode equal to the propagation constant of free space. For the modes of interest which couple to the $TE_{01}$ mode due to curvature, the angular wave number is small compared to the longitudinal and radial wave numbers, and neglecting angular variation allows a great simplification by permitting consideration of the jacket structure in rectangular coordinates with no variations of the field in the plane of the helix wires.

For the present helix, the jacket structure consists of a thin layer of clear glass-fiber roving followed by a thicker layer of material that is an aqueous formulation of graphite bonded to continuous filaments of glass roving. A universal wrap gives a checkerboard array with the fibers oriented at about $\pm 45°$ with respect to the $z$ axis of the guide. These layers are enclosed in a steel jacket and are impregnated with epoxy resin. We assume an equivalent transmission line circuit for this jacket, as shown in Fig. 3. $Z_c$ is the shunt capacitance due to the helix wires, region 1 is the lossless layer, region 2 the lossy layer, and the line is terminated by the short circuit of the steel jacket. We calculate the impedance and propagation constant of region 2 from a model used by S. E. Miller,[8] which assumes an infinite lossless region of fiber glass in which are imbedded parallel resistive fibers of zero cross section, spaced uniformly a distance $t$ from each other. The conductance of the fibers is deduced from the dc resistance of the carbon-coated glass roving. We multiply the measured conductance by $\cos^2 45°$ to include the effect of the $\pm 45°$ wrap. The formulas for the impedances and propagation constants are:

($i$) helix-wire capacitance:[6]

$$Z_c = \frac{1}{j\omega\epsilon\, d\left(\dfrac{d}{D-d} - \dfrac{\ln 4}{\pi}\right)} \tag{6}$$

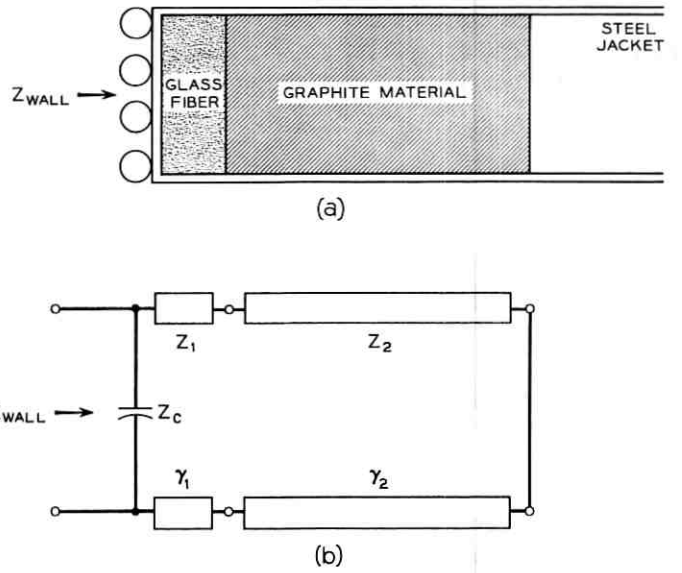


Fig. 3 — (a) Physical structure of helix waveguide used in experiment; (b) equivalent transmission line circuit.

where we have

$D = 0.0055$ inch (diameter of helix wires including insulation)

$d = 0.0045$ inch (diameter of helix wires)

$\epsilon = 2.77\epsilon_0$ (permittivity of Formvar insulation)

(*ii*) lossless layer: $\epsilon = 4\epsilon_0$, $l_1 = 0.005$ inch

$$Z_1 = \sqrt{\frac{\mu_0}{4\epsilon_0}} = \frac{1}{2} Z_0$$

$$\gamma_1 = \omega\sqrt{4\mu_0\epsilon_0}$$

(*iii*) lossy layer:[8] $\epsilon = 4\epsilon_0$; $l_2 = 0.033$ inch

$$Z_2 = \frac{1}{\dfrac{\cos^2 45°}{2R} + \left\{\dfrac{4\epsilon_0}{\mu_0} + \dfrac{1}{4}\left[\dfrac{\cos^2 45°}{R}\right]^2 - j\,\dfrac{\cos^2 45°\sqrt{\dfrac{4\epsilon_0}{\mu_0}}}{R\tan\omega\sqrt{4\mu_0\epsilon_0}\,t}\right\}^{\frac{1}{2}}}$$

$$\gamma_2 = (1/t)\ln\left\{\cos\omega\sqrt{4\mu_0\epsilon_0}\,t\left[1 + \frac{\cos^2 45°}{R\left(\dfrac{1}{Z_2} - \dfrac{\cos^2 45°}{R}\right)}\right.\right.$$

$$+ \frac{j\sqrt{\frac{4\epsilon_0}{\mu_0}}\tan \omega\sqrt{4\mu_0\epsilon_0}\, t}{\left(\frac{1}{Z_2} - \frac{\cos^2 45°}{R}\right)}\Bigg]\Bigg\}$$

where $t = 0.0081$ inch (average spacing of resistive fibers) and $R = 234$ ohms/square (from dc resistance measurement). These parameters correspond to a wall impedance at 55.5 kmc of

$$\frac{Z_{\text{wall}}}{Z_0} = 0.41 \angle -32°$$

where $Z_0 = \sqrt{\mu_0/\epsilon_0}$.

From the wall impedance the coupling coefficients and propagation constants of the spurious modes can be determined.[7] Then the $TE_{01}$ loss due to mode conversion can be computed approximately using the method of Picard developed by Rowe and Warters.[1] Summing over all spurious modes $k$, we have:

$$A(\text{nepers}) = \sum_{k=1}^{\infty} \text{Re}\left[\int_0^L e^{\Delta\Gamma_k u}\, du \int_0^{L-u} c_k(x)c_k(x+u)\, dx\right] \quad (8)$$

where

$$\Delta\Gamma_k = \Delta\alpha_k + j\Delta\beta_k = \Gamma_{TE_{01}} - \Gamma_k$$

and

$$\begin{aligned} c_k(x) &= (c_k R)y''(x) \\ &= (c_k' + jc_k'')Ry''(x). \end{aligned}$$

The sum is over the infinite number of coupling modes; however convergence of the coupling coefficients is rapid, and only three modes contribute significantly, $TE_{11}$, $TE_{12}$, and $TM_{11}$. Since each $\Delta\alpha$ is large and negative and the slowness of the bends makes the inner integral (8) a slowly varying function of $u$, we have,

$$\int_0^L e^{\Delta\Gamma_k u}\, du \int_0^{L-u} c_k(x)c_k(x+u)\, dx \approx \int_0^L e^{\Delta\Gamma_k u}\, du \int_0^L c_k^2(x)\, dx. \quad (9)$$

With this approximation and the fact that $e^{\Delta\alpha_k L} \ll 1$, (8) becomes

$$A \approx \int_0^L [y''(x)]^2 dx \sum_{k=1}^{\infty} \frac{P_k(-\Delta\alpha_k) - Q_k\Delta\beta_k}{(\Delta\alpha_k)^2 + (\Delta\beta_k)^2} \quad (10)$$

where

$$P_k = (c_k'^2 - c_k''^2)R^2$$
$$Q_k = +2c_k'c_k''R^2.$$

To evaluate

$$\int_0^L [y''(x)]^2 dx$$

we need the second derivative of (4):

$$\int_0^L [y''(x)]^2 dx = N\left[\frac{\omega_0}{2EI}\right]^2 \frac{l^5}{5} - \frac{\omega_0 l^4}{4EI}\sum_{i=1}^{N} A_i + \frac{l^3}{3}\sum_{i=1}^{N}\left(A_i^2 - \frac{\omega_0}{EI}B_i\right) + l^2\sum_{i=1}^{N} A_iB_i + l\sum_{i=1}^{N} B_i^2. \tag{11}$$

The $A_i$ and $B_i$ are obtained from the mechanical boundary conditions. After eliminating the $C_i$ and $D_i$ from (4), we have the following $2N - 1$ equations in $A_i$ and $B_i$ :

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_N \end{bmatrix} = \frac{1}{l}\begin{bmatrix} B_2 - 0 + \frac{\omega_0}{2EI}l^2 \\ B_3 - B_2 + \frac{\omega_0}{2EI}l^2 \\ \vdots \\ 0 - B_N + \frac{\omega_0}{2EI}l^2 \end{bmatrix} \tag{12}$$

$$\begin{bmatrix} 4100\ \ldots\ldots\ldots \\ 1410\ \ldots\ldots\ldots \\ 0141\ \ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ 1410 \\ \ldots\ldots\ldots\ 0141 \\ \ldots\ldots\ldots\ 014 \end{bmatrix}\begin{bmatrix} B_2 \\ B_3 \\ \cdot\cdot \\ \cdot\cdot \\ \cdot\cdot \\ B_{N-2} \\ B_{N-1} \\ B_N \end{bmatrix} = \frac{6}{l^2}\begin{bmatrix} y_3 - 2y_2 + y_1 - \frac{\omega_0}{12EI}l^4 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{N+1} - 2y_N + y_{N-1} - \frac{\omega_0}{12EI}l^4 \end{bmatrix} \tag{13}$$

TABLE II — WALL IMPEDANCES

| | $\Delta\alpha$(nepers/ft) | $\Delta\beta$(radians/ft) | $P$ | $Q$ |
|---|---|---|---|---|
| $TE_{11}$ | −0.072 | −1.87 | 45.5 | 2.3 |
| $TM_{11}$ | −0.560 | +2.11 | 31.4 | 22.3 |
| $TE_{12}$ | −0.133 | +2.94 | 65.6 | −24.6 |

where $y_i$ is the measured displacement of the guide at the $i$th support. The inverse of the square matrix in (11) may be found by solution of difference equations and symmetry properties. The inverse matrix has the following elements:

$$a_{ik} = c_1\alpha_1^{k-i+1}\frac{1-\alpha_1^{2i}}{1-\alpha_1^{2}} + c_2\alpha_2^{k-i+1}\frac{1-\alpha_2^{2i}}{1-\alpha_2^{2}} \qquad (14)$$

$$\text{for} \quad i \leqq k \leqq N-i+1$$

$$i < N/2$$

where

$$\alpha_1 = -2+\sqrt{3}$$

$$\alpha_2 = -2-\sqrt{3}$$

$$c_1 = \frac{\alpha_1^{N}}{\alpha_2^{N}-\alpha_1^{N}}$$

$$c_2 = \frac{\alpha_2^{N}}{\alpha_1^{N}-\alpha_2^{N}}$$

and the remaining elements are obtained from the symmetry about the main diagonal and the off diagonal.

The wall impedance $0.41 \angle -32°$ corresponds to the values given in Table II. The computed loss and the experimentally measured loss are listed in Table III.

TABLE III — COMPUTED AND EXPERIMENTALLY MEASURED LOSSES

| Case | Period | Max. Deflection | Added Exp. Loss | Added Cal. Loss |
|---|---|---|---|---|
| 1 | 40 feet | 2.4 inches | 0.65 db/mile | 0.69 db/mile |
| 2 | 200 feet | 30.0 inches | 0.23 db/mile | 0.20 db/mile |
| 3 | — | 2.4 inches | 0.30 db/mile | 0.27 db/mile |

## IV. RESULTS AND CONCLUSIONS

The experimental results agree favorably with the computed results, considering the various approximations in the theoretical model. The important result is that mechanical deviations of long wavelengths, such as we might expect from random laying errors, contribute very little to the transmission loss. Thus the tolerance on guide axis wiggles changes several orders of magnitude when the mechanical wavelength changes from two feet to two hundred feet.

## V. ACKNOWLEDGMENTS

The author would like to thank C. W. Curry, who made the experimental measurements, and H. E. Rowe, J. A. Young, and C. F. P. Rose for their assistance and interest.

## REFERENCES

1. Rowe, H. E., and Warters, W. D., Transmission in Multimode Waveguides with Random Imperfections, B.S.T.J., *41*, May, 1962, pp. 1031–1170.
2. Warters, W. D., unpublished work.
3. Rowe, H. E., unpublished work.
4. Beck, A. C., and Rose, C. F. P., Waveguide for Circular Electric Mode Transmission, Proc. IEE, *106*, Part B, Suppl. No. 13, Sept., 1959.
5. Miller, S. E., and Beck, A. C., Low-Loss Waveguide Transmission, Proc. IRE, *41*, March, 1953, pp. 348–358.
6. Unger, H. -G., Helix Waveguide Theory and Applications, B.S.T.J., *37*, Nov., 1958, pp. 1599–1647.
7. Unger, H. -G., Normal Modes and Mode Conversion in Helix Waveguides, B.S.T.J., *40*, Jan., 1961, pp. 255–280.
8. Miller, S. E., unpublished work.

# Losses Suffered by Coherent Light Redirected and Refocused Many Times in an Enclosed Medium

By O. E. DeLANGE

(Manuscript received September 30, 1964)

*If a beam of light is to be transmitted for any considerable distance along the surface of the earth, it will have to be redirected at intervals in order to follow the terrain and focused repeatedly to counteract diffraction. The directing and focusing elements, whether lenses or mirrors, will introduce some loss in addition to that produced by the transmission medium itself. The experiment described in this paper was performed to determine the magnitude of the total loss encountered with such transmission and to determine how much of this loss is due to each of the contributing factors.*

*A beam of light, enclosed in a metal pipe, was redirected many times by confocally-spaced spherical mirrors, and the loss as a function of the distance over which the beam had been transmitted was determined. At the operating wavelength of 6328 Å these losses, which were found to be almost entirely due to mirror deficiencies, amounted to about 1 per cent per reflection. As a result of the loss being largely in the mirrors the loss per mile depends to a considerable extent upon the spacing between these optical elements. The expected loss for a number of assumed spacings is tabulated.*

*The experimental results encourage the belief that beams of coherent light can be redirected and focused many times without excessive loss, and that the mechanical stability required can be obtained — in the laboratory at least.*

## I. PURPOSE OF EXPERIMENT

The advent of the optical maser as a source of coherent light has stimulated considerable interest in the possibility of employing light beams as extremely broadband carriers of information. If a beam of light is to be transmitted along the surface of the earth it will be necessary to redirect and focus it at intervals by means of lenses or mirrors in order to follow the terrain. By employing a sufficient number of redirectors a long-distance transmission system can be built up. A number of such

systems have been proposed by Kompfner.[1] Goubau and Christian have also described such a transmission system.[2] The experiment described in this paper was performed to determine the amount of loss suffered by a beam of light due to transmission through one such system. The experiment has also provided an indication of some of the other problems involved in light transmission, such as that of obtaining the necessary alignment, stability, freedom from vibration, etc.

## II. DESCRIPTION OF THE EXPERIMENT

Since the loss through a light-transmission system can be very low, it is desirable when measuring loss to employ a long path in order to obtain accurate measurements. One means of obtaining a long transmission path in a limited space is to use a single path repeatedly, thus making the effective path length many times the actual length. Since mirrors are ideal for folding a transmission path back upon itself, they were chosen as the redirecting elements for this experiment. They have the additional advantages of being simple and available; if spherical mirrors are employed, they can be made to refocus the beam at each reflection.

To isolate the transmission medium from the surrounding environment, it was enclosed in an aluminum pipe 6 inches in diameter. Isolation was the only purpose served by the pipe — it played no part in the actual transmission. Also, since the maximum beam diameter was less than $\frac{1}{2}$ inch, a much smaller pipe would have been satisfactory. Fig. 1 shows the experimental setup in schematic form. The pipeline, which was approximately 330 feet long, was light-tight and was treated

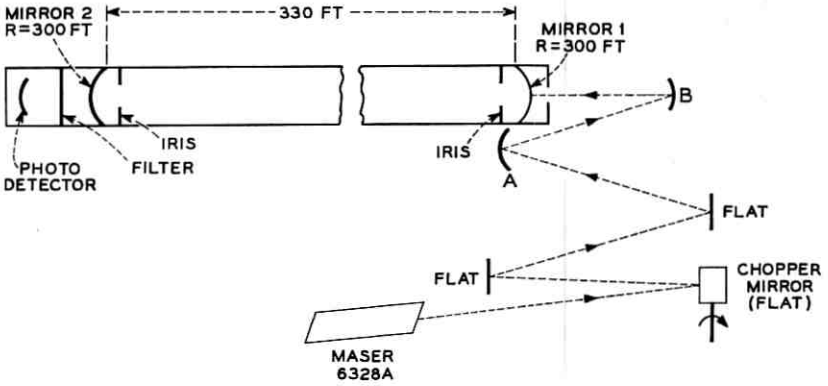


Fig. 1 — Light propagation experiment, system arrangement.

on the inside to minimize reflections. Each end was provided with a spherical mirror with a radius of curvature of approximately 300 feet. This radius was chosen in order to make the configuration nearly, but not exactly, confocal. Each mirror was provided with an iris which made it possible to adjust its effective diameter from practically zero up to 1.63 inches. The mirrors were coated with dielectric layers to produce a reflectivity of about 99.5 per cent at the operating wavelength of 6328 angstroms.

The mirrors 1 and 2 arranged in the configuration shown on Fig. 1 constitute a confocal resonator as described by Boyd and Gordon[3] and by Fox and Li.[4] Some consideration was given to the idea of applying light to the line continuously and determining losses by the usual method of measuring the "$Q$" of the resonator. However, because of the extremely high $Q$, the requirements imposed on mechanical stability would have been formidable, so a more feasible method was sought. The stability requirements were reduced to reasonable values by applying light to the line in short pulses. The pulse duration was made less than the round-trip transit time through the line so that there would be no overlapping of pulses and hence no critical relationship between the light wavelength and the mirror spacing.

With this configuration, when a pulse of light strikes a mirror some of it is lost but most is refocused and reflected to the opposite mirror. Here it is refocused and reflected back to the first mirror and so on indefinitely. In this way each single pulse of light applied to the line results in a train of pulses decaying in intensity, with the rate of decay providing a measure of the transmission losses.

Part of the light lost at each reflection was transmitted through the dielectric coating of the mirror. The part transmitted through the mirror at the far end of the line was applied to a photomultiplier, the output of which was, in turn, applied to an oscilloscope or other measuring equipment. We thus obtained an output from the line without increasing the losses, since this light would have been dissipated in any case. Since most of the light striking a mirror was reflected, and some of the remainder absorbed, the loss from the line into the measuring equipment was high, being about 26 db for the mirrors employed. For the same reason we were able to get light into the line without increasing losses by going in through the back of the mirror at the near end. Here, again, the loss was about 26 db.

The light-signal source was a dc-excited helium-neon gas maser about one meter long, of the type described by White and Rigden.[5] One of the maser mirrors was stopped down with an iris to such an extent that

it could oscillate at only the fundamental transverse mode. The several longitudinal modes present caused no difficulty since there were no frequency sensitive elements involved. The output power was approximately 1 milliwatt.

The output beam was chopped into short pulses by means of a small, flat, rotating mirror which could be driven at rates up to 40,000 rpm (see Fig. 1). Pulses as short as 0.2 microsecond could be obtained, because the beam was effective in exciting the line only during the very short time that it was accurately aligned with the axis of the pipe. Since the round-trip transit time for the line was nearly 0.7 $\mu$sec, pulses of 0.5 $\mu$sec duration were sufficiently short, and the chopper could be run at considerably less than its maximum speed. The mirrors A and B shown in Fig. 1 between the chopper mirror and the end of the line served the purpose of focusing the beam to get it launched properly, as discussed in the next section. Sweeping a light beam across the mirrors of a system such as this undoubtedly produces many higher-order modes. However, these modes are generated when the beam is out of alignment with the axis of the line, are off-axis modes, and die out very rapidly.

## III. BEAM LAUNCHING

Refs. 3 and 4 show that after many reflections in a confocal system the light has a very definite distribution of intensity at each part of the resonator. Further, this distribution is the one which provides the lowest losses. If light is launched into the line with this distribution, losses and starting transient effects will be minimized. Using equations (19) and (24) of Boyd and Gordon[3] we calculate a beam diameter of 0.35 inch at each end mirror and 0.25 inch at the center of the line. If the beam is launched into the line in such a way as to fit these dimensions there will be a minimum of loss. Launching conditions were controlled by two spherical mirrors, A and B, shown in Fig. 1 mounted between the chopper mirror and the end of the line. These mirrors were spaced so as to be nearly confocal, with the second mirror having twice the focal length of the first in order to provide a two-to-one increase in beam diameter. To minimize the distortion of beam shape produced by the spherical mirrors the launching arrangement was set up to provide as nearly as possible normal incidence on these mirrors. By proper adjustment of the spacing between the focusing mirrors the beam was made slightly convergent as it entered the line. It converged to the center of the line, where it reached a minimum diameter and then diverged slowly until it reached the mirror at the far end of the line. Here it was reflected and again made converging, thus repeating the process.

The beam was photographed at various points in the line in order to determine its cross section. Some of the photographs, which were taken with the beam not being chopped, are shown in Fig. 2. Fig. 2(a) shows the beam as it entered the line. It was 0.37 inch in diameter in comparison to the calculated value of 0.35 inch. The ring segments directly above the spot were part of an interference pattern produced by reflections from the back surface of the output mirror of the maser. Fortunately the reflected light left the maser at an angle slightly different from that of the transmitted beam, and as a result the two beams were fairly well separated in space at the center of the pipe line. This is seen in Fig. 2(b), which shows the beam diameter to have decreased to 0.28 inch at this point. Fig. 2(c) shows that by the time the beam reached the far end of the line its diameter had increased to 0.4 inch. Figs. 2(d), (e) and (f) show the beam after it had been reflected the first time from the mirror at the far end of the line. The focusing effect of this mirror is quite evident. Although the measured beam diameters are somewhat different

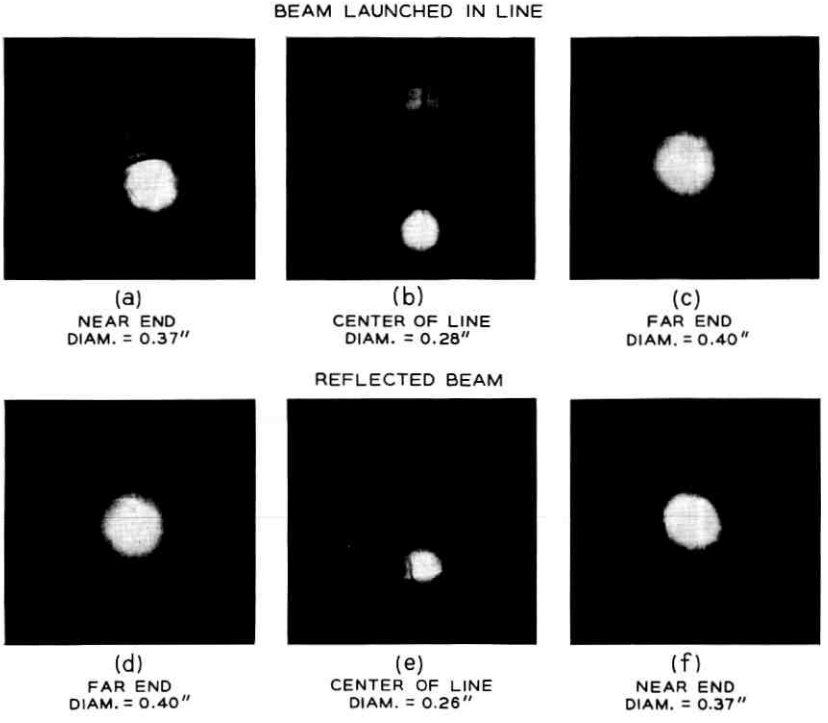


Fig. 2 — Beam cross sections.

from the calculated values, they differ by only about 12 per cent in the worst case.

## IV. DETERMINATION OF LOSSES

### 4.1 *Two-Mirror Shuttle-Pulse*

Fig. 3 illustrates the performance of a shuttle-pulse experiment described above and shown in Fig. 1. For this figure, which illustrates the decay of light power with successive reflections, each pulse on the oscilloscope trace represents one round trip of the light beam through the transmission line, there being a total of 75 pulses shown. For Fig. 4 the oscilloscope sweep was expanded to show individual pulses. Fig. 4(a) shows the first 13 round trips. Fig. 4(b) shows pulses corresponding to 300 to 310 round trips for a distance of 37.5 to 39 miles. Pulses which have made 400 round trips for a total distance of 50 miles have been detected with little difficulty. The first pulse in the group shown on Fig. 4(a) is the one applied to the line. It has a peak power of only $5 \times 10^{-9}$ watts when it arrives at the cathode of the photomultiplier tube.

Fig. 5 is a typical plot of power loss versus number of trips through the line. After about 40 trips the loss is seen to remain constant at the rate of 0.046 db per trip, which corresponds to a power loss of 1 per cent per trip. The fact that the loss was somewhat higher for the early trips may be due to higher-order modes, present because of imperfect launch-

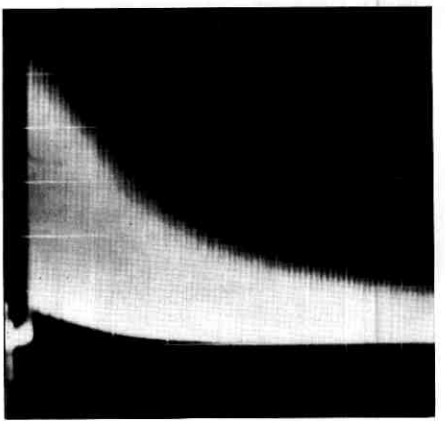

Fig. 3 — First 75 round trips; illustrates decay of light energy produced by successive reflections.

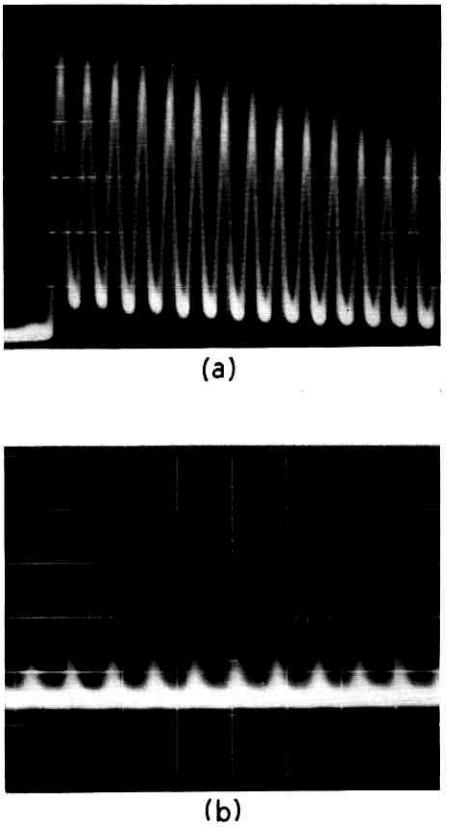


Fig. 4 — Individual pulses: (a) first 13 round trips, (b) round trips number 300 to 310.

ing of the beam. These results were obtained with effective mirror diameters of 0.87 inch, which corresponds to a Fresnel number $N$ of 2, where $N$ is equal to $a^2/b\lambda$; $a$ is the mirror radius and $b$ the spacing between mirrors.

In order to determine the effects of diffraction on the measured losses the value of $N$ for the system was varied, in steps, from 0.5 to 4 by adjusting the iris in front of each mirror, thus changing the effective mirror diameter. The photographs of Fig. 6 illustrate the effect of $N$ upon losses. It is interesting to compare Fig. 6(e), for small mirrors at both ends, with Fig. 6(f), where the diameter of the mirror at the receiving end has been increased. When making comparisons involving an arrangement of mirrors with different diameters the losses should be considered

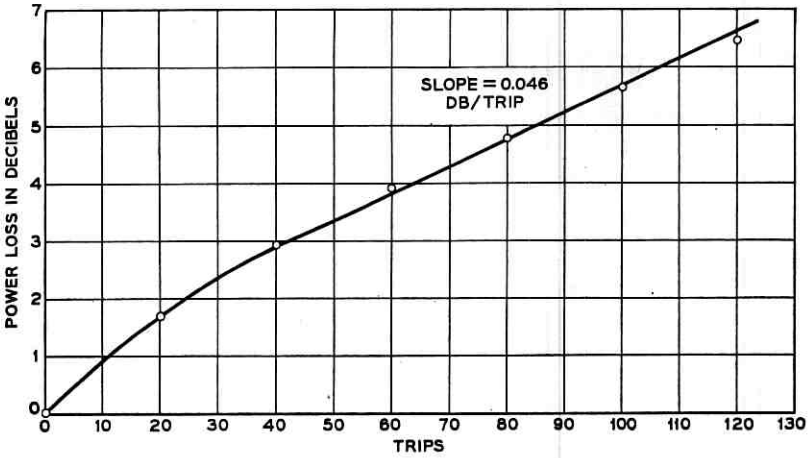


Fig. 5 — Power loss versus number of trips through the line.

on a round-trip basis so as to include one reflection from each mirror. For the case shown by Fig. 6(e) the loss was 14.5 per cent at each mirror. For the conditions of Fig. 6(f) one might expect the losses to be 14.5 per cent for the small mirror and 1.5 per cent for the larger mirror, to give a total round-trip loss of 16 per cent. Fig. 6(f) shows this loss to be only 6 per cent. The lower loss resulted from the fact that the light redistributed itself to match the changed configuration, there being a smaller beam diameter at the small mirror and a larger beam diameter at the large mirror.

Fig. 7 shows a plot of measured loss versus $N$ and also diffraction loss as calculated by Fox and Li plotted against this same parameter. It is evident that for values of $N$ of 0.6 or less diffraction losses predominate, whereas for values of $N$ greater than 1.0 other losses are more important. The curve labeled "expected loss" was obtained by adding the known mirror loss of 0.5 per cent to the calculated diffraction loss.

It can be seen that in the region of high losses the measured losses are close to the diffraction losses as calculated by Fox and Li. The data of Fig. 7 are somewhat inaccurate for two reasons. For the lower values of $N$ the mirror diameters were small and it was much more difficult to obtain accurate system alignment, so that there are possibly some alignment losses included. Also, the losses become so high for low values of $N$ that we have only a few trips through the line before the signal becomes comparable to the noise. As a result losses must be measured in a region where higher-order modes are important. More accurate results could be

obtained if enough power were available to make it possible to determine the rate of decay after 50 to 100 trips where the higher-order modes have had time to die out. The presence of higher-order modes probably accounts for the fact that losses are still decreasing with increase of $N$ for values of $N$ as great as 4.

### 4.2 *Periscope System*

If mirrors are to be used as beam directors in a practical system they will need to be used in pairs. Kompfner[1] has suggested pairs of cylindrical mirrors; however, a plane mirror to direct the beam in combination with a spherical mirror for focusing also appears to be a satisfactory arrangement. A transmission path made up of such pairs is shown schematically in Fig. 8(a). It is evident that this combination allows the direction of the beam to be changed at any pair of mirrors.

In order to simulate the use of mirror pairs, two plane mirrors were inserted near the center of the transmission path as shown in Fig. 8(b). Except for the addition of the two plane mirrors the system was operated

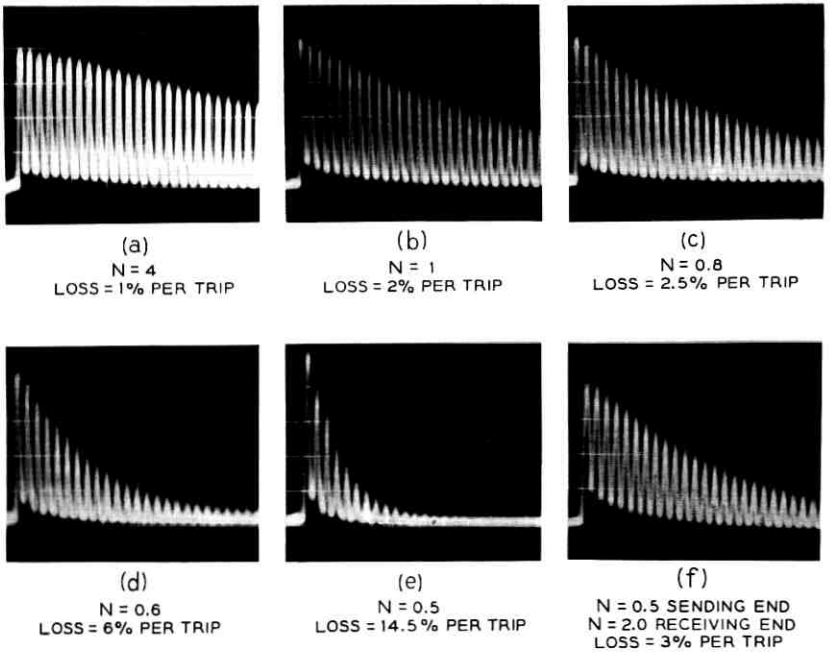


Fig. 6 — The effect of mirror diameter on loss.

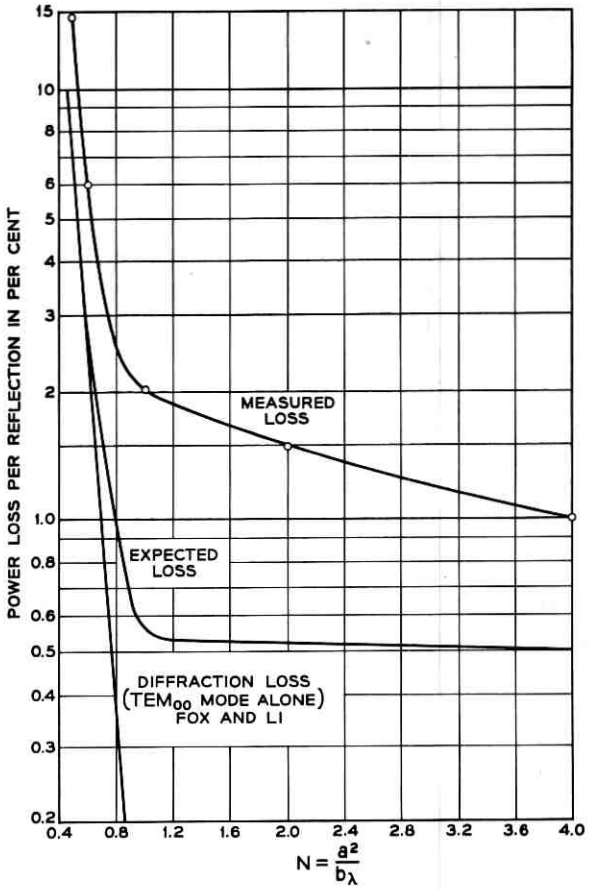


Fig. 7 — Comparison of measured loss with expected loss.

just as before, with pulses of light shuttled back and forth between the two spherical mirrors at the ends of the line. With this configuration there were three reflections per trip in comparison to one reflection for the two-mirror case. For this arrangement the total loss was measured to be 0.08 db, or 2 per cent per trip in comparison to 1 per cent for the two-mirror system. The loss was thus increased by 0.5 per cent per reflection from the flat mirrors, which is just the reflection loss of these mirrors.

### 4.3 *Multimirror Experiment*

The two-mirror shuttle pulse experiment differs from a practical transmission line in one other respect — the same two mirrors are used re-

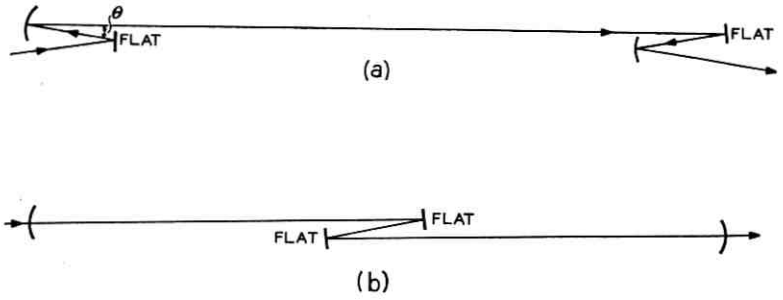


Fig. 8 — (a) Use of mirror pairs to change beam direction; (b) experimental equivalent with flat mirrors at the center of the line.

peatedly, whereas in an actual system each reflection would involve a different mirror.

In order to obtain a better simulation of an actual line a multimirror experiment was planned. Mirrors were purchased for this experiment but, unfortunately, upon delivery were found to be defective. The best we could do was to set up a four-mirror experiment with the four good mirrors available. These mirrors, which were of excellent quality, were ground by the Schutte Optical Company, Incorporated of Rochester, New York, and coated by W. L. Bond of the Murray Hill, N. J., Bell Laboratories. The mirrors were first set up in a four-mirror shuttle pulse experiment as indicated in Fig. 9(a). For this arrangement the light traversed the path $M_1$, $M_2$, $M_3$, $M_4$, $M_3$, $M_2$, $M_1$, $M_2$, etc. The same four mirrors were also arranged in a circulating loop as shown in Fig. 9(b). Here the light path was $M_1$, $M_2$, $M_3$, $M_4$, $M_1$, $M_2$, etc.

Fig. 10 is a plot of the losses measured for the four-mirror shuttle

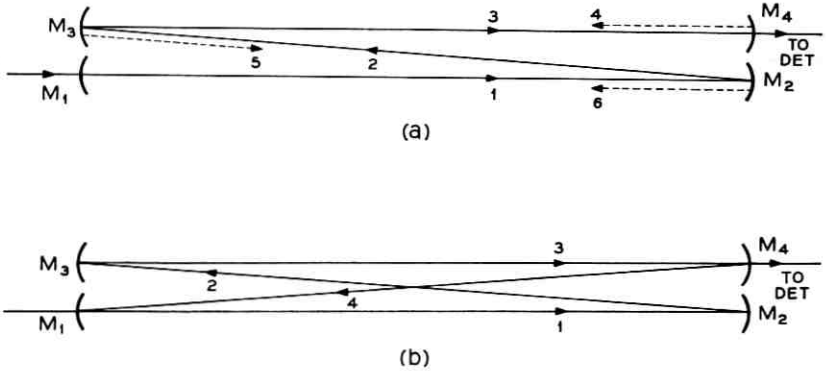


Fig. 9 — Some four-mirror experiments.

pulse, the four-mirror circulating loop and a two-mirror shuttle-pulse system. After it has reached a steady value the loss is seen to be the same, 0.05 db per trip, for all three systems.* A number of comparisons have yielded no measurable difference between the losses for a two-mirror system and a four-mirror experiment. This tends to indicate that the shuttle-pulse data can be applied to a long, straight-through system. Obviously more conclusive results could be obtained from an arrangement using at least ten mirrors, set up either as a circulating loop or as a shuttle-pulse experiment.

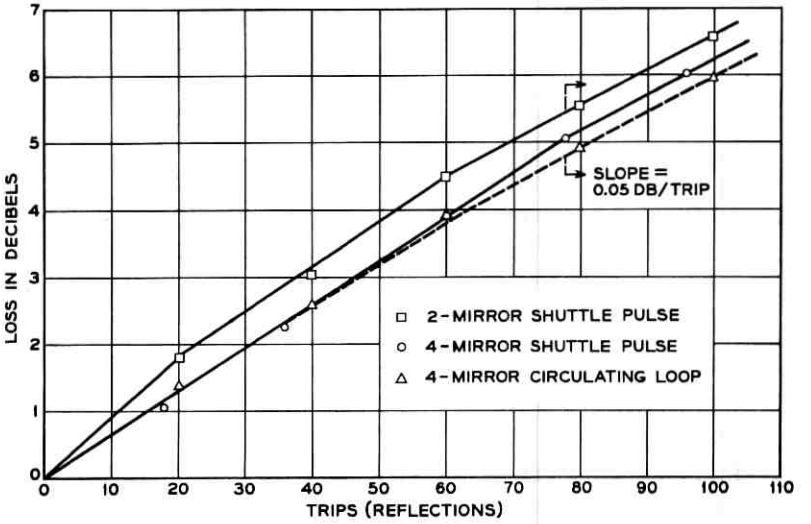


Fig. 10 — Comparison of four-mirror system losses with two-mirror system losses.

## V. LOSSES

According to the best data available from other measurements our mirrors have reflectivity of 99.5 per cent or a reflection loss of 0.5 per cent. The measurements described here yield a loss of 1 per cent per trip through the line when there is one reflection per trip. This leaves a 0.5 per cent loss per trip to be accounted for. We know, both from theory and experiment, that for an $N$ of 2 diffraction loss is negligible. The remaining half per cent of loss, amounting to 0.023 db per trip, must have

* This is slightly greater than the 0.046 db per trip for the two-mirror shuttle pulse as plotted in Fig. 5. The two sets of data were taken at different times, which could well account for the discrepancy.

been due to the atmosphere contained in the pipe, to lack of perfect mirror alignment and to beam spreading and scattering produced by mirror imperfections, dust particles etc. All of these except atmospheric losses should decrease with increasing mirror diameter. The plot of measured loss of Fig. 7 shows this loss to decrease with increasing mirror diameter for values of $N$ up to 4 or more. This would indicate that at least part of the loss is due to misalignment and the beam spreading which results from the presence of higher-order modes.

In general, atmospheric losses result from absorption by water vapor, carbon dioxide and oxygen, and from scattering by small particles of dust etc. The losses measured for the 330-foot line are slightly smaller than those determined by the author from a similar shuttle-pulse experiment in which the mirrors were only 110 feet apart. Any atmospheric losses should, of course, be greater for the longer line. This indicates that small, undetermined differences in mirror loss* were more important than the atmospheric loss and that the latter is too small to measure accurately by this experiment. From a study of data on the solar spectrum after transmission through the Earth's atmosphere, Long and Lewis[6] conclude that atmospheric absorption losses should be negligible at our operating wavelength of 6328 angstroms.

At first thought the conclusions stated above might be considered inconsistent with experimental data obtained by Taylor and Yates.[7] For a path over water they measured a loss of 1.1 db per mile for a 3.4-mile path and 0.63 db per mile for a 10.1-mile path at our operating wavelength. Even the smaller of these losses is considerably greater than the 0.37 db per mile which we are attempting to account for. The results obtained by Taylor and Yates probably do not apply to our setup for several reasons. In the first place, part of the losses they measured may have been due to haze, water droplets, dust, etc. which we do not have in the pipe line. Also, for their experiment, each determination of loss was made over a band of wavelengths, whereas for the experiment described here we are dealing with monochromatic light. Burch, Howard and Williams[8] have pointed out that results measured for the wideband case do not necessarily apply to the transmission of monochromatic light.

Taking all of the above factors into consideration we are led to the conclusion that atmospheric loss in the pipe is much less than the 0.37 db per mile which is unaccounted for.

Of particular interest to a system designer is the loss per mile which can be expected. For a light-transmission system where the beam is re-

* The mirrors employed in the two experiments were supposed to be identical except for radius of curvature.

directed at intervals this loss will depend to a great extent upon the spacing between directors. For the 330-foot spacing and using the periscope arrangement we have measured 0.08 db per trip for mirrors of 0.875-inch diameter. This corresponds to 1.28 db per mile. This figure is pessimistic, since three reflections per trip were involved in the experiment, whereas in an actual line involving pairs of mirrors there would be only two reflections per link. This should reduce the loss to 0.96 db per mile.

From the experimental data it is possible to determine what the approximate value of loss would be for various spacings between directors. Some of these values are listed in Table I along with the assumed conditions. This table is based on the assumption that each director consists of two mirrors, each with a reflection loss of 0.5 per cent, and, unless otherwise stated, the loss produced by the atmosphere in the line is assumed to be 0.1 db per mile. An additional 0.5 per cent is added for each mirror pair to represent the loss still unaccounted for.

For all of the cases listed in the table the mirror diameter is great enough to make diffraction losses negligible. Although decreasing the mirror diameter to values somewhat less than those shown in the table would not cause an appreciable increase in loss, such a decrease would make alignment more critical and decrease stability.

## VI. PROBLEMS ENCOUNTERED

A number of the problems encountered in this experiment will also be present in any practical system. For this reason some of these will be discussed briefly.

### 6.1 *Alignment*

If losses are to be kept small it is obviously necessary to obtain rather accurate alignment. Not only must the beam be launched along the axis

TABLE I — LOSSES FOR VARIOUS SPACINGS BETWEEN DIRECTORS

| Spacing between Mirror Pairs | N | Mirror Diameter (inches) | Assumed Atmospheric Loss, db/mile | Total Loss, db/mile |
|---|---|---|---|---|
| 165 feet | 1.3 | 0.5 | 0.1 | 2.02 |
| 165 feet | 1.3 | 0.5 | 0 | 1.92 |
| 330 feet | 2.6 | 1 | 0.1 | 1.06 |
| 330 feet | 2.6 | 1 | 0 | 0.96 |
| 1,320 ft (0.25 mile) | 2.6 | 2 | 0.1 | 0.34 |
| 1,320 ft (0.25 mile) | 2.6 | 2 | 0 | 0.24 |
| 2,640 ft (0.5 mile) | 1.3 | 2 | 0.1 | 0.22 |
| 2,640 ft (0.5 mile) | 1.3 | 2 | 0 | 0.12 |

of the line but each reflection must be such as to keep the beam on this axis. Simple calculations show that, for the 330-foot line, a tilt of 20 seconds of arc of one of the mirrors results in the displacement of the beam by a full beam diameter at the opposite end of the line; greater mirror spacings would be still more critical. In spite of the stringent requirements it has been feasible to obtain satisfactory alignment for the mirror spacings used in this experiment.

### 6.2 *Stability*

Both the shuttle-pulse experiment and the circulating loop described here are affected to a much greater extent by mirror movement than a carefully designed straight-through system of the same length would be. For the experimental systems, rotation of a mirror through an angle $\theta$ would shift the beam, for a single reflection, through an angle $2\theta$. For a straightaway system employing pairs of mirrors, as shown on Fig. 8, the situation would be quite different. If the two mirrors of a pair are parallel and are mounted close together on a rigid mount, small rotations of the mount will produce no angular deviation of the transmitted beam, since both mirrors rotate together. At those locations where it is desired to change the direction of propagation, the mirrors in a pair will not be parallel and rotation of the mount will result in some angular deviation. However, for small departures from parallelism the combination will still be very superior to a single mirror and deviations will be small.

A system using mirrors in this way is sensitive to rotation of either mirror with respect to the other mirror of a pair. Such relative motion can be minimized by mounting each pair of reflectors on a very rigid mount, made of material with a low coefficient of expansion. A pair could be enclosed in a comparatively small volume, which would make it practical to control the temperature of the complete mount and thus assure additional stability.

The only precautions taken to provide stability in the experimental set-up consisted of employing rugged mirror mounts and isolating the mirrors from mechanical motions of the pipe line. The latter was accomplished by mounting the end mirrors on rigid platforms which were only very loosely attached to the pipe. In spite of these shortcomings of the experimental system, it would operate for hours without readjustment. Only large temperature changes produced serious deflections of the beam. It is evident that the inherently superior stability of a straight-through system would go far toward compensating for its much greater length.

### 6.3 *Atmospheric Turbulence*

In spite of the fact that our experimental line is inside a building there were point-to-point differences in temperature sufficient to produce air currents inside the line. The resulting variations in index of refraction as masses of air at different temperatures moved through the beam caused random fluctuations of the position and shape of the beam arriving at the end of the line. This difficulty was overcome by applying a one-inch layer of insulation over the line, which reduced fluctuations to the point where they were no longer discernible.

The shuttle-pulse experiment is considerably less susceptible to turbulence effects than a straight-through system of the same length would be. Any movement of air masses will take place in times long in comparison to transit time through the line. Hence any displacement of the beam by passage through a refractive region will be almost exactly canceled by passage through the same region, but in the reverse direction, on the return trip. Conversely, air turbulence has a greater effect on the circulating loop of Fig. 9(b) than on a straightaway system. In the loop the beam passes many times in the same direction through any discontinuity and each time is deflected in the same way whereas, for the straight-through system, there would be only one passage through any one discontinuity and the resultant deflections would be random. Insulating the line reduced turbulence effects to the point where they were insignificant — even with the circulating loop.

In a practical system air currents could be made negligible by partial evacuation of the line or possibly reduced to a sufficiently low value by other means. In any case the beam enclosure would most likely be installed underground, where temperature variations are much smaller than they are out in the open air.

## VII. THE LINE AS A RESONANT CAVITY

Some experiments have been performed with the chopper mirror stopped in such a position that the beam was continuously lined up with the axis of the line. There was evidence of resonance in that the intensity of the beam transmitted through the line increased very noticeably when the mirrors were properly aligned. This increase was very evident to the eye even though it responds only to average intensity. The photomultiplier output as recorded in Fig. 11(a) shows that, as expected, the line was continuously going in and out of resonance in a very random manner. In this picture the most negative pulses correspond to the greatest light intensity and result when the system is nearest to reso-

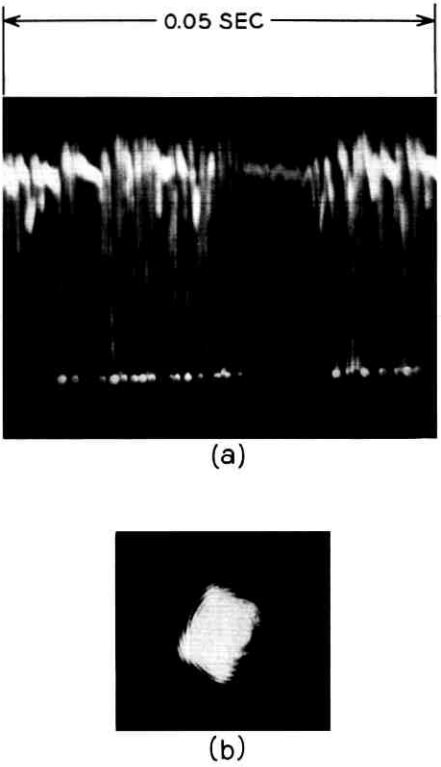


Fig. 11 — Continuous light input, line near resonance: (a) photomultiplier output, (b) beam cross section at one mirror. Mirror diameter 1.63 inches.

nance. It is seen that during the 50-millisecond period represented by the picture the system was out of resonance most of the time but went into, or through, resonance for short intervals. A very slight tapping of the maser produced a large increase in the number of pulses obtained, probably by causing the frequency to sweep back and forth through the resonant value. The peaks shown do not represent maximum buildup, for two reasons. First, the output amplifier was obviously overloading and, second, the true resonant condition probably never lasted long enough for the intensity to build up to its maximum value. Peak intensities during resonance have been measured to be as much as 100 times the intensity for the nonresonant condition. The calculated value of the $Q$ for this system is $9.9 \times 10^{10}$ based on a loss of 1 per cent per reflection. It is not surprising that adjustments are very critical.

Fig. 11(b) illustrates the effect of misalignment of the mirrors. The

photograph shows the pattern formed on the mirror at the far end of the line with light applied continuously at the near end and after an attempt had been made to adjust the system to resonance. The rectangular shape of the pattern indicates that there were higher-order modes present. More careful adjustment resulted in a pattern corresponding to only the lowest-order mode, i.e., a beam with a circular cross section, similar to those shown in Fig. 2.

The photograph of Fig. 12 shows an example of exaggerated misalignment. For this case the mirrors were deliberately tilted out of adjustment and the beam was applied off axis. As a result the beam did not double back on itself but followed a different return path, and therefore made a different spot on the mirror for each round trip. The picture

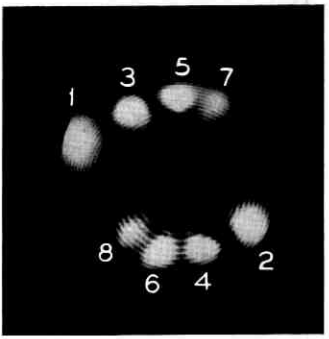


Fig. 12 — Multipath transmission: 1.25-inch diameter tilted confocal mirrors.

shows that there were seven round trips through the line before the beam was finally lost. It should be possible to obtain an adjustment which produces a pattern of the type shown but which is repetitive and continues indefinitely without the beam becoming lost.[9] It is evident that even the degree of misalignment shown on Fig. 12 does not cause excessive loss as long as the beam is confined to the surface of the mirrors, but that the greater the degree of misalignment the larger the mirrors must be to meet this requirement.*

## VIII. CONCLUSIONS

Results of the experiment described above show that, under proper conditions in an enclosed path, coherent light can be transmitted with

* Figs. 11 and 12 were obtained from the 110-foot line, which accounts for the small beam diameters. The circular lines running through the beams result from interference between reflections from the back and front surfaces of the mirrors.

low loss even when it is necessary to redirect the beam at relatively short intervals. Since a large part of the transmission loss is in the directors, in this case mirrors, the loss per mile depends upon how close together these directors must be placed. If future development produces better directors, then losses should be even lower than those measured in this experiment. Also, other devices may well prove to be superior to mirrors as beam directors.

In a practical transmission line the spacing between directors will be dictated to a considerable degree by the terrain being traversed. Another very important factor involves the difficulty of keeping the directors properly aligned in the presence of vibration and temperature changes. There is also the problem of air currents in the line.

For the experimental system the vibration problem was solved by ruggedizing all components. Insulation of the pipeline reduced air currents to the point where they produced no measurable effects. Although temperature changes produced displacements of the beam, these deviations were small enough, in the laboratory environment, to be tolerable even though no other steps were taken to provide temperature stability. The question as to whether or not these problems can be solved in a practical transmission line has not been answered by this experiment. The difficulties will be greater in the practical line and will call for more sophisticated design; however, the fact that the problems were solved with relative ease for the experimental set-up is encouraging.

The work described here represent only a fraction of the interesting and informative experiments which could be performed using this technique.

## IX. ACKNOWLEDGMENTS

The writer wishes to thank R. Kompfner for his interest and encouragement in this project. He is also indebted to D. J. Brangaccio, E. I. Gordon and A. D. White for supplying maser tubes for the transmitter. The mirrors were coated by W. L. Bond. C. P. Frazee made important contributions to the mechanical design and construction of the system. F. E. Guilfoyle assisted in alignment and operation.

## REFERENCES

1. Kompfner, R., unpublished work.
2. Goubau, G., and Christian, J. R., Some Aspects of Beam Waveguide for Long Distance Transmission at Optical Frequencies, IEEE Trans. MTT, *MTT-12*, March, 1964, p. 212.
3. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J., *40*, March, 1961, pp. 498–508.

4. Fox, A. G., and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., *40*, March, 1961, pp. 453–488.
5. White, A. D., and Rigden, J. D., Continuous Gas Maser Operation in the Visible, Proc. IRE, *50*, July, 1962, p. 1697.
6. Long, R. K. and Lewis, T. H., Water Vapor Absorption Studies With a Helium-Neon Optical Maser, Report from the Antenna Laboratory, Ohio State University, Contract AF33(616)-7081, Project 5237, Task No. 523704, Nov., 1962.
7. Taylor, J. H., and Yates, H. W., Atmospheric Transmission in the Infrared, J. Opt. Soc. Am., *47*, 1957, p. 225.
8. Burch, D. E., Howard, J. N., and Williams, D., Infrared Transmission of Synthetic Atmospheres, J. Opt. Soc. Am., *46*, 1956, p. 452.
9. Herriot, D., Kogelnik, H., and Kompfner, R., Off Axis Paths in Spherical-Mirror Interferometers, Appl. Opt., *3*, April, 1964, p. 523.

# The Structure and Properties of Binary Cyclic Alphabets

By JESSIE MACWILLIAMS

(Manuscript received July 8, 1964)

*A code which is to be used for error control on a real data system is necessarily restricted by the nature of the transmitting equipment. These restrictions have no connection with the primary function of the code; indeed, they frequently eliminate most of the codes about which anything is known at present.*

*For example, the code to be used for error control by detection and retransmission on the trunks between data switching centers is required to be a cyclic (or truncated cyclic) code with 744 information places and 20 parity check bits. The computational problem in this case is to locate those cyclic codes which have exactly 20 parity checks and a block length of 764 or greater, and pick the one which is best suited for error control over a particular channel.*

*This paper outlines a procedure for attacking such problems. It describes how to locate the cyclic codes with a fixed block length and a fixed number of parity checks, if any such exist, and gives some methods of finding the number of code words of each weight in a particular code. If one knows the statistics of the channel it is then possible to estimate the error control properties of the code.*

*The procedure depends on an analysis of the algebraic structure of cyclic codes, which is given in Section II of this paper. Section I contains step-by-step instructions with no mathematical justification. It is hoped that the theory presented in Section II may be useful in other applications.*

## INTRODUCTION

In this paper the word alphabet denotes a systematic code — one in which each code word contains a certain fixed number, $k$, of *information* places, the contents of which are arbitrary, and a fixed number, $n - k$, of *parity check* places. Each parity check digit is the sum of the contents of a particular subset of the information places. The number $n$ is called the *block length* of the alphabet. The individual members of the alphabet are called *letters*.

It is well known[1] that the letters of an alphabet of block length $n$ form a subspace of the vector space of all possible rows of $n$ binary symbols. This large space is denoted by $V^n$, $V$ being the field 0, 1. The number of information places, $k$, is also the dimension of the subspace occupied by the alphabet.[1] A cyclic alphabet has the additional property that, if it contains a letter $\alpha$, it contains as well every vector of $V^n$ which is a cyclic permutation of $\alpha$.

Cyclic alphabets are popular for error control for several good reasons. First, it is easy and relatively inexpensive to encode a cyclic alphabet. Second, the "best" known alphabets are cyclic alphabets.* Third, the cyclic property introduces a great deal of algebraic structure, which may be used to predict the error-detecting properties of the alphabet and to find alphabets with appropriate properties.

An alphabet to be used in a data transmission system must satisfy certain requirements. There will certainly be restrictions on the size of $n$ and $k$, and one naturally requires also that the alphabet should be of some use for error control. These restrictions cannot be completely arbitrary; for a given pair of integers $n$, $k$ there is likely to be no cyclic alphabet at all, let alone one with desirable error control properties.

The Hamming distance between two vectors is the number of coordinate places in which they differ. The distance between $v_1$ and $v_2$ is thus the minimum number of changes one would have to make in $v_1$ in order to convert it into $v_2$. The usual strategy for choosing an alphabet is to place its members as far apart as possible in terms of the Hamming distance. It would then require a relatively large number of errors to change a letter of the alphabet into another letter of the same alphabet.

The weight of a vector of $V^n$ is its distance from the origin, which is the same as the number of ones it contains. Let $\alpha$, of weight $s$, be a letter of an alphabet $\mathfrak{A}$. If $\beta$ is another letter of $\mathfrak{A}$, so also is $\alpha + \beta$, since $\mathfrak{A}$ is a vector space; $\alpha + \beta$ is at distance $s$ from $\beta$. Let $A(s)$ denote the number of letters of $\mathfrak{A}$ of weight $s$. $A(s)$ is then the number of letters of $\mathfrak{A}$ which are at distance $s$ from an arbitrary letter of $\mathfrak{A}$.

The set of numbers $A(0), \cdots, A(n)$ is called the spectrum of $\mathfrak{A}$. The spectrum of $\mathfrak{A}$, combined with the statistics of the channel, may be used to obtain an approximate estimate of the error control performance of the alphabet.[2]

An alphabet used for error detection will fail to detect an error pattern which is itself a letter of the alphabet. If $A(i) = 0$, the alphabet will

* The reason for this is very possibly that no other class of alphabets has been so systematically studied.

detect all $\binom{n}{i}$ patterns of $i$ errors in a block of length $n$. If $A(i) \neq 0$, the alphabet will fail to detect $A(i)$ of these $\binom{n}{i}$ patterns. It is usual to require that $A(1)$ and $A(2)$ should be zero; it can be seen from analysis of the available data[2] that this makes good sense even on the telephone network. For larger values of $i$ it would be fortunate if the letters of weight $i$ were not the same as the most common error pattern, which is usually assumed to be a "burst." This assumption leads to the vaguely formulated requirement that letters of small weight should have their nonzero digits spread out as much as possible. A cyclic alphabet satisfies this requirement to a certain extent, since the letters of smallest weight must spread over at least $n - k + 1$ adjacent places.

Since it may actually become necessary to choose particular alphabets for error control purposes, and since that the requirements which these alphabets will have to satisfy are not yet known, it is desirable to be able to obtain some rather detailed information about available alphabets. This paper describes a computer-assisted procedure by which one may locate the cyclic alphabets which have values of $n$ and $k$ within certain bounds, and find the spectra of these alphabets. A considerable library of computer programs which are useful in this procedure has been developed.

The plan of the paper is as follows:

Section I contains step-by-step instructions for locating cyclic alphabets and finding their spectra.

Section II contains the mathematical justification for the procedures of Section I, and is in fact a fairly complete account of the structure of cyclic alphabets.

It is not necessary to read Section II in order to follow the recipes given in Section I. However, in a troublesome case — which means a case that involves a large expenditure of computer time — the material in Section II may suggest a way out of the difficulty.

## I. PROPERTIES OF CYCLIC ALPHABETS

In this section we outline a procedure to attack the following problem:

Given that the block length, $n$, and the number of parity checks, $m$, of a binary cyclic alphabet are required to lie in the ranges

$$N_1 \leqq n \leqq N_2, \qquad M_1 \leqq m \leqq M_2,$$

find the alphabet (or alphabets) which have the greatest minimum distance.

It is assumed throughout that $n$ *is an odd number*. Many of the propositions quoted in this section, and justified in Section II, are not true for even values of $n$.

Let $\mathfrak{R}_n$ be the ring of polynomials mod $x^n - 1$ over the binary field. $\mathfrak{R}_n$ consists of all polynomials of degree $\leqq n - 1$ with coefficients in the binary field. Addition of polynomials is done as usual; to multiply two polynomials, we multiply in the normal way and then reduce exponents of $x$ mod $n$.

A cyclic alphabet of block length $n$ may be regarded as a set $\mathfrak{A}$ of polynomials of $\mathfrak{R}_n$, with the property that every polynomial of $\mathfrak{A}$ is divisible (mod $x^n - 1$) by a fixed polynomial $a(x)$. $a(x)$ may, and will, be taken to be a factor of $x^n - 1$; then the number of parity checks for $\mathfrak{A}$ is the degree of $a(x)$. $a(x)$ will be called the *generating factor* of $\mathfrak{A}$. We write $\mathfrak{A} = \mathfrak{R}_n \cdot a(x)$.

Let $\omega$ stand for one of the numbers $0, 1, \cdots, n - 1$. Denote by $\Sigma_2(n)$ the permutation $\omega \rightarrow 2\omega$ mod $n$. $\Sigma_2(n)$ divides the integers $0, 1, \cdots, n - 1$ into a number of disjoint cycles; the cycles of $\Sigma_2(63)$, for example, are shown in Table I.

Let $f_0(x), f_1(x), \cdots, f_{t-1}(x)$ be the irreducible factors of $x^n - 1$. Since $n$ is odd, these factors are all distinct. Let $\zeta$ be a primitive $n$th root of unity. The cycles of $\Sigma_2(n)$ and the polynomials $f_i(x)$ are associated in the following way: The zeros of $f_i(x)$ in a suitable* extension field of the binary field are $\zeta^{r_1}, \zeta^{r_2}, \cdots, \zeta^{r_k}$, where $(r_1, r_2, \cdots, r_k)$ is a cycle of $\Sigma_2(n)$; and each cycle represents in this way the zeros of one of the $f_i(x)$. The number of irreducible factors of $x^n - 1$ is, of course, the same as the number of cycles of $\Sigma_2(n)$. We say that the polynomial $f_i(x)$ with zeros $\zeta^{r_1}, \zeta^{r_2}, \cdots, \zeta^{r_k}$ is associated with the cycle $(r_1, r_2, \cdots, r_k)$.

Let $S$ be a set of cycles of $\Sigma_2(n)$; let $f_{i_1}, \cdots, f_{i_r}$ be the irreducible factors of $x^n - 1$ which are associated with the cycles of $S$. Let

$$a(x) = f_{i_1}(x) \cdot f_{i_2}(x) \cdot \cdots \cdot f_{i_r}(x)$$

be the generating factors of an alphabet $\mathfrak{A}$. We say that the cyclic alphabet $\mathfrak{A} = \mathfrak{R}_n \cdot a(x)$ is associated with the set $S$.

Let $1 < r_1 < r_2 < \cdots < n$ be a list of the factors of $n$. Attach to each cycle of $\Sigma_2(n)$ an *exponent* $e_i = n/r_i$ defined by the property that each member of the cycle is divisible by $r_i$ mod $n$, and that $r_i$ is the largest factor of $n$ for which this is true.

A great deal of information about the cyclic alphabets of block length $n$ can be obtained by looking at the cycles of $\Sigma_2(n)$.

* For example, the Galois field of order $2^t$, consisting of the roots of $y^{2^t} = y$, where $t$ is the length of the cycle of $\Sigma_2(n)$ which contains 1. A proof of this "well-known" correspondence is given in Section II.

TABLE I — CYCLES OF $\Sigma_2(63)$

| Cycles | | | | | | Exponent |
|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 8 | 16 | 32 | 63 |
| 3 | 6 | 12 | 24 | 33 | 48 | 21 |
| 5 | 10 | 17 | 20 | 34 | 40 | 63 |
| 7 | 14 | 28 | 35 | 49 | 56 | 9 |
| 9 | 18 | 36 | | | | 7 |
| 11 | 22 | 25 | 37 | 44 | 50 | 63 |
| 13 | 19 | 26 | 38 | 41 | 52 | 63 |
| 15 | 30 | 39 | 51 | 57 | 60 | 21 |
| 21 | 42 | | | | | 3 |
| 23 | 29 | 43 | 46 | 53 | 58 | 63 |
| 27 | 45 | 54 | | | | 7 |
| 31 | 47 | 55 | 59 | 61 | 62 | 63 |
| 0 | | | | | | 1 |

*Proposition I: Let* $\eta_0, \eta_1, \cdots, \eta_{t-1}$ *be the cycles of* $\Sigma_2(n)$ *and let* $m_i$ *be the length of* $\eta_i$. *The number* of cyclic alphabets of block length* $n$ *is* $2^t$. *The alphabet associated with a set* $S$ *of cycles has* $m = \sum_{\eta_i \in S} m_i$ *parity checks.*

*Proposition II: Let* $e$ *be the least common multiple of the exponents of the cycles contained in* $S$. *If* $e < n$ *the alphabet associated with* $S$ *has minimum distance 2. If* $e = n$ *the minimum distance of the alphabet is at least 3.*

*Proposition III* (Bose-Chaudhuri Bound)*: If* $S$ *contains the numbers* $1, 2, 3, \cdots, d-1$, $d$ *among its cycles, the minimum distance of the alphabet associated with* $S$ *is* $\geqq d + 1$.

It should be noted that the minimum distance may be, and often is, larger than the lower bounds given in propositions 2 and 3.

At this point one may, of course, be forced to conclude that there are no satisfactory cyclic alphabets of block length $n$. The main purpose of propositions 1 and 2 is to eliminate useless values of $n$. Suppose, however, that we have a value of $n$ for which there exist alphabets with the required number of parity checks and of minimum distance at least 3. It is then useful to establish a 1-1 correspondence between the cycles of $\Sigma_2(n)$ and the irreducible factors of $x^n - 1$.

The exponent of a polynomial $f(x)$ is the least value of $e$ for which $f(x)$ divides $x^e - 1$. We find the irreducible factors of $x^n - 1$,† and of $x^{e_i} - 1$, $(e_i = n/a_i)$ for each factor $a_i$ of $n$. Some of the irreducible fac-

* This number includes three "trivial" alphabets: the alphabet consisting of all of $\mathfrak{B}_n$, the alphabet containing only zero, and the alphabet containing only zero and the vector of weight $n$.

† This has, in fact, been done for all odd values of $n \leqq 1023$.

tors of $x^n - 1$ have exponent $e_i$ ; these appear among the irreducible factors of $x^{e_i} - 1$, and can be identified by inspection.

Any irreducible factor of $x^n - 1$ which has exponent $n$ can be chosen to correspond to the cycle of $\Sigma_2(n)$ which contains 1. Let $f_1(x)$ be this polynomial; $f_1(x)$ has $\zeta$ as a zero. If $r$ is a proper factor of $n$, $r$ is the least member of a cycle of exponent $e = n/r$. The polynomial associated with this cycle also has exponent $e$. Let $g_1 , g_2 , \cdots , g_s$ be the irreducible factors of $x^n - 1$ with exponent $e$. By picking $f_1(x)$ to correspond to the cycle containing 1, we have implicitly chosen which of the $g_i(x)$ corresponds to the cycle containing $r$. The choice can be made explicit in the following way:

*Proposition IV: $g_i(x^r)$ is exactly divisible by $f_1(x)$ if and only if it corresponds to the cycle containing $r$.**

We can now assign to each factor $r_i$ of $n$ an irreducible factor $f_i$ of $x^n - 1$, which will have exponent $e_i = n/r_i$ . We have not yet matched every cycle of $\Sigma_2(n)$ with an irreducible factor of $x^n - 1$; the remaining work will be done by a different method. Before describing this we illustrate the procedure so far.

Suppose that the restrictions on $n$, $m$ are $52 \leqq n \leqq 64$, $m = 9$. It is found that

$\Sigma_1(53)$ has two cycles, lengths 1, 52
$\Sigma_2(55)$ has five cycles, lengths 1, 4, 10, 20, 20
$\Sigma_2(57)$ has five cycles, lengths 1, 2, 18, 18, 18
$\Sigma_2(59)$ has two cycles, lengths 1, 58
$\Sigma_2(61)$ has two cycles, lengths 1, 60
$\Sigma_2(63)$ has thirteen cycles, lengths 1, 2, 3, and 6.

By Proposition I, 63 is the only possible block length, since the lengths of the cycles of the other numbers do not add up to nine. The factors of 63 are 3, 7, 9, 21. The cycles of $\Sigma_2(63)$ and their exponents are shown in Table I. The nine parity checks are obtained by taking a cycle of length 6 and a cycle of length 3 or a cycle of length 6 and the cycles of length 2 and 1. By Proposition II, the least common multiple of the exponents of the cycles should be 63; hence the cycle of length 6 should have exponent 63 or 9.† The Bose-Chaudhuri bound provides no information; a minimum distance of three is guaranteed by Proposition II, and we cannot assemble a collection of cycles containing the numbers 1, 2, 3 with only nine parity checks. Hence we are faced with the possibility of having to compute the spectra of 18 different alphabets. (It will be shown later that this is not necessary.)

* This elegant and time-saving device was suggested by Mr. R. L. Graham.
† This case will be omitted because the author did not notice it in time.

Suppose that $m$ is allowed to be 12; we pick the first and second cycles, because Proposition III then guarantees a minimum distance of at least five.*

Table II contains a list of irreducible factors of $x^{63} - 1$ and their exponents. Associate the first polynomial (714) with the first cycle. One of the polynomials of exponent 21 then corresponds to the second cycle; by Proposition IV we find that 534 is the correct choice. For the sake of completeness we use Proposition IV again to ascertain that the polynomial 640 corresponds to the cycle beginning with 9. The unique polynomials of exponent 9 and 3 must, of course, correspond to the cycles beginning with 7 and 21.

TABLE II — IRREDUCIBLE FACTORS OF $x^{63} - 1$

| | Factor | Exponent | Associated Cycle |
|---|---|---|---|
| $f_1$ | 714 | 63 | 1, 2, 4, 8, 16, 32 |
| $f_2$ | 414 | 63 | |
| $f_3$ | 700 | 3 | 21, 42 |
| $f_4$ | 554 | 63 | |
| $f_5$ | 534 | 21 | 3, 6, 12, 24, 33, 48 |
| $f_6$ | 634 | 63 | |
| $f_7$ | 444 | 9 | 7, 14, 28, 35, 49, 56 |
| $f_8$ | 664 | 63 | |
| $f_9$ | 724 | 21 | |
| $f_{10}$ | 604 | 63 | |
| $f_{11}$ | 600 | 1 | |
| $f_{12}$ | 540 | 7 | |
| $f_{13}$ | 640 | 7 | 9, 18, 36 |

The polynomials are in octal, which stands for a binary number denoting the positions of the nonzero coefficients. The least exponent is on the left, e.g.

$$714 = 111001100 = 1 + x + x^2 + x^5 + x^6.$$

To explain the next steps it is necessary to introduce some more definitions.

Let $q$ be an integer prime to $n$. The mapping $\sigma_q : x^j \to x^{qj}$ (exponents mod $n$) is an automorphism of $\mathfrak{R}_n$. The effect of $\sigma_q$ on a cyclic alphabet is to change it into an equivalent[1] cyclic alphabet; $\mathfrak{A}\sigma_q = \mathfrak{A}'$, and $\mathfrak{A}$, $\mathfrak{A}'$ have the same spectrum. The number of $\sigma_q$ which have a different effect on $\mathfrak{A}$ is rather small; if $q_1$, $q_2$ are in the same cycle of $\Sigma_2(n)$, then

$$\mathfrak{A}\sigma_{q_1} = \mathfrak{A}\sigma_{q_2}.$$

(In particular if $q$ is in the cycle which contains 1, $\mathfrak{A}\sigma_q = \mathfrak{A}$.)

* It is not established mathematically that a different choice cannot give a greater minimum distance. To be completely safe we should calculate the spectra of all alphabets with 12 parity checks and exponent 63.

We select one $q$ out of each cycle of $\Sigma_2(n)$ which contains numbers prime to $n$. For $n = 63$ we choose the numbers underlined in Table I; this choice is computationally advantageous, since $5^2 = 25 \bmod 63$, $5^3 = 62 \bmod 63$, $5^4 = 58 \bmod 63$, $5^5 = 38 \bmod 63$.

Every cyclic alphabet of $\mathfrak{A}$ contains a unique polynomial $c(x)$, the *idempotent* of $\mathfrak{A}$ which has the useful property that $\mathfrak{A}\sigma_q = \mathfrak{A}'$ if and only if $c(x)\sigma_q = c'(x)$. For computational purposes it is much better to know the idempotent of $\mathfrak{A}$ than the generating factor of $\mathfrak{A}$. The idempotent of the alphabet $\mathfrak{R}_n \cdot f_i(x)$, where $f_i(x)$ is an irreducible factor of $x^n - 1$, is denoted by $1 + \theta_i(x)$. The polynomials $\theta_i(x)$, $i = 1, 0, \cdots, t - 1$ are called the primitive idempotents of $\mathfrak{R}_n$, and have several useful properties:

(*i*) The $\theta_i(x)$ are easy to compute, and in fact have been computed for all odd values of $n \leqq 1023$. (The method by which this is done is described in the next section.)

(*ii*) The idempotent of the alphabet with generating factor $f_{i_1}(x)\, f_{i_2}(x) \cdot \cdots \cdot f_{i_r}(x)$ is

$$1 + \theta_{i_1}(x) + \theta_{i_2}(x) + \cdots + \theta_{i_r}(x).$$

(*iii*) The $\theta_i(k)$ are permuted among themselves by the automorphisms $\sigma_q$.

The alphabet with idempotent $\theta_i(x)$ is a *minimal* alphabet of $\mathfrak{R}_n$ ( t contains no subalphabet except 0). Its generating factor is $(x^n - 1)/f_i(x)$. The alphabet with generating factor $f_i(x)$ has generating idempotent $1 + \theta_i(x)$ and is a maximal alphabet of $\mathfrak{R}_n$.

In the future the cyclic alphabet $\mathfrak{A}$ will be identified by a sum of primitive idempotents of $\mathfrak{R}_n$ rather than by a product of irreducible factors of $x^n - 1$.

*Proposition V: If $f_1(x)$, $f_2(x)$ are irreducible factors of $x^n - 1$ with the same exponent, then $\theta_1(x)\sigma_q = \theta_2(x)$ for some automorphism $\sigma_q$ of $R_n$. Hence the minimal alphabets generated by $\theta_1(x)$, $\theta_2(x)$ are equivalent, and the maximal alphabets generated by $1 + \theta_1(x)$, $1 + \theta_2(x)$ are also equivalent. Conversely, if two minimal (maximal) alphabets have the same spectrum, they are equivalent under one of the automorphisms $\sigma_q$.*

*Proposition VI: The alphabet with idempotent $(1 + \theta_{i_1} + \cdots + \theta_{i_r})$ is equivalent to the alphabet with idempotent $(1 + \theta_{i_1}\sigma_q + \cdots + \theta_{i_r}\sigma_q)$.*

*Proposition VII: Let $1 + \theta_i(x)$ be the idempotent associated with the cycle of $\Sigma_2(n)$ which contains 1. Let $u$, $v$ be integers prime to $n$ such that $u \cdot v \equiv 1 \bmod n$. Then $1 + \theta_i(x)\sigma_u$ is the idempotent associated with the cycle which contains $v$.*

We illustrate again for the case $n = 63$. Table III contains a list of primitive idempotents of $\mathfrak{R}_{63}$. This list is parallel to the list in Table II.

TABLE III — PRIMITIVE IDEMPOTENTS OF $\mathfrak{R}_{63}$

| | | | Associated Cycle | |
|---|---|---|---|---|
| | | | From Table II | From Prop. 6 |
| $\theta_1$ | 321026251170 | 156307227 | 1, 2, 4, 8, 16, 32 | |
| $\theta_2$ | 010305172162 | 267315277 | | 11, 22, 25, 37, 44, 50 |
| $\theta_3$ | 333333333333 | 333333333 | 21, 42 | |
| $\theta_4$ | 044160277124 | 317353233 | | 5, 10, 17, 20, 34, 40 |
| $\theta_5$ | 012231301223 | 130122313 | 3, 6, 12, 24, 33, 48 | |
| $\theta_6$ | 375343166036 | 225150213 | | 31, 47, 55, 56, 61, 62 |
| $\theta_7$ | 044044044044 | 044044044 | 7, 14, 28, 35, 49, 56 | |
| $\theta_8$ | 331327363052 | 375016044 | | 22, 29, 43, 46, 53, 58 |
| $\theta_9$ | 323112032311 | 203231120 | | 15, 30, 39, 51, 57, 60 |
| $\theta_{10}$ | 375263355116 | 136243020 | | 13, 19, 26, 38, 41, 52 |
| $\theta_{11}$ | 777777777777 | 777777777 | 0 | |
| $\theta_{12}$ | 456271345627 | 134562713 | | 27, 45, 54 |
| $\theta_{13}$ | 723516472351 | 647235164 | 9, 18, 36 | |

The $i$th factor, $f_i(x)$ of Table II, is the generating factor of the alphabet with idempotent $1 + \theta_i(x)$ where $\theta_i(x)$ is the $i$th primitive idempotent of Table III. We associate some of the $\theta_i$ with a cycle of $\Sigma_2(63)$ by copying from Table II.

The automorphism $\sigma_5$ produces the following permutation of the set of primitive idempotents of $\mathfrak{R}_{63}$

$$(\theta_1 , \theta_{10} , \theta_8 , \theta_6 , \theta_2 , \theta_4) (\theta_5 , \theta_9) (\theta_{12} , \theta_{13} )(\theta_3) (\theta_7) (\theta_{11}).$$

The other automorphisms, as already noted, produce powers of this permutation; for example $\sigma_{62} = \sigma_5^3$ gives

$$(\theta_1 , \theta_6) (\theta_{10} , \theta_2) (\theta_8 , \theta_4) (\theta_5 , \theta_9) (\theta_{12} , \theta_{13}) (\theta_3) (\theta_7) (\theta_{11}).$$

Consider now the alphabet with nine parity checks which is associated with the cycles (1, 2, 4, 8, 16, 32) and (9, 18, 36). By Table II the generating factor of this alphabet is $f_1(x) \cdot f_{13}(x)$; its idempotent is $(1 + \theta_1 + \theta_{13})$. The idempotents which can be obtained from this by the permutations $\sigma_5$ and its powers are

$$1 + \theta_{10} + \theta_{12} , \ 1 + \theta_8 + \theta_{13} , \ 1 + \theta_6 + \theta_{12} , \ 1 + \theta_2 + \theta_{13} , \ 1 + \theta_4 + \theta_{12} .$$

The generating factors of the corresponding alphabets are (including the original alphabet),

$$f_1 \cdot f_{13} , \ f_{10} \cdot f_{12} , \ f_8 \cdot f_{13} , \ f_6 \cdot f_{12} , \ f_2 \cdot f_{13} , \ f_4 \cdot f_{12} .$$

By Proposition VI these six alphabets are all equivalent. Similarly, the alphabet associated with cycles (1, 2, 4, 8, 16, 32) and (27, 45, 54) has

idempotent $1 + \theta_1 + \theta_{12}$, and is equivalent to the alphabets with idempotents

$$1 + \theta_{10} + \theta_{13}, \quad 1 + \theta_8 + \theta_{12}, \quad 1 + \theta_6 + \theta_{13}, \quad 1 + \theta_2 + \theta_{12}, \quad 1 + \theta_4 + \theta_{13}.$$

The third possibility for nine parity checks consists of the cycles (1, 2, 4, 16, 32), (21, 42), (0). The associated idempotent is $1 + \theta_1 + \theta_3 + \theta_{11}$; equivalent alphabets are given by the idempotents

$$1 + \theta_{10} + \theta_3 + \theta_{11}, \quad 1 + \theta_8 + \theta_3 + \theta_{11}, \quad 1 + \theta_6 + \theta_3 + \theta_{11},$$

$$1 + \theta_2 + \theta_3 + \theta_{11}, \quad 1 + \theta_4 + \theta_3 + \theta_{11}.$$

Hence, among the 18 alphabets with nine parity checks and minimum distance $\geqq 3$, there are actually at most three different spectra.

We observed before that the alphabet with twelve parity checks associated with cycles (1, 2, 4, 8, 16, 32) and (3, 6, 12, 24, 33, 34) has minimum distance at least 5. The idempotent of this alphabet is $1 + \theta_1 + \theta_5$. There are at least* five equivalent alphabets with idempotents

$$1 + \theta_{10} + \theta_9, \quad 1 + \theta_8 + \theta_5, \quad 1 + \theta_6 + \theta_9, \quad 1 + \theta_2 + \theta_5, \quad 1 + \theta_4 + \theta_9.$$

It may very well happen that one of these alphabets is easier to instrument than our original choice.

The 1-1 correspondence between cycles of $\Sigma_2(63)$ and primitive idempotents of $\mathfrak{R}_{63}$ is completed by Proposition VII, and entered in Table IV. For example, $5 \cdot 38 = 190 = 1 \bmod 63$ ($38 = 5^5 \bmod 63$); hence

$$\theta_1 \sigma_5 = \theta_{10}$$

corresponds to the cycle (13, 19, 26, 38, 41, 52).

It is now necessary to face the problem of actually computing the spectrum of a cyclic alphabet.

For a small alphabet this can be done by counting, without too large an expenditure of computer time. An alphabet of block length 765 with $2^{20}$ letters can be examined, a letter at a time, in 0.32 hours on a 7094. This alphabet has 745 parity checks. Typically, however, one wishes to know the spectrum of the alphabet with $2^{745}$ letters and 20 parity checks; to compute this by counting would take over a million computer years. Fortunately there is a way out of this dilemma.

Let $a(x)$, of degree $m$, be a factor of $x^n - 1$, and let

$$b(x) = (x^n - 1)/a(x).$$

* It is entirely possible that alphabets not contained in this list also have the same spectrum, and are perhaps equivalent to the first alphabet under a permutation which is not an automorphism of $\mathfrak{R}_n$.

TABLE IV — SPECTRA OF SMALL ALPHABETS OF $R_{63}$

| $\theta_1 + \theta_{12}$ $2^9$ letters |
|---|
| $B(0) = 1$ |
| $B(28) = 189$ |
| $B(32) = 252$ |
| $B(36) = 7$ |
| $B(40) = 63$ |
| $\theta_1 + \theta_{13}$ $2^9$ letters |
| $B(0) = 1$ |
| $B(28) = 252$ |
| $B(32) = 63$ |
| $B(36) = 196$ |
| $\theta_1 + \theta_3 + \theta_{11}$ $2^9$ letters |
| $B(0) = 1$ |
| $B(25) = 3$ |
| $B(26) = 63$ |
| $B(29) = 126$ |
| $B(31) = 63$ |
| $B(32) = 63$ |
| $B(34) = 126$ |
| $B(37) = 63$ |
| $B(42) = 3$ |
| $B(63) = 1$ |
| $\theta_1 + \theta_5$ $2^{12}$ letters |
| $B(0) = 1$ |
| $B(24) = 210$ |
| $B(28) = 1512$ |
| $B(32) = 1071$ |
| $B(36) = 1176$ |
| $B(40) = 126$ |

The alphabets $\mathfrak{A} = \mathfrak{R}_n \cdot a(x)$ and $\mathfrak{B} = \mathfrak{R}_n \cdot b(x)$ are called dual or orthogonal alphabets.* Let $A(s)$, $B(s)$ be the number of letters of weight $s$ in $\mathfrak{A}$, $\mathfrak{B}$. We suppose that $\mathfrak{A}$ with $m$ parity checks is a large alphabet, whose spectrum we wish to find; $\mathfrak{B}$ contains $2^m$ letters, and its spectrum can be found by counting or by more sophisticated procedures. The $A(s)$ can be found from the $B(s)$ by the following proposition.[3]

*Proposition VIII: The quantities $A(s)$, $B(s)$ are related by the expression*

$$2^m \sum_{s=0}^{n} A(s) z^s = \sum_{s=0}^{n} B(s) \, (1 + z)^{n-s} (1 - z)^s.$$

* This is not quite the usual definition; the usual dual alphabet of $\mathfrak{A}$ is equivalent to $\mathfrak{B}$, so has the same spectrum. The difference is explained fully in Section II.

We now describe methods which are sometimes useful for finding the spectra of small cyclic alphabets.

Let $\alpha$ be a letter of a cyclic alphabet $\mathfrak{A}$, and let $\alpha T$ be the letter obtained from $\alpha$ by one cyclic permutation to the right. For example, for $n = 7$ one might have

$$\alpha = (0\ 1\ 1\ 0\ 1\ 1\ 1), \qquad \alpha T = (1\ 0\ 1\ 1\ 0\ 1\ 1),$$
$$\alpha T^2 = (1\ 1\ 0\ 1\ 1\ 0\ 1) \text{ etc.}$$

The letters $\alpha T^\nu$ all belong to $\mathfrak{A}$. The set of distinct letters $\alpha T^\nu$ for fixed $\alpha$ is called a cycle of $\mathfrak{A}$; $\alpha$ is a representative of this cycle; the number of distinct letters is $\pi(\alpha)$, the period of $\alpha$ or the length of the cycle. Knowing the length of each cycle of $\mathfrak{A}$ and the weight of a letter from each cycle, we can at once compute the spectrum of $\mathfrak{A}$.

If $\mathfrak{A}$, $\mathfrak{B}$ are dual alphabets, and the idempotent of $\mathfrak{A}$ is $1 + c(x)$, then $c(x)$ is the idempotent of $\mathfrak{B}$. The alphabet $\mathfrak{M}_i$ with idempotent $\theta_i(x)$ is the dual of the maximal alphabet with irreducible generating factor $f_i(x)$. $\mathfrak{M}_i$ is called a minimal alphabet. The alphabet with generating factor $f_i(x)f_j(x)$ has idempotent $1 + \theta_i(x) + \theta_j(x)$; its dual alphabet is the union of $\mathfrak{M}_i$ and $\mathfrak{M}_j$ and has idempotent $\theta_i(x) + \theta_j(x)$. The procedure is to find cycle representatives for the $\mathfrak{M}_i$ and then put them together to get cycle representatives for $\mathfrak{M}_i \cup \mathfrak{M}_j$. This is done by the following propositions.

*Proposition IX: Every cycle (except that containing the zero letter) of $\mathfrak{M}_i$ has length $\pi(\theta_i)$; further $\pi(\theta_i)$ is the exponent, $e_i$, of the irreducible polynomial $f_i(x)$.*

For example, for $n = 63$, $\mathfrak{M}_1$ has one cycle of length 63. This cycle contains the letter corresponding to $\theta_1(x)$, which has weight 32; the spectrum of $\mathfrak{M}_1$ is $B(0) = 1$, $B(32) = 63$. The spectrum of the maximal alphabet $\mathfrak{R}_{63} \cdot f_1(x)$ is given by

$$2^6 \sum_{s=0}^{63} A(s)z^s = (1 + z)^{63} + 63(1 + z)^{31}(1 - z)^{32}.$$

Similarly $\mathfrak{M}_{12}$ has one cycle of length 7 which contains the letter corresponding to $\theta_{12}(x)$, of weight 36. The spectrum of $\mathfrak{M}_{12}$ is $B(0) = 1$, $B(36) = 7$, and the dual alphabet $\mathfrak{R}_{63} \cdot f_{12}(x)$ has the spectrum $A(s)$ given by

$$2^3 \sum_{s=0}^{63} A(s)z^s = (1 + z)^{63} + 7(1 + z)^{27}(1 - z)^{36}.$$

We note that

$$8A(2) = \binom{63}{2} + 7\left[\binom{27}{2} - 27\cdot 36 + \binom{36}{2}\right] = 2016,$$

agreeing with the statement of Proposition II that this particular alphabet will contain letters of weight 2.

The alphabet $\mathfrak{M}_5$ contains three cycles of length 21. It is possible to check by hand that $\theta_5$, $\theta_5 + \theta_5 T$ and $\theta_5 + \theta_5 T^2$ are in different cycles; their weights are 24, 36 and 36, respectively; the spectrum of $\mathfrak{M}_5$ is $B(0) = 1, B(24) = 21, B(36) = 42$.

The technique is useful only if $\mathfrak{M}_i$ contains a rather small number of different cycles; otherwise the process of finding cycle representatives becomes extremely laborious.

Once the cycle representatives for $\mathfrak{M}_i$ and $\mathfrak{M}_j$ are known, one constructs cycle representatives for the alphabet $\mathfrak{M}_i \cup \mathfrak{M}_j$ (with idempotent $\theta_i + \theta_j$) by the following proposition.

*Proposition X: Let $\mathfrak{M}_i$ have cycle representatives $m_1, m_2, \cdots, m_\alpha$, of period $e_i$, and $\mathfrak{M}_j$ have cycle representatives $n_1, n_2, \cdots, n_\beta$, of period $e_j$. Let $H$, $h$ be the least common multiple and highest common factor of $e_i, e_j$. Then $\mathfrak{M}_i \cup \mathfrak{M}_j$ has cycle representatives $m_1, \cdots, m_\alpha, n_1, \cdots, n_\beta$, and in addition, for each pair $i, j$, cycle representatives $m_i + n_j T^\nu$, $\nu = 0, 1, \cdots, h - 1$ of period $H$.*

For example, for $n = 63$ the alphabet $\mathfrak{M}_1 \cup \mathfrak{M}_{12}$ has one cycle representative $\theta_1$ (period 63), one cycle representative $\theta_{12}$ (period 7), and 7 cycle representatives $\theta_1 + \theta_{12} T^\nu$, $\nu = 0, 1, \cdots, 6$ of period 63. The alphabet $\mathfrak{M}_1 \cup \mathfrak{M}_{13}$ is constructed similarly. The spectrum of the alphabet

$$\mathfrak{M}_1 \cup \mathfrak{M}_3 \cup \mathfrak{M}_{11}$$

is obtained by constructing that of $\mathfrak{M}_1 \cup \mathfrak{M}_3$ (cycle representatives $\theta_1$, $\theta_3$, $\theta_1 + \theta_3$, $\theta_1 + \theta_3 T$) and adding the letter of weight 63 represented by $\theta_{11}$. The spectra of these three alphabets and their duals are given in Tables IV and V. The dual alphabets are the three nonequivalent alphabets of block length 63 and with nine parity checks which we set out to find.

The alphabet $\mathfrak{M}_1 \cup \mathfrak{M}_5$ has cycle representatives $\theta_1$, period 63, $\theta_5$, $\theta_5 + \theta_5 T$, $\theta_5 + \theta_5 T^2$, period 21, and $\theta_1 + \theta_5 T^\nu$, $\theta_1 + (\theta_5 + \theta_5 T) T^\nu$, $\theta_1 + (\theta_5 + \theta_5 T^2) T^\nu$, $\nu = 0, 1, \cdots, 20$. The spectra of this alphabet and its dual are given in Tables IV and V; the dual alphabet has minimum distance 5 as predicted.

We now give a summary of the procedure:

(1) Obtain a list of the cycles of $\Sigma_2(n)$ for each allowable value of $n$, and check to see whether an allowable number of parity checks can be

TABLE V — SPECTRAL PROBABILITIES* OF LARGE ALPHABETS OF $R_{63}$

| $1 + \theta_1 + \theta_{12}$ (9 parity checks) |
|---|
| $a(0) = 1 = a(63)$<br>$a(1) = a(2) = a(3) = 0;$<br>$a(4) = 0.21153 \times 10^{-2},$<br>$a(5) = 0.20973 \times 10^{-2},$<br>$a(6) = 0.19243 \times 10^{-2},$<br>$a(7) = 0.19571 \times 10^{-2},$<br>$a(8) = 0.19526 \times 10^{-2},$<br>$a(s) = a(n - s)$<br>$a(s) = 2^{-9}$ for other values of $s$. |
| $1 + \theta_1 + \theta_{13}$ (9 parity checks) |
| $a(0) = 1 = a(63)$<br>$a(1) = a(2) = 0$<br>$a(3) = a(4) = 0.15865 \times 10^{-2},$<br>$a(5) = a(6) = 0.20077 \times 10^{-2},$<br>$a(7) = a(8) = 0.19451 \times 10^{-2},$<br>$a(9) = a(10) = 0.19544 \times 10^{-2},$<br>$a(11) = a(12) = 0.19528 \times 10^{-2},$<br>$a(s) = a(n - s)$<br>$a(s) = 2^{-9}$ for other values of $s$. |
| $1 + \theta_1 + \theta_3 + \theta_{11}$ (9 parity checks) |
| $a(0) = 1$<br>$a(1) = a(2) = a(3) = 0,$<br>$a(4) = 0.19634 \times 10^{-2}$<br>$a(6) = 0.19626 \times 10^{-2}$<br>$a(8) = 0.19502 \times 10^{-2}$<br>$a(i) = 0$ all odd values of $i$<br>$a(2i) = 2^{-9}$ other values of $i$. |
| $1 + \theta_1 + \theta_5$ (12 parity checks) |
| $a(0) = 1 \quad a(63) = 1$<br>$a(1) = a(2) = a(3) = a(4) = 0,$<br>$a(5) = a(6) = 0.26889 \times 10^{-3},$<br>$a(7) = a(8) = 0.24119 \times 10^{-3},$<br>$a(9) = a(10) = 0.24461 \times 10^{-3},$<br>$a(11) = a(12) = 0.24404 \times 10^{-3},$<br>$a(13) = a(14) = 0.24416 \times 10^{-3},$<br>$a(s) = a(n - s)$<br>$a(s) = 2^{-12}$ other values of $s$. |

* The spectral probability $a(s)$ is $A(s)/\binom{n}{s}$; the number $A(s)$ is frequently too large for the computer.

obtained as the sum of lengths of distinct cycles. Discard the values of $n$ for which this is not possible.

(2) Attach an exponent to each cycle of $\Sigma_2(n)$.

Let $S$ be a set of cycles of suitable lengths. Find the least common multiple of the exponents of the cycles in $S$. Discard the sets $S$ for which this number is less than $n$.

(3) It is now necessary to set up a correspondence between the cycles of $S$ and the primitive idempotents of $\mathfrak{R}_n$. This is done in two steps, as follows. Obtain a list of the irreducible factors of $x^n - 1$, and of $x^{e_i} - 1$, $e_i = n/r_i$ for each proper factor $r_i$ of $n$. Let $f_0(x), f_1(x), \cdots, f_{t-1}(x)$ be the irreducible factors of $x^n - 1$. The irreducible factors of $x^{e_i} - 1$ will be among the $f_i(x)$. Starting with the smallest value of $e_i$, attach an exponent to each of the $f_i(x)$ by comparing lists. Pick any $f_i(x)$ of exponent $n$ to correspond to the cycle beginning with 1, and using Proposition IV, find the polynomial of exponent $e_i$, which then corresponds to the cycle beginning with $r_i$.

(4) Obtain a parallel list of primitive idempotents of $\mathfrak{R}_n$, and transfer to this the cycles whose position in the correspondence has been found. Pick an integer $q$ from each cycle of $\Sigma_2(n)$ which contains numbers prime to $n$, and find the effect of the permutation $\sigma_q$ on the set of primitive idempotents. Use Proposition VII to complete the correspondence between cycles and primitive idempotents.

(5) Let $s_1, s_2, \cdots, s_\nu$ be the cycles of an allowable set $S$, $f_1(x), \cdots, f_\nu(x)$ the corresponding irreducible factors of $x^n - 1$, and $\theta_1(x), \cdots, \theta_\nu(x)$, the corresponding primitive idempotents. The desired alphabet has the generating factor $f(x) = f_1(x)f_2(x) \cdots f_\nu(x)$. The orthogonal alphabet has the generating idempotent

$$\theta(x) = \theta_1(x) + \theta_2(x) + \cdots + \theta_\nu(x).$$

Divide the allowable alphabets into automorphism classes by looking at the effect of the automorphisms $\sigma_q$ on the idempotent $\theta(x)$. Alphabets in the same automorphism class have the same spectrum.

(6) The orthogonal alphabet $\mathfrak{R}_n \cdot \theta(x)$ is frequently much smaller than the desired alphabet $\mathfrak{R}_n \cdot f(x)$. In this case it is advantageous to compute the spectrum of $\mathfrak{R}_n \cdot \theta(x)$ and to obtain the spectrum of $\mathfrak{R}_n \cdot f(x)$ from this by Proposition VIII. If $\theta(x)$ is the sum of two or three primitive idempotents, its spectrum may be built up in the way described in Proposition X. Otherwise the alphabet may be generated by the vectors corresponding to polynomials $\theta(x), x\theta(x), \cdots, x^m\theta(x)$ [$m$ = degree of $f(x)$], and the spectrum obtained by counting.

## II. PROOFS

In this section we give the proofs of the propositions of the first section.

Let $V$ be the binary field, and $V^n$ the set of all possible rows of $n$ binary symbols. $V^n$ is a vector space of dimension $n$ over $V$. Let $\mathfrak{R}_n$ be, as before, the set of polynomials mod $x^n - 1$ over $V$. $\mathfrak{R}_n$ is a commutative ring.

We may relate $V^n$ and $\mathfrak{R}_n$ by (1-1) mapping

$$\alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1}x^{n-1} \leftrightarrows \alpha_0 , \alpha_1 , \cdots , \alpha_{n-1} .$$

This mapping clearly preserves addition in both $\mathfrak{R}_n$ and $V^n$.

A subset $\mathfrak{A}$ of polynomials of $\mathfrak{R}_n$ is an ideal if

(*i*) $g_1 , g_2 \in \mathfrak{A} \Rightarrow g_1 + g_2 \in \mathfrak{A}$,

(*ii*) $g \in \mathfrak{A} \Rightarrow rg \in \mathfrak{A}$ for any $r \in \mathfrak{R}_n$ .

An ideal in $\mathfrak{R}_n$ corresponds by property (*i*) to a linear subspace of $V^n$. By property (*ii*), with $r = x$, this subspace in invariant under a cyclic permutation of coordinates, hence is a cyclic alphabet in $V^n$. Conversely a cyclic alphabet in $V^n$ is an ideal in $\mathfrak{R}_n$ . We represent both ideal and alphabet by the same symbol, $\mathfrak{A}$.

*Lemma 2.0: An ideal $\mathfrak{A}$ of $\mathfrak{R}_n$ consists of all multiples (in $\mathfrak{R}_n$) of a polynomial $a(x)$ which divides $x^n - 1$.* $a(x)$ is the unique polynomial of least degree in $\mathfrak{A}$.*

The proof of this lemma can be found in Peterson,[4] section 6.4.

$a(x)$ will be called the *generating factor* of $\mathfrak{A}$. The polynomial $b(x) = (x^n - 1)/a(x)$ will be called the *reciprocal factor* of $\mathfrak{A}$. This notation is used throughout; the ideal named $\mathfrak{A}$ always has a generating factor named $a(x)$ and a reciprocal factor named $b(x)$. The degree of $a(x)$ will be denoted by $m$, and that of $b(x)$ by $k$; of course $m + k = n$.

*Lemma 2.1: The dimension of $\mathfrak{A}$ as a vector space of $V^n$ is k; the number of parity checks for the alphabet $\mathfrak{A}$ is m.* For proof, see Peterson,[4] theorem 6.11.

The number of different alphabets of $\mathfrak{R}_n$ is the number of different factors of $x^n - 1$; the dimension of alphabet $\mathfrak{A}$ is the degree of its reciprocal factor. However, if $n$ is odd, (which we always assume) one can find which dimensions are available in block length $n$ without going to the considerable trouble of finding all the factors of $x^n - 1$.

Let $\omega$ stand for one of the numbers $0, 1, \cdots, n - 1$. Let $\Sigma_2(n)$ denote the mapping $\omega \rightarrow 2\omega \bmod n$. Since $n$ is odd, this mapping is a permutation of the numbers $0, 1, \cdots, n - 1$.

The permutation $\Sigma_2(n)$ on $0, 1, \cdots, n - 1$ factors into a number of cycles; the cycles of $\Sigma_2(63)$ are shown in Table I, Section I. It is a fairly trivial matter to find these cycles.

The relation between the cycles of $\Sigma_2(n)$ and the factors of $x^n - 1$ over $V$ is a well-known part of Galois theory. It is described in detail here only because of the difficulty of finding a concise reference.

* $a(x)$ divides $x^n - 1$ in the ring $V[x]$ of all polynomials over $V$. It is meaningless to say that something divides $x^n - 1$ in $\mathfrak{R}_n$ .

*Lemma 2.2: Let $S$ be a subset of the integers $0, 1, \cdots, n-1$. $S$ is invariant under $\Sigma_2(n)$ if and only if it is the union of a number of cycles of $\Sigma_2(n)$.*

*Proof:* If $S$ is such a union it is invariant under $\Sigma_2(n)$, since each separate cycle is invariant.

Suppose $S$ to be invariant under $\Sigma_2(n)$ and let $r$ belong to $S$. Then $2^\nu r$ also belongs to $S$ for any value of $\nu$. $S$ contains with $r$ the whole cycle containing $r$. Thus $S$ is a union of cycles of $\Sigma_2(n)$.

*Lemma 2.3: Let $S$ be invariant under $\Sigma_2(n)$, and let $S_t$ be the set of all sums $r_{s_1} + r_{s_2} + \cdots + r_{s_\tau}$, $r_{s_i} \in S$, $r_{s_i} \neq r_{s_j}$. Then $S_\tau$ is invariant under $\Sigma_2(n)$.*

*Proof:* We need show only that $\Sigma_2(n)$ maps $S_\tau$ into itself; the mapping must then be 1-1. Let $r_{s_1} + r_{s_2} + \cdots + r_{s_\tau} \in S_\tau$; applying $\Sigma_2(n)$ we obtain $2r_{s_1} + 2r_{s_2} + \cdots + 2r_{s_\tau}$, which is again in $S_\tau$. Hence the lemma is proved.

Let $(1,2,2^2, \cdots, 2^{m_1-1})$ be the cycle of $\Sigma_2(n)$ which contains 1. $2^{m_1} \equiv 1 \bmod n$, or $n$ divides $2^{m_1} - 1$. Set $N = 2^{m_1} - 1$. Every $n$th root of unity is also an $N$th root of unity. Let $V(2^{m_1})$ be the Galois field of the $N$th roots of unity over the prime field $V$. $x^n - 1$ factors into linear factors over $V(2^{m_1})$ and these factors are of the form $x - \zeta^\nu$, where $\zeta$ is a primitive $n$th root of unity. ($\zeta$ is not a primitive $N$th root of unity unless $n = N$.)

The automorphisms of $V(2^{m_1})$ over $V$ are given by $\alpha \to \alpha^2$ and its powers, where $\alpha \in V(2^{m_1})$; further, $\alpha = \alpha^2$ if and only if $\alpha \in V$.[5]

The explicit connection between the cycles of $\Sigma_2(n)$ and the factors of $x^n - 1$ is as follows:

*Lemma 2.4:* Let $S = r_1, r_2, \cdots, r_m$ be a set of integers invariant under $\Sigma_2(n)$, with $r_i \neq r_j$. The polynomial $f(x) = \pi(x - \zeta^{r_i})$ has coefficients in $V$, and is a factor of $x^n - 1$ over $V$.*

*Let $f(x) = (x - \zeta^{r_1}) \cdots (x - \zeta^{r_m})$ be the factorization over $V(2^{m_1})$ of a polynomial $f(x)$ which divides $x^n - 1$ over $V$. The set $r_1, r_2, \cdots, r_m$ is then invariant under $\Sigma_2(n)$.*

*Proof:* Let $S = r_1, r_2, \cdots, r_m$ be a set of distinct integers which is invariant under $\Sigma_2(n)$. $f(x) = (x - \zeta^{r_1}) \cdots (x - \zeta^{r_m})$ divides $x^n - 1$ over $V(2^{m_1})$ since each linear factor divides $x^n - 1$, and $r_i \neq r_j$. Let $a_{n-\tau}$ be the coefficient of $x^{n-\tau}$ in $f(x)$. $a_{n-\tau}$ is the $\tau$th symmetric function of $\zeta^{r_1}, \cdots, \zeta^{r_m}$, or

$$a_{n-\tau} = \sum_{\substack{s_i \in S \\ s_i \neq s_j}} \zeta^{r_{s_1} + \cdots + r_{s_\tau}}.$$

* Note again that we are working in $V[x]$, not in $\mathfrak{R}_n$.

$$(a_{n-r})^2 = \sum_{\substack{s_i \in S \\ s_i \neq s_j}} \zeta^{2r s_1 + \cdots + 2r s_r} = a_{n-r} \text{ by 2.3.}$$

Thus the coefficients of $f(x)$ are in $V$, and $f(x)$ divides $x^n - 1$ over $V$.

Suppose that $f(x)$ divides $x^n - 1$ over $V$. The zeros of $f(x)$ in $V(2^{m_1})$ are $\zeta^{r_1}, \zeta^{r_2}, \cdots, \zeta^{r_m}$ where $\zeta$ is a primitive $n$th root of unity, and $r_1, \cdots, r_m$ are integers mod $n$. Since by Lemma 2.3 all the symmetric functions of $\zeta^{r_1}, \zeta^{r_2}, \cdots, \zeta^{r_m}$ are in $V$, the transformation $\zeta \rightarrow \zeta^2$ preserves $f(x)$, and must be simply a permutation of the zeros of $f(x)$. Thus the set $r_1, r_2, \cdots, r_m$ is invariant under $\Sigma_2(n)$.

The smallest sets which are invariant under $\Sigma_2(n)$ are the individual cycles of $\Sigma_2(n)$. Each such cycle determines, in the way described above, the zeros of an irreducible factor of $x^n - 1$; each irreducible factor of $x^n - 1$ corresponds in this way to a cycle of $\Sigma_2(n)$.

*Proof of Proposition I*

The number of cycles of $\Sigma_2(n)$ is $t$, and, by the above, $t$ is also the number of irreducible factors of $x^n - 1$. These irreducible factors are all different [$(x^n - 1)$ has no multiple roots over $V$ if $n$ is odd], and can be combined by multiplication to give $2^t$ different factors of $x^n - 1$. Further, these are all the factors of $x^n - 1$. Hence there are $2^t$ cyclic alphabets of block length $n$. Let $a(x) = f_1(x) \cdots f_\nu(x)$ be the generating factor of the cyclic alphabet $\mathfrak{A}$. Let $m_i$ be the degree of $f_i(x)$. $m_i$ is the length of the cycle of $\Sigma_2(n)$ corresponding to $f_i(x)$. By Lemma 2.1 the number of parity checks for $\mathfrak{A}$ is $m = \sum_{i=1}^{\nu} m_i$.

The exponent of a polynomial $a(x)$ is the least integer $e$ such that $a(x)$ divides $x^e - 1$. Let

$$a(x) = (x - \zeta^{r_1}) \cdots (x - \zeta^{r_m}),$$

where $\zeta$ is a primitive $n$th root of unity, and $r_1, \cdots, r_m$ is a set of cycles of $\Sigma_2(n)$. The exponent of $a(x)$ is then the least value of $e$ such that

$$(\zeta^{r_i})^e = 1, \quad \text{or} \quad e r_i \equiv 1 \bmod n, \qquad i = 1, \cdots, m.$$

$e = n/\alpha$ where $\alpha$ is the greatest common factor of $r_1, \cdots, r_m$ and $n$.

If $a(x)$ is an irreducible factor of $x^n - 1$ [$r_1, \cdots, r_m$ is a single cycle of $\Sigma_2(n)$], the quantity $\alpha$ is the largest factor of $n$ which divides each member of the cycle $r_1, \cdots, r_m$. $e = n/\alpha$ is said to be the exponent of the cycle as well as of the polynomial $a(x)$.

The exponent of a union of cycles or of a product of irreducible polynomials is the least common multiple of their individual exponents.

*Proof of Proposition II*

The ideal $\mathfrak{A}$ with generating factor $a(x)$ contains the polynomial $x^e - 1$ $(= x^e + 1)$, where $e$ is the exponent of $a(x)$. If $e = n$, this polynomial is the zero of $\mathfrak{A}$; if $e < n$, it corresponds to a letter of weight 2 in the alphabet $\mathfrak{A}$.

If $\mathfrak{A}$ contains a letter of weight 2, the ideal $\mathfrak{A}$ contains, by suitable cyclic permutation, a polynomial $x^e - 1$, $e < n$, which is divisible by $a(x)$; the exponent of $a(x)$ is then less than $n$.

Thus $\mathfrak{A}$ contains letters of weight 2 if and only if its generating factor has exponent less than $n$.

Proposition III is a restatement of the Bose-Chaudhuri theorem; a proof can be found in Peterson,[4] Theorem 9.1.

There is considerable freedom of choice in setting up an exact correspondence between cycles of $\Sigma_2(n)$ and irreducible factors of $x^n - 1$. This occurs because there are several primitive $n$th roots of unity; if $\zeta$ is one such, then so also is $\zeta^\nu$, where $\nu$ is any integer prime to $n$.

We pick any irreducible polynomial of exponent $n$ to correspond to the cycle $(1, 2, \cdots, 2^{m-1})$. If this is to make sense, the alphabets generated by irreducible polynomials with the same exponent should be indistinguishable for our purposes. In fact they are equivalent;[1] this will be proved later.

The choice of a polynomial to correspond to the cycle $(1, 2, \cdots, 2^{m-1})$ implicitly fixes the exact correspondence between cycles of $\Sigma_2(n)$ and irreducible factors of $x^n - 1$. It remains to make this correspondence explicit, preferably by calculations involving only numbers in the prime field $V$. This is done in two stages, the first of which is given by Proposition IV.

*Proof of Proposition IV*

Let $f_1(x)$ be the polynomial chosen to correspond to the cycle $(1, 2, \cdots, 2^{m-1})$. Over the field $V(2^m)$ $f_1(x)$ factors into $(x - \zeta)(x - \zeta^2) \cdots (x - \zeta^{2^{m-1}})$. Let $r$ be a factor of $n$ and $\{g_i(x)\}$ the set of irreducible factors of $x^n - 1$ of exponent $e = n/r$. One of the $g_i(x)$ has $\zeta^r$ as a zero over $V(2^m)$, and corresponds to the cycle containing $r$. This $g_i(x)$ can be identified by the following lemma.

*Lemma 2.5: $g_i(x^r)$ is divisible by $f_1(x)$ over $V$ if and only if $g_i(x)$ has $\zeta^r$ as a zero over $V(2^m)$.*

*Proof:* Let $g(x)$ be any polynomial of exponent $e$. Since $g(x)$ divides $x^e - 1$ over $V$, $g(x^r)$ divides $x^{er} - 1 = x^n - 1$. $g(x^r)$ is a product of irreducible factors of $x^n - 1$.

Let $\alpha_0, \alpha_1, \cdots, \alpha_{s-1}$ be the cycle associated with $g(x)$, so that a typical factor of $g(x^r)$ is $(x^r - \zeta^{\alpha_i})$. The cycle $\beta_0, \beta_1, \cdots, \beta_{m-1}$ is associated with $g(x^r)$ if and only if $r\beta_j = \alpha_i$ $[(\zeta^{\beta_j})^r = \zeta^{\alpha_i}]$ for a suitable choice of $i,j$.

Suppose now that $g_i(x^r)$ is divisible for $f_1(x)$ over $V$. The cycle 1, 2, $\cdots$, $2^{m-1}$ is then associated with $g_i(x^r)$, and $\zeta^r = \zeta^{\alpha_i}$ for some $i$; thus $\zeta^r$ is a zero of $g_i(x)$.

Suppose that $\zeta^r$ is a zero of $g_i(x)$. The cycle associated with $g_i(x)$ is then $r, 2r, \cdots, 2^{s-1}r$. Clearly, 1, 2, $\cdots$, $2^{m-1}$ is a cycle associated with $g_i(x^r)$, and $f_1(x)$ divides $g_i(x^r)$.

It may be noted that the proof of this theorem provides a way of finding the factors of $g(x^r)$ which is useful in other applications.

*Automorphisms and Idempotents of* $\mathfrak{R}_n$

Let $q$ be an integer prime to $n$, and let $\sigma_q$ be the mapping of $\mathfrak{R}_n$ onto itself defined by $h(x) \to h(x^q)$, exponents reduced mod $n$ where necessary. $\sigma_q$ clearly preserves addition and multiplication in $\mathfrak{R}_n$, and is 1-1, since with $q$ prime to $n$, $x^{iq} = x^{jq}$ implies $iq \equiv jq \bmod n$, implies $i \equiv j \bmod n$. $\sigma_q$ is an automorphism of $\mathfrak{R}_n$, and $\mathfrak{A}\sigma_q$ is again an ideal.

In $V^n$, $\sigma_q$ is a permutation of coordinate places, described by $\omega \to q\omega \bmod n$ [$\Sigma_2(n)$ is the special case $\sigma_2$]. Thus $\sigma_q$ changes alphabets of $V^n$ into equivalent alphabets, and in particular changes cyclic alphabets into equivalent cyclic alphabets.

The automorphisms $\sigma_q$ are useful because it is easy to compute their effect on the ideals of $\mathfrak{R}_n$.

*Lemma 2.6:** *Every ideal* $\mathfrak{A}$ *of* $\mathfrak{R}_n$ *contains a unique polynomial* $c(x)$ *with the following properties:*

(*i*) $c(x) = c(x)^2$; $c(x)$ *is idempotent*

(*ii*) $\mathfrak{A} = \mathfrak{R}_n \cdot c(x)$; $c(x)$ *generates* $\mathfrak{A}$

(*iii*) $c(x)$ *is a unit for* $\mathfrak{A}$.

(*iv*) $c(x)\sigma_q$ *is the idempotent of* $\mathfrak{A}\sigma_q$.

*Proof:* Let $a(x)$, $b(x)$ be the generating factor and reciprocal factor of $\mathfrak{A}$. Since $n$ is odd, they are relatively prime. There exist polynomials $h_1(x)$, $h_2(x)$ such that $h_1(x)a(x) + h_2(x)b(x) = 1$, and $h_1(x)$, $h_2(x)$ are relatively prime to $b(x)$, $a(x)$, respectively. We show that $c(x) = h_1(x)a(x)$ is the idempotent of $\mathfrak{A}$.

(*i*) $c(x)^2 + c(x)h_2(x)b(x) = c(x)$. The second term on the left is zero since it contains the factor $x^n - 1$. Hence $c(x)$ is idempotent.

* In other words, $\mathfrak{R}_n$ is a commutative, semisimple ring. It is, of course, the group algebra over $V$ of the cyclic group of order $n$; $n$ odd implies that it is semisimple.[6]

(*ii*) The generating factor of the ideal $\mathfrak{R}_n \cdot c(x)$ is the highest common factor of $c(x)$ and $x^n - 1$. This is $a(x)$ by the construction of $c(x)$. Hence $\mathfrak{R}_n \cdot c(x) = \mathfrak{A}$.

(*iii*) If $\alpha(x) \in \mathfrak{A}$, $\alpha(x) = \alpha'(x)c(x)$ by (*ii*). Then $\alpha(x)c(x) = \alpha'(x)c(x)^2 = \alpha'(x)c(x)$ [by (*i*)] $= \alpha(x)$. Hence $c(x)$ is a unit for $\mathfrak{A}$. $c(x)$ is then necessarily unique, since the commutative ring $\mathfrak{A}$ cannot have two unities.

(*iv*) $c(x)\sigma_q$ is idempotent because $\sigma_q$ is an automorphism of $\mathfrak{R}_n$, and is the unique idempotent of the ideal $\mathfrak{R}_n c(x)\sigma_q = \mathfrak{A}\sigma_q$.

We now associate with each ideal $\mathfrak{A}$ a third polynomial $c(x)$, the generating idempotent of $\mathfrak{A}$.

*Corollary 2.7:* $\mathfrak{A}\sigma_q = \mathfrak{A}$ *if and only if* $c(x)\sigma_q = c(x)$.

*Corollary 2.8:* $\mathfrak{A}\sigma_2 = \mathfrak{A}$ *for every ideal* $\mathfrak{A}$ *of* $\mathfrak{R}_n$; *equivalently, the permutation* $\Sigma_2(n)$ *preserves every cyclic alphabet of* $V^n$.

Two vectors $(\alpha_0, \alpha_1, \cdots, \alpha_{n-1})$, $(\beta_0, \beta_1, \cdots, \beta_{n-1})$ are said to be orthogonal if

$$\sum_{i=0}^{n-1} \alpha_i \cdot \beta_i = 0 \text{ (multiplication and addition in } V).$$

The orthogonal complement (dual alphabet) $\mathfrak{A}^\perp$ of $\mathfrak{A}$ consists of the vectors of $V^n$ which are orthogonal to every vector of $\mathfrak{A}$. For our purposes it is convenient to say that cyclic alphabets $\mathfrak{A}$, $\mathfrak{B}$ are orthogonal if $\mathfrak{B}$ is generated by $b(x) = (x^n - 1)/a(x)$. This is justified by the following lemma.

*Lemma 2.9:* $\mathfrak{A}^\perp$ *is equivalent to the ideal generated by* $b(x)$ *and is obtained from it by the transformation* $x \to x^{-1}$.

The proof of this lemma can be found in Peterson[4] (6.12).

*Lemma 2.10: If* $\mathfrak{A}$ *has idempotent* $c(x)$, *the ideal* $\mathfrak{B} = \mathfrak{R}_n \cdot b(x)$ *has idempotent* $1 + c(x)$.

*Proof:* By 2.6 the idempotent of $\mathfrak{B}$ is

$$h_2(x)b(x) = 1 + h_1(x)a(x) = 1 + c(x).$$

Since we have agreed to say that $\mathfrak{A}$, $\mathfrak{B}$ are orthogonal ideals, we may also say that $c(x)$, $1 + c(x)$ are orthogonal idempotents. This is fortunate, since it is a well-established convention in the theory of algebras to say that two idempotents are orthogonal if their product is zero.[6] [$c(x)(1 + c(x)) = c(x) + c(x) = 0$.] We shall adopt this convention. It is to be noted that orthogonality for ideals is still not the same as orthogonality for idempotents. The idempotents $c_1(x)$, $c_2(x)$ are orthogonal if $c_1(x) \cdot c_2(x) = 0$. The ideals they generate are not orthogonal unless also $c_1(x) + c_2(x) = 1$.

*Lemma 2.11:* (*i*) *The ideal* $\mathfrak{a}_1 \cap \mathfrak{a}_2$ *has idempotent* $c_1c_2$. (*ii*) *The ideal* $\mathfrak{a}_1 \cup \mathfrak{a}_2$ *has idempotent* $c_1 + c_2 + c_1c_2$.

*Proof:*

(*i*) $\mathfrak{a}_1 \cap \mathfrak{a}_2$ is generated by the least common multiple of $a_1(x)$, $a_2(x)$, say $\bar{a}(x)$. $\bar{a}(x)$ is the highest common factor of $c_1(x)c_2(x)$ and $x^n - 1$; hence $c_1(x)c_2(x)$ is the idempotent of the ideal $\mathfrak{R}_n \cdot \bar{a}(x)$.

(*ii*) Set $d(x) = c_1(x) + c_2(x) + c_1(x)c_2(x)$. The $c_1(x)d(x) = c_1(x)$, $c_2(x)d(x) = c_2(x)$. Thus $d(x)$ is idempotent, and the ideal $\mathfrak{R}_n d(x)$ contains $\mathfrak{a}_1$ and $\mathfrak{a}_2$.

Let $\bar{\mathfrak{a}}$ be any ideal which contains $\mathfrak{a}_1$ and $\mathfrak{a}_2$, and let $\bar{c}(x)$ be the idempotent of $\bar{\mathfrak{a}}$. Since $\bar{c}(x)$ is a unit for $\bar{\mathfrak{a}}$, $c_i(x)\bar{c}(x) = c_i(x)$, $i = 1, 2$. Then $d(x)\bar{c}(x) = d(x)$, and $\mathfrak{R}_n d(x)$ is contained in every ideal $\bar{\mathfrak{a}}$. Hence $\mathfrak{R}_n \cdot d(x) = \mathfrak{a}_1 \cup \mathfrak{a}_2$.

An ideal of $\mathfrak{R}_n$ is said to be a minimal ideal if it contains no subideal other than (0). A minimal ideal of $\mathfrak{R}_n$ will be denoted by $\mathfrak{M}_i$, its generating factor by $m_i(x)$, its reciprocal factor by $f_i(x)$, and its generating idempotent by $\theta_i(x)$. The idempotent of a minimal ideal is called a *primitive* idempotent.

*Lemma 2.12:*

(*i*) $\mathfrak{M}_i$ *is a minimal ideal if and only if* $f_i(x)$ *is an irreducible factor of* $x^n - 1$.

(*ii*) $\mathfrak{M}_i \cap \mathfrak{M}_j = 0$ *if* $i \neq j$; *the dimension of* $\mathfrak{M}_i \cup \mathfrak{M}_j$ *is the sum of the dimensions of* $\mathfrak{M}_i$ *and* $\mathfrak{M}_j$.

(*iii*) *Any ideal* $\mathfrak{a}$ *is the union of the minimal ideals continued in* $\mathfrak{a}$. *In particular,* $\mathfrak{R}_n$ *is the union of all its minimal ideals.*

*Proof:*

(*i*) follows from 2.1, since the dimension of a minimal ideal is as small as possible.

(*ii*) The generating factor of the ideal orthogonal to $\mathfrak{M}_i \cap \mathfrak{M}_j$ is the highest common factor of $f_i(x)$ and $f_j(x)$, which is 1. Hence $\mathfrak{M}_i \cap \mathfrak{M}_j$ is equivalent to $\mathfrak{R}_n{}^{\perp}$ and is zero. The second statement follows immediately.

(*iii*) Let $b(x)$ be the reciprocal polynomial of $\mathfrak{a}$, and let $b(x) = f_1(x)f_2(x) \cdots f_\nu(x)$ where (since $n$ is odd) the $f_i(x)$ are distinct irreducible factors of $x^n - 1$. $\mathfrak{a}$ contains the polynomials $(x^n - 1)/f_i(x)$, hence contains the minimal ideals $\mathfrak{M}_i$, $i = 1, \cdots, \nu$, hence contains their union $\mathfrak{M}_1 \cup \mathfrak{M}_2 \cup \cdots \cup \mathfrak{M}_\nu$. By (*ii*) the dimension of this union is the sum of the degrees of $f_1(x), \cdots, f_\nu(x)$ which by 2.1 is the dimension of $\mathfrak{a}$. Thus $\mathfrak{a} = \mathfrak{M}_1 \cup \mathfrak{M}_2 \cup \cdots \cup \mathfrak{M}_\nu$.

We note that this theorem is not true for even $n$.

Let $\theta_0, \theta_1, \cdots, \theta_{t-1}$ be the set of primitive idempotents of $\mathfrak{R}_n$.

*Corollary 2.13:*

(*i*) $\theta_i \cdot \theta_j = 0 \quad i \neq j.$

(*ii*) *Every idempotent of* $\mathfrak{R}_n$ *is of the form*

$$\sum_{i=0}^{t-1} \epsilon_i \theta_i ,$$

*where* $\epsilon_i$ *belongs to* $V$. *In particular,*

$$\sum_{i=0}^{t=1} \theta_i = 1.$$

*Proof:*

(*i*) Follows from 2.12 (*ii*) and 2.11 (*i*).

(*ii*) Since any ideal in $\mathfrak{R}_n$ is the union of minimal ideals, any idempotent can be obtained from the $\theta_i$ by repeated applications of 2.11 (*ii*). The product terms disappear by part (*i*) of this lemma. In particular

$$\mathfrak{R}_n \cdot 1 = \mathfrak{R}_n \cdot \left( \sum_{i=0}^{t-1} \theta_i \right).$$

*Lemma 2.14:** *If* $\mu_1$ , $\mu_2$ *belong to the minimal ideal* $\mathfrak{M}$, *and* $\mu_1\mu_2 = 0$, *then either* $\mu_1 = 0$ *or* $\mu_2 = 0$.

*Proof:* Suppose that $\mu_2 \neq 0$. Consider the set $\Lambda$ of elements $m$ in $\mathfrak{M}$ such that $m\mu_2 = 0$. If $m_1$ , $m_2 \in \Lambda$, so does $m_1 + m_2$ ; if $m \in \Lambda$ and $\mu \in \mathfrak{R}_n$ , then $\mu m \in \Lambda$. Hence $\Lambda$ is a subideal of $\mathfrak{M}$, so is either all of $\mathfrak{M}$ or the zero ideal. Let $\theta$ be the idempotent of $\mathfrak{M}$; then $\theta \cdot \mu_2 = \mu_2 \neq 0$; hence $\theta \in \Lambda$, and $\Lambda \neq \mathfrak{M}$. We must then have $\Lambda = 0$; consequently $\mu_1 = 0$.

It is clear that it will be advantageous to find the explicit forms of the primitive idempotents $\theta_i(x)$. Indeed if this were not easy the above theoretical results would have little practical value; however it is easy, and has in fact been done for all odd values of $n$ through 1023. The method used is due to Prange,[7] and is described below.

Let $r = r_1 , r_2 , \cdots , r_m$ be a cycle of $\Sigma_2(n)$ and let $\eta_r$ denote the polynomial $x^{r_1} + x^{r_2} + \cdots + x^{r_m}$ $\eta_r$ is an idempotent, since squaring it simply rearranges the numbers which occur as exponents of $x$.

*Lemma 2.14: The polynomial*

$$\sum_{j=0}^{n-1} a_j x^a , \qquad a_j \in V,$$

*is an idempotent if and only if it can be written as a sum of the* $\eta_r$ .

* Alternatively we might quote the well known theorem[4,5] that the minimal ideal $\mathfrak{M}$ is isomorphic to the Galois field $V[y]/f(y)$.

*Proof:* Clearly any sum of the $\eta_r$ is idempotent. The "if" part of the lemma follows immediately from 2.2.

*Lemma 2.15: The number of primitive idempotents of $\mathfrak{R}_n$ is the same as the number of cycles of $\Sigma_2(n)$.*

*Proof:* Let $s$ be the number of primitive idempotents. By 2.12 (*iii*) the number of ideals in $\mathfrak{R}_n$ is $2^s$. Hence $s$ is the number of cycles $\Sigma_2(n)$.

Any idempotent may be expressed as a linear combination of the $\eta_r$ (which we can find easily) or as a linear combination of the primitive idempotents $\theta_j$. The $\theta_j$ have the additional property that they are mutually orthogonal. In particular, each $\eta_r$ is the sum of a subset of the $\theta_j$; the problem is to split it into its components.

We observe that if $S$, $T$ are nonempty subsets of the indices 0, 1, $\cdots$, $t-1$, $S \neq T$, then

$$\left(\sum_{j\in S} \theta_j\right) \cdot \left(\sum_{j\in T} \theta_j\right) = \sum_{j\in S\cap T} \theta_j .$$

The product of two idempotents will contain fewer primitive idempotents than either factor.

Let $t$ be the number of primitive idempotents. Then

$$1 = \sum_{j=0}^{t-1} \theta_j , \quad \text{and if} \quad 1 = \sum_{j=0}^{t-1} \xi_j$$

where the $\xi_j$ are orthogonal idempotents, then the $\xi_j$ are, except possibly in order, the same as the $\theta_j$. We use this fact to set up an algorithm as follows:

Suppose that we have at some stage a decomposition of 1 into $\tau < t$ mutually orthogonal idempotents;

$$1 = \sum_{j=0}^{\tau-1} \xi_j , \ \xi_j^2 = 1, \ \xi_i\xi_j = 0 \qquad i \neq j.$$

Let $\xi$ be an idempotent; set

$$\xi_j = \xi_j\xi + \xi_j(1 + \xi) = \xi_{j1} + \xi_{j2}j = 0, 1, \cdots, \tau - 1.$$

$\xi_{j1}$, $\xi_{j2}$ are idempotent, and the new idempotents are mutually orthogonal. If the splitting is genuine (it may happen that $\xi_j = \xi$ or $\xi_j = 1 + \xi$, in which case no splitting takes places) the result is a decomposition of 1 into more than $\tau$ mutually orthogonal idempotents.

To start the algorithm we set $1 = \eta_1 + (1 + \eta_1)$; the other $\eta_j$ provide successive candidates for $\xi$. The computation is finished when there are $t$ components in the decomposition of 1. Since the $\eta_r$ are also a base for the idempotents of $\mathfrak{R}_n$, this stage must be reached by the time the set of $\eta_r$ is exhausted.

The primitive idempotent $\theta_i(x)$ is the generating idempotent of a minimal ideal $\mathfrak{M}_i$; the orthogonal idempotent $1 + \theta_i(x)$ is the generating idempotent of a maximal ideal $\mathfrak{M}_i$: the generating factor $f_i(x)$ of $\mathfrak{M}_i$ is an irreducible factor of $x^n - 1$, and is the greatest common factor of $1 + \theta_i(x)$ and $x^n - 1$. In this way we can produce the parallel lists of primitive idempotents and irreducible factors of $x^n - 1$ referred to in Section I.

We return now to the automorphisms $\sigma_q$ of $\mathfrak{R}_n$.

The set of automorphisms $\sigma_q$ is an Abelian group, with $\sigma_{q_1}\sigma_{q_2} = \sigma_{q_1 q_2}$ defined in the usual way. It is isomorphic to the (multiplicative) group of integers mod $n$ which are prime to $n$. Since $\sigma_2$ and its powers leave the idempotents of $\mathfrak{R}_n$ unchanged, we may, for our purposes, factor out this subgroup. In practice we choose one $q$ from each cycle of $\Sigma_2(n)$ which contains integers prime to $n$. These $q$ (and the associated $\sigma_q$) form a rather small Abelian group, whose structure may be found by hand, as illustrated for $n = 63$. It is worthwhile to find a set of generators for the group. One need only compute the effect of these generators on the set of primitive idempotents of $\mathfrak{R}_n$; it is then simple to calculate the effect of any automorphism on any ideal. Proposition VI is now established.

*Proof of Proposition VII:* Let $f_1(x)$ be the irreducible factor of $x^n - 1$ associated with the cycle $(1, 2, \cdots, 2^{m-1})$. $v$ is an integer prime to $n$, and we wish to identify the polynomial $f_r(x)$ associated with the cycle $(v, 2v, \cdots, 2^{m-1}v)$. Since $v$ is prime to $n$ the two cycles will be the same length. $f_1(x)$ is the highest common factor of $1 + \theta_1(x)$ and $x^n - 1$. $1 + \theta_1(x)$ is thus divisible by the polynomial $(x - \zeta)(x - \zeta^2) \cdots (x - \zeta^{2^m})$. Let $u$, prime to $n$, be such that $uv \equiv 1 \bmod n$. Then $(1 + \theta_1(x))\sigma_u = 1 + \theta_1(x^u)$ is divisible by

$$(x^u - \zeta)(x^u - \zeta^2) \cdots (x^u - \zeta^{2^m}) = (x^u - \zeta^{uv})(x^u - \zeta^{2uv}) \cdots (x^u - \zeta^{2^m uv}),$$

which is divisible by $(x - \zeta^v)(x - \zeta^{2v}) \cdots (x - \zeta^{2^m v})$.

Thus $f_r(x)$ divides $(1 + \theta_1(x))\sigma_u$ over $V(2^m)$, and since both polynomials have coefficients in $V$, $f_r(x)$ divides $(1 + \theta_1(x))\sigma_u$ over $V$. Hence $f_r(x)$ is the highest common factor of $(1 + \theta_1(x))\sigma_u$ and $x^n + 1$.

*Spectra of Cyclic Alphabets*

Let $a(x)$, $b(x)$ be the generating factor and reciprocal factor of an ideal $\mathfrak{A}$ in $\mathfrak{R}_n$. Let $b(x)$ belong to exponent $e$, where $n = e\alpha$, $\alpha > 1$. Let $\mathfrak{A}'$ be the ideal in $\mathfrak{R}_e$* with reciprocal polynomial $b(x)$.

* $\mathfrak{R}_e$ is the ring of polynomials mod $x^e - 1$.

*Lemma 2.16: Every letter of $\mathfrak{A}$ consists of $\alpha$ repetitions of a letter of $\mathfrak{A}'$.*

*Proof:* Let $a'(x) = (x^e - 1)/b(x)$ be the generating polynomial of $\mathfrak{A}'$. Then

$$a(x) = (x^n - 1)/b(x) = \frac{x^n - 1}{x^e - 1} \cdot a'(x).$$

$$= \left(\sum_{i=0}^{\alpha} x^{n-ie}\right) a'(x).$$

Let $r(x)a'(x) = \sum_{i=0}^{\alpha} \alpha_i x^i$ (multiplication in $\mathfrak{R}_e$) be a letter of $\mathfrak{A}'$. With multiplication in $\mathfrak{R}_n$, $\mathfrak{A}$ contains

$$r(x)a(x) = \left(\sum_{i=0}^{\alpha} x^{n-ie}\right)\left(\sum_{i=0}^{e} \alpha_i x^i\right).$$

Hence each letter of $\mathfrak{A}'$ gives rise to a letter of $\mathfrak{A}$, which consists of $\alpha$ repetitions of the letter of $\mathfrak{A}'$. It is evident that different letters of $\mathfrak{A}'$ give rise to different letters of $\mathfrak{A}$. Since the dimensions of $\mathfrak{A}$ and $\mathfrak{A}'$ are both equal to the degree of $b(x)$, all of $\mathfrak{A}$ is obtained in this way.

*Corollary 2.17: Let the spectrum of $\mathfrak{A}'$ be $A'(i)$ $i = 0, \cdots, e$. The spectrum of $\mathfrak{A}$ is given by the equations $A(\alpha i) = A'(i)$, $i = 0, \cdots, e$.*

For example, let $n = 15$, and $b(x) = 1 + x + x^2$. $b(x)$ has exponent 3; $a'(x) = (x^3 + 1)/b(x) = 1 + x$; $a(x) = (1 + x^3 + x^6 + x^9 + x^{12})(1 + x)$. The ideals $\mathfrak{A}'$, $\mathfrak{A}$ are tabulated below.

| 0 | 1 | 2 |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |

Let $T$ denote the cycle permutation $\omega \to \omega + 1 \pmod{n}$ of the numbers $0, 1, \cdots, n - 1$. $T$ shall also denote the mapping $h(x) \to xh(x)$ (exponents mod $n$) of $\mathfrak{R}_n$ onto itself. Clearly $T^n$ is the identity mapping. If $\alpha \in \mathfrak{A}$, the polynomials (or vectors) $\alpha T$, $\alpha T^2$, $\cdots$, $\alpha T^{n-1}$ also belong to $\mathfrak{A}$. The letters of $\mathfrak{A}$ are divided into a number of nonoverlapping cycles; to construct $\mathfrak{A}$ we need to know only one element from each cycle.

In fact it would seldom be useful to construct a picture of $\mathfrak{A}$ in this way. We restrict ourselves to finding the spectrum of $\mathfrak{A}$.

The set $\alpha, \alpha T, \cdots, \alpha T^{n-1}$ does not always contain $n$ different letters. We denote by $\pi(\alpha)$ the number of different letters in this set; $\pi(\alpha)$ is called the *period* of $\alpha$. The set

$$\alpha, \alpha T, \cdots, \alpha T^{\pi(\alpha)-1}$$

is then a complete cycle of $\mathfrak{A}$, and the length of the cycle is $\pi(\alpha)$.

Let $r(x) \in \mathfrak{R}_n$ ; let $a(x)$ be the highest common factor of $r(x)$ and $x^n - 1$, and let $b(x) = (x^n - 1)/a(x)$.

*Lemma 2.18: The period of $r(x)$ is the exponent of $b(x)$.*

*Proof:* Suppose $b(x)$ belongs to exponent $e$; set $a'(x) = [(x^e - 1)/b(x)]$, $r(x) = h(x)a(x)$, where $h(x)$ is relatively prime to $x^n - 1$. $r(x)(x^e - 1) = h(x)\cdot a(x)\cdot b(x)a'(x) = h(x)a'(x)(x^n - 1) = 0$. Hence $x^e r(x) = r(x)$, and the period of $r(x)$ is $\leqq e$.

Suppose that $e'$ is the period of $r(x)$. Then $e' \leqq n$, and $r(x)(x^{e'} - 1) = 0$; in $V[x]$, $h(x)a(x)(x^{e'} - 1) = i(x)(x^n - 1) = i(x)a(x)b(x)$ where $i(x)$ is a polynomial in $V[x]$. $b(x)$ and $h(x)$ are relatively prime since $b(x)$ is a factor of $x^n - 1$. Thus $b(x)$ divides $(x^{e'} - 1)$, and $e' \geqq e$.

*Proof of Proposition IX:* $\pi(\theta_i)$ is the period of $\theta_i$, and $\pi(m)$ the period of $m \in \mathfrak{R}_n \cdot \theta_i$. Then $mx^{\pi(\theta_i)} = m\theta_i x^{\pi(\theta_i)} = m\theta_i = m$; hence $\pi(m) \leqq \pi(\theta_i)$. Also $0 = m\cdot(x^{\pi(m)} + 1) = m\theta_i\cdot(x^{\pi(m)} + 1) = m\theta_i\cdot\theta_i(x^{\pi(m)} + 1)$. By 2.13, since $m\theta_i \neq 0$, we must have $\theta_i(x^{\pi(m)} + 1) = 0$. Thus $\pi(\theta_i) \leqq \pi(m)$, so that $\pi(\theta_i) = \pi(m)$. By 2.18, $\pi(\theta_i)$ is the exponent of the irreducible polynomial $f_i(x)$.

If $n = 2^m - 1$, an irreducible polynomial $f(x)$ of exponent $n$ has degree $m$, and a minimal ideal of period $n$ contains just one cycle besides the zero cycle. The maximal ideal with generating factor $f(x)$ is a Hamming code (a close-packed code of minimum distance 3).[3] If $n$ is not of this form, the minimal ideals of period $n$ contain more than one cycle; it is then necessary to find several cycle representatives. No shortcut for doing this has been developed; the particular cases which have been studied have been solved by brute force.

If we have found a cycle representative for each cycle of $\mathfrak{R}\theta_i$ and $\mathfrak{R}\theta_j$, we can construct cycle representatives for $\mathfrak{R}(\theta_i + \theta_j)$ with the help of the following lemmas.

Let $m \in \mathfrak{R}\theta_i$, $n \in \mathfrak{R}\theta_j$.

*Lemma 2.19: $mT^\mu + nT^\nu = nT^{\mu'} + nT^{\nu'}$ if and only if $mT^\mu = mT^{\mu'}$ and $nT^\nu = nT^{\nu'}$.*

*Proof:* The equation above may be written

$$mT^\mu - mT^{\mu'} = nT^\nu - nT^{\nu'}.$$

The left-hand side belongs to $\mathfrak{R}\theta_i$ and the right-hand side to $\mathfrak{R}\theta_j$. The intersection of these ideals is zero.

Let $\pi(m)$, $\pi(n)$ be the periods of $m$, $n$. Let $H$, $h$ be respectively the least common multiple and highest common factor of these numbers.

*Lemma 2.20:* (Proof of Proposition X)*: The $\pi(m)\cdot\pi(n)$ elements $mT^\mu + nT^\nu$ are partitioned into $h$ cycles of period $H$. The vectors $mT^\mu + n$, $\mu = 0, 1, \cdots, h - 1$ are in different cycles, and may be taken as cycle representatives.*

*Proof:* Let $\lambda$ be the period of the vector $mT^{\mu} + nT^{\nu}$. Then

$$(mT^{\mu} + nT^{\nu})\, T^{\lambda} = mT^{\mu} + nT^{\nu},$$

and by 2.19 $\nu + \lambda \equiv \nu \bmod \pi(m)$ and $\mu + \lambda \equiv \mu \bmod \pi(n)$. Thus $\lambda$ is divisible by both $\pi(m)$ and $\pi(n)$ and $\lambda = qH$, $q$ an integer $\geqq 1$.

$mT^{\mu} + n$ and $mT^{\mu'} + n$ are in the same cycle if and only if

$$(mT^{\mu} + n)T^{\rho} = mT^{\mu'} + n,$$

or $\mu + \rho \equiv \mu' \bmod \pi(m)$ and $\rho \equiv 0 \bmod \pi(n)$. $\rho$ and $\pi(m)$ are both divisible by $h$; hence $\mu - \mu' \equiv \rho \bmod \pi(m)$ implies that $\mu - \mu'$ is divisible by $h$. The $h$ vectors $mT^{\mu} + n$, $\mu = 0, 1, \cdots, h - 1$ must be in different cycles.

Thus there are at least $h$ different cycles, and the period of each is $\geqq H$. Since there are only $\pi(m)\pi(n) = hH$ elements altogether, the only possibility is that there are $h$ cycles of period $H$.

We now return to Proposition V, which was omitted earlier. We restate the proposition as follows:

*Theorem 2.21: Let $\mathfrak{M}_1$, $\mathfrak{M}_2$ be minimal ideals of $\mathfrak{R}_n$. The following three statements are equivalent:*

*(i) $\mathfrak{M}_1$, $\mathfrak{M}_2$ have the same spectrum.*

*(ii) $\mathfrak{M}_1$, $\mathfrak{M}_2$ have the same dimension and period.*

*(iii) There exists an automorphism $\sigma_q$ of $\mathfrak{R}_n$ such that $\mathfrak{M}_1\sigma_q = \mathfrak{M}_2$.*

*Proof:* We show that *(i)* $\Rightarrow$ *(ii)* $\Rightarrow$ *(iii)* $\Rightarrow$ *(i)*.

Let $A(s)$ be the number of letters of weight $s$ in $\mathfrak{M}_i$. We prove that the period of $\mathfrak{M}_i$ is the highest common factor of $A(s)$, $s > 0$.

Suppose first that the period of $\mathfrak{M}_i$ is $n$; let $2^{k_i}$ be the total number of letters in $\mathfrak{M}_i$. The orthogonal complement of $\mathfrak{M}_i$ can contain no letters of weight 1 since it is a nontrivial cyclic alphabet. By Proposition VIII we obtain

$$\sum_{s=1}^{n} A(s) = 2^{k_i} - 1$$

$$\sum_{s=1}^{n} sA(s) = 2^{k_i - 1}n.$$

By the first equation, $k_1 = k_2$, so the dimensions of $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are equal. Since every cycle of $\mathfrak{M}_i$, except that containing the zero letter, is of length $n$, $n$ divides each $A(s)$ for $s > 0$. By the second equation, any other common factor of $A(s)$ is a power of 2. By the first equation, there can be no such factor.

Suppose now that the period of $\mathfrak{M}_i$ is $e_i$, and $n/e_i = \alpha_i > 1$. By 2.16

and 2.17 there is a minimal ideal $\mathfrak{M}_i'$ in $\mathfrak{R}_{e_i}$ with period $e_i$ and spectrum $A'(s)$ such that the spectrum of $\mathfrak{M}_i$ is given by $A(\alpha_i s) = A'(s)$. By the first part of the proof, $e_i$ is the highest common factor of $A'(s)$, $s > 0$; hence $e_i$ is the highest common factor of $A(\alpha_i s)$, $\alpha_i s > 0$.

*(ii)* $\Rightarrow$ *(iii)*. Suppose first that the period of $\mathfrak{M}_1$ and $\mathfrak{M}_2$ is $n$. Let $\mathfrak{M}_1$ correspond to the cycle $(1, 2, \cdots, 2^{m-1})$ of $\Sigma_2(n)$, and $\mathfrak{M}_2$ correspond to the cycle $(v, 2v, \cdots, 2^{m-1}v)$. $v$ must be prime to $n$, since the irreducible polynomial associated with $\mathfrak{M}_2$ has exponent $n$. Choose $u$, prime to $n$ so that $uv \equiv 1 \bmod n$. As in the proof of Proposition VII, $\mathfrak{M}_1\sigma_u = \mathfrak{M}_2$.

Suppose now that $\mathfrak{M}_i$ has exponent $e$, $i = 1, 2$, where $n/e = r > 1$. Let $\mathfrak{M}_1$, $\mathfrak{M}_2$ be associated with cycles $(r, 2r, \cdots, 2^{m'-1}r)$ and $(s, 2s, \cdots, 2^{m'-1}s)$. The lengths of these cycles are the same because $\mathfrak{M}_1$ and $\mathfrak{M}_2$ have the same dimension.

As in the proof of Lemma 2.5, $s = qr$, where $q$ is prime to $n$. Applying again the proof of Proposition VII, we see that $\mathfrak{M}_1\sigma_q = \mathfrak{M}_2$.

*(iii)* $\Rightarrow$ *(i)*. If $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are equivalent they clearly have the same spectrum.

We have thus shown that minimal and maximal cyclic alphabets of $\mathfrak{R}_n$ which have the same spectrum are equivalent. It is not known whether this is true for other cyclic alphabets. However many cases have been found of cyclic alphabets which have the same spectrum but which are definitely not related by one of the automorphism $\sigma_q$.

## CONCLUSION

Up to this time much of the theoretical work on binary cyclic alphabets has been concentrated on alphabets with block lengths of the form $n = 2^m - 1$. Such numbers become rather sparse as $n$ increases. On the other hand, alphabets of long block length are important for actual use on the telephone network, and for such applications the block length, though large, is likely to be restricted to a narrow range. It is therefore expedient to develop economical procedures which will pick out the alphabets with preassigned properties if any such exist. The amount of information presented in this paper about the structure of the polynomial ring $\mathfrak{R}_n$ is no doubt formidable; it has, however, very practical applications.

## REFERENCES

1. Slepian, D., A Class of Binary Signaling Alphabets, B.S.T.J., *35*, January, 1956, p. 203.
2. Elliott, E. O., Estimates of Error Rates for Codes on Burst Noise Channels, B.S.T.J., *42*, Sept., 1963, p. 1977.

3. MacWilliams, Jessie, A Theorem on the Distribution of Weights in a Systematic Code, B.S.T.J., *42*, Jan., 1963, p. 79.
4. Peterson, W. W., *Error Correcting Codes*, John Wiley and Sons, Inc., New York, 1961.
5. Van de Waerden, B. L., *Modern Algebra*, Julius Springer, Berlin, 1937.
6. Curtis, C. W., and Reiner, I., *Representation Theory of Finite Groups and Associative Algebras*, John Wiley and Sons, Inc., New York, 1962.
7. Prange, E., An Algorithm for Factoring $x^n - 1$ over a Finite Field, Air Force Cambridge Research Center, AFCRC-TN-59-775.

# Eigenmodes of a Symmetric Cylindrical Confocal Laser Resonator and Their Perturbation by Output-Coupling Apertures

By D. E. McCUMBER

(Manuscript received October 26, 1964)

*Using a numerical technique which is different from the iteration method of Fox and Li and which is more suitable for the analysis of high-order modes, we have calculated the diffraction losses and the field distributions at the reflectors of the low-loss modes of a symmetric confocal resonator for Fresnel numbers* $0.6 \leqq N_m \leqq 2.0$. *We have also computed the modifications which result when the two end reflectors are perturbed by circular output-coupling apertures centered on the cavity axis. For a range of small but useful aperture Fresnel numbers* $N_0$ *the aperture diffraction losses can be estimated by first-order perturbation theory from the finite-*$N_m$ *results appropriate to* $N_0 = 0$. *Such estimates fail for those larger Fresnel numbers* $N_0$ *for which the mode intensity patterns are significantly distorted at the reflectors by the finite coupling apertures.*

## I. INTRODUCTION

Fox and Li[1] demonstrated by numerical iteration that modes in the sense of self-reproducing field patterns exist for open Fabry-Perot resonators. Using a numerical technique which is different from that of Fox and Li and which is more suitable than iteration for the analysis of high-order modes, we have calculated the diffraction losses and the field distributions at the reflectors of the low-loss modes of a symmetric cylindrical confocal resonator for Fresnel numbers $0.6 \leqq N_m \leqq 2.0$. The results are discussed below.

An axial section of the symmetric confocal resonator under examination is illustrated in Fig. 1. The cavity is bounded at each end by identical spherical (parabolic) mirrors whose perfectly reflecting surfaces extend over the annular region $a_0 \leqq \rho \leqq a_m$. While a comparison of

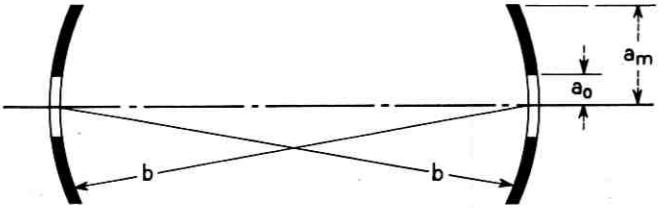


Fig. 1 — Axial section of cylindrical confocal laser cavity. The system is symmetric about the cavity midpoint; the two identical reflectors have radii of curvature $b$; and the reflecting surfaces are confined to the annular region between the two radii ($a_0$ , $a_m$).

the $a_0 = 0$ and $a_0 \neq 0$ eigenmodes is instructive as an example of the perturbation of eigenmodes by mirror imperfections, this particular geometry also has relevance to an aperture output coupling scheme proposed by Patel et al.[2] Asymmetric resonators in which, for example, only one reflector is pierced by a coupling aperture will be treated in a subsequent article. Boyd and Gordon[3] have derived closed-form expressions for the eigenvalues and eigenfunctions of symmetric *rectangular* confocal resonators in terms of the angular and radial prolate spheroidal wave functions. These results were extended to asymmetric rectangular confocal systems with output coupling *slits* by Boyd and Kogelnik.[4] Generalized prolate spheroidal functions relevant to the cylindrical confocal geometry have been defined by Slepian.[5] Basic expressions are summarized in a review article by Kogelnik.[6]

Assuming that the dimensions of the resonator in Fig. 1 are large compared to the wavelength $\lambda$ of light in the cavity, we define the resonator eigenmodes from the same scalar formulation of Huygens' principle used by other authors.[1,3] For the cylindrical confocal geometry the field amplitude at the reflectors for a typical mode can be written in the form

$$F_{lp}(\rho,\varphi) = f_{lp}(\rho) \exp\ (-il\varphi), \tag{1}$$

where $(\rho,\varphi)$ are radial and angular coordinates in a plane perpendicular to the resonator axis and where $(l,p)$ are integral quantum numbers. For a symmetric system with identical mirrors, the field amplitude at one reflector must be a constant multiple of that at the other. This self-reproducing requirement together with Huygens' principle gives the following integral equation which must be satisfied by the radial function $f_{lp}(\rho)$:

$$\kappa_{lp} f_{lp}(\rho) = \frac{2\pi}{b\lambda} \int_{a_0}^{a_m} d\rho' \rho' J_l \left(\frac{2\pi\rho\rho'}{b\lambda}\right) f_{lp}(\rho'). \tag{2}$$

Here $J_l(z)$ is the Bessel function of order $|l|$, $(a_0, a_m)$ are the $\rho$ radii limiting the reflecting surfaces (Fig. 1), and $b$ is the mirror separation and radius of curvature.

The magnitude of the eigenvalue $\kappa_{lp}$ determines the diffraction loss of the $(lp)$ mode:

$$\text{power loss/pass} = 1 - |\kappa_{lp}|^2. \tag{3}$$

The phase of the eigenvalue determines the resonant wavelength:

$$\text{resonant } \lambda = 4\pi b\{(l+1)\pi - 2 \operatorname{Arg} \kappa_{lp} - 2\pi n\}^{-1}, \tag{4}$$

where $n$ is an arbitrary integer.

If we normalize the functions $f_{lp}(\rho)$ over the surface area of the mirrors, then

$$\frac{2\pi}{b\lambda} \int_{a_0}^{a_m} d\rho\, \rho f_{lp}(\rho) f_{lq}(\rho) = \delta_{pq}, \tag{5}$$

where $\delta_{pq}$ is the Kronecker delta symbol ($\delta_{pq} = 1$ if $p = q$, and $\delta_{pq} = 0$ if $p \neq q$). The orthogonality indicated in (5) for $p \neq q$ follows immediately from (2) when the eigenvalues $\kappa_{lp}$, $\kappa_{lq}$ are nondegenerate and can be imposed if they are degenerate. We choose the arbitrary sign of the function $f_{lp}(\rho)$ such that $f_{lp}(0^+) > 0$.

For numerical calculations it is useful to replace the radial variable $\rho$ by a dimensionless variable $r$ defined such that

$$N(\rho) \equiv r^2 \equiv \rho^2/\lambda b \tag{6}$$

is the Fresnel number appropriate to the radius $\rho$. We characterize the hole and mirror radii $(a_0, a_m)$ by Fresnel numbers

$$N_0 = r_0^2 = \frac{a_0^2}{\lambda b}, \qquad N_m = r_m^2 = \frac{a_m^2}{\lambda b}. \tag{7}$$

In place of the function $f_{lp}(\rho)$ we introduce a function

$$g_{lp}(r) = f_{lp}(r\sqrt{\lambda b}) \tag{8}$$

for which (2) and (5) become:

$$\kappa_{lp} g_{lp}(r) = 2\pi \int_{r_0}^{r_m} dr'\, r' J_l(2\pi r r') g_{lp}(r'); \tag{9}$$

$$\delta_{pq} = 2\pi \int_{r_0}^{r_m} dr\, r g_{lp}(r) g_{lq}(r). \tag{10}$$

The sign convention $f_p(0^+) > 0$ requires $g_{lp}(0^+) > 0$.

The eigenvalue equation (9) for the confocal geometry is atypical

in the sense that it can be transformed to an equation having a Hermitian kernel: $2\pi(rr')^{\frac{1}{2}}J_l(2\pi rr')$. This fact implies that the eigenvalues $\kappa_{lp}$ are real and that the largest eigenvalue (characteristic of the mode with lowest loss) can be computed by a variational method. These two features do not obtain in other laser geometries for which the integral-equation kernel is generally symmetric but not real — that is, not Hermitian.[7] While we do therefore expect some qualitative differences between the properties of confocal and nonconfocal geometries, we can infer from the work of Fox and Li[1] and other authors[8,9] that many features are similar. Both the iterative technique of Fox and Li and the kernel-expansion-truncation technique we describe below can be applied to nonconfocal as well as confocal systems.

If we assume that the set of functions $g_{lp}(r)$ is complete, we can understand the iterative method of Fox and Li as follows. Given an arbitrary initial field amplitude $g^{(0)}(r)$, we express it in the form

$$g^{(0)}(r) = \Sigma_p C_p g_{lp}(r). \tag{11}$$

Substituting this expression into the right-hand side of (9), we obtain on the left-hand side

$$g^{(1)}(r) = \Sigma_p C_p \kappa_{lp} g_{lp}(r),$$

the field amplitude after one transit of the optical cavity. Using this function on the right-hand side of (9) and repeating this iterative procedure, we obtain after $n$ iterations

$$g^{(n)}(r) = \Sigma_p C_p \kappa_{lp}{}^n g_{lp}(r), \tag{12}$$

the field amplitude after $n$ transits. In the limit of large $n$ only terms belonging to the eigenvalue of largest magnitude represented ($C_p \neq 0$) on the right-hand side of (11) will remain. All other terms will be reduced in proportion to $(|\kappa_{lp}|/|\kappa_{lp}|_{\max})^n$. If the two largest eigenvalues are sufficiently different, this procedure conveniently yields for each angular quantum number $l$ the eigenvalue of largest magnitude, the eigenvalue of second-largest magnitude (through the rate of convergence of the iteration), and the two amplitude functions belonging to these eigenvalues. Results for the cylindrical confocal resonator are given by Fox and Li.[1]

In our analysis of (9) we have chosen to apply a different technique from that outlined above. Briefly, we expand the Bessel-function kernel in (9) as a power series

$$J_l(z) = \left(\frac{z}{2}\right)^l \sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{(m+l-1)!(m-1)!}\left(\frac{z}{2}\right)^{2(m-1)}, \tag{13}$$

truncate the series after a finite number $M$ of terms, and reduce the integral eigenvalue equation (9) to an $M$-dimensional-matrix eigenvalue equation which we can easily solve numerically with standard matrix-diagonalization routines. [The reduction of (9) to a matrix equation is described in the Appendix.] A similar technique has been used for the plane cylindrical geometry by She and Heffner.[8] For the confocal system for which real-number algebra is sufficient, our computations require slightly less computer time than the iterative method. Moreover, they give the eigenvalues and eigenfunctions of higher-order modes, whereas for each $l$ the iterative scheme of Fox and Li is practical only for the two largest eigenvalues and their amplitude functions.

Related methods have been utilized in limited calculations by other authors.[9, 10] In place of the power-series expansion (13) it has been suggested[10] that one utilize an expansion in associated Laguerre functions. Because the associated Laguerre functions are the exact eigenfunctions of the infinite-mirror problem ($a_0 = 0$, $a_m = \infty$), one might expect that fewer terms are required than for the power-series expansion (13) and that one can thereby simplify the solution of the matrix eigenvalue equation. While such considerations may indeed be relevant for Fresnel numbers so large that the matrix eigenvalue problem based upon (13) becomes prohibitive, the advantages are largely offset for small Fresnel numbers ($N_m \lesssim 4$) by the increased effort required to compute the necessary overlap integrals. A similar remark applies to the Fourier-Bessel expansion used for the plane cylindrical geometry by Bergstein and Schachter.[9]

One other numerical technique, different from both the iterative and the expansion-truncation techniques, deserves brief mention. If one approximates the integral in (9) by a sum over small but finite radial intervals, one has in effect reduced the integral equation to a matrix eigenvalue equation. If the number of intervals is not too large ($\lesssim 50$) it is practical to solve this problem directly, although for small Fresnel numbers considerably less effort is required with the iteration or expansion-truncation techniques.

In the following section we present results appropriate to the symmetric cylindrical confocal geometry in the absence of coupling apertures ($N_0 = 0$). We compare those finite-$N_m$ results with expressions derived by first-order perturbation theory from the infinite-$N_m$ eigenfunctions and find significant discrepancies. In Section III we indicate how finite coupling apertures ($N_0 \neq 0$) modify these results and derive simple mathematical expressions which approximate the machine-computed results in useful regions. In Section IV we briefly discuss the far-field

output patterns of the aperture-coupled resonator. In the final section we briefly recapitulate some of our conclusions.

## II. EIGENVALUES AND EIGENFUNCTIONS WITH NO MIRROR APERTURES ($N_0 = 0$)

Using the kernal expansion-truncation technique outlined in the preceding section and in the Appendix, we have computed the eigenvalues of (9) for Fresnel numbers $N_m$ in the range $0.6 \leqq N_m \leqq 2.0$. Where they overlap, our results agree with those of Fox and Li[1] and other authors.[10]

In our calculation we retained $M = \max(10N + 1,10)$ terms of the truncated series (13), where, if $r \geqq r_m$ is the maximum radius of interest in the field amplitudes $g_{lp}(r)$, we define $N = r_m r/\lambda b \geqq N_m$. This choice insures that the remainder

$$\left| J_l(2\pi rr') - \sum_{m=1}^{M} \frac{(-1)^{m-1}(\pi rr')^{l+2(m-1)}}{(m+l-1)!(m-1)!} \right| \tag{14}$$

will never be greater than 0.001 for the relevant radii. We have indicated in Table I for $N_m = 0.8$ and $N_0 = 0$ how the eigenvalues of the three lowest-loss modes converge as the number $M$ of terms increases from 1 to 10.

In Fig. 2 we have plotted the power loss/pass $= 1 - |\kappa_{lp}|^2$ of the least lossy modes for $0.6 \leqq N_m \leqq 2.0$, $N_0 = 0$. It is noteworthy that no modes have less than 1 per cent loss/pass for $N_m \leqq 0.7$, that only two modes have such losses for $N_m \leqq 1.0$, but that *ten* modes have less than 1 per cent loss/pass for $N_m \leqq 2.0$. The number of low-loss modes increases very rapidly for $N_m > 1.0$ so that, whereas $N_m \leqq 1.0$ is in one sense a small Fresnel number, $N_m = 2.0$ is already rather large.

In Figs. 3–8 we have indicated for various low-order modes how the field amplitude and intensity varies with radius on the end reflecting surfaces for $N_0 = 0$ and $N_m = 0.8, 1.6$. From the intensity plots it is clear that the power loss/pass increases as the mode order increases because the higher-order eigenfunctions have more intensity lying outside the reflecting mirrors (and hence lost) than do the low-order eigenfunctions whose intensity is more concentrated near the mirror center.

From (9) and (13) it follows [compare (32) in the Appendix] that as $r \to 0$

$$g_{lp}(r) \to G_1(lp) \left[\frac{\pi^l}{l!}\right]^{\frac{1}{2}} r^l, \tag{15}$$

TABLE I

A. DEPENDENCE OF EIGENVALUES ON NUMBER $M$ OF TERMS IN SERIES (13)*

| $M$ | $\kappa_{00}$ | $\kappa_{01}$ | $\kappa_{02}$ |
|---|---|---|---|
| 1 | 1.25853309 | — | — |
| 2 | 0.90311672 | −3.6815805 | — |
| 3 | 1.7198455 | −0.32162843 | 0.83714554 |
| 4 | 0.98489118 | −1.1885874 | 0.24997641 |
| 5 | 1.0010976 | −0.71086465 | 0.20030462 |
| 6 | 0.99744669 | −0.77432308 | 0.12014589 |
| 7 | 0.99780308 | −0.76531222 | 0.13087265 |
| 8 | 0.99777117 | −0.76618958 | 0.12962865 |
| 9 | 0.99777343 | −0.76612108 | 0.12973537 |
| 10 | 0.99777330 | −0.76612543 | 0.12972805 |

B. CHANGE IN EIGENVALUES AS NUMBER OF TERMS IN SERIES (13) INCREASES*

| $M$ | $\kappa_{00}$ | $\kappa_{01}$ | $\kappa_{02}$ |
|---|---|---|---|
| 2-1 | −0.35541637 | −3.6815805 | — |
| 3-2 | +0.81672878 | +3.35995207 | +0.83714554 |
| 4-3 | −0.73495432 | −0.86695897 | −0.58716913 |
| 5-4 | +0.01620642 | +0.47772275 | −0.04967179 |
| 6-5 | −0.00365091 | −0.06345843 | −0.08015873 |
| 7-6 | +0.00035639 | +0.00901086 | +0.0172676 |
| 8-7 | −0.00003191 | −0.00087736 | −0.00124400 |
| 9-8 | +0.00000226 | +0.00006850 | +0.00010672 |
| 10-9 | −0.00000013 | −0.00000435 | −0.00000732 |

* Tables IA and IB are computed for the case $N_m = 0.8$, $N_0 = 0$.

where $G_1(lp)$ is a constant. This $r^l$ dependence of the field amplitude and a corresponding $r^{2l}$ dependence of the intensity is apparent in Figs. 3–8. Because only the angular-independent ($l = 0$) modes have nonzero intensity at $r = 0$, we anticipate that the $l = 0$ modes are much more sensitive to a coupling aperture centered at $r = 0$ than are the $l \neq 0$ modes, a fact confirmed by the finite-$N_0$ calculations to be discussed in the following section.

For infinite mirrors without apertures ($N_m \to \infty$, $N_0 = 0$),

$$\kappa_{lp} = (-1)^p \tag{16a}$$

and

$$g_{lp}(r) = \left[\frac{2p!}{(l+p)!}\right]^{\frac{1}{2}} (2\pi r^2)^{l/2} e^{-\pi r^2} L_p{}^l(2\pi r^2), \tag{16b}$$

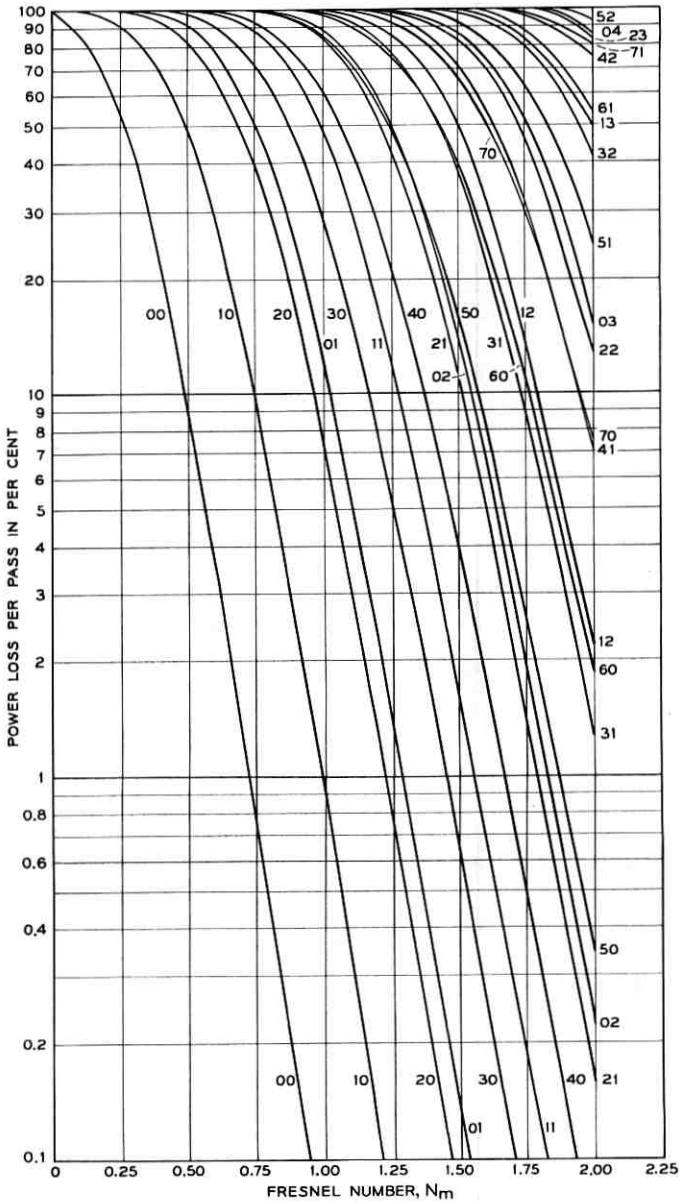


Fig. 2 — Power loss/pass versus mirror Fresnel number $N_m$ for low-loss modes of resonator having no output-coupling aperture ($N_0 = 0$).

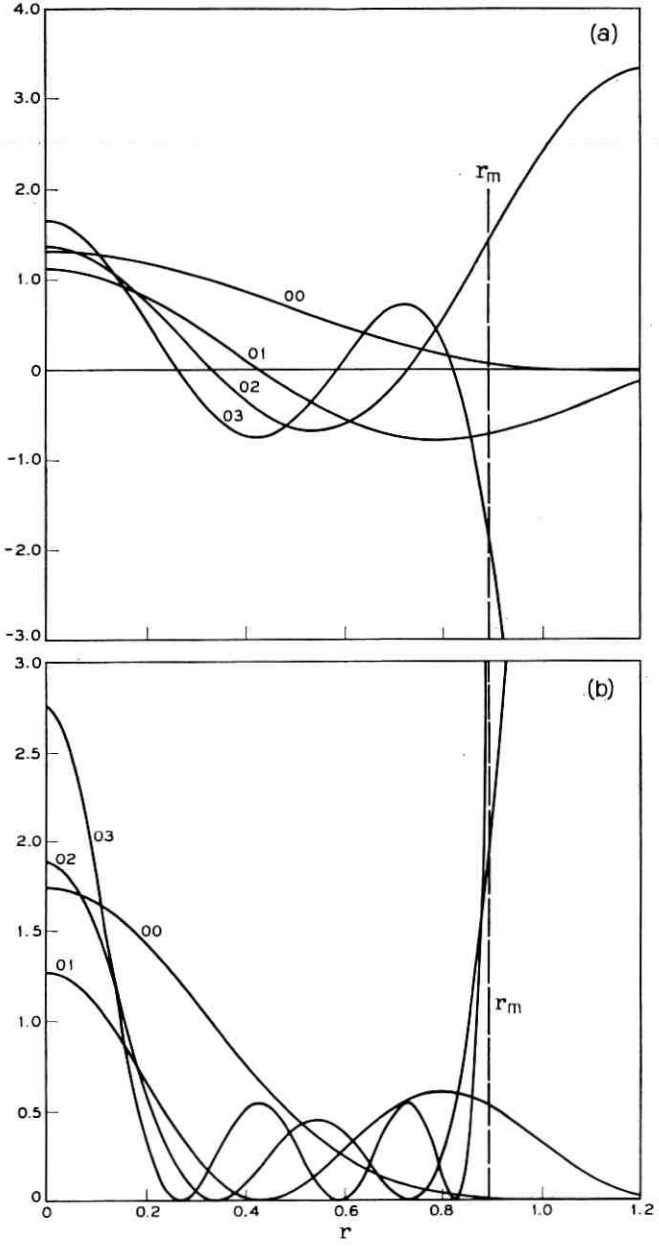


Fig. 3 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (00)$, (01), (02), and (03) with $N_m = 0.8$ and $N_0 = 0$.

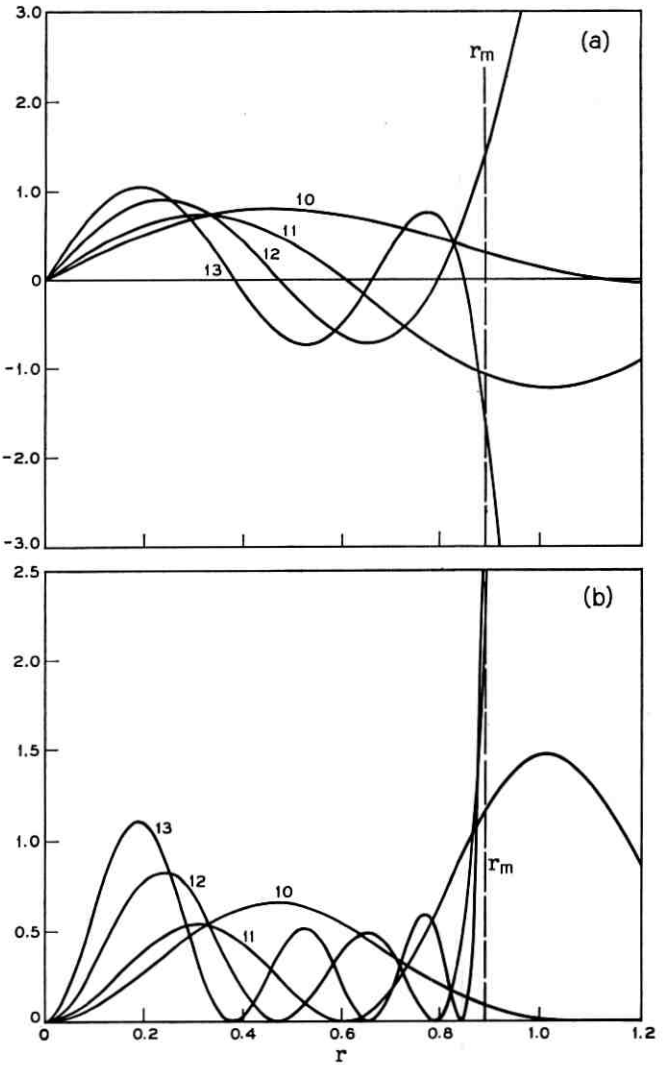


Fig. 4 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (10)$, (11), (12), and (13) with $N_m = 0.8$ and $N_0 = 0$.

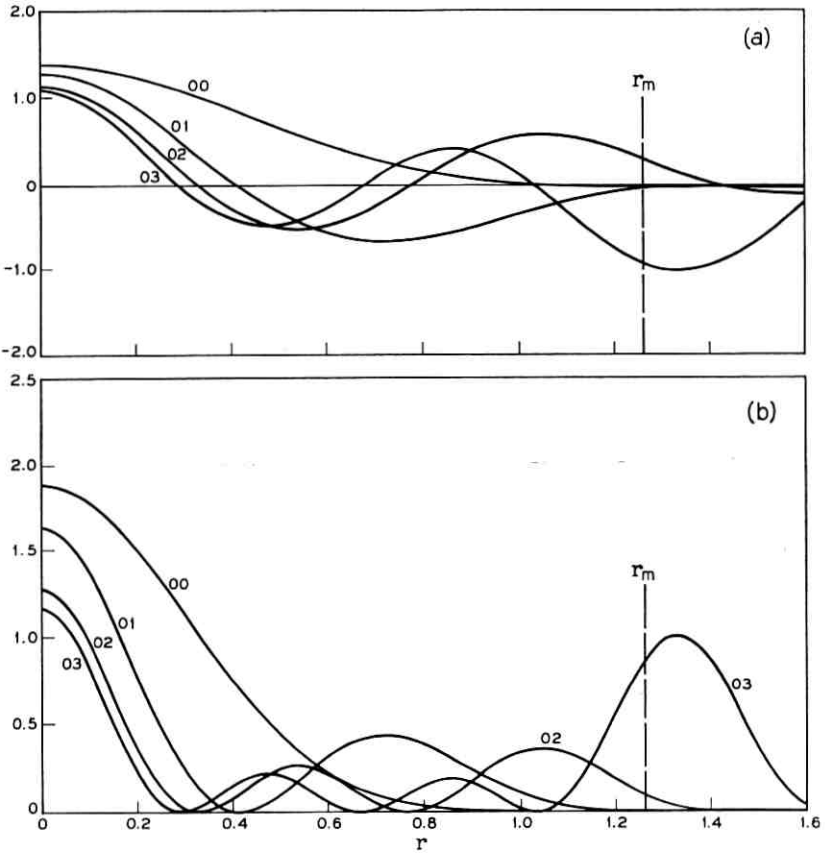


Fig. 5 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (00)$, (01), (02), and (03) with $N_m = 1.6$ and $N_0 = 0$.

where $L_p^{\,l}(z)$ is the associated Laguerre polynomial[11]

$$L_p^{\,l}(z) = \frac{e^z z^{-l}}{p!}\frac{d^p}{dz^p}(e^{-z}z^{p+l}) = \sum_{m=0}^{p}\frac{(p+l)!(-z)^m}{(p-m)!(l+m)!m!}. \quad (17)$$

Low-order Laguerre polynomials are

$$L_0^{\,l}(z) = 1; \qquad L_1^{\,l}(z) = l + 1 - z;$$

$$L_2^{\,l}(z) = \tfrac{1}{2}[(l+2)(l+1) - 2z(l+2) + z^2].$$

The finite-$N_m$ results we have computed transform continuously into the solutions (16) as $N_m$ increases. If for $N_0 = 0$ and arbitrary $N_m$ the

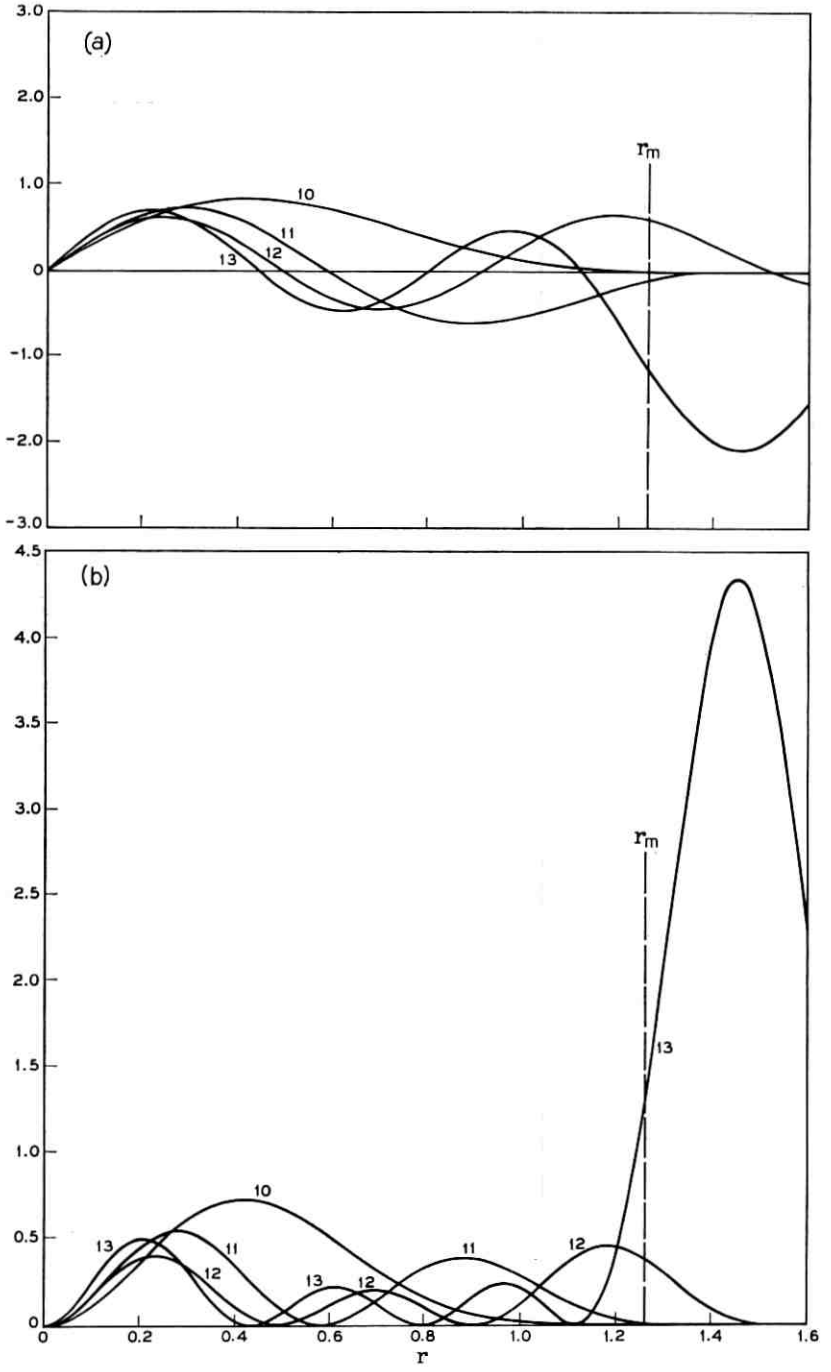


Fig. 6 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (10)$, (11), (12), and (13) with $N_m = 1.6$ and $N_0 = 0$.

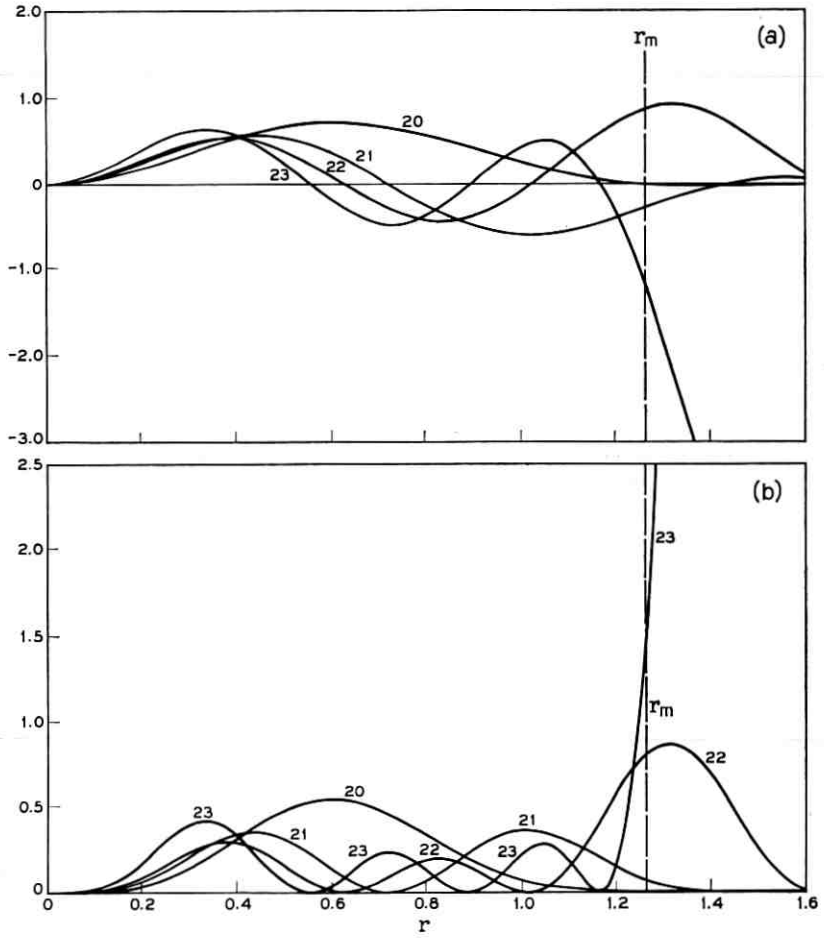


Fig. 7 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (20)$, $(21)$, $(22)$, and $(23)$ with $N_m = 1.6$ and $N_0 = 0$.

integer $p \geqq 0$ orders the eigenmodes of a given angular quantum number $l$ with respect to increasing power loss/pass, then the eigenvalue

$$\kappa_{lp} = (-1)^p \, | \, \kappa_{lp} \, | \tag{18}$$

and the amplitude function has $p$ zeros in the interval $0 < r < r_m$.

If we assume for finite $N_m$ that the low-loss eigenfunctions approximate the limiting expressions (16b), we can estimate the deviation of $\kappa_{lp}$ from its infinite-$N_m$ value (16a) by first-order perturbation theory:

$$1 - (-1)^p \kappa_{lp} \text{ (pert)} = \int_{2\pi N_m}^{\infty} dx \, x^l e^{-x} [L_p{}^l(x)]^2. \tag{19}$$

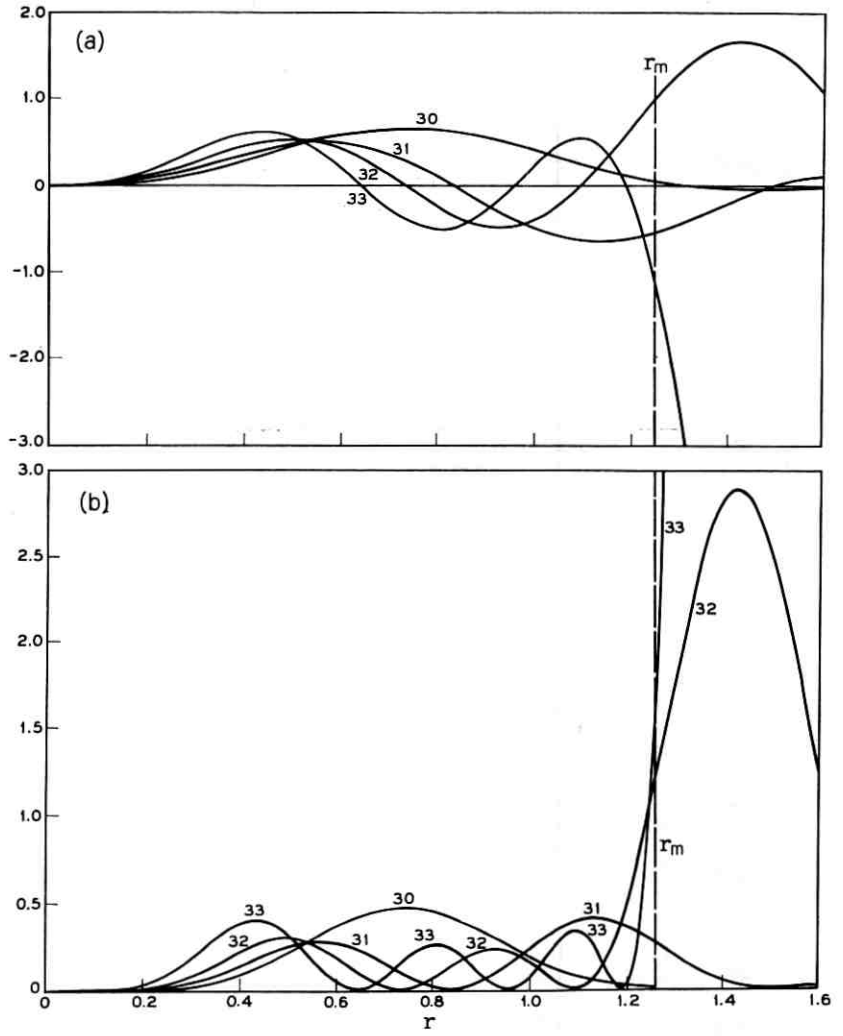


Fig. 8 — (a) Field amplitude $g_{lp}(r)$ and (b) field intensity $g_{lp}^2(r)$ for modes $(lp) = (30)$, $(31)$, $(32)$, and $(33)$ with $N_m = 1.6$ and $N_0 = 0$.

Two special cases are

$$1 - \kappa_{00}\ (\text{pert}) = e^{-2\pi N_m}, \qquad 1 + \kappa_{01}\ (\text{pert}) = [1 + (2\pi N_m)^2]e^{-2\pi N_m}.$$

We compare these estimates with computed values in Table II. The errors between the computed and estimated values in Table II, while small relative to the eigenvalues themselves, are nevertheless significant

TABLE II — DEVIATION OF EIGENVALUES FROM INFINITE-MIRROR VALUES (16b)

| $N_m$* | $1 - \kappa_{00}$ | $1 - \kappa_{00}$(pert)† | $1 - \kappa_{00}$(asym)†† | $1 + \kappa_{01}$ | $1 + \kappa_{01}$(pert)† | $1 + \kappa_{01}$(asym)†† |
|---|---|---|---|---|---|---|
| 0.6 | $1.82 \times 10^{-2}$ | $2.31 \times 10^{-2}$ | $2.52 \times 10^{-2}$ | 0.545 | 0.351 | 5.73 |
| 0.8 | $2.23 \times 10^{-3}$ | $6.56 \times 10^{-3}$ | $2.72 \times 10^{-3}$ | 0.234 | 0.172 | 1.099 |
| 1.0 | $2.38 \times 10^{-4}$ | $1.87 \times 10^{-3}$ | $2.75 \times 10^{-4}$ | $6.37 \times 10^{-2}$ | $7.56 \times 10^{-2}$ | 0.174 |
| 1.2 | $2.38 \times 10^{-5}$ | $5.31 \times 10^{-4}$ | $2.68 \times 10^{-5}$ | $1.20 \times 10^{-2}$ | $3.07 \times 10^{-2}$ | $2.43 \times 10^{-2}$ |
| 1.4 | $2.24 \times 10^{-6}$ | $1.51 \times 10^{-4}$ | $2.53 \times 10^{-6}$ | $1.80 \times 10^{-3}$ | $1.19 \times 10^{-2}$ | $3.13 \times 10^{-3}$ |
| 1.6 | $\sim 2.7 \times 10^{-7}$ ‡ | $4.31 \times 10^{-5}$ | $2.34 \times 10^{-7}$ | $2.40 \times 10^{-4}$ | $4.39 \times 10^{-3}$ | $3.79 \times 10^{-4}$ |

* $N_0 = 0$.
† Estimated values computed from (19) of first-order perturbation theory.
†† Estimated values computed from asymptotic (20).
‡ The accuracy of $\kappa_{00}$ for $N_m = 1.6$ is limited by machine rounding errors.

when compared to the difference $1 - (-1)^p \kappa_{lp} = 1 - |\kappa_{lp}|$ which for these eigenvalues is roughly one-half the power loss/pass (13). The errors arise because the real eigenfunctions are not identical to the limiting expressions (16a). Slepian[5] has derived more accurate asymptotic results appropriate to $N_m$ large:

$$1 - (-1)^p \kappa_{lp}(\text{asym}) = \frac{\pi (8\pi N_m)^{l+2p+1} e^{-4\pi N_m}}{p!(l+p)!} \left[1 + 0\left(\frac{1}{N_m}\right)\right]. \quad (20)$$

Two special cases are

$$1 - \kappa_{00}\,(\text{asym}) = 8\pi^2 N_m e^{-4\pi N_m}; \qquad 1 + \kappa_{01}(\text{asym}) = \pi (8\pi N_m)^3 e^{-4\pi N_m}.$$

Values computed from these expressions are also listed in Table II.

In Table III we have listed values at $r = 0$ of $g_{lp}(r)$ for $(lp) = (00)$, (01), and (02). These values are consistently less than the values predicted from the infinite-$N_m$ functions (16b) renormalized to the finite interval $(0, r_m)$:

$$g_{lp}(r) = \left[\frac{2p!}{(l+p)!}\right]^{\frac{1}{2}} (2\pi r^2)^{l/2} e^{-\pi r^2} L_p^{\,l}(2\pi r^2) \times \left\{1 - \frac{p!}{(l+p)!} \int_{2\pi N_m}^{\infty} dx\, x^l\, e^{-x} [L_p^{\,l}(x)]^2\right\}^{-\frac{1}{2}} \quad (21)$$

for which

$$g_{0p}(0) = 2^{\frac{1}{2}} \left\{1 - \int_{2\pi N_m}^{\infty} dx\, e^{-x} [L_p^{\,0}(x)]^2\right\}^{-\frac{1}{2}}. \quad (22)$$

The differences between the calculated and estimated results again reflect the distortion appropriate to finite $N_m$ of the eigenfunctions (16b).[5] For a given angular quantum number $l$, this distortion is generally less

TABLE III — FIELD AMPLITUDE AT MIRROR CENTER FOR $l = 0$ MODES

| $N_m$* | $g_{00}(0)$ | $g_{01}(0)$ | $g_{02}(0)$ |
|---|---|---|---|
| 0.6 | 1.277 | 1.225 | 1.616 |
| 0.8 | 1.321 | 1.125 | 1.373 |
| 1.0 | 1.346 | 1.151 | 1.193 |
| 1.2 | 1.360 | 1.209 | 1.089 |
| 1.4 | 1.369 | 1.253 | 1.081 |
| 1.6 | 1.375 | 1.281 | 1.130 |
| 2.0 | 1.384 | 1.315 | 1.220 |
| $\infty$ | $\sqrt{2} = 1.414$ | 1.414 | 1.414 |

* $N_0 = 0$.

significant in low-order modes than it is in high-order modes. (The former are usually the only modes relevant to laser oscillators.) That this is so can be understood if we recall that for the cylindrical confocal geometry $g_{lp}(r)$ will have $p$ zeros in the interval $r_0 < r < r_m$. As the interval $(r_0, r_m)$ is compressed within the interval $(0, \infty)$ appropriate to (16a,b), those functions $g_{lp}(r)$ whose zeros initially lay outside $(r_0, r_m)$ will clearly be more distorted by the compression than will those functions which have no zeros ($p = 0$) or those whose zeros lie well within $(r_0, r_m)$. In Table IV we have tabulated the zeros within $(r_0, r_m)$ of several low-loss eigenfunctions for different $N_m$.

## III. EIGENMODES FOR FINITE MIRROR APERTURES ($N_0 \neq 0$)

In this section, in order to distinguish the $N_0 = 0$ and the $N_0 \neq 0$ results, we mark the eigenvalues and eigenfunctions for $N_0 = 0$ by a superscript "0": $\kappa_{lp}^{0}, g_{lp}(r)^{0}$. As in the preceding section, we assign the integer $p \geqq 0$ to the $N_0 = 0$ modes in the order of their increasing power loss/pass: $|\kappa_{lp}^{0}| > |\kappa_{lp+1}^{0}|$. We identify the $N_0 \neq 0$ modes by

TABLE IV — ZEROS OF $g_{lp}(r)$ IN THE DOMAIN $0 = r_0 < r < r_m$

| $N_m$* | $r_m$‡ | $l = 0$ | 0 | | 1 | 1 | | 2 | 2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $p$† = 1 | $2_1$ | $2_2$ | 1 | $2_1$ | $2_2$ | 1 | $2_1$ | $2_2$ |
| 0.6 | 0.775 | 0.434 | 0.318 | 0.662 | 0.573 | 0.432 | 0.699 | 0.641 | 0.502 | 0.718 |
| 0.8 | 0.894 | 0.433 | 0.339 | 0.735 | 0.604 | 0.474 | 0.792 | 0.705 | 0.560 | 0.821 |
| 1.0 | 1.000 | 0.425 | 0.345 | 0.772 | 0.606 | 0.496 | 0.859 | 0.734 | 0.599 | 0.903 |
| 1.2 | 1.095 | 0.418 | 0.342 | 0.780 | 0.598 | 0.504 | 0.897 | 0.737 | 0.621 | 0.965 |
| 1.4 | 1.183 | 0.414 | 0.336 | 0.774 | 0.590 | 0.501 | 0.909 | 0.729 | 0.628 | 1.002 |
| 1.6 | 1.265 | 0.412 | 0.330 | 0.766 | 0.585 | 0.493 | 0.905 | 0.721 | 0.626 | 1.016 |
| $\infty$ | $\infty$ | 0.399 | 0.305 | 0.707 | 0.564 | 0.449 | 0.868 | 0.691 | 0.564 | 0.977 |

* $N_0 = 0$.
† The function $g_{lp}(r)$ has $p$ zeroes in $r_0 < r < r_m$.
‡ $r_m = N_m$‡.

the indices that those modes would carry if they deformed continuously from $N_0 = 0$. As we shall see, it is not necessary that $|\kappa_{lp}| > |\kappa_{lp+1}|$ when $N_0 \neq 0$; however, the $N_0 \neq 0$ modes do have the properties (15) and (18) of the $N_0 = 0$ modes. In addition, the field amplitude $g_{lp}(r)$ will continue to have $p$ zeros in the reflecting interval $r_0 < r < r_m$ ($r_0 = 0$ for $N_0 = 0$).

Using perturbation theory to express the $N_0 \neq 0$ eigenfunctions in terms of the $N_0 = 0$ functions, we have to first order:[12]

$$g_{lp}(r) = g_{lp}(r)^0 \left\{1 + \pi \int_0^{r_0} dr'\, r'[g_{lp}(r')^0]^2\right\} - \sum_{q\neq p}{}' \frac{g_{lq}(r)^0}{\kappa_{lp}{}^0 - \kappa_{lq}{}^0} \kappa_{lq}{}^0 2\pi \int_0^{r_0} dr'\, r' g_{lp}(r')^0 g_{lq}(r')^0. \tag{23}$$

To second order, the eigenvalues are

$$\kappa_{lp} = \kappa_{lp}{}^0 \left\{1 - 2\pi \int_0^{r_0} dr\, r[g_{lp}(r)^0]^2\right\} + \sum_{q\neq p}{}' \frac{\kappa_{lp}{}^0 \kappa_{lq}{}^0}{\kappa_{lp}{}^0 - \kappa_{lq}{}^0} \left[2\pi \int_0^{r_0} dr\, r g_{lp}(r)^0 g_{lq}(r)^0\right]^2. \tag{24}$$

The factor multiplying $g_{lp}(r)^0$ in (23) is a normalization correction compensating for the fact that the $g_{lp}(r)$ are normalized in (10) over the interval $(r_0, r_m)$, whereas the $g_{lp}(r)^0$ are normalized over the larger interval $(0, r_m)$. The first-order correction to the eigenvalue in the first term of (24) decreases the unperturbed eigenvalue $\kappa_{lp}{}^0$ by that fraction of the unperturbed field intensity which falls on the aperture.

The second terms in (23) and (24) describe eigenfunction mixing by the aperture. The amount of mixing depends upon the eigenvalue difference as well as upon the strength of the perturbative coupling. The circular apertures, centered on the resonator axis, do not mix modes with different angular quantum numbers. Because the signs of the eigenvalues alternate as in (18), mode mixing in the symmetric identical-mirror cavity is strongest among modes with the same $p$ parity $(-1)^p$. The situation is somewhat different in resonators with dissimilar mirrors such as obtains in the apparatus of Patel et al.[2] where only one mirror is pierced by the output-coupling aperture. In such systems there is significant mixing between even-$p$ and odd-$p$ modes.[4]

Whereas mode mixing will preclude two eigenvalues from actually crossing (if the two modes are coupled by the perturbation), there is, because of the sign property (18), no such restriction on the absolute values $|\kappa_{lp}|$ and $|\kappa_{lp+1}|$ or, equivalently, on the diffraction losses of the $(lp)$ and $(lp + 1)$ modes. For some special values of $(N_0, N_m)$ one can

in fact reverse the power-loss progression of the $N_0 = 0$ case to give $|\kappa_{lp+1}| > |\kappa_{lp}|$; however, it is always true for the identical-mirror cylindrical confocal system that

$$|\kappa_{lp}| > |\kappa_{lp+2}| . \tag{25}$$

In more general geometries with more complicated eigenvalue phase relations than (18) even this restricted condition can be violated.

Using the kernel expansion-truncation method outlined in the Appendix, we have computed the effects of finite coupling apertures on the properties of cavity eigenmodes. In Fig. 9 we have indicated for a Fresnel number $N_m = 0.8$ how a finite coupling aperture with Fresnel number $N_0 \neq 0$ affects the loss/pass of the lowest-order modes. In the confocal geometry the finite aperture affects only the magnitude of the eigenvalues; their signs (phases) are still given by (18). In no case do the eigenvalues belonging to the same angular quantum number $l$ cross

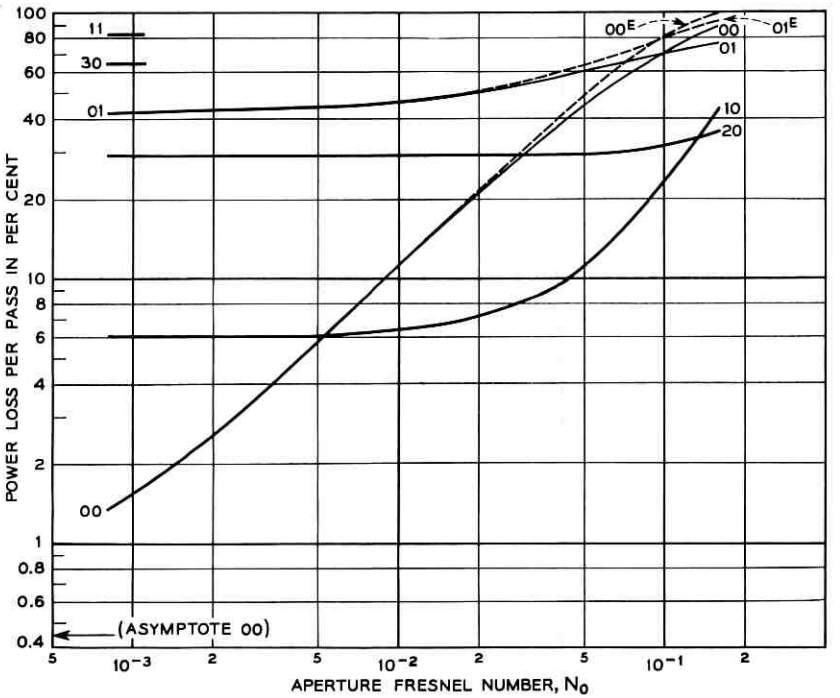


Fig. 9 — Power loss/pass versus aperture Fresnel number $N_0$ for low-loss modes with $N_m = 0.8$. Dashed curves ($00^E$) and ($01^E$) are estimates based on (26b).

(the circular-aperture perturbations do not couple modes with different $l$), although in Fig. 9 we do see for the (00) and (01) modes a reversal of the $N_0 = 0$ sequence $|\kappa_{00}| > |\kappa_{01}|$ in the interval $0.08 < N_0 < 0.16$.

Fig. 9 confirms our previous conjecture that the modes with angular quantum number $l = 0$ are more sensitive to deleterious aperture loss than are the $l \neq 0$ modes. When for $N_m = 0.8$ the area of the aperture is only 0.6 per cent of the total mirror area ($N_0/N_m = 0.006$), the loss/pass of the (00) mode has increased to the point where it equals the loss/pass of the (10) mode. For this same hole size the losses of the (10) mode are virtually unaffected by the aperture.

In Fig. 10 we have indicated the intensity distribution of the (00) mode for $N_m = 0.8$ and $N_0 = 0, 0.01$. Except for the normalization correction implicit in the first term of (23), the intensity distribution for $N_0 = 0.01$ is nearly identical to that for $N_0 = 0$. We conclude for $N_m = 0.8$ and $N_0 \lesssim 0.01$ that eigenfunction mixing in (23) is unimportant and, as a consequence, that the eigenvalues are accurately given by the first-order term of (24). Using the infinite-$N_m$ functions (16b) to approximate $g_{lp}(r)^0$ in the first term of (24), we can estimate the ratio $\kappa_{lp}/\kappa_{lp}^0$ analytically. For $l = 0$ this estimate can be considerably improved if we renormalize the infinite-$N_m$ functions (16b) by the factor $g_{0p}(0)/\sqrt{2}$ computed from Table III. Doing this, we estimate

$$\kappa_{0p} = \kappa_{0p}^0 \left\{ 1 - \frac{1}{2} [g_{0p}(0)^0]^2 \int_0^{2\pi N_0} dx\, e^{-x} [L_p^0(x)]^2 \right\}. \qquad (26a)$$

where $\kappa_{lp}^0$ and $g_{lp}(0)^0$ are implicitly dependent upon $N_m$. In the limit

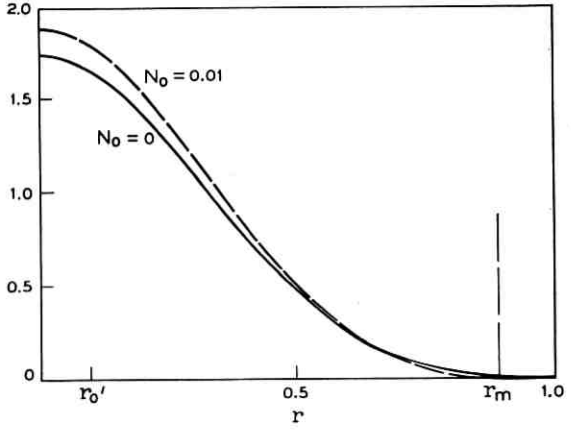


Fig. 10 — Field intensity $g_{lp}^2(r)$ of the mode $(lp) = (00)$ for $N_m = 0.8$ with $N_0 = 0$ (solid curve) and $N_0 = 0.01$ (dashed curve). ($N_m \equiv r_m^2$; $N_0 \equiv r_0^2$.)

of small $N_0$, for which we can in effect replace $g_{lp}(r)^0$ by $g_{lp}(0)^0$ in the right-hand side of (24), this gives

$$\kappa_{0p} = \kappa_{0p}{}^0\{1 - \pi N_0[g_{0p}(0)^0]^2\}. \tag{26b}$$

To the same accuracy, the $r^l$ dependence of $g_{lp}(r)$ noted in (15) implies that $\kappa_{lp} = \kappa_{lp}{}^0$ for $l \neq 0$. The approximation (26b) has been used to compute the dashed curves in Fig. 9. For $N_0 \lesssim 0.02$ the fit to the machine-computed curves is excellent.

The effect of a finite aperture ($N_0 \neq 0$) on the losses of the low-loss modes for $N_m = 1.6$ is indicated in Fig. 11. More modes are shown than

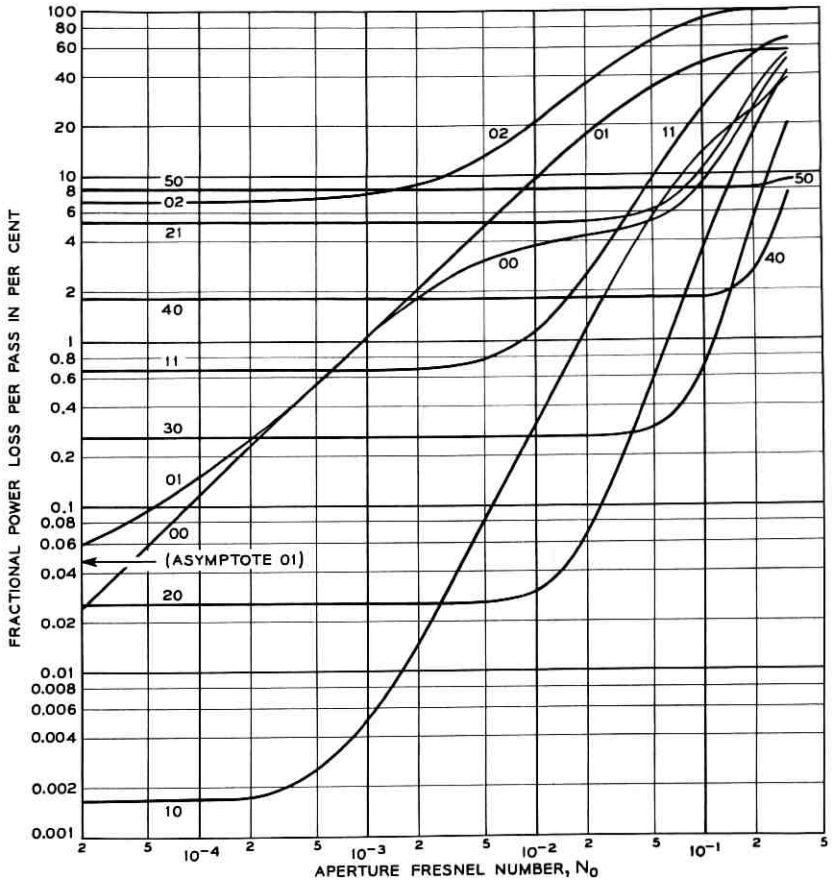


Fig. 11 — Power loss/pass versus aperture Fresnel number $N_0$ for low-loss modes with $N_m = 1.6$.

in Fig. 9 because at the larger Fresnel number more modes have low loss (cf. Fig. 2). Observe for any fixed radial quantum number $p$ that, consistent with the $r^l$ behavior noted in (15), modes with low angular quantum number $l$ are more sensitive to a small aperture than are the modes with higher angular quantum number. The sequence in Fig. 11 of upward breaks in the losses of the modes (00), (10), (20), $\cdots$ is particularly striking, as is that for the modes (01), (11), (21), $\cdots$.

In Fig. 12 we have redrawn those curves of Fig. 11 which pertain to the angular-invariant $l = 0$ modes. The dashed curves in Fig. 12 derive from the approximation (26b), which is here valid only for $N_0 \lesssim 0.0005$ in the (00) and (02) modes and for $N_0 \lesssim 0.006$ in the (01) mode. [For $N_m = 0.8$ it applied to all $N_0 \lesssim 0.02$.] The approximation (26b), based upon first-order perturbation theory, fails when eigenfunction mixing becomes significant. Mixing is strong for the (00) mode when the losses of that mode due to the finite aperture approximate the edge

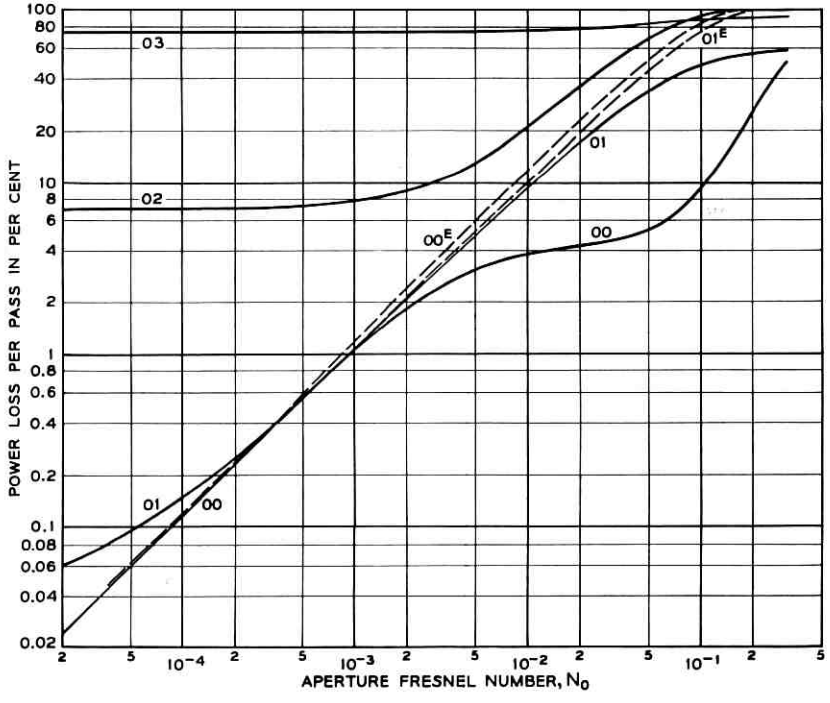


Fig. 12 — Power loss/pass versus aperture Fresnel number $N_0$ for the low-loss $l = 0$ modes with $N_m = 1.6$. Dashed curves ($00^E$) and ($01^E$) are estimates based on (26b).

losses of the (02) mode — that is, when $\kappa_{00} \approx \kappa_{02}$. If we recall that a variational principle applies to the present eigenvalue problem [because, in contradistinction to more general resonator kernels, the kernel in the integral equation (9) is Hermitian], we can view mode mixing as an attempt by the field in the low-loss (00) mode to reduce its intensity at the aperture and to reduce thereby the total (00) loss. Because edge as well as aperture losses contribute to the total loss, the deleterious edge losses of the (02) mode preclude appreciable (00)-(02) mixing until the aperture losses of the undistorted (00) mode approximate the edge losses of the (02) mode. Because edge losses decrease rapidly with increasing $N_m$ ($N_m = 1.6$ is already quite large), the aperture losses required for appreciable mixing decrease rapidly with increasing $N_m$. The relevance of mode mixing to the breakdown of (26b) is clearly illustrated in Fig. 13, where we show the intensity distribution of the (00) mode for $N_m = 1.6$ and $N_0 = 0$, 0.01, and 0.02. [The same aperture Fresnel numbers $N_0$ gave insignificant mode distortion for $N_m = 0.8$ (Fig. 10).]

In Fig. 14 is shown the intensity distribution for three other low-order $l = 0$ modes besides the (00) mode for $N_m = 1.6$ and $N_0 = 0.01$. This figure should be compared with Fig. 5b, which shows the intensity distribution of the same modes for $N_m = 1.6$ and $N_0 = 0$. Note that, whereas the intensity at $r = 0$ of the (00) mode decreased as a result of aperture mode mixing, the intensity at $r = 0$ of the (02) mode increased. This increase is reflected in Fig. 12 in the sharp rise of (02) losses as the (00) and (02) eigenvalues "repel" for $N_0 \gtrsim 0.003$.

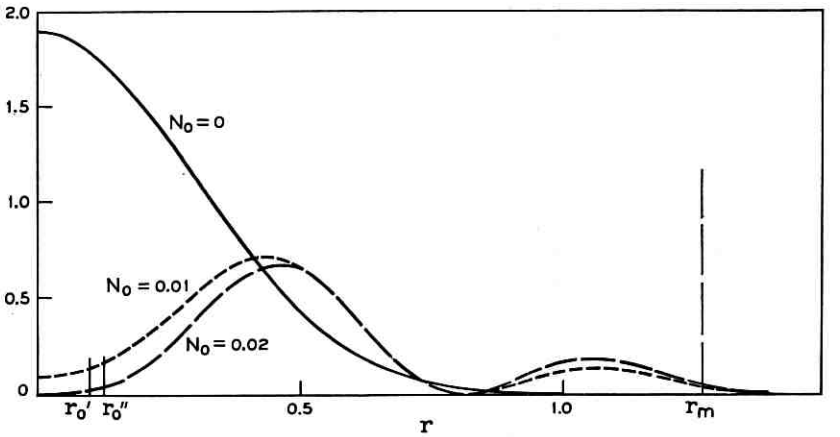


Fig. 13 — Field intensity $g_{lp}^2(r)$ of the mode $(lp) = (00)$ for $N_m = 1.6$ with $N_0 = 0$ (solid curve) and $N_0 = 0.01$, 0.02 (dashed curves). ($N_m \equiv r_m^2$; $N_0 \equiv r_0^2$.)

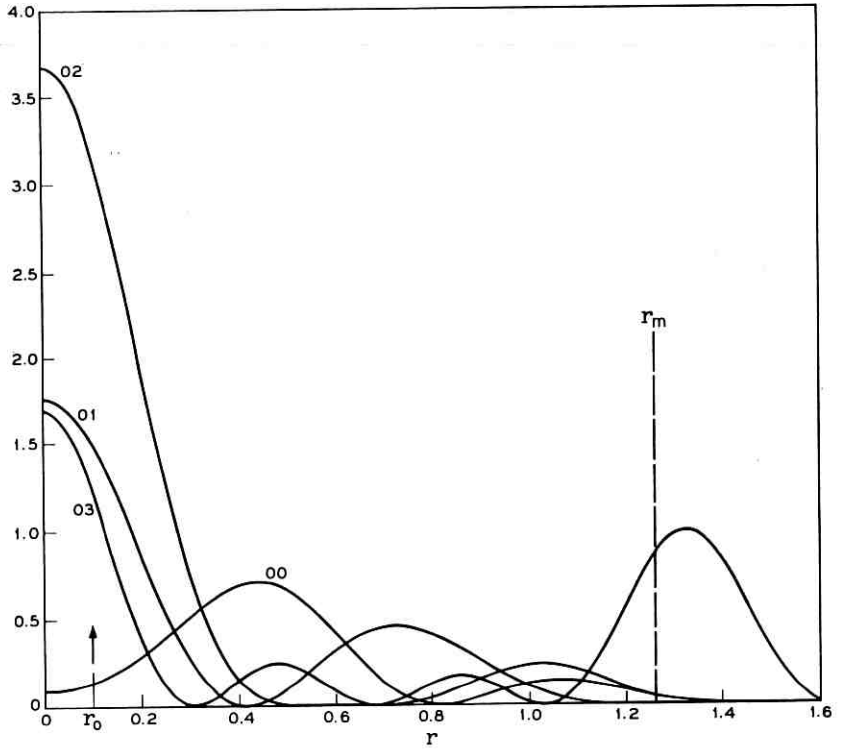


Fig. 14 — Field intensity $g_{lp}^2(r)$ of the low-loss $l = 0$ modes for $N_m = 1.6$ and $N_0 = 0.01$. ($N_m \equiv r_m^2$; $N_0 \equiv r_0^2$.) Notice how mode mixing has changed the intensity distribution near the aperture ($0 \leqq r < r_0$) from that in Fig. 5(b) where $N_0 = 0$.

In Figs. 15, 16, and 17 we have indicated how the power loss/pass of the low-loss modes varies when for fixed aperture size $N_0$ the Fresnel number $N_m$ changes. Notice in the typical Fig. 15 that the losses of the (00) mode decrease as $N_m$ increases from $N_0$ until for $N_m \approx 0.7$ those losses saturate at about 11 per cent/pass, approximately the loss predicted from (26b) with $g_{00}(0)^0 = \sqrt{2}$ and $\kappa_{00}{}^0 = 1$. As $N_m$ increases beyond 1.3, mode mixing reduces the losses of the (00) mode as the modified intensity distribution avoids both the aperture and the reflector edges. While by $N_m = 1.6$ the (00) mode again has the lowest loss of the $l = 0$ modes, its total loss is greater than that of certain $l \neq 0$ modes and its intensity distribution (Fig. 13) is considerably different from the simple Gaussian of (16b).

A quantity of interest in the design of lasers with aperture output

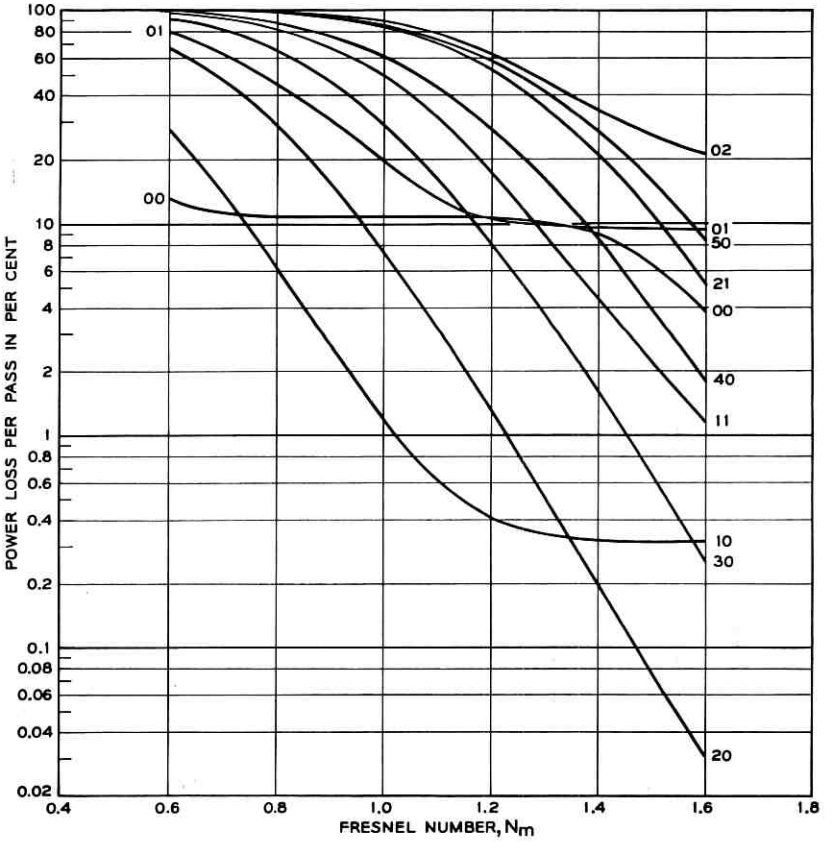


Fig. 15 — Power loss/pass of low-loss modes versus Fresnel number $N_m$ for aperture Fresnel number $N_0 = 0.01$.

coupling[2] is that value $N_{0c}$ of $N_0$ for which the losses of the (00) mode equal the losses of the (10) mode. All other things being equal, the laser will oscillate in the (00) mode for $N_0 < N_{0c}$, whereas for $N_0 > N_{0c}$ it will operate in the (10) mode or, for large values of $N_0$, in still another mode (cf. Fig. 11). In Fig. 18 we have plotted $N_{0c}$ as a function of $N_m$. For $N_m > 0.6$ this curve can be accurately reproduced by the following expression based upon (26b):

$$N_{0c} = (\kappa_{00}^{0} - \kappa_{10}^{0})/\pi\kappa_{00}^{0}[g_{00}(0)^{0}]^2. \tag{27}$$

This result obtains even for large $N_m$ because $N_{0c}$ decreases so rapidly with increasing $N_m$ that mode mixing is never relevant.

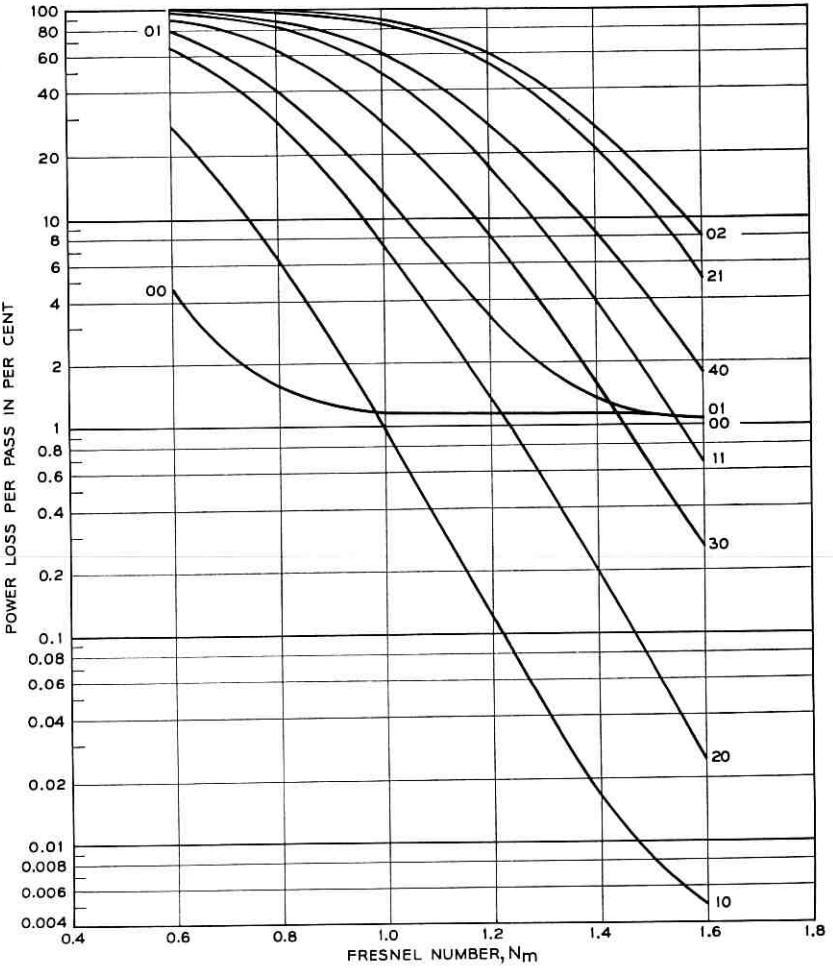


Fig. 16 — Power loss/pass of low-loss modes versus Fresnel number $N_m$ for aperture Fresnel number $N_0 = 0.001$.

## IV. FAR-FIELD PATTERNS, APERTURE OUTPUT COUPLING

If we assume that the useful output coupling of the mode $(lp)$ is exclusively through one of the small reflector apertures of Fresnel number $N_0$, then at a large distance $d$ from the relevant output aperture and in a direction making an angle $\theta$ with the cavity axis (see Fig. 19) the field amplitude will be proportional to

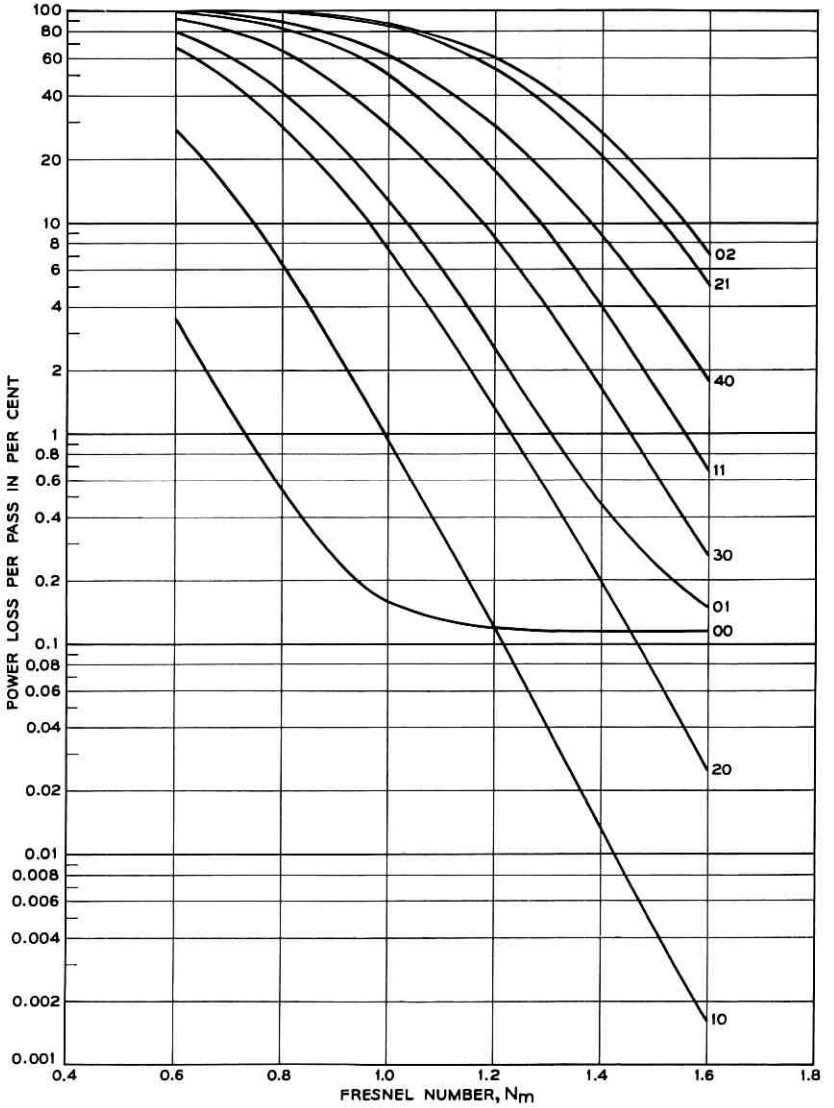


Fig. 17 — Power loss/pass of low-loss modes versus Fresnel number $N_m$ for aperture Fresnel number $N_0 = 0.0001$.

$$A_{lp}(\theta,\varphi) = e^{-il\varphi}\frac{2\pi b}{d}\int_0^{r_0} dr\, rJ_l[2\pi r(b/\lambda)^{\frac{1}{2}}\sin\theta]g_{lp}(r). \quad (28)$$

Here $\varphi$ is the azimuthal angle used in (1); $r$ is the radial variable defined in (6); and $g_{lp}(r)$ is the mirror-field amplitude function (8). The derivation of (28) from the Fraunhofer formula parallels that of (2) and (9).[1] The basic approximation used is

$$[d^2 + \rho^2 - 2\rho d\sin\theta\cos\varphi\,]^{\frac{1}{2}} \approx d - \rho\sin\theta\cos\varphi. \quad (29)$$

This approximation is suitable when $d \gg a_0 \geqq \rho$ and $a_0^2/\lambda d = bN_0/d \ll 2$.

In the important case for which $r_0$ is so small ($N_0 \lesssim N_{0c}$ is sufficient)

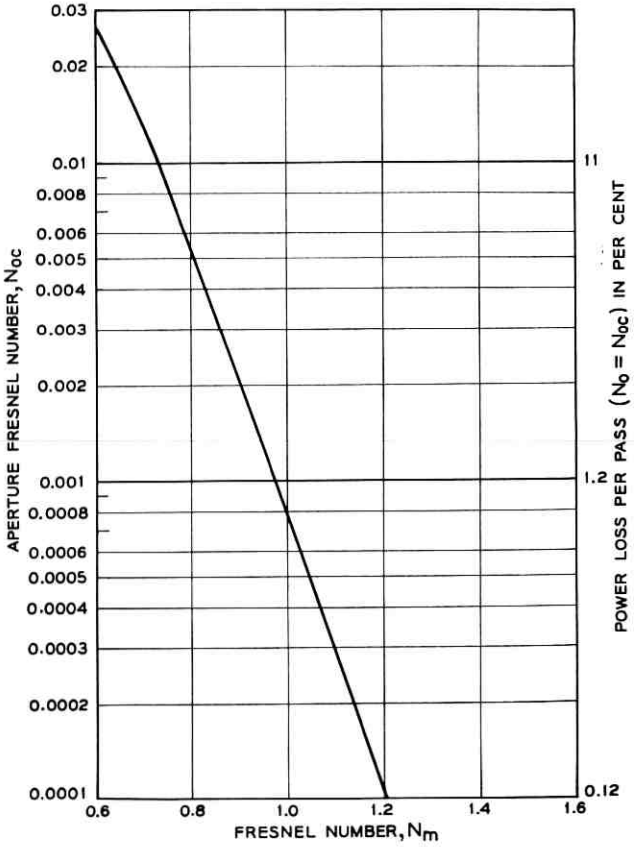


Fig. 18 — Critical aperture Fresnel number $N_{0c}$ for which diffraction losses of (00) mode equal those of (10) mode versus Fresnel number $N_m$. Also shown for each Fresnel number $N_m$ is the loss/pass when $N_0 = N_{0c}$. This loss is approximately, but not exactly, proportional to $N_{0c}$.

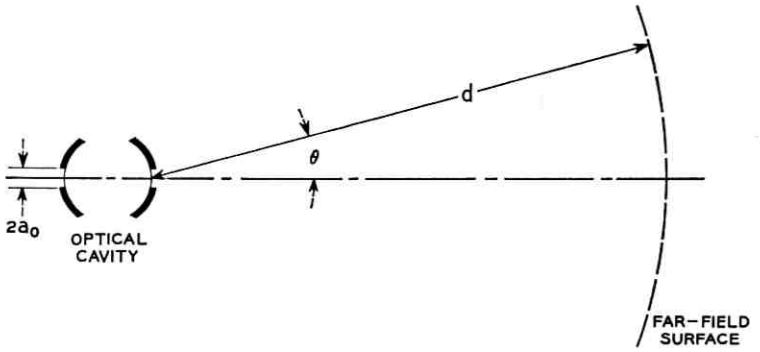


Fig. 19 — Geometry appropriate to the analysis of the far-field pattern for aperture output coupling of cavity.

that we can approximate $g_{lp}(r)$ by the lowest-order term (15), we obtain from (28):

$$A_{lp}(\theta,\varphi) = \frac{1}{l!}\frac{d^l g_{lp}}{dr^l}\bigg|_{r=0} e^{-il\varphi}\frac{a_0 N_0^{\,l/2}}{d\sin\theta} J_{l+1}\left(\frac{2\pi a_0 \sin\theta}{\lambda}\right). \tag{30}$$

The observed intensity pattern for a pure $l$ mode will be proportional to

$$I_{lp}(\theta,\varphi) = \left[\frac{1}{l!}\frac{d^l g_{lp}}{dr^l}\bigg|_{r=0}\frac{a_0 N_0^{\,l/2}}{d\sin\theta}\right]^2 \cos^2[l(\varphi-\delta_l)] J_{l+1}{}^2\left(\frac{2\pi a_0\sin\theta}{\lambda}\right), \tag{31}$$

where $\delta_l$ is an appropriate phase angle. For $l = 0$ this gives the celebrated Airy diffraction pattern.[13]

## V. SUMMARY REMARKS

We have computed the eigenmodes of a symmetric cylindrical confocal laser cavity of Fresnel number $N_m \leqq 2.0$ and have determined how those modes change when a small circular element centered on the axis is removed from each reflector. The calculation methods can easily be adapted to cylindrical confocal resonators for which the two mirrors and mirror apertures have different sizes. (Results relevant to a coupling aperture in only one end reflector will be published in a subsequent article.) The basic expansion-truncation methods outlined in Section I and in the Appendix are quite general[10] and can usefully be applied to nonconfocal geometries for which complex-number computations are required when the mirror surfaces are not surfaces of constant phase.

For the cylindrical confocal geometry the results reported above indicate that, while the infinite-$N_m$ functions with appropriate normalization do approximate the low-order finite-$N_m$ eigenfunctions,

there are significant differences which influence the calculation of both aperture and edge diffraction losses. For sufficiently small aperture Fresnel numbers $N_0$ the aperture diffraction losses can be estimated by first-order perturbation theory based upon the finite-$N_m$ eigenfunctions (or upon the infinite-$N_m$ functions renormalized to the finite-$N_m$ amplitude at $r = 0$). The value of $N_0$ for which such first-order calculations are valid decreases rapidly as the Fresnel number $N_m$ increases, because for large apertures the field distributions distort (higher-order perturbation theory) to avoid the aperture. This distortion occurs at approximately those values of $N_0$ and $N_m$ for which an observer at one reflector, using light of the relevant wavelength and optics limited by the radius $r_m = N_m^{\frac{1}{2}}$, can resolve the aperture of radius $r_0 = N_0^{\frac{1}{2}}$ at the opposite reflector.[14]

## VI. ACKNOWLEDGMENTS

I am grateful to W. L. Faust, C. G. B. Garrett, and R. A. McFarlane for suggesting the calculation of properties of perturbed cylindrical cavities, for discussions regarding the physical relevance of the results, and for comments concerning the preceding presentation. I am also grateful to T. Li and D. Slepian for references to the published literature and for several discussions correlating these results and methods with those of previous calculations.

## APPENDIX

### *Reduction of Integral Equation* (9) *to a Matrix Equation*

Truncating the series (13) after $M$ terms and substituting the result into (9), we obtain[15] ($l = |l|$ in this Appendix)

$$\begin{aligned}\kappa_{lp} g_{lp}(r) &= 2\pi \int_{r_0}^{r_m} dr'\, r' \sum_{m=1}^{M} \frac{(-1)^{m-1}(rr'\pi)^{l+2(m-1)}}{(m+l-1)!(m-1)!}\, g_{lp}(r') \\ &= 2\pi \sum_{m=1}^{M} \frac{(-1)^{m-1}(\pi r)^{l+2(m-1)}}{(m+l-1)!(m-1)!} \\ &\quad \cdot \int_0^{r_m} dr'(r')^{l+2m-1}\, g_{lp}(r') \\ &= \left[\frac{(\pi r^2)^l}{l!}\right]^{\frac{1}{2}} \sum_{m=1}^{M} \frac{(-1)^{m-1}(\pi r^2)^{m-1}}{[(m-1)!(m+l-1)!/l!]^{\frac{1}{2}}} G_m(lp),\end{aligned} \tag{32}$$

where

$$G_m(lp) \equiv \frac{2\pi}{[(m+l-1)!(m-1)!]^{\frac{1}{2}}} \cdot \int_{r_0}^{r_m} dr'\, r'(\pi r'^2)^{m-1+l/2} g_{lp}(r'). \tag{33}$$

Solving (32) for $g_{lp}(r)$ and substituting this expression into (33), we obtain after simple manipulations

$$\kappa_{lp} G_m(lp) = \sum_{k=1}^{M} \frac{(-1)^{k-1}}{[(m-1)!(m+l-1)!(k-1)!(k+l-1)!]^{\frac{1}{2}}} \times \frac{[(\pi N_m)^{l+m+k-1} - (\pi N_0)^{l+m+k-1}]}{(l+m+k-1)} G_k(lp). \tag{34}$$

We have used the definitions (7) to replace $(r_0^2, r_m^2)$ by the Fresnel numbers $(N_0, N_m)$.

Equation (34) is a matrix equation which must be solved for the eigenvalue $\kappa_{lp}$ and for the $M$ vector components $G_m(lp)$ appropriate to that eigenvalue. When these latter components are used in (32), one obtains the eigenfunction $g_{lp}(r)$ appropriate to the $M$-term truncation of (13). The normalization condition (10) on $g_{lp}(r)$ is equivalent to the condition

$$\kappa_{lp}\delta_{pq} = \sum_{m=1}^{M} (-1)^{m-1} G_m(lp) G_m(lq) \tag{35}$$

on the real vector components $G_m(lp)$. The sign condition Re $g_{lp}(0^+) > 0$ becomes $G_1(lp) \geqq 0$ where, if $G_1(lp) = 0$, $G_2(lp) \leqq 0$, etc.

In programming the above equations for electronic-computer solution, one must insure that at each stage the computations maintain sufficient numerical accuracy. The relevance of this remark is clearly evident from the fact that, while the Bessel function $J_l(z)$ is of order unity for all real $z \geqq 0$, some terms of the series (13) will for $z \gg 1$ be of order $(e/2)^{2z}/2\pi z \gg 1$. That is, $J_l(z)$ will be the small difference of large numbers and care must be taken to insure that such small differences are accurately represented.

The program utilized to compute the results reported in this paper requires a nominal 0.0042 hr. of IBM 7094 running time to compute the $M$ different eigenvalues and eigenvectors of (34) for $M = 20$. Timing for other values of $M$ varies roughly as $M^3$.

REFERENCES

1. Fox, A. G., and Li, T., B.S.T.J., *40*, 1961, p. 453.
2. Patel, C. K. N., Faust, W. L., McFarlane, R. A., and Garrett, C. G. B., Appl. Phys. Letters, *4*, 1964, p. 18.

3. Boyd, G. D., and Gordon, J. P., B.S.T.J., *40*, 1961, p. 489.
4. Boyd, G. D., and Kogelnik, H., B.S.T.J., *41*, 1962, p. 1347.
5. Slepian, D., B.S.T.J., *43*, Nov., 1964, p. 3009.
6. Kogelnik, H., to be published in *Advances in Lasers*, ed. A. K. Levine.
7. Fox, A. G., Li, T., and Morgan, S. P., Appl. Opt., *2*, 1963, p. 544; Morgan, S. P., IEEE Trans. Microwave Theory and Techniques, *MTT-11*, 1963, p. 191; Kaplan, S., and Morgan, S. P., IEEE Trans. Microwave Theory and Techniques, *MTT-12*, 1964, p. 254; Newman, D. J., and Morgan, S. P., B.S.T.J., *43*, 1964, p. 113; and Cochran, J. Alan, B.S.T.J., *44*, Jan., 1965, p. 77.
8. She, C. Y., and Heffner, H., Appl. Opt., *3*, 1964, p. 703.
9. Bergstein, L., and Schachter, H., J. Opt. Soc. Am., *54*, 1964, p. 887. Cf. comments on this paper by Morgan, S. P., and Li, T., to be publ. in J. Opt. Soc. Am.
10. Goubau, G., and Schwering, F., IEEE Trans. Antennas and Propagation, *AP-9*, 1961, p. 248; Beyer, J. B., and Scheibe, E. H., IEEE Trans. Antennas and Propagation, *AP-10*, 1962, p. 349.
11. Erdelyi, A., et al., *Higher Transcendental Functions*, Vol. 2, New York, McGraw-Hill Book Co., 1953, pp. 188–192.
12. Li, T., to be published; cf. also Gloge, D., Arch. elektrischen Übertragung, *18*, 1964, p. 197.
13. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, pp. 394–397.
14. Faust, W. L., private communication.
15. Tricomi, F. G., *Integral Equations*, Interscience Publishers, Inc., New York, 1957, p. 55 ff.

# Contributors to This Issue

A. A. Bergh, M.S. (Phys. Chem.), 1952, University of Szeged, Hungary; Ph.D. (Phys. Chem.), 1959, University of Pennsylvania; Bell Telephone Laboratories, 1959—. He was first engaged in process development and surface studies in the applied chemistry area. His work has included studies on metal semiconductor and oxide semiconductor interfaces, and on the effects of atomic hydrogen on surfaces and semiconductor devices. He is presently responsible for a group concerned with planar transistor technology and npn and pnpn transistor development.

O. E. De Lange, B.S. in E.E., 1930, University of Utah; M.A. (Physics), 1937, Columbia University; Bell Telephone Laboratories, 1930—. He was involved in studies of FM up to the start of World War II. The war years were spent on development and design of naval radar. The following period was devoted to studies of broadband pulse systems with emphasis on PCM. He was responsible for the satellite tracking radar employed at the Holmdel, N. J., Bell Laboratories for the Echo I Experiment. Recent years have been devoted to studies of light propagation and light transmission systems. Senior member, IEEE.

Richard W. Hamming, B.S., University of Chicago, 1937; M.A., University of Nebraska, 1939; Ph.D., University of Illinois, 1942. Mathematics Instructor, University of Illinois, 1942–44; Professor, University of Louisville, 1944–45; Member of Staff at Los Alamos, 1945–46. In 1946, he joined Bell Telephone Laboratories, where he works in the general area of computing and computing machines. Member, ACM, SIAM, AMS, MAA, AAAS and IEEE.

R. D. Heidenreich, B.S., 1938, M.S., 1940, Case Institute of Technology; Bell Telephone Laboratories, 1945—. His work has been chiefly in the areas of electron microscopy and electron diffraction. He developed the thin metal section methods for transmission electron microscopy now widely used for studying defects in solids. His early application of electron methods to semiconductors resulted in chemical polishing techniques and long surface lifetime treatments for germanium. He has

conducted extensive joint research programs on magnetic materials which have correlated structure with magnetic anisotropy in both hard and soft permanent magnets. His more recent theoretical studies concerning elastic and inelastic scattering of electrons has led to his present interest in high-resolution electron imaging aimed toward resolving atomic configurations. Member, AAAS; Fellow, American Physical Society; Past President, Electron Microscope Society of America.

Ludwik Kurz, B.E.E., 1951, M.E.E., 1955, College of the City of New York; Eng. Sc. D., 1961, New York University. He taught graduate and undergraduate courses at City College between 1951 and 1959. He was awarded the National Science Foundation Science Faculty Fellowship for the period 1959–1961. Since 1961, he has been a member of the faculty of New York University, where he is currently associate professor of electrical engineering.

Jessie MacWilliams, B.A., 1939, M.A., 1941, Cambridge, England; Ph.D., 1962, Harvard University; Bell Telephone Laboratories 1956—. Mrs. MacWilliams has been concerned with network analysis and synthesis, and with data systems studies, particularly in the field of error control. She is at present on leave of absence as a visiting professor in the department of mathematics of Cambridge University, England. Member, Mathematical Association of America and American Mathematical Society.

D. E. McCumber, B.E., 1952, and M.E., 1955, Yale University; A.M., 1956, and Ph.D., 1960, Harvard University; National Science Foundation Postdoctoral Fellow 1959–61; Bell Telephone Laboratories 1961—. His research as a theoretical physicist has been concerned with the analysis of the optical spectra of impurities in solids and with features of optical masers. Member, American Physical Society.

Michael Rappeport, B.S., 1957, Rensselaer Polytechnic Institute; M.E.E., 1958, Yale University; Bell Telephone Laboratories, 1959—. Since joining Bell Laboratories, he has worked on various analytical approaches to data transmission systems, stressing simulation approaches to studying such systems. Member, IEEE, and Institute of Mathematical Statistics.

Burton R. Saltzberg, B.E.E., 1954, New York University; M.S., 1955, University of Wisconsin; Eng. Sc. D., New York University,

1964; Bell Telephone Laboratories, 1957—. He has been engaged in the development of digital data communications systems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

D. T. YOUNG, B.S., 1956, M.E.E., 1960, University of Oklahoma; Bell Telephone Laboratories, 1960—. He initially worked on mode conversion problems in multimode waveguide. At present he is working on a solid-state repeater for a waveguide transmission system. Member, IEEE, Tau Beta Pi, Eta Kappa Nu and Sigma Xi.

# B.S.T.J. BRIEFS

## A Silicon Diode Microwave Oscillator

By R. L. JOHNSTON, B. C. DE LOACH, Jr., and B. G. COHEN

(Manuscript received December 28, 1964)

Microwave oscillations have been obtained on a pulse basis from silicon diodes. This brief reports fabrication details and performance data. The similarities of this device to that proposed by Read[1] are discussed.

The diodes were made by diffusing boron to a depth of approximately 25 $\mu$ into one face of a slice of 0.15 ohm-cm n-type silicon, and then lapping the other face to a final slice thickness of 175 $\mu$. Electroless nickel was then applied to both faces and sintered at 800°C for 5 minutes in $N_2$. Nickel was then replated and followed by a final plate of gold. The slice was then ultrasonically sectioned into squares 125 $\mu$ × 125 $\mu$. To remove cutting damage the "wafers" were etched for about 10 seconds in CP-8 (3 parts $HNO_3$, 1 part HF), and were then incorporated into a microwave encapsulation. A sketch of a wafer before encapsulation is shown in Fig. 1.

Oscillations are observed when a critical reverse voltage is applied to the diode. This voltage has been observed on a variety of samples to correspond to that required to produce enough reverse current to create an electric field on the order of 2 kv/cm in the 150-$\mu$ n-type region. A typical reverse $V$-$I$ characteristic obtained on a pulse basis using a sampling oscilloscope is shown in Fig. 1. On samples which were lapped to reduce the drift region length from 150 $\mu$ to 75 $\mu$, the required voltage in excess of the avalanche voltage was halved. Voltages in excess of threshold produce more output until a maximum is reached. Some lower frequencies in the 1-2-gc region exhibited several maxima, but the higher frequencies (12 and 24 gc) had but one.

The fact that voltages considerably in excess of the "breakdown" voltage are employed tends to deemphasize the role of microplasmas and nonuniformities in the junction and thus contributes to the ease of fabrication of these devices.

The diodes were tested in either a coaxial system, for the lower frequencies ($f \lesssim 12$ gc), or in a reduced-height waveguide for the higher frequencies (8–24 gc). The mounts incorporated a bypass capacitor which allowed the introduction of a video pulse to power the diode.

Microwave oscillations (when present) were coupled to a spectrum analyzer.

The operation of a particular sample of the geometry shown in Fig. 1 will be described. It was placed in the coaxial circuit and driven with a 2 $\mu$sec pulse at a 10-kc repetition rate. When the applied pulse voltage reached a critical value, microwave power output was observed. Power was obtainable over a wide range of frequencies, but some frequencies had particularly high amplitudes. A plot of some of these "high points" is presented in Fig. 2. Included in this figure are two points obtained from a similar sample operated in the waveguide mount. Fig. 2 can also be interpreted as a rough plot of efficiency versus frequency, since pulse powers between 15 and 30 watts were employed for all these points. The 80 mw obtained at 12 gc represents 0.5 per cent efficiency. Similar samples have been operated with duty cycles of 25 per cent (to burnout).

The higher-frequency operation for which the efficiencies are on the order of 0.5 per cent is most likely an oscillation involving primarily the space charge depletion width for the drift space. The extent of this region is of the order of that predicted by Read for this frequency of operation. The requirement of 2 kv/cm in the 150-$\mu$ region for this operation most likely assures that fields greater than this exist across

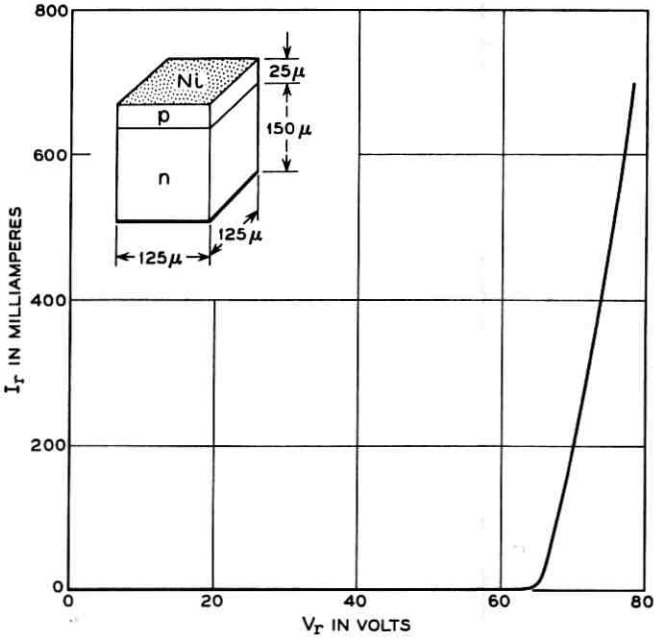


Fig. 1 — A typical sample and its $V$-$I$ curve.

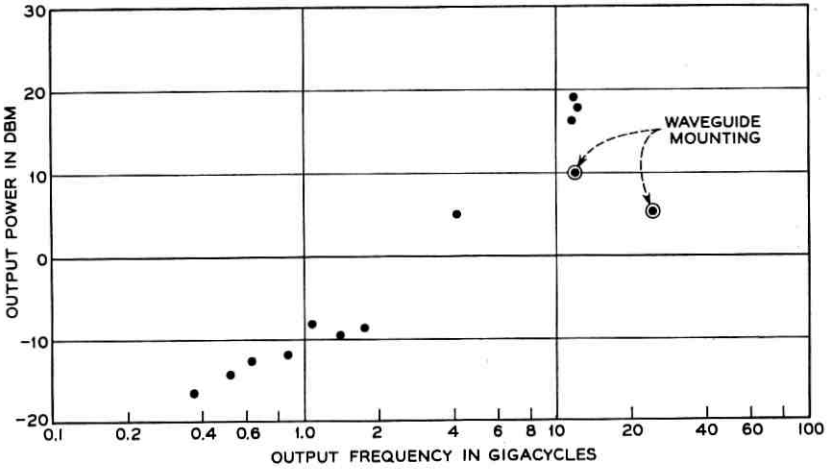


Fig. 2 — Output power vs frequency for a diode in a coaxial circuit and one in a rectangular waveguide circuit.

the depletion region. This added length should then be removable without changing the character of these oscillations and indeed should enhance the efficiency of operation. The circuitry employed in obtaining the higher-frequency microwave power was crude, and thus the power obtained in the X and K band regions (12 and 24 gc) should be taken as a poor lower limit of that available.

We are uncertain at present as to the explanation for the lower-frequency points exhibited in Fig. 2. The 2 kv/cm field required is high enough that the mobility has decreased in the 150-$\mu$ region. The usual dielectric relaxation time of this material of $10^{-13}$ sec is thus increased and some bunching of charge is preserved in the region, with charge transport ($1 \times 10^7$ cm/sec) becoming significant. This "stiffening" of the conductive region could allow it to function as a drift region in the same manner as does the swept region in the above. The approximately linear increase in output power with frequency in this region could then be due to a redistribution of ac field between the space charge depletion layer capacity and the ac impedance of the drift region. That this mechanism is not very effectual can be deduced from the efficiencies of some $1 \times 10^{-3}$ per cent in the region up to 2 gc.

Many helpful discussions with R. M. Ryder and J. C. Irvin are gratefully acknowledged.

*Note Added in Proof:*

Subsequent to the pulsed microwave operations described in this

Brief, Lee et al.[2] have obtained low-frequency cw operation in a silicon diode with a $np\pi p$ structure closely approximating the structures described by Read.[1] Still more recently, continuous microwave oscillations have been obtained by Johnston and De Loach[3] in structures similar to those described herein.

REFERENCES

1. Read, W. T., A Proposed High-Frequency Negative-Resistance Diode, B.S.T.J., *37*, pp. 401–446.
2. Lee, C. A., Batdorf, R. L., Wiegmann, W., and Kaminsky, G., to be published.
3. Johnston, R. L., and De Loach, B. C., to be published.