THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 July–August 1974 Number 6

Scattering of a Plane Electromagnetic Wave by Axisymmetric Raindrops

By J. A. MORRISON and M.-J. CROSS

(Manuscript received January 9, 1974)

This paper gives details of the analytical and numerical procedures used to solve the basic problem of the scattering of a plane electromagnetic wave by an axisymmetric raindrop. A nonperturbative solution is obtained by expanding the scattered and transmitted fields in terms of spherical vector wave functions, so that Maxwell's equations are satisfied exactly in the regions exterior and interior to the raindrop, and by combining point matching with least-squares fitting to satisfy the boundary conditions on the surface of the raindrop with sufficient accuracy.

Numerical results are presented for scattering by oblate spheroidal raindrops, with eccentricity depending on (and increasing with) drop size, for two orthogonal polarizations of the incident wave. The calculations were made at 4, 11, 18.1, and 30 GHz, in the case in which the direction of propagation of the incident wave is perpendicular to the axis of symmetry of the raindrop, which is of interest for terrestrial microwave relay systems. At 30 GHz, the calculations were also made for the case in which the angle between the direction of propagation and the axis of symmetry is 70° and 50°, since different elevation angles are of interest for satellite systems. These basic results were summed earlier over the drop-size distribution to calculate the differential attenuation and differential phase shift caused by rain, which are of importance in the investigation of cross polarization in radio communication systems.

We also derive the first-order perturbation approximation to the scattering by axisymmetric raindrops that are nearly spherical, which generalizes Oguchi's results for spheroidal raindrops with small eccentricity. Some simplifications that may be made in his formulas are pointed out. The perturbation results serve as a useful check on the least-squares-fitting procedure applied to spheroidal raindrops with small eccentricity. In addition, considerable improvement is obtained in the closeness of the perturbation results to the least-squares-fitting ones, in particular for the larger drop sizes, by perturbing about an equivolumic spherical raindrop, with appropriate perturbation parameter, rather than perturbing about an inscribed spherical raindrop, as did Oguchi. Similar comparisons were also made earlier for the rain-induced differential attenuation and differential phase shift, and these quantities were calculated approximately at frequencies up to 100 GHz, using the results corresponding to perturbation about the equivolumic spherical raindrop. The perturbation results are obtained quite inexpensively, whereas the least-squares-fitting procedure is very costly.

I. INTRODUCTION

In a recent short note, the authors and Chu[1] gave calculated results of differential attenuation and differential phase shift caused by rain, based on scattering of a plane electromagnetic wave by oblate spheroidal raindrops. These results are of importance in the investigation of cross polarization in radio communications systems. In this paper, we give details of the analytical and numerical procedures used to solve the problem of scattering by a single raindrop, which were only outlined in the note. Although the results given in this paper are for oblate spheroidal raindrops, the procedures are applicable for axisymmetric raindrops that are not too nonspherical, and calculations could be made for raindrops that are more flattened on the bottom than on the top, such as for the shapes determined by Pruppacher and Pitter.[2]

Two polarizations of the incident wave, designated I and II, are considered, as depicted in Fig. 1. The factor $e^{-i\omega t}$ has been suppressed. In the first polarization, the electric field is parallel to the plane containing the axis of symmetry of the raindrop and the direction of propagation of the incident wave. In the second polarization, the electric field is perpendicular to this plane. The angle between the direction of propagation and the axis of symmetry is denoted by α. In terrestrial microwave relay systems, $\alpha = 90°$ is of main concern, but other values of α are of interest for satellite systems.

The incident wave induces a transmitted field in the interior of the raindrop and a scattered field. In the far field, the quantities of primary interest are the complex forward scattering amplitudes,[3] $S_{\mathrm{I}}(0)$ and $S_{\mathrm{II}}(0)$. In the two polarizations considered, the polarization of the far scattered field in the forward direction is the same as that of the incident wave. Also of basic interest are the cross sections of the raindrop. The total cross sections Q_t^{I} and Q_t^{II} are given in terms of the real parts of the forward scattering amplitudes by eq. (39), where k_0 is the free space wave number. We also calculate the scattering cross sections Q_s^{I} and Q_s^{II}. The absorption cross sections Q_a^{I} and Q_a^{II} are given in terms of the total and scattering cross sections by (38).

In Section II, we discuss the formulation of the problem of the scattering of a plane electromagnetic wave by a single raindrop. Spherical coordinates are chosen with polar axis along the axis of symmetry of the raindrop, and origin interior to it, as in Fig. 2. The scattered field is expanded in terms of solutions of the vector wave equation, with wave number k_0, which satisfy the radiation condition. An analogous expansion is assumed for the transmitted field inside the raindrop, in terms of vector wave functions, with wave number $k_1 = Nk_0$, which are finite at the origin. Here, N is the complex refractive index of the raindrop. The complex coefficients in the expansions are to be determined by satisfying the boundary conditions, namely, the continuity of the tangential components of the total electric and magnetic fields across the surface of the raindrop.

In Section III, the incident field is expanded in a (complex) Fourier series in the azimuthal angle φ. Because of the axial symmetry of the raindrop, the problem can be decomposed and the boundary conditions satisfied independently for each term of the Fourier series. In Section III, expressions are also given for the forward scattering amplitudes and the scattering cross sections in terms of the coefficients in the expansion of the scattered field. In addition, we express the total and scattering cross sections for an elliptically polarized incident wave in terms of those for the two linearly polarized incident waves under consideration.

In Section IV, we give expressions for the first-order approximations to the coefficients in the expansions of the scattered and transmitted fields for axisymmetric raindrops that are nearly spherical. These results generalize those given by Oguchi[4] for spheroidal raindrops with small eccentricity. Since our derivation follows closely the one given by Oguchi, we omit most details. However, we point out some simplifications that may be made in his expressions.

In Section V, we discuss an approximate nonperturbative solution to the problem of scattering by an axisymmetric raindrop. For each term in the Fourier series expansion in the azimuthal angle φ, the four boundary conditions should be satisfied on the cross-sectional boundary curve $r = R(\theta)$, $0 \leqq \theta \leqq \pi$, which defines the shape of the raindrop. Only a finite number of coefficients in the expansions of the scattered and transmitted fields is considered. These coefficients are determined approximately by requiring the boundary conditions to be satisfied in a least-squares sense at a total number of points on the cross-sectional curve that is greater than the number of unknown coefficients.

In Section V, we also discuss the advantage of using least-squares fitting rather than collocation, in which the total number of fitting points is equal to the number of unknown coefficients that are then determined by solving a system of simultaneous linear equations. After we had completed the calculations for scattering by oblate spheroidal raindrops at 4, 18.1, and 30 GHz, a paper was published by Oguchi[5] in which he carried out similar calculations for $\alpha = 90°$ at 19.3 and 34.8 GHz and used collocation for the expansions in terms of spherical vector wave functions. At 34.8 GHz, he also used an expansion in terms of spheroidal wave functions and truncated the infinite system of equations which he derived from the boundary conditions.

In Section VI, the least-squares-fitting program and some subsidiary programs are discussed. The numerical routines used for calculating the special functions that enter into the boundary conditions are also described. In addition, some indication of the running times involved and the storage requirements are given.

In Section VII, we first discuss the checks that were made on the least-squares-fitting program. These include comparison with the Mie theory[6] of the results of scattering by spherical raindrops at different angles of incidence. We also compare extrapolated results for oblate spheroidal raindrops with small eccentricity to the first-order perturbation results.

We then discuss our calculations of the scattering by oblate spheroidal raindrops, for which the ratio of minor to major semiaxis depends linearly on the radius $\bar{a}$ (in centimeters) of the equivolumic spherical drop; specifically,

$$a/b = (1 - \bar{a}), \qquad ab^2 - \bar{a}^3. \tag{1}$$

This relationship is similar to that used by Oguchi.[4,5] The rain-induced attenuation and phase shift were calculated[1] for both polarizations by

summing the real and imaginary parts of the forward scattering amplitudes over the Laws and Parsons drop-size distribution, as quoted by Setzer.[7] Thus, for rain rates up to 150 milimeters per hour, there are 14 different drop sizes, $\bar{a} = 0.025(0.025)0.35$, to be considered.

The calculations were done for wavelengths of 7.5, 2.727, 1.6575, and 1.0 cm, corresponding to frequencies of approximately 4, 11, 18.1, and 30 GHz. The refractive indices N at 20°C were obtained from an elaborate fitting equation in a recently published survey[8] of available measured data (except at 4 GHz, for which older data were used, since the calculations at that frequency were made at an earlier date). The angle of incidence α was taken to be 90° at 4, 11, and 18.1 GHz, while at 30 GHz the calculations were done for $\alpha = 70°$ and $\alpha = 50°$ also. The calculated values of the forward scattering amplitudes $S_{\mathrm{I}}(0)$ and $S_{\mathrm{II}}(0)$ are given in Tables II to VII, and those of the total cross sections Q_t^{I} and Q_t^{II} and the scattering cross sections Q_s^{I} and Q_s^{II} are given in Tables VIII to XIII. Section VII concludes by discussing some calculations using collocation and mentioning some checks on Oguchi's calculations at 19.3 and 34.8 GHz.

In Section VIII, we compare three sets of first-order perturbation results with those obtained by least-squares fitting for oblate spheroidal raindrops. One set of results corresponds to perturbation about a spherical raindrop with radius equal to the length a of the minor semiaxis of the spheroidal raindrop, which was the procedure used by Oguchi.[4] The other two sets correspond to perturbations about the equivolumic spherical raindrop, with different perturbation parameters that are consistent for small drop sizes. Considerable improvement is obtained in the closeness of the perturbation results to the least-squares-fit results by perturbing about the equivolumic spherical raindrop with the appropriate perturbation parameter. The comparisons are presented graphically in Figs. 3 to 14. The values of the forward scattering amplitude $S(0)$ and the total and scattering cross sections Q_t and Q_s for the equivolumic spherical raindrops are given in Tables XIV to XVII. These values are independent of the polarization of the incident wave and of the angle of incidence α.

The three sets of perturbation results for the differential quantities $Q_{t_1}^{\mathrm{II}} - Q_t^{\mathrm{I}}$ and $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ are compared in Figs. 15 to 23 with those obtained by least-squares fitting for oblate spheroidal raindrops. Again, considerable improvement is obtained by perturbing about the equivolumic spherical raindrop. In a recent short note,[9] similar comparisons were made for the rain-induced differential attenuation and differential phase shift. Moreover, these quantities were calculated

approximately at frequencies up to 100 GHz, using the two sets of results corresponding to perturbations about the equivolumic spherical raindrop. The perturbation results are obtained quite inexpensively, whereas the least-squares-fitting procedure is very costly.

The appendices contain some details that it was considered desirable to omit from the main text.

II. FORMULATION OF PROBLEM

We now consider the problem of scattering of a plane electromagnetic wave by a single raindrop. Suppressing the factor $e^{-i\omega t}$, where ω is the angular frequency, the divergenceless electric and magnetic fields **E** and **H** satisfy Maxwell's equations[10]

$$\nabla \times \mathbf{E} = i\omega\mu_0\mathbf{H}, \qquad \nabla \times \mathbf{H} = (\sigma - i\omega\epsilon)\mathbf{E}, \tag{2}$$

where μ_0 is the constant permeability, σ is the conductivity, and ϵ is the dielectric constant. Exterior to the raindrop $\sigma = 0$ and $\epsilon = \epsilon_0$, while interior to it $\sigma = \sigma_1$ and $\epsilon = \epsilon_1$. The appropriate boundary conditions[11] are that the tangential components of the total electric and magnetic fields be continuous across the surface of the raindrop. Let

$$k^2 = \omega\mu_0(\omega\epsilon + i\sigma), \tag{3}$$

with $\mathrm{Re}(k) > 0$. Then the free space wave number is $k_0 = \omega(\mu_0\epsilon_0)^{\frac{1}{2}}$ and the wave number in the raindrop is

$$k_1 = Nk_0, \tag{4}$$

where N is the complex index of refraction of water.

We consider two polarizations of the incident wave depicted in Fig. 1. We choose Cartesian coordinates (x, y, z) with origin interior to the

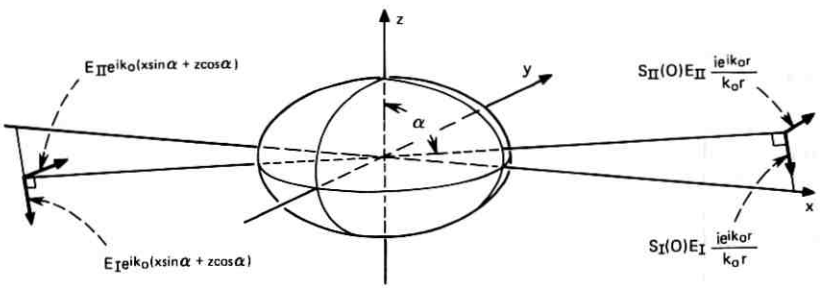

Fig. 1—Two polarizations of the incident wave.

raindrop and z-axis coinciding with the axis of symmetry of the raindrop. The direction of propagation of the incident wave is perpendicular to the y-axis and inclined at an angle α to the z-axis. In the first polarization, the magnetic field is assumed parallel to the y-axis and the incident fields are given by

$$\begin{aligned} \mathbf{E}_{\mathrm{I}}^{i} &= E_{\mathrm{I}}(\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) \exp\left[ik_0(x\sin\alpha + z\cos\alpha)\right], \\ \mathbf{H}_{\mathrm{I}}^{i} &= \frac{k_0}{\omega\mu_0} E_{\mathrm{I}}\mathbf{j} \exp\left[ik_0(x\sin\alpha + z\cos\alpha)\right], \end{aligned} \tag{5}$$

where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ denote unit vectors parallel to the coordinate axes. In the second polarization, the electric field is assumed parallel to the y-axis and the incident fields are given by

$$\mathbf{E}_{\mathrm{II}}^{i} = E_{\mathrm{II}}\mathbf{j} \exp\left[ik_0(x\sin\alpha + z\cos\alpha)\right]$$

and (6)

$$\mathbf{H}_{\mathrm{II}}^{i} = \frac{-k_0}{\omega\mu_0} E_{\mathrm{II}}(\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) \exp\left[ik_0(x\sin\alpha + z\cos\alpha)\right].$$

We now consider the problem of representing the scattered and transmitted fields induced by the incident wave. It is convenient to introduce spherical coordinates (r, θ, φ) with corresponding unit vectors $\mathbf{i}_1$, $\mathbf{i}_2$, and $\mathbf{i}_3$ as depicted in Fig. 2. Then the equations

$$\nabla \times \mathbf{M} = k\mathbf{N}, \qquad \nabla \times \mathbf{N} = k\mathbf{M} \tag{7}$$

are satisfied by the spherical vector wave functions,[12]

$$\mathbf{M}_{mn}(k) = z_n(kr)e^{im\varphi}\left[\frac{im}{\sin\theta} P_n^{|m|}(\cos\theta)\mathbf{i}_2 - \frac{dP_n^{|m|}(\cos\theta)}{d\theta}\mathbf{i}_3\right] \tag{8}$$

and

$$\begin{aligned} \mathbf{N}_{mn}(k) = e^{im\varphi}\Bigg\{ & n(n+1)\frac{z_n(kr)}{kr} P_n^{|m|}(\cos\theta)\mathbf{i}_1 + \left[\frac{z_n(kr)}{kr} + z_n'(kr)\right] \\ & \times \left[\frac{dP_n^{|m|}(\cos\theta)}{d\theta}\mathbf{i}_2 + \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta)\mathbf{i}_3\right]\Bigg\}. \end{aligned} \tag{9}$$

Here z_n denotes a spherical Bessel function[13] of order n and $P_n^{|m|}$ denotes the associated Legendre function[14] (of the first kind) of degree n and order $|m|$, where m is a positive or negative integer, and n is an integer with $n \geqq |m|$ and $n \neq 0$. The prime denotes derivative with respect to the argument. As a matter of convenience, we have chosen to use complex linear combinations of the even and odd spherical vector wave functions.[12]

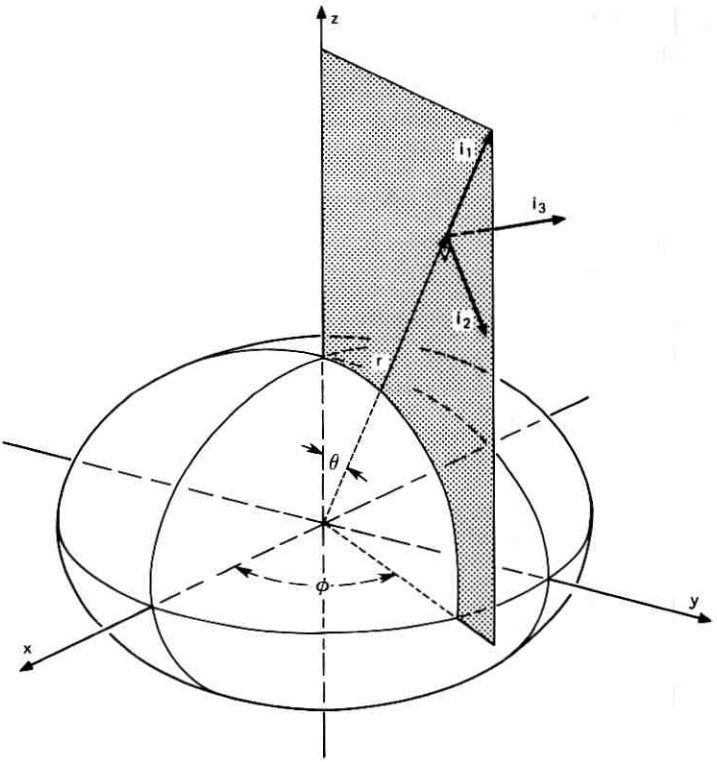

Fig. 2—Cartesian and spherical coordinates.

Outside the raindrop, the total electromagnetic field is the sum of the incident field of the plane wave and the scattered field. The scattered field must satisfy the radiation condition and, consequently, in view of eqs. (2), (3), and (7), we assume expansions of the form

$$\mathbf{E}^s = -\sum_{m=-\infty}^{\infty} \sum_{\substack{n \geqq |m| \\ n \neq 0}} [a_{mn}\mathbf{M}_{mn}^{(3)}(k_0) + b_{mn}\mathbf{N}_{mn}^{(3)}(k_0)] \tag{10}$$

and

$$\mathbf{H}^s = \frac{ik_0}{\omega\mu_0} \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geqq |m| \\ n \neq 0}} [a_{mn}\mathbf{N}_{mn}^{(3)}(k_0) + b_{mn}\mathbf{M}_{mn}^{(3)}(k_0)], \tag{11}$$

where the superscript 3 denotes that spherical Bessel functions of the third kind, i.e., spherical Hankel functions of the first kind, are used.

Thus, in (8) and (9), $z_n(k_0 r) = h_n^{(1)}(k_0 r)$. For $k_0 r \gg 1$,

$$h_n^{(1)}(k_0 r) \sim \frac{(-i)^{n+1}}{k_0 r} e^{ik_0 r}, \tag{12}$$

so that the expansions in (10) and (11) involve outgoing waves.

Analogous expansions are assumed for the transmitted field inside the raindrop except that, since the origin of the coordinate system is interior to the raindrop, spherical Bessel functions of the first kind must be used so that the field remains finite at $r = 0$. Also, the wave number inside the raindrop is k_1, as given by (4). Thus, we assume expansions of the form

$$\mathbf{E}^t = -\sum_{m=-\infty}^{\infty} \sum_{\substack{n \geqq |m| \\ n \neq 0}} [c_{mn}\mathbf{M}_{mn}^{(1)}(k_1) + d_{mn}\mathbf{N}_{mn}^{(1)}(k_1)] \tag{13}$$

and

$$\mathbf{H}^t = \frac{ik_1}{\omega\mu_0} \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geqq |m| \\ n \neq 0}} [c_{mn}\mathbf{N}_{mn}^{(1)}(k_1) + d_{mn}\mathbf{M}_{mn}^{(1)}(k_1)], \tag{14}$$

where the superscript 1 indicates that $z_n(k_1 r) = j_n(k_1 r)$ in (8) and (9).

The unknown (complex) coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn} in (10), (11), (13), and (14) must be determined from the boundary conditions. The surface of the raindrop is given by

$$r = R(\theta), \qquad 0 \leqq \theta \leqq \pi, \qquad 0 \leqq \varphi \leqq 2\pi, \tag{15}$$

where it is assumed that $R(\theta)$ is a single-valued, continuously differentiable function of θ. The continuity of the tangential components of the total electric and magnetic fields across the surface of the raindrop implies that, for $r = R(\theta)$,

$$E_3^i + E_3^s = E_3^t, \tag{16}$$

$$H_3^i + H_3^s = H_3^t, \tag{17}$$

$$E_2^i + E_2^s + \frac{1}{R}\frac{dR}{d\theta}(E_1^i + E_1^s) = E_2^t + \frac{1}{R}\frac{dR}{d\theta}E_1^t, \tag{18}$$

$$H_2^i + H_2^s + \frac{1}{R}\frac{dR}{d\theta}(H_1^i + H_1^s) = H_2^t + \frac{1}{R}\frac{dR}{d\theta}H_1^t, \tag{19}$$

where $E_j = \mathbf{E}\cdot\mathbf{i}_j$ and $H_j = \mathbf{H}\cdot\mathbf{i}_j$ and the incident fields $\mathbf{E}^i$ and $\mathbf{H}^i$ are given by (5) or (6).

III. FAR-FIELD QUANTITIES

Because of the axial symmetry of the raindrop, it is convenient to expand the incident plane wave in a (complex) Fourier series in the azimuthal angle φ, and we write

$$\mathbf{E}^i = \sum_{m=-\infty}^{\infty} \mathbf{e}_m(r, \theta)e^{im\varphi}, \qquad \mathbf{H}^i = \sum_{m=-\infty}^{\infty} \mathbf{h}_m(r, \theta)e^{im\varphi}. \tag{20}$$

It follows from (5) and (6) that

$$\mathbf{e}_m^{\mathrm{I}}(r, \theta) = E_{\mathrm{I}}\mathbf{f}_m(r, \theta), \qquad \mathbf{h}_m^{\mathrm{I}}(r, \theta) = E_{\mathrm{I}} \frac{k_0}{\omega\mu_0} \mathbf{g}_m(r, \theta) \tag{21}$$

and

$$\mathbf{e}_m^{\mathrm{II}}(r, \theta) = E_{\mathrm{II}}\mathbf{g}_m(r, \theta), \qquad \mathbf{h}_m^{\mathrm{II}}(r, \theta) = -E_{\mathrm{II}} \frac{k_0}{\omega\mu_0} \mathbf{f}_m(r, \theta), \tag{22}$$

where expressions for $\mathbf{f}_m(r, \theta)$ and $\mathbf{g}_m(r, \theta)$ are derived in Appendix A and are given by eqs. (79) and (80).

If the boundary conditions (16) to (19) are multiplied by $e^{-im\varphi}$ and integrated with respect to φ from 0 to 2π, then a set of four equations involving the unknown coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn} is obtained for each m. These equations are given by (81) to (84) in Appendix A, where we have used the notations $e_{mj} = \mathbf{e}_m \cdot \mathbf{i}_j$ and $h_{mj} = \mathbf{h}_m \cdot \mathbf{i}_j$. It follows readily from (21), (22), and (79) to (84) that, for the first polarization of the incident wave,

$$\begin{aligned} a_{-mn}^{\mathrm{I}} &= -a_{mn}^{\mathrm{I}}, & b_{-mn}^{\mathrm{I}} &= b_{mn}^{\mathrm{I}}, \\ c_{-mn}^{\mathrm{I}} &= -c_{mn}^{\mathrm{I}}, & d_{-mn}^{\mathrm{I}} &= d_{mn}^{\mathrm{I}}, \end{aligned} \tag{23}$$

and for the second polarization

$$\begin{aligned} a_{-mn}^{\mathrm{II}} &= a_{mn}^{\mathrm{II}}, & b_{-mn}^{\mathrm{II}} &= -b_{mn}^{\mathrm{II}}, \\ c_{-mn}^{\mathrm{II}} &= c_{mn}^{\mathrm{II}} & d_{-mn}^{\mathrm{II}} &= -d_{mn}^{\mathrm{II}}. \end{aligned} \tag{24}$$

Thus, it is sufficient to consider only nonnegative values of m.

It is worth noting that, if the raindrop is symmetrical about the plane $\theta = \pi/2$, i.e., $R(\pi - \theta) = R(\theta)$, $0 \leqq \theta \leqq \pi/2$, then some further reductions may be made. In particular, in the case $\alpha = \pi/2$, it is found that

$$\begin{aligned} a_{m,|m|+2s+1}^{\mathrm{I}} &= 0 = c_{m,|m|+2s+1}^{\mathrm{I}}, \\ b_{m,|m|+2s}^{\mathrm{I}} &= 0 = d_{m,|m|+2s}^{\mathrm{I}}, \end{aligned} \tag{25}$$

and

$$\begin{aligned} a_{m,|m|+2s}^{\mathrm{II}} &= 0 = c_{m,|m|+2s}^{\mathrm{II}}, \\ b_{m,|m|+2s+1}^{\mathrm{II}} &= 0 = d_{m,|m|+2s+1}^{\mathrm{II}}, \end{aligned} \tag{26}$$

for $s = 0, 1, 2, \cdots$, so that alternate coefficients vanish. Reductions

may still be made in the case of $\alpha \neq \pi/2$ by considering the sum and the difference of the boundary conditions corresponding to α and to $\hat{\alpha} = (\pi - \alpha)$, as shown in Appendix B. We have not utilized these reductions in the program for calculating the unknown coefficients, since we wanted the program to be applicable to raindrops without a meridional plane of symmetry, that is, those raindrops flattened more on the bottom than on the top. However, (25) and (26) served as a useful check on the program in the calculations for spheroidal raindrops with $\alpha = \pi/2$.

We describe in Section V how we obtain approximate values of (a finite number of) the coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn}, but we now turn to the quantities of physical interest. We consider only the far scattered field, so that $k_0 r \gg 1$. Thus, we restrict our attention to the leading term in the asymptotic expansion of the spherical Bessel function of the third kind, as given by (12). Also, it follows that

$$h_n^{(1)'}(k_0 r) \sim \frac{(-i)^n}{k_0 r} e^{ik_0 r}. \tag{27}$$

Then, from (8) to (11), it is found that

$$k_0 r e^{-ik_0 r}\mathbf{E}^s \sim \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geq |m| \\ n \neq 0}} (-i)^{n+1} \left\{ a_{mn} \left[\frac{dP_n^{|m|}(\cos\theta)}{d\theta} \mathbf{i}_3 - \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta)\mathbf{i}_2 \right] - ib_{mn} \left[\frac{dP_n^{|m|}(\cos\theta)}{d\theta} \mathbf{i}_2 + \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta)\mathbf{i}_3 \right] \right\} e^{im\varphi} \tag{28}$$

and

$$\omega\mu_0 \mathbf{H}^s \sim k_0 \mathbf{i}_1 \times \mathbf{E}^s. \tag{29}$$

Of particular interest are the scattered fields in the forward direction, corresponding to $\theta = \alpha$, $\varphi = 0$, for which, from (76),

$$(\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) = \mathbf{i}_2, \qquad \mathbf{j} = \mathbf{i}_3. \tag{30}$$

From (5), (6), (23), (24), and (28) to (30), it follows that the far scattered field in the forward direction has the same polarization as the incident wave for either polarization. The forward scattering amplitudes are[3]

$$S_{\mathrm{I}}(0) = \frac{1}{E_{\mathrm{I}}} (\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) \cdot \lim_{r\to\infty} \{ -ik_0 r e^{-ik_0 r} \mathbf{E}_{\mathrm{I}}^s |_{\theta=\alpha,\ \varphi=0} \} \tag{31}$$

and

$$S_{\mathrm{II}}(0) = \frac{1}{E_{\mathrm{II}}} \mathbf{j} \cdot \lim_{r\to\infty} \{ -ik_0 r e^{-ik_0 r} \mathbf{E}_{\mathrm{II}}^s |_{\theta=\alpha,\ \varphi=0} \}. \tag{32}$$

Thus, for the first polarization of the incident wave,

$$E_{\mathrm{I}}S_{\mathrm{I}}(0) = \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geq |m| \\ n \neq 0}} (-i)^{n-1} \times \left[a_{mn}^{\mathrm{I}} \cdot \frac{m}{\sin \alpha} P_n^{|m|}(\cos \alpha) + b_{mn}^{\mathrm{I}} \frac{dP_n^{|m|}(\cos \alpha)}{d\alpha} \right] \quad (33)$$

and for the second polarization,

$$E_{\mathrm{II}}S_{\mathrm{II}}(0) = \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geq |m| \\ n \neq 0}} (-i)^{n+2} \times \left[a_{mn}^{\mathrm{II}} \frac{dP_n^{|m|}(\cos \alpha)}{d\alpha} + b_{mn}^{\mathrm{II}} \cdot \frac{m}{\sin \alpha} P_n^{|m|}(\cos \alpha) \right]. \quad (34)$$

The energy scattered by the raindrop is[15]

$$W_s = \frac{1}{2} \operatorname{Re} \int_0^{2\pi} \int_0^{\pi} [E_2^s(H_3^s)^* - E_3^s(H_2^s)^*] r^2 \sin \theta d\theta d\varphi, \quad (35)$$

where the asterisk denotes complex conjugate. The calculation of W_s, using the asymptotic form of the scattered fields given by (28) and (29) and letting $r \to \infty$, is outlined in Appendix C. It is found that

$$W_s = \frac{2\pi}{\omega \mu_0 k_0} \sum_{m=-\infty}^{\infty} \sum_{n \geq |m|} \frac{n(n+1)(n+|m|)!}{(2n+1)(n-|m|)!} (|a_{mn}|^2 + |b_{mn}|^2). \quad (36)$$

The scattering cross section Q_s is defined as the ratio of the scattered energy flow to the mean energy flow of the incident wave per unit area. Thus,[15]

$$Q_s^{\mathrm{I}} = \frac{2\omega\mu_0 W_s^{\mathrm{I}}}{k_0 E_{\mathrm{I}} E_{\mathrm{I}}^*}, \qquad Q_s^{\mathrm{II}} = \frac{2\omega\mu_0 W_s^{\mathrm{II}}}{k_0 E_{\mathrm{II}} E_{\mathrm{II}}^*}. \quad (37)$$

The total (extinction) cross section is the sum of the scattering and absorption cross sections, so that

$$Q_t^{\mathrm{I}} = Q_s^{\mathrm{I}} + Q_a^{\mathrm{I}}, \qquad Q_t^{\mathrm{II}} = Q_s^{\mathrm{II}} + Q_a^{\mathrm{II}}. \quad (38)$$

(We note that van de Hulst[16] uses the notations C_{ext}, C_{sca}, and C_{abs} for Q_t, Q_s, and Q_a, respectively.) It is "known"[17] that

$$Q_t^{\mathrm{I}} = \frac{4\pi}{k_0^2} \operatorname{Re} S_{\mathrm{I}}(0), \qquad Q_t^{\mathrm{II}} = \frac{4\pi}{k_0^2} \operatorname{Re} S_{\mathrm{II}}(0), \quad (39)$$

so that (38) may be used to determine the absorption cross sections Q_a^{I} and Q_a^{II}. The relations (39) which are consistent with the optical

theorem may be verified directly from the relations

$$Q_t^{\mathrm{I}} = \frac{2\omega\mu_0 W_t^{\mathrm{I}}}{k_0 E_{\mathrm{I}} E_{\mathrm{I}}^*}, \qquad Q_t^{\mathrm{II}} = \frac{2\omega\mu_0 W_t^{\mathrm{II}}}{k_0 E_{\mathrm{II}} E_{\mathrm{II}}^*}, \tag{40}$$

and the expression for the total energy[15]

$$W_t = \frac{1}{2}\,\mathrm{Re}\int_0^{2\pi}\int_0^{\pi} [E_3^i(H_2^s)^* + E_3^s(H^i)^* - E_2^i(H_3^s)^* - E_2^s(H_3^i)^*] r^2 \sin\theta d\theta d\varphi, \tag{41}$$

and a few of the details are given in Appendix C.

Let us now consider an (in general) elliptically polarized incident wave that is the sum of the two linearly polarized incident waves under consideration; i.e.,

$$\mathbf{E}^i = \mathbf{E}_{\mathrm{I}}^i + \mathbf{E}_{\mathrm{II}}^i \tag{42}$$

Then the scattered electric field is

$$\mathbf{E}^s = \mathbf{E}_{\mathrm{I}}^s + \mathbf{E}_{\mathrm{II}}^s, \tag{43}$$

and, as shown in Appendix C,

$$Q_s = \frac{(E_{\mathrm{I}} E_{\mathrm{I}}^* Q_s^{\mathrm{I}} + E_{\mathrm{II}} E_{\mathrm{II}}^* Q_s^{\mathrm{II}})}{(E_{\mathrm{I}} E_{\mathrm{I}}^* + E_{\mathrm{II}} E_{\mathrm{II}}^*)} \tag{44}$$

and

$$Q_t = \frac{(E_{\mathrm{I}} E_{\mathrm{I}}^* Q_t^{\mathrm{I}} + E_{\mathrm{II}} E_{\mathrm{II}}^* Q_t^{\mathrm{II}})}{(E_{\mathrm{I}} E_{\mathrm{I}}^* + E_{\mathrm{II}} E_{\mathrm{II}}^*)}. \tag{45}$$

Since for polarizations I and II the far scattered field in the forward direction has the same polarization as the incident wave, it follows from (31), (32), and (43) that the far scattered field in the forward direction for the elliptically polarized incident wave is given by

$$\mathbf{E}^s|_{\theta=\alpha,\ \varphi=0} \sim \frac{ie^{ik_0 r}}{k_0 r}[E_{\mathrm{I}} S_{\mathrm{I}}(0)(\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) + E_{\mathrm{II}} S_{\mathrm{II}}(0)\mathbf{j}]. \tag{46}$$

Thus, from (39) and (44) to (46), it suffices to calculate $S_{\mathrm{I}}(0)$, $S_{\mathrm{II}}(0)$, Q_S^{I}, and Q_S^{II}. The relation between the polarizations of the incident field and the far scattered field in the forward direction may be determined from (42) and (46), using (5) and (6).

IV. FIRST-ORDER PERTURBATION THEORY

Oguchi considered spheroidal raindrops with small eccentricity and carried out a perturbation expansion originally determining the first-

order approximation[4] and later the second-order one.[18] We have calculated the first-order approximation for axisymmetric raindrops that are nearly spherical, so that the surface of the raindrop is given by $r = R(\theta)$, $0 \leqq \theta \leqq \pi$, where

$$R(\theta) = a[1 + \nu\sigma_1(\theta) + \cdots], \qquad |\nu| \ll 1. \tag{47}$$

Our derivation follows closely the one given by Oguchi.[4] Since the calculations are somewhat lengthy and involved, we merely outline the procedure, state the results, and point out some simplifications that may be made in the expressions given by Oguchi.

The incident wave may be expanded in terms of spherical vector wave functions,[4,19] and the expansions are given by eqs. (116) and (117) in Appendix D. These expansions are consistent with those given by Oguchi, but the reader should bear in mind that, in addition to using the even and odd spherical vector wave functions, Oguchi has assumed the time factor $e^{+i\omega t}$, and his waves propagate in the opposite direction to ours. Corresponding to (47), the coefficients in the expansions (10), (11), (13), and (14) are expanded in the form

$$a_{mn} = a_{mn}^{(0)} + \nu a_{mn}^{(1)} + \cdots, \quad b_{mn} = b_{mn}^{(0)} + \nu b_{mn}^{(1)} + \cdots, \tag{48}$$

$$c_{mn} = c_{mn}^{(0)} + \nu c_{mn}^{(1)} + \cdots, \quad d_{mn} = d_{mn}^{(0)} + \nu d_{mn}^{(1)} + \cdots. \tag{49}$$

Appendix D indicates how these coefficients may be determined from the boundary conditions (81) to (84).

The zero-order approximation, with $\nu = 0$ in (47), corresponds to a spherical raindrop of radius a. We have, for $n \geqq |m|$ and $n \neq 0$,

$$a_{mn}^{(0)} = \alpha_{mn} a_n, \qquad b_{mn}^{(0)} = \beta_{mn} b_n \tag{50}$$

and

$$c_{mn}^{(0)} = \alpha_{mn} c_n, \qquad d_{mn}^{(0)} = \beta_{mn} d_n, \tag{51}$$

where

$$\frac{\alpha_{mn}^{\mathrm{I}}}{E_{\mathrm{I}}} = \frac{-i^{n+1}(2n+1)(n-|m|)!}{n(n+1)(n+|m|)!} \frac{m}{\sin\alpha} P_n^{|m|}(\cos\alpha) = \frac{i\beta_{mn}^{\mathrm{II}}}{E_{\mathrm{II}}}, \tag{52}$$

$$\frac{\beta_{mn}^{\mathrm{I}}}{E_{\mathrm{I}}} = \frac{-i^{n+1}(2n+1)(n-|m|)!}{n(n+1)(n+|m|)!} \frac{dP_n^{|m|}(\cos\alpha)}{d\alpha} = \frac{i\alpha_{mn}^{\mathrm{II}}}{E_{\mathrm{II}}}, \tag{53}$$

and expressions for the quantities a_n, b_n, c_n, and d_n (which do not depend on the polarization) are given by eqs. (119) to (122) in Appendix D, where $\rho = k_0 a$ and the functions $F_n(\xi)$ and $G_n(\xi)$ are defined in (118). For $\alpha = 0$, the coefficients vanish unless $|m| = 1$, and the well-known Mie solution[6] is recovered.

The first-order corrections to the coefficients are given by

$$a_{mn}^{(1)} = j_n(N\rho)X_{mn}, \qquad c_{mn}^{(1)} = h_n^{(1)}(\rho)X_{mn}, \tag{54}$$

$$b_{mn}^{(1)} = G_n(N\rho)Y_{mn} + j_n(N\rho)Z_{mn}, \tag{55}$$

and

$$d_{mn}^{(1)} = F_n(\rho)Y_{mn} + \frac{1}{N}h_n^{(1)}(\rho)Z_{mn}, \tag{56}$$

where

$$X_{mn} = (1 - N^2)\rho^3 c_n \sum_{\substack{l \geq |m| \\ l \neq 0}} [d_{ml}^{(0)} G_l(N\rho)J_{nl}^m + ic_{ml}^{(0)} j_l(N\rho)I_{nl}^m], \tag{57}$$

$$Y_{mn} = (N^2 - 1)\rho^3 d_n \sum_{\substack{l \geq |m| \\ l \neq 0}} [c_{ml}^{(0)} j_l(N\rho)J_{nl}^m - id_{ml}^{(0)} G_l(N\rho)I_{nl}^m], \tag{58}$$

and

$$Z_{mn} = -i(N^2 - 1)\rho d_n \sum_{\substack{l \geq |m| \\ l \neq 0}} d_{ml}^{(0)} j_l(N\rho)H_{nl}^m. \tag{59}$$

The quantities H_{nl}^m, I_{nl}^m, and J_{nl}^m in (57) to (59) involve integrals over the perturbation of the raindrop surface from the sphere. Specifically,

$$\frac{2(n + |m|)!}{(2n + 1)(n - |m|)!} H_{nl}^m = \mathfrak{IC}_{nl}^m$$

$$\equiv l(l + 1)\int_0^\pi P_l^{|m|}(\cos\theta)P_n^{|m|}(\cos\theta)\sin\theta\sigma_1(\theta)d\theta, \tag{60}$$

$$\frac{2n(n + 1)(n + |m|)!}{(2n + 1)(n - |m|)!} I_{nl}^m = \mathfrak{g}_{nl}^m$$

$$\equiv \int_0^\pi \left[\frac{dP_l^{|m|}(\cos\theta)}{d\theta}\frac{dP_n^{|m|}(\cos\theta)}{d\theta} + \frac{m^2}{\sin^2\theta}P_l^{|m|}(\cos\theta)P^{|m|}(\cos\theta)\right]\sin\theta\sigma_1(\theta)d\theta, \tag{61}$$

and

$$\frac{2n(n + 1)(n + |m|)!}{(2n + 1)(n - |m|)!} J_{nl}^m = \mathfrak{J}_{nl}^m$$

$$\equiv m\int_0^\pi \left[P_l^{|m|}(\cos\theta)\frac{dP_n^{|m|}(\cos\theta)}{d\theta} + \frac{dP_l^{|m|}(\cos\theta)}{d\theta}P_n^{|m|}(\cos\theta)\right]\sigma_1(\theta)d\theta. \tag{62}$$

The above results were obtained after considerable algebra and after using the simplifications given in eqs. (125) to (131) in Appendix

D. The expressions in (54) to (56) hold for both polarizations of the incident wave, and the only quantities therein that depend on the polarization are the zero-order coefficients $c_{ml}^{(0)}$ and $d_{ml}^{(0)}$, which enter into the expressions given in (57) to (59) and are given by (51) to (53). In general, there are infinitely many terms in the sums in (57) to (59), but in particular cases there are only a finite number of terms.

Thus, for a spheroidal raindrop with

$$R(\theta) = a(1 - \nu \sin^2 \theta)^{-\frac{1}{2}} = a(1 + \tfrac{1}{2}\nu \sin^2 \theta + \cdots), \tag{63}$$

we have $\sigma_1(\theta) = \frac{1}{2} \sin^2 \theta$, from (47). The integrals in (60) to (62) may be evaluated explicitly, and it is found that H_{nl}^m and I_{nl}^m vanish unless $l = n, n + 2$, or $n - 2$, and J_{nl}^m vanishes unless $l = n - 1$ or $n + 1$. The explicit expressions for these quantities are given by eqs. (137), (139), and (142) in Appendix E, in which δ_{ln} denotes the Kronecker delta, i.e., $\delta_{ln} = 1$ for $l = n$, and 0 otherwise. We have verified that our results for the spheroid are consistent with those of Oguchi,[4] provided that simplifications corresponding to those in eqs. (123) and (127) to (130) are made in his expressions, and due allowance is made for the differences in notation. We remark that similar simplifications may be made in Oguchi's expressions even in the case in which the permeability of the spheroid differs from the free space value.

As is seen later, it is advantageous to obtain the first-order approximation for a spheroidal raindrop by perturbing about the equivolumic sphere, rather than the inscribed sphere as Oguchi[4] did. If $\bar{a}$ is the radius (in centimeters) of the equivolumic sphere, then, from (1),

$$a = \bar{a}(1 - \bar{a})^{\frac{1}{3}}, \qquad \nu = \bar{a}(2 - \bar{a}). \tag{64}$$

Hence, from (63),

$$\begin{aligned} R(\theta) &= \bar{a}[1 + 2\bar{a}(\tfrac{1}{2} \sin^2 \theta - \tfrac{1}{3}) + 0(\bar{a}^2)] \\ &= \bar{a}[1 + \nu(\tfrac{1}{2} \sin^2 \theta - \tfrac{1}{3}) + 0(\nu^2)]. \end{aligned} \tag{65}$$

We must now replace ρ by $\bar{\rho} = k_0\bar{a}$ and add terms to the expressions in (60) to (62) corresponding to $\bar{\sigma}_1(\theta) = -\frac{1}{3}$. It is readily found, using the orthogonality relation for the Legendre functions[14] and equations (102) and (103), that these additional terms correspond to

$$\tilde{H}_{nl}^m = -\tfrac{1}{3}n(n + 1)\delta_{ln}, \quad \tilde{I}_{nl}^m = -\tfrac{1}{3}\delta_{ln}, \quad \tilde{J}_{nl}^m = 0. \tag{66}$$

We also remark that the use of the perturbation parameter $\bar{\nu} = 2\bar{a}$, rather than ν as given by (64), generally gives better results for the larger drop sizes.

V. LEAST-SQUARES-FITTING PROCEDURE

We now discuss the calculation of an approximate nonperturbative solution of the scattering problem. As mentioned in Section III, it is sufficient to determine the unknown coefficients for nonnegative values of m and then to use the relationships in (23) or (24). For $m = 0, 1, 2, \cdots$, the boundary conditions in (81) to (84) take the form

$$K_{mq}(\theta) - \sum_{\substack{n \geqq m \\ n \neq 0}} [a_{mn}A_{mnq}(\theta) + b_{mn}B_{mnq}(\theta) + c_{mn}C_{mnq}(\theta) + d_{mn}D_{mnq}(\theta)] = 0. \quad (67)$$

for $q = 1, 2, 3, 4$ and $0 \leqq \theta \leqq \pi$, where

$$K_{m1}(\theta) = e_{m3}(R(\theta), \theta), \qquad K_{m2}(\theta) = \frac{i\omega\mu_0}{k_0} h_{m3}(R(\theta), \theta), \quad (68)$$

$$K_{m3}(\theta) = e_{m2}(R(\theta), \theta) + \frac{1}{R(\theta)} \frac{dR}{d\theta} e_{m1}(R(\theta), \theta), \quad (69)$$

and

$$K_{m4}(\theta) = \frac{i\omega\mu_0}{k_0} \left[h_{m2}(R(\theta), \theta) + \frac{1}{R(\theta)} \frac{dR}{d\theta} h_{m1}(R(\theta), \theta) \right]. \quad (70)$$

The function $R(\theta)$ describes the shape of the raindrop. The functions $e_{mj} = \mathbf{e}_m \cdot \mathbf{i}_j$ and $h_{mj} = \mathbf{h}_m \cdot \mathbf{i}_j$ are given by (21) or (22), depending on the polarization of the incident wave, where expressions for $\mathbf{f}_m(r, \theta)$ and $\mathbf{g}_m(r, \theta)$ are given by (79) and (80). The functions $A_{mnq}(\theta)$, $B_{mnq}(\theta)$, $C_{mnq}(\theta)$, $D_{mnq}(\theta)$, which do not depend on the polarization, involve the spherical Bessel functions of the first and third kinds and the associated Legendre functions and the derivatives of each of these functions. In view of (4), the argument of the spherical Bessel functions of the first kind is complex.

For each m there are infinitely many unknown coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn}. To obtain an approximate solution, only a finite number of coefficients is considered. One procedure is to truncate the sum in (67) at $n = N_0$, say, and then to satisfy the boundary conditions at the points $\theta = \theta_{lm}$, $l = 1, \cdots, (N_0 - m + 1 - \delta_{m0})$, which are appropriately selected, e.g., uniformly spaced in the interval 0 to π. This was the procedure adopted by Oguchi,[5] and it leads to a system of simultaneous linear equations for the coefficients. We refer to this procedure, in which the total number of fitting points is equal to the number of unknown coefficients, as collocation.

The method of collocation was used by Mullin et al.[20] for the much simpler two-dimensional scalar problem of scattering by a perfectly

conducting cylinder of smooth contour, which is not a gross perturbation from the circular. In this problem, there is only one set of unknown coefficients to be determined, namely, that occurring in the expansion of the scattered field. Mullin et al. checked the results for the circular cylinder obtained by collocation with the known analytical results and also considered elliptic cylinders (with ratio of minor to major axis of $\frac{2}{3}$, in particular). Before tackling the raindrop problem, we considered the same problems as Mullin et al., but combined point matching with least-squares fitting.

Thus, instead of using collocation, we satisfied the (single) boundary condition in a least-squares sense at a larger number of points than the number of unknown coefficients in the truncated expansion of the scattered field. We found that a significant improvement could be obtained in the overall fit of the boundary condition, although the far field quantities were not affected as significantly. This is because the higher-order coefficients are more significant on the boundary than in the far field. However, the accuracy of the lower-order coefficients is affected by the goodness of fit of the boundary condition. With collocation, there were much larger errors in the boundary condition (in between the fitting points) than with least-squares fitting with a sufficiently large number of points.

Since the fit of the boundary condition for the elliptic cylinder becomes poorer with increasing eccentricity, we considered it desirable to use least-squares fitting rather than collocation for the raindrop problem. Thus, in order to approximately satisfy the boundary conditions (67), we minimized the quantities

$$\Delta_m \equiv \sum_{q=1}^{4} w_{mq} \sum_{l=1}^{\lambda_m} |K_{mq}(\theta_{lm}) - \sum_{\substack{n=m \\ n\neq 0}}^{N_m} [a_{mn}A_{mnq}(\theta_{lm}) + b_{mn}B_{mnq}(\theta_{lm}) + c_{mn}C_{mnq}(\theta_{lm}) + d_{mn}D_{mnq}(\theta_{lm})]|^2 \quad (71)$$

for each $m = 0, \cdots, M$, with respect to the (complex) coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn}, where $w_{mq} > 0$ are appropriate weights and θ_{lm} are appropriate points in the interval 0 to π. It is assumed that

$$\lambda_m \geqq N_m - m + 1 - \delta_{m0}, \quad (72)$$

so that the total number of fitting points is not less than the number of unknown coefficients to be determined. In the case of equality in (72), least-squares fitting is equivalent to collocation.

The programs for carrying out least-squares fitting and for calculating the special functions occurring in the functions in (71) with argu-

ment θ_{lm} are described in the next section. Actually, more flexibility was built into the least-squares-fitting program, allowing for truncation of the sums in (67) at different limits for each of the coefficients a_{mn}, b_{mn}, c_{mn}, and d_{mn}, and for λ_m and θ_{lm} in (71) to depend on q, so that each of the four boundary conditions could be fit at different points, and in particular at a different number of them. It was anticipated that the least-squares-fit subroutine might become overloaded, in which case it would be desirable to hold the number of coefficients to a minimum. It turned out, however, that the subroutine was able to handle almost 100 (complex) coefficients without difficulty. Similarly, it might be desirable to keep the total number of fitting points to a minimum, and hence to use fewer fitting points for those of the four boundary conditions that are easier to fit. Again, it was not found necessary to do this for the calculations carried out so far.

For the calculations of this paper we took the weights to be independent of m and q, i.e., $w_{mq} \equiv 1$, since it was generally found that the difference was tolerable between the magnitudes of the maximum error in the fit of each of the four boundary conditions. We at first considered factoring out $(\sin\theta)^{m-1}$, for $m \geqq 2$, from the boundary conditions (67), but decided against it since we felt that the absolute, rather than the relative, error in the fit of the boundary conditions was important. However, because of the presence of this factor, we did experiment with unequally spaced fitting points which were closer together in the neighborhood of $\pi/2$. We decided that equally spaced fitting points would suffice, provided that enough were used. The total number of fitting points was usually taken to be slightly more than twice the number of unknown coefficients, i.e., $\lambda_m > 2(N_m - m + 1 - \delta_{m0})$.

Generally, N_m, the upper limit of n in (71), was taken to be independent of m, i.e., $N_m \equiv N_0$, $m = 0, \cdots, M$. The choice of N_0 and M depended both on drop size and on frequency. The choice of M was based on the rate of convergence of the outer sums in the expressions in (33), (34), and (36) for the far field quantities. On the other hand, to ensure the accuracy of the lower-order terms in the inner sums, it is necessary to take more terms in n than are really needed in the calculation of the far field quantities. A convergence test was carried out by doing the calculations for $N_m = \hat{N}_0$, $(\hat{N}_0 + 2)$ and $(\hat{N}_0 + 4)$.

VI. NUMERICAL ROUTINES

The program to compute the complex coefficients a_{mn}, b_{mn}, c_{mn}, d_{mn}, the scattering cross section Q_s, and the forward scattering amplitude

$S(0)$ for the two polarizations of the incident wave is written in Fortran IV for a Honeywell 6070 computer. It uses the complex arithmetic and math routines (such as sin, cos) written for that system. The program is written as a three-part package; the first part consists of a driver routine which sets up core storage for the least-squares matrix and associated vectors and the following subroutines:

L2FIT—The main subroutine which computes the elements of the complex matrix for input to the least-squares-fitting procedure and controls the other subroutines.

CLSTSQ—Computes a least-squares fit for a linear system with complex coefficients; the algorithm and Fortran routine were written by P. Businger of Bell Laboratories.

BJYNC—Computes the Bessel functions $J_n(z)$, $Y_n(z)$, for z complex, n a nonnegative integer; the algorithm and Fortran subroutine package were written by E. Sonnenblick of Bell Laboratories.

SBES—Computes the spherical Bessel functions $j_n(x)$, $y_n(x)$, x real, n a nonnegative integer.

CSB—Computes the spherical Bessel function $j_n(z)$, z complex, n a nonnegative integer.

SLEG—Computes the associated Legendre functions $P_n^m(\cos\theta)$, m, n nonnegative integers.

The second part of the package is the routine

SQS—Computes the forward scattering amplitude $S(0)$ and the scattering and total cross sections Q_s, Q_t for both polarizations from the least-squares-fit solutions.

The third part is a computational check on the least-squares fit. It consists of a driver routine as in the first part, the function subroutines BJYNC, SBES, CBS, SLEG, and the main subroutine

CHECK—Checks the accuracy of the least-squares computation of a_{mn}, b_{mn}, c_{mn}, d_{mn} by using these coefficients to compute the boundary fit at points in addition to those used to obtain the coefficients.

The internal computations of all the subroutines except CSB (see detailed description below) are done in double-precision arithmetic; on the Honeywell 6070 this means 18 digits are used for all computations. The function values (Bessel, spherical Bessel, Legendre), how-

ever, are returned as single-precision numbers (8 digits), since the accuracy of any more digits could not always be guaranteed.

The main subroutine, L2FIT, sets up the complex matrix to minimize the quantities (71). Instead of the limit N_m in (71), the program actually breaks the summation on n into four parts, using limits α_m, β_m, γ_m, δ_m for the coefficients a_{mn}, b_{mn}, c_{mn}, d_{mn}, respectively; further, it replaces the limit λ_m in the sum on l with four limits, λ_{mq}. Thus, for each $m = 0, 1, \cdots, M$ the matrix equation to be minimized takes the form $\|\mathbf{A}\cdot\mathbf{x} - \mathbf{b}\| \approx 0$, where $\mathbf{A}$ is an L-by-N matrix with L, N, which are both dependent on m, defined as follows:

$$L = \sum_{q=1}^{4} \lambda_{mq}$$

and

$$N = \alpha_m + \beta_m + \gamma_m + \delta_m - 4(m - 1 + \delta_{m0}).$$

Although the basic functions (Bessel, Legendre, and derivatives of these) are returned to L2FIT in single precision, the calculations of the elements A_{mnq}, B_{mnq}, C_{mnq}, D_{mnq} in the least-squares matrix and K_{mq} in the vector of constants, $\mathbf{b}$, are carried out and left in double precision for input to the least-squares-minimization program. To facilitate changing the raindrop shape and spacing of points on the boundary, the quantities θ_{lm} and $R(\theta_{lm})$ are computed in subroutines called by this routine.

The routine CLSTSQ uses elementary Hermitian (or Householder) transformations to compute a linear least-squares solution to the equation $\|\mathbf{A}\cdot\mathbf{x} - \mathbf{b}\| = \min$. The algorithm is an adaptation of an algorithm for a real matrix written by P. Businger and G. H. Golub.[21]

The routine CHECK uses the coefficients a_{mn}, b_{mn}, c_{mn}, d_{mn} from CLSTSQ as input into the expressions in (67) to compute the fit of the boundary condition in between the fitting points as well as at the fitting points. The goodness of this fit provides a check on the accuracy of the computed coefficients.

The computation of the elements for the least-squares matrix and constant vector requires values of $J_n(x)$, $j_n(x)$, $y_n(x)$, $j_n(z)$, $P_n^m(\cos\theta)$, where x and θ are real, z is complex, and n and m are nonnegative integers. The routine to compute $J_n(x)$, BJYNC[22] uses a downward recursion scheme (compute J_N, J_{N-1}, $\cdots$, J_0) for $|x| > 0.1$ and uses the power series expansion of $J_n(x)$ for $|x| \leqq 0.1$. Comparison of the

results of this routine with tables in Abramowitz and Stegun[23] for x values ranging from 0 to 5 yielded complete agreement in all cases.

The real spherical Bessel functions, $j_n(x)$ and $y_n(x)$, are computed recursively in the routine SBES using the algorithm of D. S. Drumheller,[24] an improved version of Miller's algorithm.

Miller's algorithm for $j_n(x)$ is: Let $g_{N+1}(x) \equiv 0$, $g_N(x) = 10^{-20}$ ⟨some small fixed number⟩; generate $g_{N-1}(x), \cdots, g_0(x)$ from the recurrence relations, $g_{n-1}(x) = (2n+1)x^{-1}g_n(x) - g_{n+1}(x)$, satisfied by the spherical Bessel functions;[25] compute $j_0(x) = x^{-1}\sin x$; normalize $j_l(x) = g_l(x)\cdot\dfrac{j_0(x)}{g_0(x)}$, $l = 0, 1, \cdots, N$.

Drumheller's algorithm generates ascending orders of $y_n(x)$ recursively, starting from Rayleigh's formulas[25] for $y_0(x)$, $y_1(x)$ up to some order N, where N is strictly limited only by the relation

$$y_N(x) \leqq \max - \text{the largest number the computer can handle (for the Honeywell 6070, max} \approx 10^{38}).$$

Letting

$$f_N(x) \equiv 0, \qquad f_{N-1}(x) = -(x^2 y_N(x))^{-1},$$

the algorithm then generates $f_{N-2}(x), \cdots, f_0(x)$ recursively, using the recurrence relations satisfied by the spherical Bessel functions. Although, as Drumheller points out, $f_0(x) = j_0(x)$ to some degree of precision (determined to a large extent by N), to ensure exactness in the lower orders of $j_n(x)$ and to shift any error to the higher-order, smaller-magnitude terms, we calculate $j_0(x)$ exactly and normalize $j_1(x), \cdots, j_N(x)$ as in Miller's algorithm. We compared these results to tables in Abramowitz and Stegun[26] and in the U. S. Math Tables Project[27] for the values of n and the range of x used in the least-squares fitting, and the agreement was excellent.

The complex spherical Bessel function, $j_n(z)$, is computed in the routine CSB, using an algorithm designed by A. E. Kaplan,[28] since Drumheller's (or Miller's) recursive algorithm yields inaccurate results for complex arguments with a significant imaginary part. Kaplan uses a Taylor series expansion to compute $j_N(z)$, $j_{N-1}(z)$, with $N \approx |z|^2$ for best results, then uses the backward recursion and normalization techniques discussed above to generate $j_{N-2}(z), \cdots, j_0(z)$. The use of the Taylor series for "large" complex arguments produces better starting values and, therefore, more accurate recursion results than either Drumheller's or Miller's algorithm. A "large" complex argument in

this case is one such that $|z| \geqq 9$; for $|z| < 9$, we use Drumheller's algorithm directly. In this routine, complex arithmetic is used for all internal computations; thus, since the Honeywell 6070 does not have double-precision complex arithmetic, all computations are in single precision. The accuracy of the results of this routine was checked by the following comparisons:

(*i*) For z real, they were compared to the spherical Bessel functions in Abramowitz and Stegun[26] and to results from SBES.

(*ii*) For z pure imaginary, they were compared (multiplied by appropriate factors) to the modified spherical Bessel functions of the first kind in Abramowitz and Stegun.[29]

(*iii*) For z any complex number, they were compared to Rayleigh's formulas[25] of order 0, 1, 2 (a straight computation of Rayleigh's formulas produces inaccurate values for higher orders).

For all values of z (from $|z| = 0$ to $|z| = 50$) and for order n as large as we could compute (or as large as we could compare to in tables or formulas), this routine produced answers agreeing in six to eight decimal places with the other results.

The associated Legendre function $P_n^m(\cos\theta)$ is computed directly from its series expansion, in powers of sines or cosines, as given by L. Robin.[30] It should be noted here that we use the following definition for $P_n^m(\cos\theta)$,

$$P_n^m(\cos\theta) = (-1)^m \sin^m\theta \frac{d^m P_n(\cos\theta)}{d(\cos\theta)^m},$$

whereas Stratton[31] omits the factor $(-1)^m$. Comparison of the results of this routine to explicit formulas for $m \leqq 4$, $n \leqq 7$ yielded eight places of agreement. Further checks of this routine against tables given by S. L. Belousov[32] and tables in the U. S. Math Tables Project[33] were done by S. Hoffberg for the values of m and n used in the least-squares fitting; her results also found complete agreement in all cases.

In the scattering problem, the matrix and vector sizes, as well as the highest order needed for the functions, depend on drop size and frequency; the sizes required by these parameters are discussed in the next section. Here, we give approximate running times for each of the three parts of the package and give some feeling for the correlation of the least-squares matrix size and the limit max m with the overall core storage and run time in the first part. In (71), we generally took $N_m \equiv N_0$, and λ_m dependent upon N_0 and decreasing with m (for $m \geqq 1$). Then the largest matrix size occurs when $m = 0$, and is thus a function

of N_0 alone. Tabulated below are some typical storage and run time figures for the least-squares fit.

max m	N_0	Largest Matrix Size	Approximate Storage	Approximate Run Time (hours)
6	13, 15, 17	164 × 68	70 K	0.13
7	15, 17, 19	180 × 76	80 K	0.2
8	19, 21, 23	212 × 92	103 K	0.39

In each of the above cases, we computed a set of coefficients a_{mn}, b_{mn}, c_{mn}, d_{mn}, $n = m, \cdots, N_0$ $(n \neq 0)$, for $m = 0, 1, \cdots,$ max m, and for each of the three values of N_0.

The second and third parts require much less storage because neither requires a large storage matrix. The total core requirement to run SQS is 12 K; it computes the quantities $S(0)$, Q_s, Q_t in typically less than 1 second. The third part uses 35 K to do the calculations for 428 rows and 92 columns (more rows are needed because the boundary fit is checked at θ values in addition to those used to obtain the coefficients); to check the L2FIT results for $N_0 = 23$ required approximately 0.05 hours.

As a final note, we save the coefficients computed from L2FIT on magnetic tape. At first, all coefficients were written on tape and the second and third parts used the tape as input; this proved very inefficient, due in part to the high cost of tape usage and in part to the fact that we were saving all data generated. We switched to writing the data from each run of L2FIT onto a high-speed disc file, using this as input to the other two parts; this change resulted in a noticeable cost reduction and allowed us to permanently save only the best data.

Table I — Raindrop parameters for different drop sizes

$\bar{a}$(cm)	a(cm)	ν
0.025	0.02458158	0.049375
0.05	0.04831913	0.0975
0.075	0.07120149	0.144375
0.1	0.09321698	0.19
0.125	0.11435330	0.234375
0.15	0.13459757	0.2775
0.175	0.15393617	0.319375
0.2	0.17235477	0.36
0.225	0.18983822	0.399375
0.25	0.20637045	0.4375
0.275	0.22193444	0.474375
0.3	0.23651206	0.51
0.325	0.25008398	0.544375
0.35	0.26262956	0.5775

Table II — Forward scattering amplitudes at 4 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$6.9215 \times 10^{-8} - 8.6909 \times 10^{-6}i$	$7.3309 \times 10^{-8} - 8.9487 \times 10^{-6}i$
0.05	$5.8893 \times 10^{-7} - 6.8473 \times 10^{-5}i$	$6.5803 \times 10^{-7} - 7.2647 \times 10^{-5}i$
0.075	$2.2523 \times 10^{-6} - 2.2822 \times 10^{-4}i$	$2.6370 \times 10^{-6} - 2.4970 \times 10^{-4}i$
0.1	$6.3705 \times 10^{-6} - 5.3582 \times 10^{-4}i$	$7.7684 \times 10^{-6} - 6.0510 \times 10^{-4}i$
0.125	$1.5414 \times 10^{-5} - 1.0399 \times 10^{-3}i$	$1.9518 \times 10^{-5} - 1.2134 \times 10^{-3}i$
0.15	$3.3841 \times 10^{-5} - 1.7918 \times 10^{-3}i$	$4.4527 \times 10^{-5} - 2.1630 \times 10^{-3}i$
0.175	$6.9433 \times 10^{-5} - 2.8484 \times 10^{-3}i$	$9.5324 \times 10^{-5} - 3.5628 \times 10^{-3}i$
0.2	$1.3555 \times 10^{-4} - 4.276 \times 10^{-3}i$	$1.9565 \times 10^{-4} - 5.552 \times 10^{-3}i$
0.225	$2.551 \times 10^{-4} - 6.154 \times 10^{-3}i$	$3.916 \times 10^{-4} - 8.318 \times 10^{-3}i$
0.25	$4.68 \times 10^{-4} - 8.59 \times 10^{-3}i$	$7.77 \times 10^{-4} - 1.212 \times 10^{-2}i$
0.275	$8.45 \times 10^{-4} - 1.170 \times 10^{-2}i$	$1.553 \times 10^{-3} - 1.735 \times 10^{-2}i$
0.3	$1.51 \times 10^{-3} - 1.57 \times 10^{-2}i$	$3.20 \times 10^{-3} - 2.46 \times 10^{-2}i$
0.325	$2.72 \times 10^{-3} - 2.08 \times 10^{-2}i$	$6.96 \times 10^{-3} - 3.47 \times 10^{-2}i$
0.35	$4.9 \times 10^{-3} - 2.7 \times 10^{-2}i$	$1.63 \times 10^{-2} - 4.8 \times 10^{-2}i$

VII. LEAST-SQUARES-FITTING RESULTS

We begin by discussing the different ways in which the least-squares-fitting program was checked. First, calculations were carried out at both 4 and 34.8 GHz for centered spherical raindrops, corresponding to $R(\theta) \equiv a$ in (15). The results were compared with the calculations based on the zero-order solution given in Section IV, corresponding to $\nu = 0$ in (47). Comparison was made for several values of a and different values of α, and excellent agreement was obtained for the far-field quantities, generally to six or seven significant figures. As expected, the far-field quantities are independent of the angle of incidence α.

Table III — Forward scattering amplitudes at 11 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$4.8189 \times 10^{-6} - 1.8194 \times 10^{-4}i$	$5.0841 \times 10^{-6} - 1.8734 \times 10^{-4}i$
0.05	$5.9423 \times 10^{-5} - 1.4657 \times 10^{-3}i$	$6.5120 \times 10^{-5} - 1.5555 \times 10^{-3}i$
0.075	$3.6151 \times 10^{-4} - 5.0737 \times 10^{-3}i$	$4.0920 \times 10^{-4} - 5.5577 \times 10^{-3}i$
0.1	$1.6675 \times 10^{-3} - 1.2532 \times 10^{-2}i$	$1.9652 \times 10^{-3} - 1.4186 \times 10^{-2}i$
0.125	$6.5377 \times 10^{-3} - 2.5280 \times 10^{-2}i$	$8.0766 \times 10^{-3} - 2.9419 \times 10^{-2}i$
0.15	$2.0109 \times 10^{-2} - 4.0292 \times 10^{-2}i$	$2.4674 \times 10^{-2} - 4.6783 \times 10^{-2}i$
0.175	$3.7203 \times 10^{-2} - 4.7914 \times 10^{-2}i$	$4.1732 \times 10^{-2} - 5.8328 \times 10^{-2}i$
0.2	$4.735 \times 10^{-2} - 5.734 \times 10^{-2}i$	$5.492 \times 10^{-2} - 8.098 \times 10^{-2}i$
0.225	$5.958 \times 10^{-2} - 7.504 \times 10^{-2}i$	$7.929 \times 10^{-2} - 1.1621 \times 10^{-1}i$
0.25	$7.67 \times 10^{-2} - 9.68 \times 10^{-2}i$	$1.173 \times 10^{-1} - 1.580 \times 10^{-1}i$
0.275	$9.89 \times 10^{-2} - 1.225 \times 10^{-1}i$	$1.725 \times 10^{-1} - 2.060 \times 10^{-1}i$
0.3	$1.29 \times 10^{-1} - 1.51 \times 10^{-1}i$	$2.51 \times 10^{-1} - 2.54 \times 10^{-1}i$
0.325	$1.67 \times 10^{-1} - 1.78 \times 10^{-1}i$	$3.54 \times 10^{-1} - 2.93 \times 10^{-1}i$
0.35	$2.1 \times 10^{-1} - 2.0 \times 10^{-1}i$	$4.8 \times 10^{-1} - 3.1 \times 10^{-1}i$

Table IV — Forward scattering amplitudes at 18.1 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$4.0158 \times 10^{-5} - 8.1579 \times 10^{-4}i$	$4.2246 \times 10^{-5} - 8.4000 \times 10^{-4}i$
0.05	$6.5425 \times 10^{-4} - 6.7265 \times 10^{-3}i$	$7.1168 \times 10^{-4} - 7.1424 \times 10^{-3}i$
0.075	$5.0674 \times 10^{-3} - 2.3476 \times 10^{-2}i$	$5.6959 \times 10^{-3} - 2.5717 \times 10^{-2}i$
0.1	$2.2608 \times 10^{-2} - 5.1254 \times 10^{-2}i$	$2.5696 \times 10^{-2} - 5.7588 \times 10^{-2}i$
0.125	$5.0374 \times 10^{-2} - 8.0587 \times 10^{-2}i$	$5.7722 \times 10^{-2} - 9.6353 \times 10^{-2}i$
0.15	$8.5403 \times 10^{-2} - 1.2201 \times 10^{-1}i$	$1.0834 \times 10^{-1} - 1.5714 \times 10^{-1}i$
0.175	$1.3950 \times 10^{-1} - 1.7392 \times 10^{-1}i$	$1.9921 \times 10^{-1} - 2.3098 \times 10^{-1}i$
0.2	$2.1700 \times 10^{-1} - 2.2871 \times 10^{-1}i$	$3.3903 \times 10^{-1} - 2.9883 \times 10^{-1}i$
0.225	$3.181 \times 10^{-1} - 2.742 \times 10^{-1}i$	$5.229 \times 10^{-1} - 3.320 \times 10^{-1}i$
0.25	$4.307 \times 10^{-1} - 2.999 \times 10^{-1}i$	$7.173 \times 10^{-1} - 3.136 \times 10^{-1}i$
0.275	$5.390 \times 10^{-1} - 3.089 \times 10^{-1}i$	$8.899 \times 10^{-1} - 2.633 \times 10^{-1}i$
0.3	$6.36 \times 10^{-1} - 3.12 \times 10^{-1}i$	$1.038 - 2.14 \times 10^{-1}i$
0.325	$7.22 \times 10^{-1} - 3.20 \times 10^{-1}i$	$1.179 - 1.84 \times 10^{-1}i$
0.35	$8.0 \times 10^{-1} - 3.4 \times 10^{-1}i$	$1.34 - 1.8 \times 10^{-1}i$

Moreover, the values of the coefficients obtained from least-squares fitting were checked against those calculated from (50) and (51), subject to (52), (53), and (119) to (122).

Next, the least-squares fit was carried out for spherical raindrops when the origin of the coordinate system was offset from the center of the raindrop, so that

$$R(\theta) = a[\delta \cos\theta + (1 - \delta^2 \sin^2\theta)^{\frac{1}{2}}]. \tag{73}$$

The calculations were done for different values of a, δ, and α, with the largest value of δ being 0.325 at 4 GHz and at 0.2 at 34.8 GHz. As ex-

Table V — Forward scattering amplitudes at 30 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$3.4513 \times 10^{-4} - 3.7481 \times 10^{-3}i$	$3.6235 \times 10^{-4} - 3.8595 \times 10^{-3}i$
0.05	$6.9873 \times 10^{-3} - 3.1187 \times 10^{-2}i$	$7.5859 \times 10^{-3} - 3.3144 \times 10^{-2}i$
0.075	$4.5783 \times 10^{-2} - 9.5267 \times 10^{-2}i$	$5.1071 \times 10^{-2} - 1.0490 \times 10^{-1}i$
0.1	$1.3415 \times 10^{-1} - 1.8677 \times 10^{-1}i$	$1.6165 \times 10^{-1} - 2.1613 \times 10^{-1}i$
0.125	$2.9755 \times 10^{-1} - 2.8388 \times 10^{-1}i$	$3.8979 \times 10^{-1} - 3.2171 \times 10^{-1}i$
0.15	$5.1727 \times 10^{-1} - 3.3781 \times 10^{-1}i$	$6.9041 \times 10^{-1} - 3.3561 \times 10^{-1}i$
0.175	$7.3731 \times 10^{-1} - 3.4278 \times 10^{-1}i$	$9.6534 \times 10^{-1} - 2.7438 \times 10^{-1}i$
0.2	$9.274 \times 10^{-1} - 3.461 \times 10^{-1}i$	$1.2001 - 2.302 \times 10^{-1}i$
0.225	$1.1122 - 3.885 \times 10^{-1}i$	$1.4661 - 2.419 \times 10^{-1}i$
0.25	$1.3309 - 4.693 \times 10^{-1}i$	$1.8221 - 2.627 \times 10^{-1}i$
0.275	$1.601 - 5.57 \times 10^{-1}i$	$2.245 - 2.21 \times 10^{-1}i$
0.3	$1.902 - 6.23 \times 10^{-1}i$	$2.662 - 1.20 \times 10^{-1}i$
0.325	$2.20 - 6.70 \times 10^{-1}i$	$3.06 - 2.4 \times 10^{-2}i$
0.35	$2.49 - 7.3 \times 10^{-1}i$	$3.50 + 4 \times 10^{-2}i$

Table VI — Forward scattering amplitudes at 30 GHz with $\alpha = 70°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$3.4679 \times 10^{-4} - 3.7610 \times 10^{-3}i$	$3.6200 \times 10^{-4} - 3.8594 \times 10^{-3}i$
0.05	$7.0267 \times 10^{-3} - 3.1417 \times 10^{-2}i$	$7.5553 \times 10^{-3} - 3.3145 \times 10^{-2}i$
0.075	$4.6185 \times 10^{-2} - 9.6618 \times 10^{-2}i$	$5.0856 \times 10^{-2} - 1.0513 \times 10^{-1}i$
0.1	$1.3701 \times 10^{-1} - 1.9131 \times 10^{-1}i$	$1.6133 \times 10^{-1} - 2.1727 \times 10^{-1}i$
0.125	$3.0867 \times 10^{-1} - 2.9154 \times 10^{-1}i$	$3.9039 \times 10^{-1} - 3.2498 \times 10^{-1}i$
0.15	$5.4041 \times 10^{-1} - 3.4346 \times 10^{-1}i$	$6.9413 \times 10^{-1} - 3.4113 \times 10^{-1}i$
0.175	$7.7133 \times 10^{-1} - 3.4313 \times 10^{-1}i$	$9.7408 \times 10^{-1} - 2.8097 \times 10^{-1}i$
0.2	$9.723 \times 10^{-1} - 3.425 \times 10^{-1}i$	$1.2143 - 2.357 \times 10^{-1}i$
0.225	$1.1720 - 3.832 \times 10^{-1}i$	$1.4838 - 2.454 \times 10^{-1}i$
0.25	$1.4123 - 4.621 \times 10^{-1}i$	$1.8412 - 2.678 \times 10^{-1}i$
0.275	$1.710 - 5.45 \times 10^{-1}i$	$2.269 - 2.34 \times 10^{-1}i$
0.3	$2.042 - 6.04 \times 10^{-1}i$	$2.699 - 1.41 \times 10^{-1}i$
0.325	$2.38 - 6.46 \times 10^{-1}i$	$3.11 - 4.8 \times 10^{-2}i$
0.35	$2.71 - 7.0 \times 10^{-1}i$	$3.56 + 1.5 \times 10^{-2}i$

pected, the far-field quantities are independent of δ, as well as α, and again excellent results were obtained. These calculations provided a nontrivial check on the programming of the boundary conditions, since $dR/d\theta \not\equiv 0$. In addition, they gave some idea of the increase in the number of terms in n that is required, a result of the ratio of maximum to minimum distance from the raindrop surface to the origin, which is necessarily greater than unity for oblate spheroidal raindrops.

As a final check on the least-squares-fitting program, calculations were carried out at 34.8 GHz for oblate spheroidal raindrops with small

Table VII — Forward scattering amplitudes at 30 GHz with $\alpha = 50°$ for different drop sizes

$\bar{a}$(cm)	$S_{\mathrm{I}}(0)$	$S_{\mathrm{II}}(0)$
0.025	$3.5101 \times 10^{-4} - 3.7936 \times 10^{-3}i$	$3.6111 \times 10^{-4} - 3.8590 \times 10^{-3}i$
0.05	$7.1265 \times 10^{-3} - 3.1998 \times 10^{-2}i$	$7.4779 \times 10^{-3} - 3.3147 \times 10^{-2}i$
0.075	$4.7204 \times 10^{-2} - 1.0005 \times 10^{-1}i$	$5.0312 \times 10^{-2} - 1.0571 \times 10^{-1}i$
0.1	$1.4431 \times 10^{-1} - 2.0284 \times 10^{-1}i$	$1.6054 \times 10^{-1} - 2.2016 \times 10^{-1}i$
0.125	$3.3714 \times 10^{-1} - 3.1102 \times 10^{-1}i$	$3.9193 \times 10^{-1} - 3.3329 \times 10^{-1}i$
0.15	$6.0012 \times 10^{-1} - 3.5745 \times 10^{-1}i$	$7.0369 \times 10^{-1} - 3.5523 \times 10^{-1}i$
0.175	$8.5967 \times 10^{-1} - 3.4211 \times 10^{-1}i$	$9.9679 \times 10^{-1} - 2.9777 \times 10^{-1}i$
0.2	$1.0888 - 3.278 \times 10^{-1}i$	$1.2518 - 2.491 \times 10^{-1}i$
0.225	$1.3247 - 3.580 \times 10^{-1}i$	$1.5306 - 2.520 \times 10^{-1}i$
0.25	$1.6145 - 4.252 \times 10^{-1}i$	$1.8902 - 2.752 \times 10^{-1}i$
0.275	$1.974 - 4.91 \times 10^{-1}i$	$2.327 - 2.58 \times 10^{-1}i$
0.3	$2.379 - 5.27 \times 10^{-1}i$	$2.785 - 1.87 \times 10^{-1}i$
0.325	$2.80 - 5.45 \times 10^{-1}i$	$3.24 - 1.03 \times 10^{-1}i$
0.35	$3.24 - 5.7 \times 10^{-1}i$	$3.72 - 4 \times 10^{-2}i$

Table VIII — Total and scattering cross sections at 4 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	1.2393×10^{-6}	1.3126×10^{-6}	8.9950×10^{-10}	9.5367×10^{-10}
0.05	1.0545×10^{-5}	1.1782×10^{-5}	5.5428×10^{-8}	6.2399×10^{-8}
0.075	4.0327×10^{-5}	4.7215×10^{-5}	6.0811×10^{-7}	7.2816×10^{-7}
0.1	1.1406×10^{-4}	1.3909×10^{-4}	3.2922×10^{-6}	4.2012×10^{-6}
0.125	2.7599×10^{-4}	3.4947×10^{-4}	1.2106×10^{-5}	1.6500×10^{-5}
0.15	6.0592×10^{-4}	7.9725×10^{-4}	3.4869×10^{-5}	5.0884×10^{-5}
0.175	1.2432×10^{-3}	1.7068×10^{-3}	8.4904×10^{-5}	1.3303×10^{-4}
0.2	2.427×10^{-3}	3.503×10^{-3}	1.830×10^{-4}	3.088×10^{-4}
0.225	4.568×10^{-3}	7.011×10^{-3}	3.599×10^{-4}	6.567×10^{-4}
0.25	8.38×10^{-3}	1.391×10^{-2}	6.60×10^{-4}	1.309×10^{-3}
0.275	1.51×10^{-2}	2.78×10^{-2}	1.15×10^{-3}	2.50×10^{-3}
0.3	2.71×10^{-2}	5.73×10^{-2}	1.93×10^{-3}	4.7×10^{-3}
0.325	4.87×10^{-2}	1.246×10^{-1}	3.2×10^{-3}	8.8×10^{-3}
0.35	8.8×10^{-2}	2.92×10^{-1}	5.2×10^{-3}	1.78×10^{-2}

eccentricity, corresponding to $\nu = 0$, 0.05, 0.1, and 0.15 in (63). The calculations were done for $\alpha = 90°$ and for $a = 0.025(0.025)0.275$. Corresponding to (48), the total cross section may be expanded in the form $Q_t = Q_t^{(0)} + \nu Q_t^{(1)} + \cdots$. Values of $Q_t^{I(1)}$ and $Q_t^{II(1)}$ were obtained from the least-squares results by extrapolation and were compared with the perturbation values given by Oguchi.[4] Unfortunately, there were significant discrepancies for the larger drop sizes, the largest error being more than 17 percent. Consequently, we did the perturbation calculations ourselves and obtained results differing from our extrap-

Table IX — Total and scattering cross sections at 11 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	1.1407×10^{-5}	1.2035×10^{-5}	5.1556×10^{-8}	5.4666×10^{-8}
0.05	1.4066×10^{-4}	1.5415×10^{-4}	3.2095×10^{-6}	3.6172×10^{-6}
0.075	8.5573×10^{-4}	9.6863×10^{-4}	3.5933×10^{-5}	4.3203×10^{-5}
0.1	3.9471×10^{-3}	4.6519×10^{-3}	2.0259×10^{-4}	2.6114×10^{-4}
0.125	1.5476×10^{-2}	1.9118×10^{-2}	8.1236×10^{-4}	1.1302×10^{-3}
0.15	4.7600×10^{-2}	5.8407×10^{-2}	2.7238×10^{-3}	4.0701×10^{-3}
0.175	8.8064×10^{-2}	9.8785×10^{-2}	7.3803×10^{-3}	1.1535×10^{-2}
0.2	1.1208×10^{-1}	1.3001×10^{-1}	1.559×10^{-2}	2.643×10^{-2}
0.225	1.4103×10^{-1}	1.8769×10^{-1}	2.815×10^{-2}	5.325×10^{-2}
0.25	1.815×10^{-1}	2.777×10^{-1}	4.60×10^{-2}	9.76×10^{-2}
0.275	2.341×10^{-1}	4.084×10^{-1}	7.21×10^{-2}	1.701×10^{-1}
0.3	3.05×10^{-1}	5.94×10^{-1}	1.09×10^{-1}	2.84×10^{-1}
0.325	3.95×10^{-1}	8.4×10^{-1}	1.57×10^{-1}	4.5×10^{-1}
0.35	5.0×10^{-1}	1.13	2.1×10^{-1}	6.5×10^{-1}

Table X — Total and scattering cross sections at 18.1 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	3.5118×10^{-5}	3.6944×10^{-5}	3.7881×10^{-7}	4.0173×10^{-7}
0.05	5.7214×10^{-4}	6.2236×10^{-4}	2.4119×10^{-5}	2.7242×10^{-5}
0.075	4.4314×10^{-3}	4.9810×10^{-3}	2.8514×10^{-4}	3.4542×10^{-4}
0.1	1.9770×10^{-2}	2.2471×10^{-2}	1.7668×10^{-3}	2.3060×10^{-3}
0.125	4.4052×10^{-2}	5.0478×10^{-2}	7.1078×10^{-3}	1.0013×10^{-2}
0.15	7.4684×10^{-2}	9.4745×10^{-2}	1.9966×10^{-2}	3.0950×10^{-2}
0.175	1.2199×10^{-1}	1.7421×10^{-1}	4.3974×10^{-2}	7.4834×10^{-2}
0.2	1.8976×10^{-1}	2.9648×10^{-1}	8.312×10^{-2}	1.5229×10^{-1}
0.225	2.782×10^{-1}	4.572×10^{-1}	1.377×10^{-1}	2.640×10^{-1}
0.25	3.766×10^{-1}	6.273×10^{-1}	1.999×10^{-1}	3.888×10^{-1}
0.275	4.714×10^{-1}	7.782×10^{-1}	2.598×10^{-1}	5.028×10^{-1}
0.3	5.56×10^{-1}	9.08×10^{-1}	3.12×10^{-1}	6.00×10^{-1}
0.325	6.32×10^{-1}	1.031	3.54×10^{-1}	6.89×10^{-1}
0.35	7.0×10^{-1}	1.17	3.9×10^{-1}	7.8×10^{-1}

olated least-squares results by at most $\frac{1}{2}$ percent, which is reasonably consistent with the error to be expected from the extrapolation. Oguchi[34] has since redone his calculations, and he agrees with our perturbation results. The same good agreement was obtained between the extrapolated and perturbation values of $S_{\mathrm{I}}^{(1)}(0)$, $S_{\mathrm{II}}^{(1)}(0)$, $Q_s^{\mathrm{I}(1)}$, and $Q_s^{\mathrm{II}(1)}$.

After the least-squares-fitting program had been checked in the above manner, we carried out calculations for oblate spheroidal raindrops corresponding to (63), with a and ν given by (64). Here $\bar{a}$ is the

Table XI — Total and scattering cross sections at 30 GHz with $\alpha = 90°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	1.0986×10^{-4}	1.1534×10^{-4}	2.8823×10^{-6}	3.0581×10^{-6}
0.05	2.2241×10^{-3}	2.4147×10^{-3}	1.9542×10^{-4}	2.2197×10^{-4}
0.075	1.4573×10^{-2}	1.6256×10^{-2}	2.5201×10^{-3}	3.1009×10^{-3}
0.1	4.2701×10^{-2}	5.1454×10^{-2}	1.4263×10^{-2}	1.9185×10^{-2}
0.125	9.4713×10^{-2}	1.2407×10^{-1}	4.3818×10^{-2}	6.2849×10^{-2}
0.15	1.6465×10^{-1}	2.1977×10^{-1}	8.9103×10^{-2}	1.2903×10^{-1}
0.175	2.3469×10^{-1}	3.0728×10^{-1}	1.3595×10^{-1}	1.9338×10^{-1}
0.2	2.9521×10^{-1}	3.8199×10^{-1}	1.7408×10^{-1}	2.4628×10^{-1}
0.225	3.5403×10^{-1}	4.6666×10^{-1}	2.0731×10^{-1}	3.0183×10^{-1}
0.25	4.236×10^{-1}	5.800×10^{-1}	2.450×10^{-1}	3.766×10^{-1}
0.275	5.095×10^{-1}	7.146×10^{-1}	2.938×10^{-1}	4.710×10^{-1}
0.3	6.05×10^{-1}	8.47×10^{-1}	3.51×10^{-1}	5.69×10^{-1}
0.325	7.01×10^{-1}	9.74×10^{-1}	4.08×10^{-1}	6.62×10^{-1}
0.35	7.9×10^{-1}	1.11	4.6×10^{-1}	7.6×10^{-1}

Table XII — Total and scattering cross sections at 30 GHz with $\alpha = 70°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	1.1039×10^{-4}	1.1523×10^{-4}	2.9029×10^{-6}	3.0582×10^{-6}
0.05	2.2367×10^{-3}	2.4049×10^{-3}	1.9851×10^{-4}	2.2196×10^{-4}
0.075	1.4701×10^{-2}	1.6188×10^{-2}	2.5867×10^{-3}	3.1000×10^{-3}
0.1	4.3613×10^{-2}	5.1354×10^{-2}	1.4859×10^{-2}	1.9213×10^{-2}
0.125	9.8252×10^{-2}	1.2427×10^{-1}	4.6292×10^{-2}	6.3156×10^{-2}
0.15	1.7202×10^{-1}	2.2095×10^{-1}	9.4694×10^{-2}	1.3017×10^{-1}
0.175	2.4552×10^{-1}	3.1006×10^{-1}	1.4482×10^{-1}	1.9599×10^{-1}
0.2	3.0950×10^{-1}	3.8653×10^{-1}	1.8634×10^{-1}	2.5071×10^{-1}
0.225	3.7305×10^{-1}	4.7231×10^{-1}	2.2389×10^{-1}	3.0787×10^{-1}
0.25	4.496×10^{-1}	5.861×10^{-1}	2.676×10^{-1}	3.837×10^{-1}
0.275	5.442×10^{-1}	7.223×10^{-1}	3.240×10^{-1}	4.796×10^{-1}
0.3	6.50×10^{-1}	8.59×10^{-1}	3.90×10^{-1}	5.81×10^{-1}
0.325	7.57×10^{-1}	9.91×10^{-1}	4.58×10^{-1}	6.79×10^{-1}
0.35	8.6×10^{-1}	1.13	5.2×10^{-1}	7.8×10^{-1}

radius (in centimeters) of the equivolumic spherical drop, and the calculations were done for $\bar{a} = 0.025(0.025)0.35$. The corresponding values of a and ν are given in Table I. The values taken for the wavelength $\lambda = 2\pi/k_0$ were (in centimeters) 7.5, 2.727, 1.6575, and 1.0, corresponding approximately to frequencies of 4, 11, 18.1, and 30 GHz. At 20°C, the refractive indices $N = 7.884 + 2.184i$ at 11 GHz, $N = 6.859 + 2.716i$ at 18.1 GHz, and $N = 5.581 + 2.848i$ at 30 GHz were obtained from an elaborate fitting equation in a recently published survey[8] of available measured data. Since the calculations at 4 GHz were made at an earlier date, the value $N = 8.77 + 0.915i$, taken from the older

Table XIII — Total and scattering cross sections at 30 GHz with $\alpha = 50°$ for different drop sizes

$\bar{a}$(cm)	Q_t^{I}(cm)^2	Q_t^{II}(cm)^2	Q_s^{I}(cm)^2	Q_s^{II}(cm)^2
0.025	1.1173×10^{-4}	1.1495×10^{-4}	2.9552×10^{-6}	3.0584×10^{-6}
0.05	2.2684×10^{-3}	2.3803×10^{-3}	2.0634×10^{-4}	2.2193×10^{-4}
0.075	1.5025×10^{-2}	1.6015×10^{-2}	2.7560×10^{-3}	3.0978×10^{-3}
0.1	4.5936×10^{-2}	5.1102×10^{-2}	1.6378×10^{-2}	1.9286×10^{-2}
0.125	1.0732×10^{-1}	1.2476×10^{-1}	5.2632×10^{-2}	6.3941×10^{-2}
0.15	1.9102×10^{-1}	2.2399×10^{-1}	1.0914×10^{-1}	1.3308×10^{-1}
0.175	2.7364×10^{-1}	3.1729×10^{-1}	1.6796×10^{-1}	2.0272×10^{-1}
0.2	3.4658×10^{-1}	3.9846×10^{-1}	2.1846×10^{-1}	2.6230×10^{-1}
0.225	4.2166×10^{-1}	4.8721×10^{-1}	2.6706×10^{-1}	3.2371×10^{-1}
0.25	5.139×10^{-1}	6.017×10^{-1}	3.253×10^{-1}	4.021×10^{-1}
0.275	6.283×10^{-1}	7.405×10^{-1}	3.996×10^{-1}	5.009×10^{-1}
0.3	7.57×10^{-1}	8.87×10^{-1}	4.87×10^{-1}	6.09×10^{-1}
0.325	8.9×10^{-1}	1.03	5.8×10^{-1}	7.2×10^{-1}
0.35	1.03	1.18	6.75×10^{-1}	8.3×10^{-1}

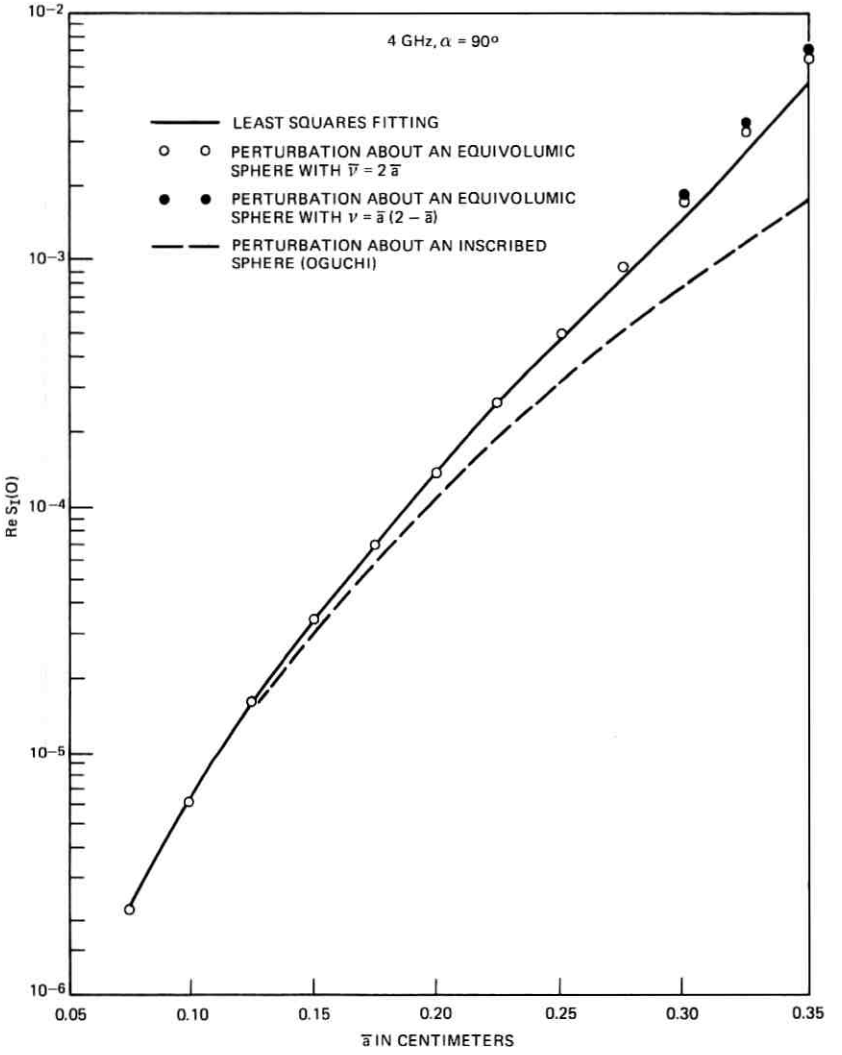

Fig. 3—Comparison of perturbation approximations to the least-squares-fitting value of Re $S_I(0)$ at 4 GHz with $\alpha = 90°$ as a function of drop size.

literature, was used, rather than $N = 8.78 + 0.977i$. The angle of incidence α was taken to be 90° at 4, 11, and 18.1 GHz, while at 30 GHz the calculations were done for $\alpha = 70°$ and $\alpha = 50°$ also.

The calculated values of the forward scattering amplitudes $S_I(0)$ and $S_{II}(0)$ are given in Tables II to VII, and those of the total cross

sections Q_t^{I} and Q_t^{II} and the scattering cross sections Q_s^{I} and Q_s^{II} are given in Tables VIII to XIII, all rounded in the last significant figure. The values of the absorption cross sections Q_a^{I} and Q_a^{II} follow from (38).

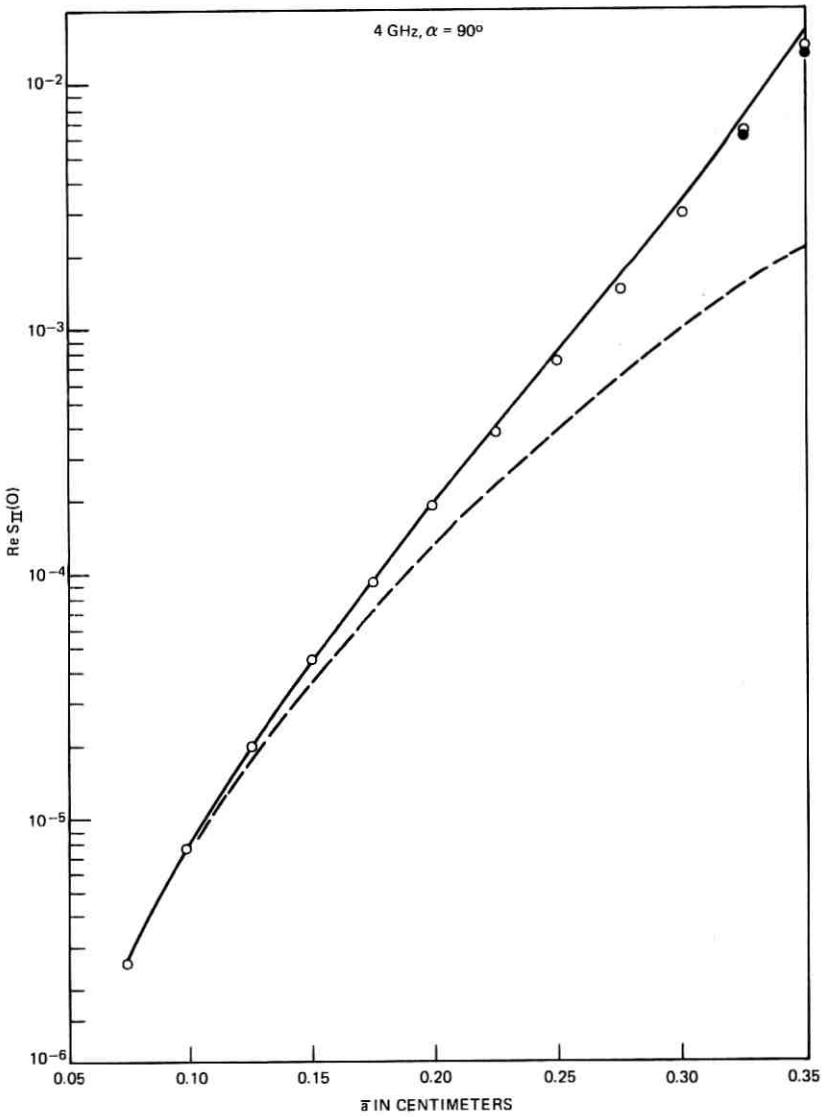

Fig. 4—Comparison of perturbation approximations to the least-squares-fitting value of Re $S_{\mathrm{II}}(0)$ at 4 GHz with $\alpha = 90°$ as a function of drop size.

The accuracy of the least-squares fit of the boundary conditions decreases with increasing drop size, because of the increase in eccentricity, which is why fewer significant figures are given in the tables for the larger drop sizes. Except for the smaller drop sizes, for which the results could be given more accurately, the number of significant figures reflects the degree of convergence of the results, as evidenced by increasing the upper limit of n in the least-squares fit by 2 and by 4. The

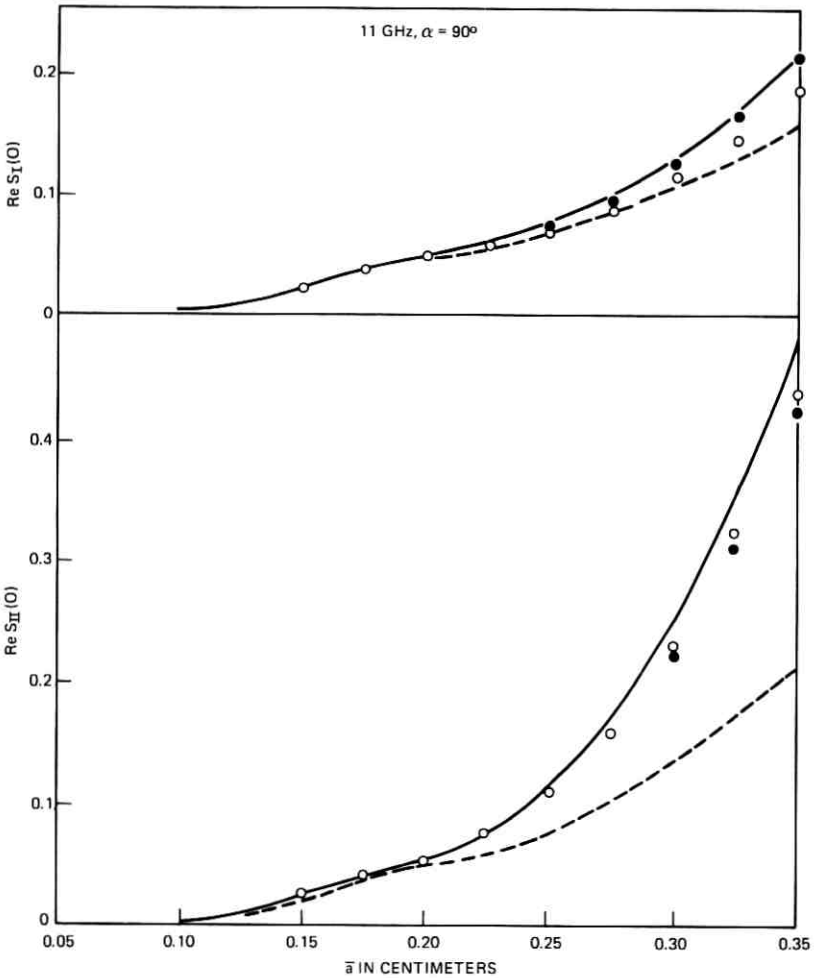

Fig. 5—Comparison of perturbation approximations to the least-squares-fitting values of Re $S_{\mathrm{I}}(0)$ and Re $S_{\mathrm{II}}(0)$ at 11 GHz with $\alpha = 90°$ as a function of drop size.

accuracy of the far-field results is generally at least one order of magnitude greater than that of the fit of the boundary conditions for the reasons discussed in Section V. We note that the largest drops occur only at the heaviest rain rates,[7] and then only a small percentage of them, so that the lower accuracy of the results for these drops is acceptable when summing over the drop size distribution.

The number of terms required to obtain the desired accuracy for the far-field quantities and to adequately satisfy the boundary conditions increases with both drop size and with frequency. At 4 GHz, it was found that max $m = 4$ and max $n = 17$ were sufficient for the largest drop size. For $\alpha = 90°$ at 30 GHz, it was necessary to take max $m = 8$ and max $n = 23$ for the largest drop size. In this latter case, more than half the capacity of the Honeywell 6070 computer was used. In some cases, it was found that max n or max m could not be increased without causing overflow in some of the subroutines, in particular SBES and L2FIT.

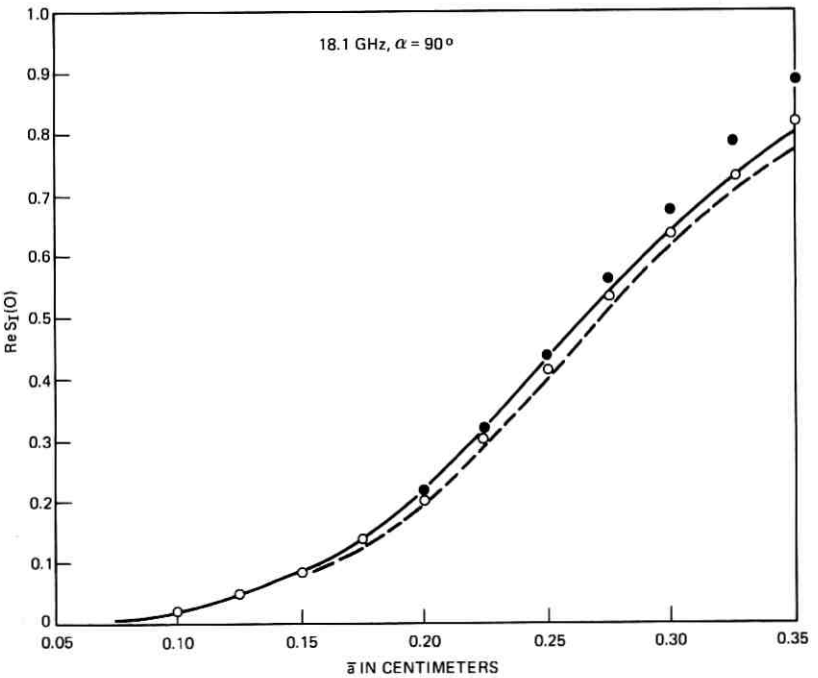

Fig. 6—Comparison of perturbation approximations to the least-squares-fitting value of Re $S_1(0)$ at 18.1 GHz with $\alpha = 90°$ as a function of drop size.

To check on the advantage of using least-squares fitting (with approximately twice as many fitting points as unknown coefficients) rather than collocation, we used collocation in several cases at different frequencies and for different drop sizes. Our general conclusion is that, for the same max m and max n, results may be obtained by collocation for the far-field quantities that are almost as accurate as those obtained by least-squares fitting. However, there are much larger errors in the boundary conditions (in between the fitting points) with collocation

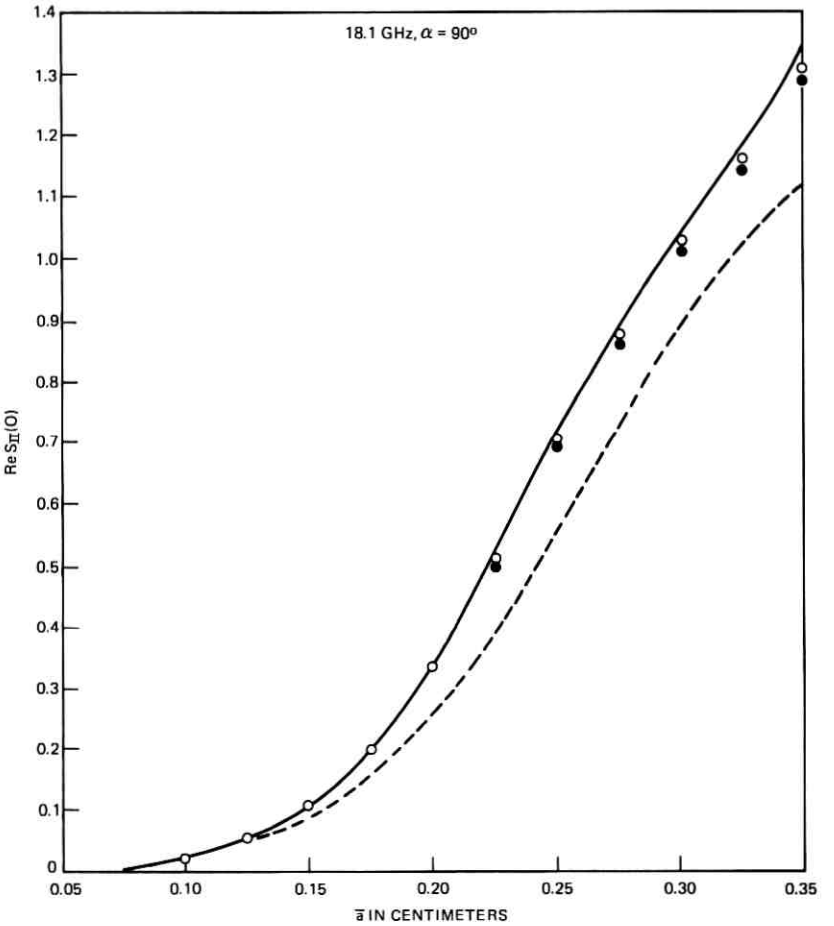

Fig. 7—Comparison of perturbation approximations to the least-squares-fitting value of Re $S_{11}(0)$ at 18.1 GHz with $\alpha = 90°$ as a function of drop size.

than with least-squares fitting. For the larger, more eccentric raindrops, some of these errors were of the order of 100 percent, which seem to be unacceptable. However, for the smaller raindrops, the errors in the boundary conditions with collocation are acceptable. Since the cost of carrying out the least-squares fit is less when fewer fitting points are used, collocation has the advantage of reducing the cost, although fewer terms are required, anyway, to obtain the desired accuracy for the smaller raindrops. It is possible that the collocation fit may be improved

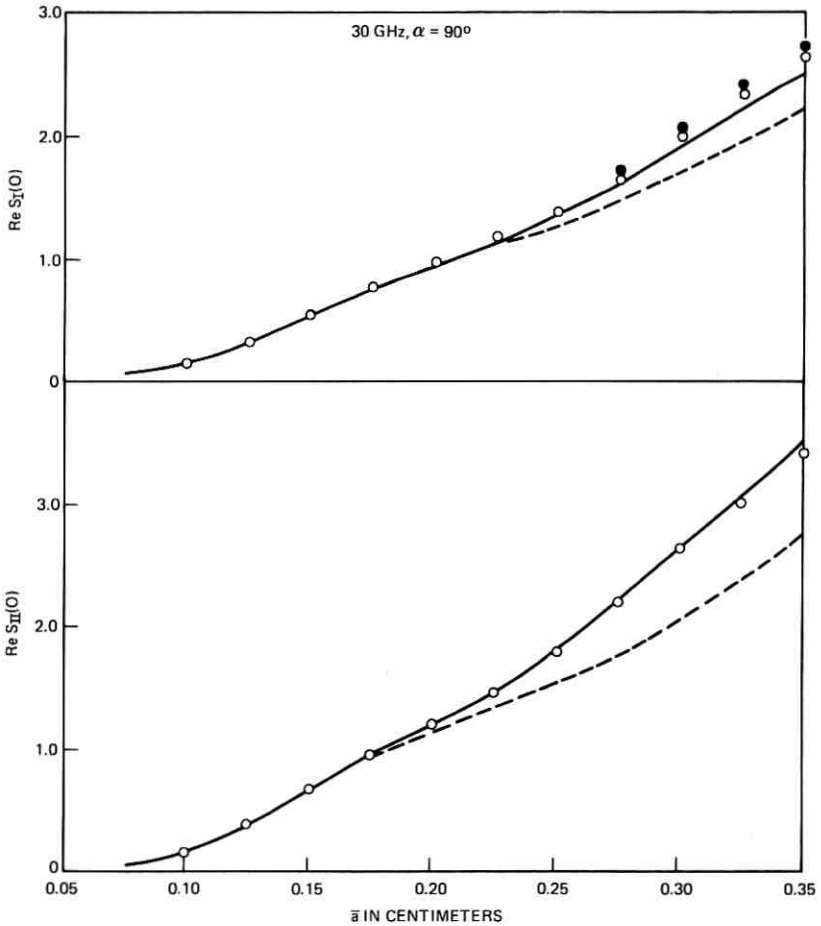

Fig. 8—Comparison of perturbation approximations to the least-squares-fitting values of Re $S_{\mathrm{I}}(0)$ and Re $S_{\mathrm{II}}(0)$ at 30 GHz with $\alpha = 90°$ as a function of drop size.

by satisfying the boundary conditions at nonuniformly spaced points, but we have not investigated this.

As a check on the point-matching (collocation) results of Oguchi,[5] we carried out the least-squares fitting for $\alpha = 90°$, at 19.3 GHz for $\bar{a} = 0.3$ and at 34.8 GHz for $\bar{a} = 0.075$, 0.15, 0.225, and 0.3, using

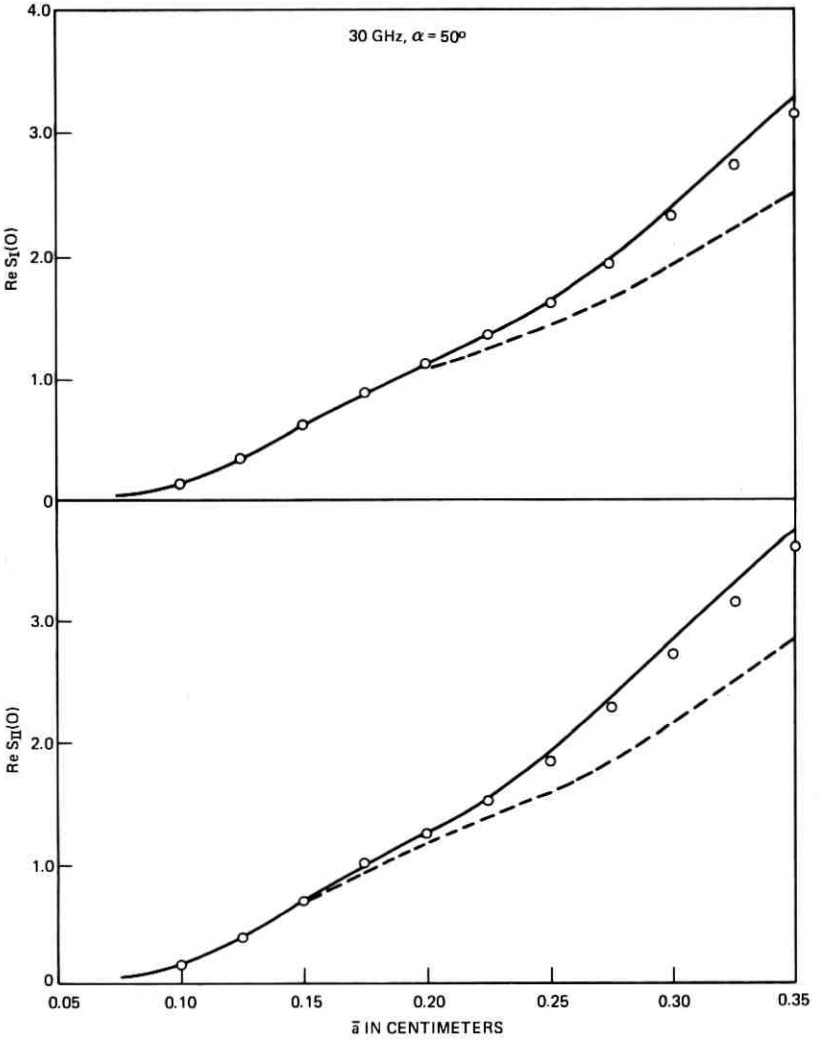

Fig. 9—Comparison of perturbation approximations to the least-squares-fitting values of Re $S_{\mathrm{I}}(0)$ and Re $S_{\mathrm{II}}(0)$ at 30 GHz with $\alpha = 50°$ as a function of drop size.

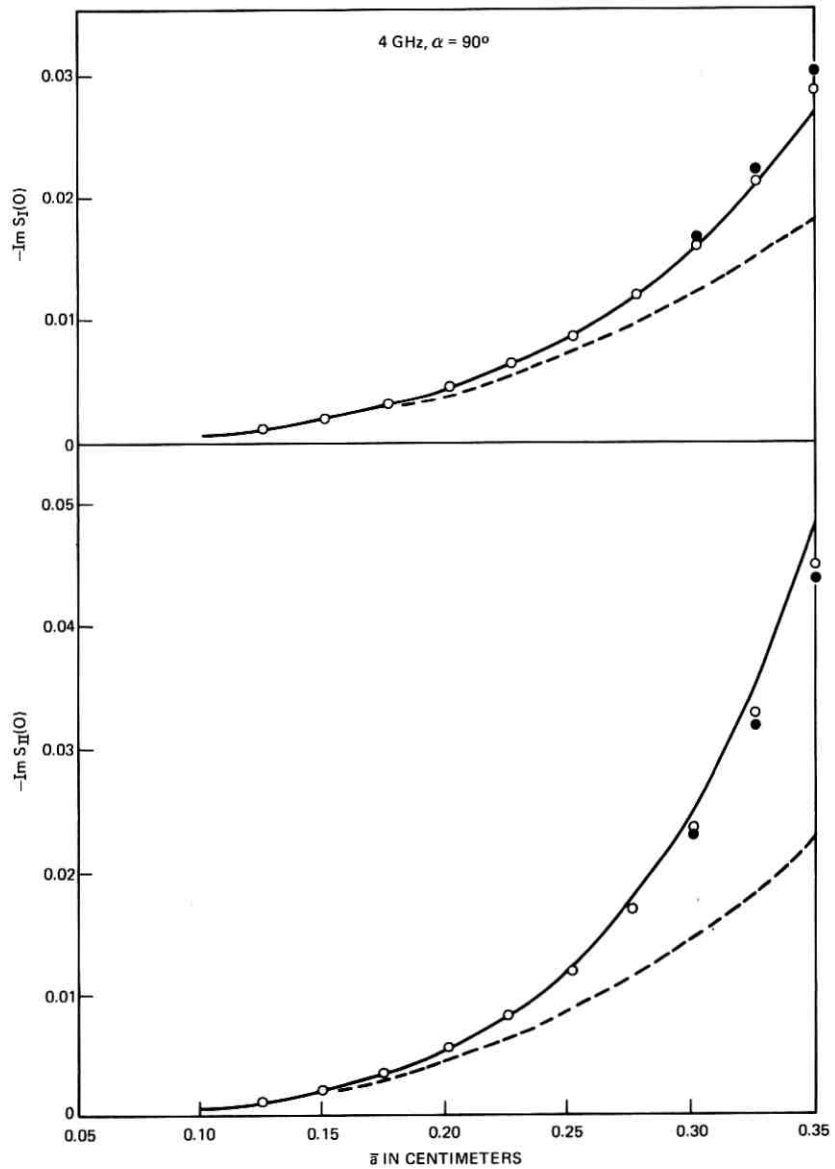

Fig. 10—Comparison of perturbation approximations to the least-squares-fitting values of Im $S_{\mathrm{I}}(0)$ and Im $S_{\mathrm{II}}(0)$ at 4 GHz with $\alpha = 90°$ as a function of drop size.

Oguchi's relationship[5]

$$\frac{a}{b} = \left(1 - \frac{4.1}{4.5}\,\bar{a}\right), \tag{74}$$

instead of the first relationship in (1). Our results for the forward scattering amplitudes are consistent with his point-matching values, but they may be given to greater accuracy. Our truncated values for $\bar{a} = 0.3$ are given below where, in Oguchi's notation,[5]

$$f^v \times 10^3 = \frac{-5\lambda}{\pi}\, iS_{\mathrm{I}}^*(0), \qquad f^h \times 10^3 = \frac{-5\lambda}{\pi}\, iS_{\mathrm{II}}^*(0), \tag{75}$$

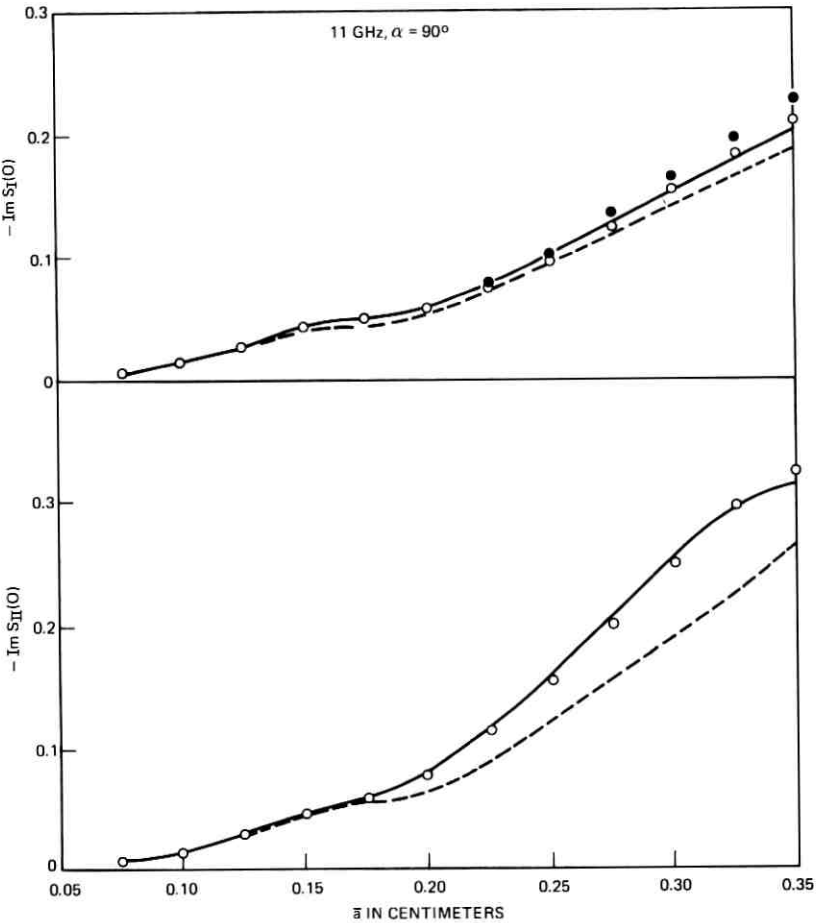

Fig. 11—Comparison of perturbation approximations to the least-squares-fitting values of Im $S_{\mathrm{I}}(0)$ and Im $S_{\mathrm{II}}(0)$ at 11 GHz with $\alpha = 90°$ as a function of drop size.

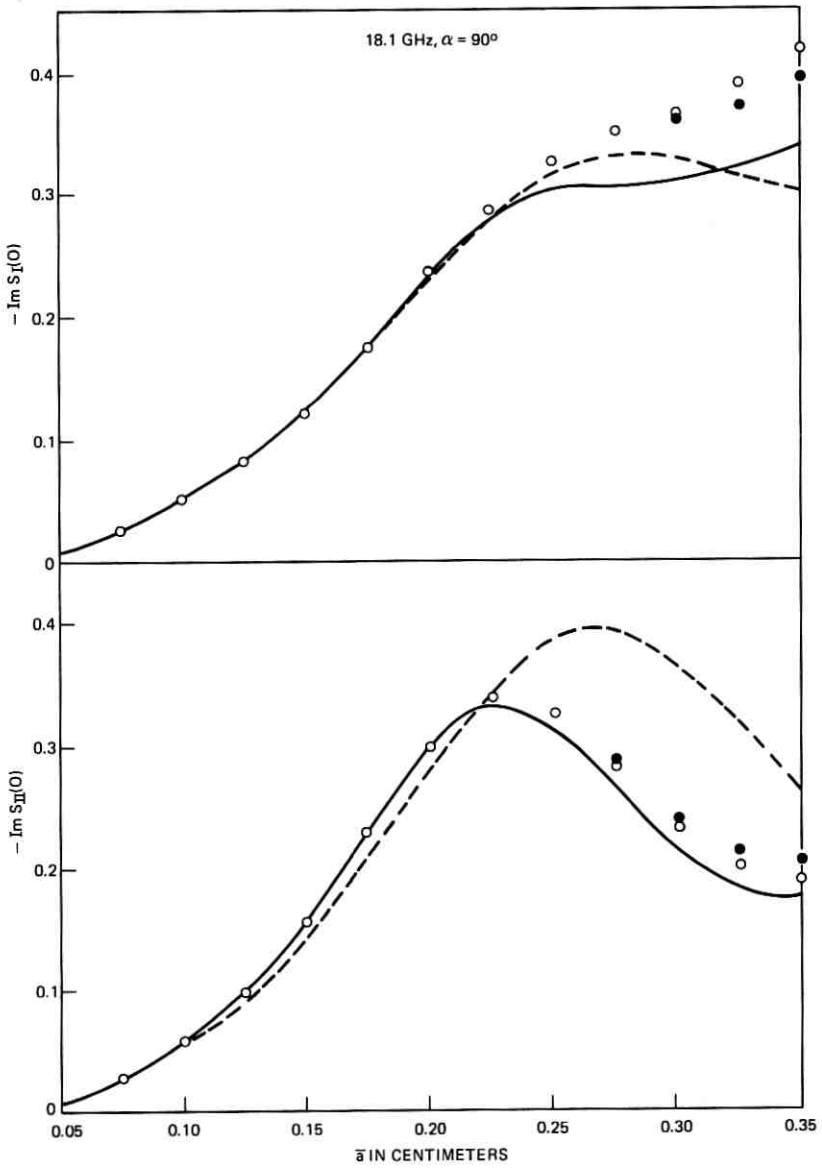

Fig. 12—Comparison of perturbation approximations to the least-squares-fitting values of Im $S_{\mathrm{I}}(0)$ and Im $S_{\mathrm{II}}(0)$ at 18.1 GHz with $\alpha = 90°$ as a function of drop size.

and the vertical lines indicate where Oguchi truncated his results:

GHz	$f^v \times 10^3$	$f^h \times 10^3$
19.3	$0.81\,|30 - 1.884\,|i$	$0.510\,|9 - 2.81\,|7i$
34.8	$0.917\,|8 - 3.73\,|0i$	$|0.0646 - 4.71\,|0i$

The values taken[34] for the wavelength λ (in centimeters) were 1.5533330 and 0.86135810, corresponding approximately to frequencies of 19.3 and 34.8 GHz, with refractive indices $N = 6.5449188 + 2.8104040i$ and $N = 5.0487284 + 2.7948416i$, respectively.

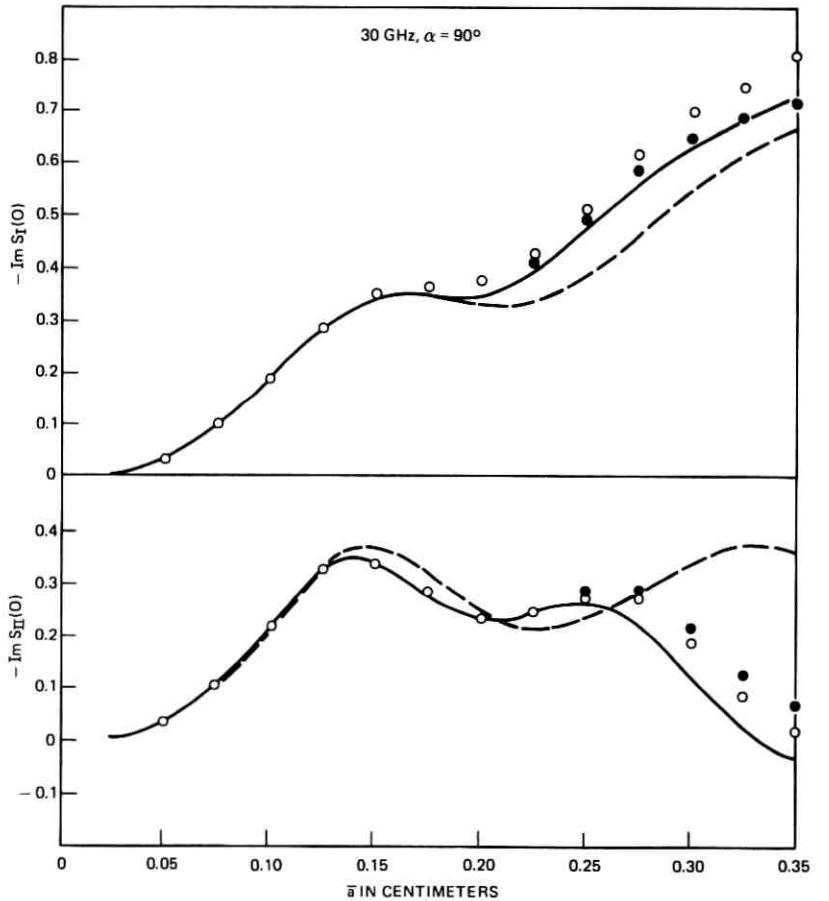

Fig. 13—Comparison of perturbation approximations to the least-squares-fitting values of Im $S_I(0)$ and Im $S_{II}(0)$ at 30 GHz with $\alpha = 90°$ as a function of drop size.

The main reason for the greater accuracy of our results is that we took larger values of max n than Oguchi, who did the point-matching at both frequencies for max $n = 12$ and 14, with max $m =$ max n. For $\bar{a} = 0.3$ we took max $m = 7$ at 19.3 GHz and max $m = 9$ at 34.8 GHz, which were sufficient, and took max $n = 21$ at both frequencies for the least-squares fitting. We also used collocation for $\bar{a} = 0.3$ and max $n = 12$, 14, and 21. For max $n = 21$ the collocation results differ by at most 1 in the last decimal place from the results given above, but errors in some boundary conditions were of the order of 10 percent, as compared with much less than 1 percent for least-squares fitting. This is consistent with our general conclusion discussed earlier in this section. We point out that the raindrops satisfying (74) are less eccentric than those satisfying (1), so that the overall errors are correspondingly smaller. For collocation with max $n = 14$, some errors in the boundary conditions were close to 100 percent, which explains why, with point-matching, Oguchi did not give any significant figures for Re f^h at 34.8 GHz, for either $\bar{a} = 0.3$ or $\bar{a} = 0.325$.

Although Oguchi[5] gives four significant figures for f^v and f^h at 34.8

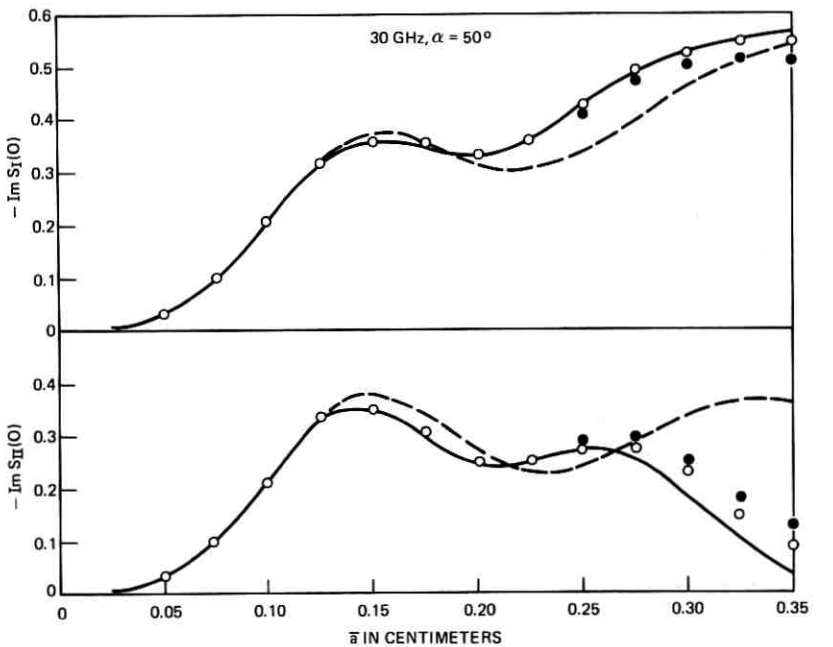

Fig. 14—Comparison of perturbation approximations to the least-squares-fitting values of Im $S_I(0)$ and Im $S_{II}(0)$ at 30 GHz with $\alpha = 50°$ as a function of drop size.

GHz for all drop sizes corresponding to his solution in terms of spheroidal wave functions (with modal sums truncated at 9), these values are not consistent with his point-matching ones for the larger drop

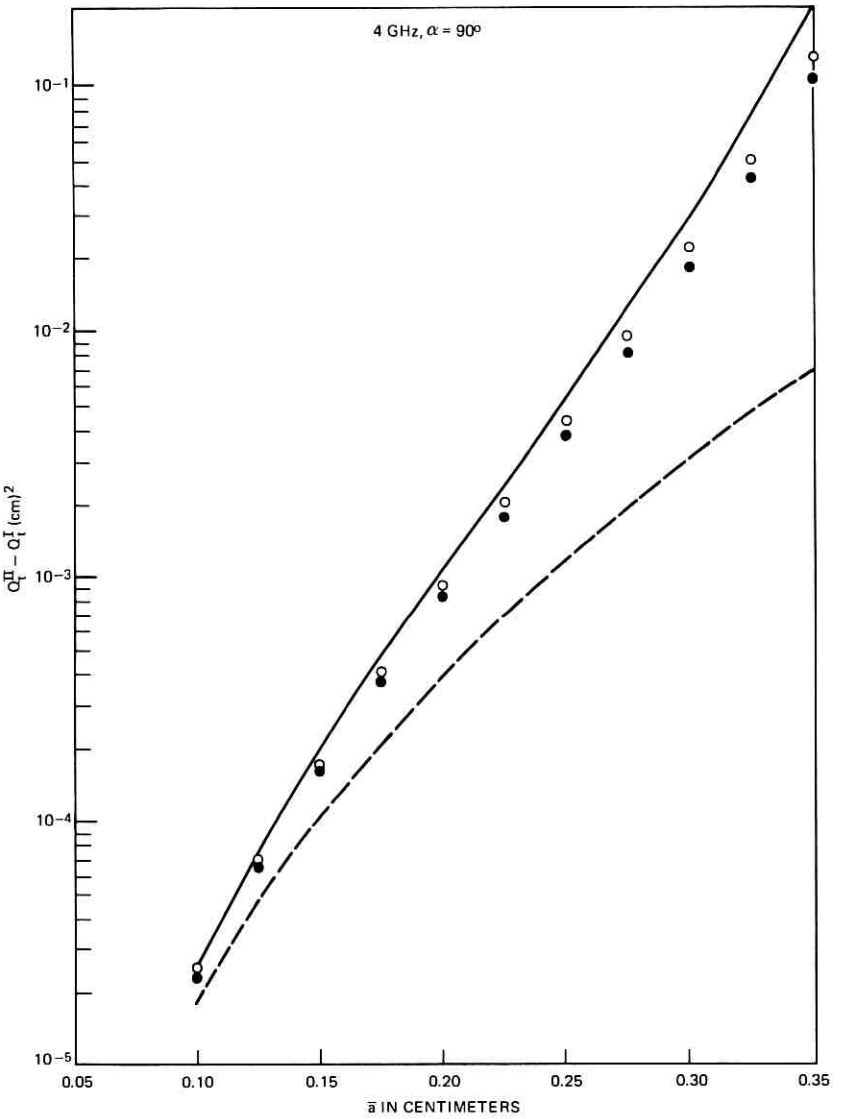

Fig. 15—Comparison of perturbation approximations to the least-squares-fitting value of $Q_t^{II} - Q_t^I$ at 4 GHz with $\alpha = 90°$ as a function of drop size.

sizes. For $\bar{a} = 0.3$, he gives $f^h \times 10^3 = 0.06470 - 4.709i$, which is in fact quite close to our value, but he also gives $f^v \times 10^3 = 0.9100 - 3.726i$, with real part differing by almost 1 percent from our value.

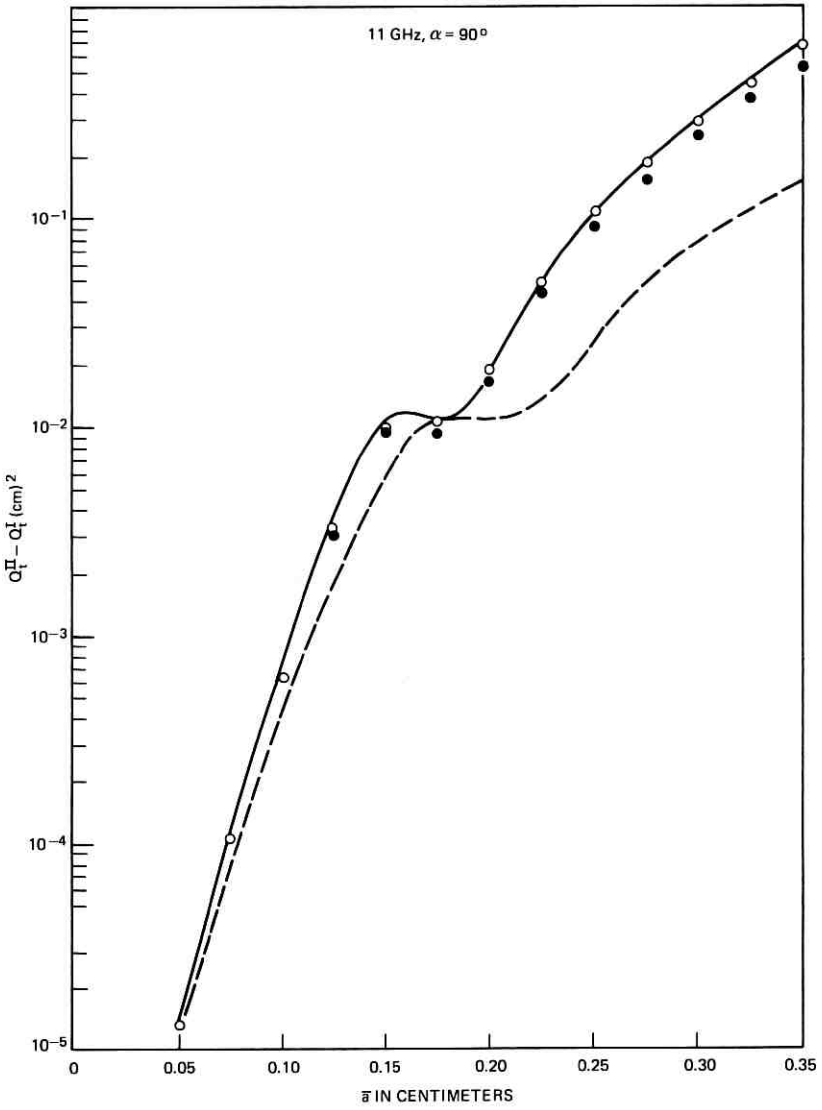

Fig. 16—Comparison of perturbation approximations to the least-squares-fitting value of $Q_t^{II} - Q_t^I$ at 11 GHz with $\alpha = 90°$ as a function of drop size.

VIII. PERTURBATION RESULTS

In this final section, we compare three sets of first-order perturbation results with those obtained by least-squares fitting. The comparisons are made graphically in Figs. 3 to 23, since this is much more revealing than tabulating the results. The solid curves correspond to least-squares fitting and the dashed curves to perturbation about the inscribed sphere of radius a, corresponding to the expansion in (63), with perturbation parameter ν given by (64). The circles and dots correspond to perturbation about the equivolumic sphere of radius $\bar{a}$, with perturbation parameters $\bar{\nu} = 2\bar{a}$ and ν, respectively, corresponding to the expansions in (65). The dots have been omitted in those cases in which they would lie very close to the corresponding circle or solid curve. Comparisons are made for $\alpha = 90°$ at 4, 11, 18.1, and 30 GHz and for $\alpha = 50°$ at 30 GHz.

The real parts of the forward scattering amplitudes $S_{\mathrm{I}}(0)$ and $S_{\mathrm{II}}(0)$ and the first-order approximations to these quantities are depicted in

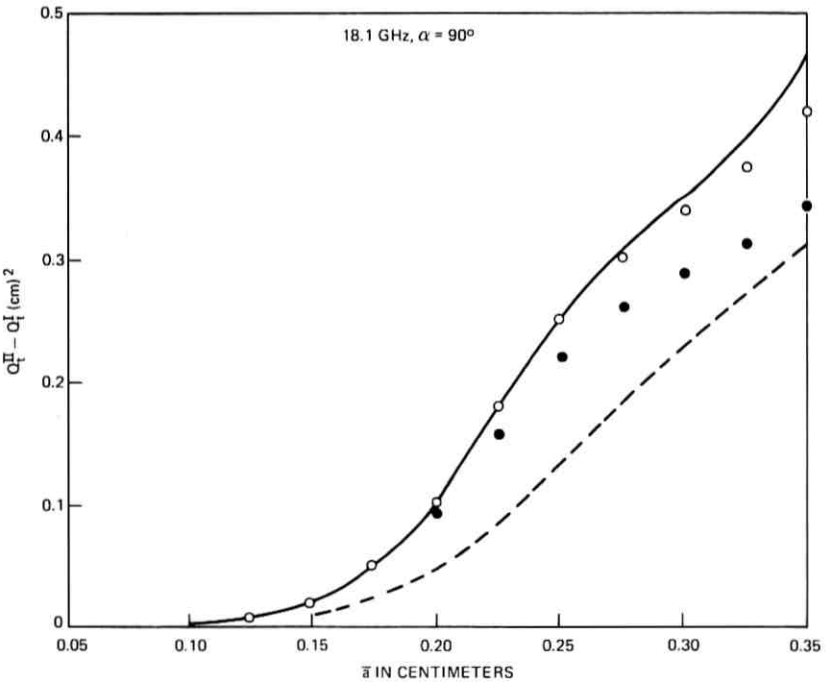

Fig. 17—Comparison of perturbation approximations to the least-squares-fitting value of $Q^{\mathrm{II}} - Q_t^{\mathrm{I}}$ at 18.1 GHz with $\alpha = 90°$ as a function of drop size.

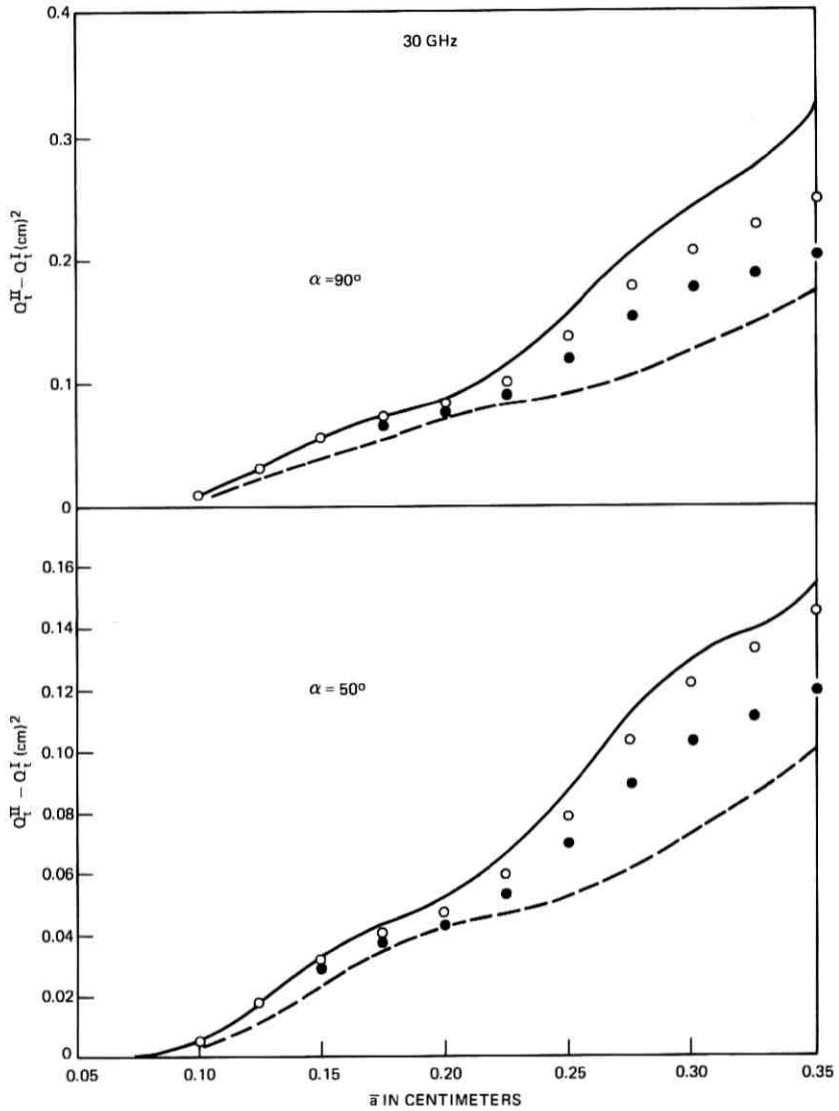

Fig. 18—Comparison of perturbation approximations to the least-squares-fitting value of $Q_t^{II} - Q_t^I$ at 30 GHz with $\alpha = 90°$ and $\alpha = 50°$ as a function of drop size.

Figs. 3 to 9, while the imaginary parts are depicted in Figs. 10 to 14. It should be noted that a logarithmic scale has been used in Figs. 3 and 4. Thus, at 4 GHz the first-order approximation to the real part of

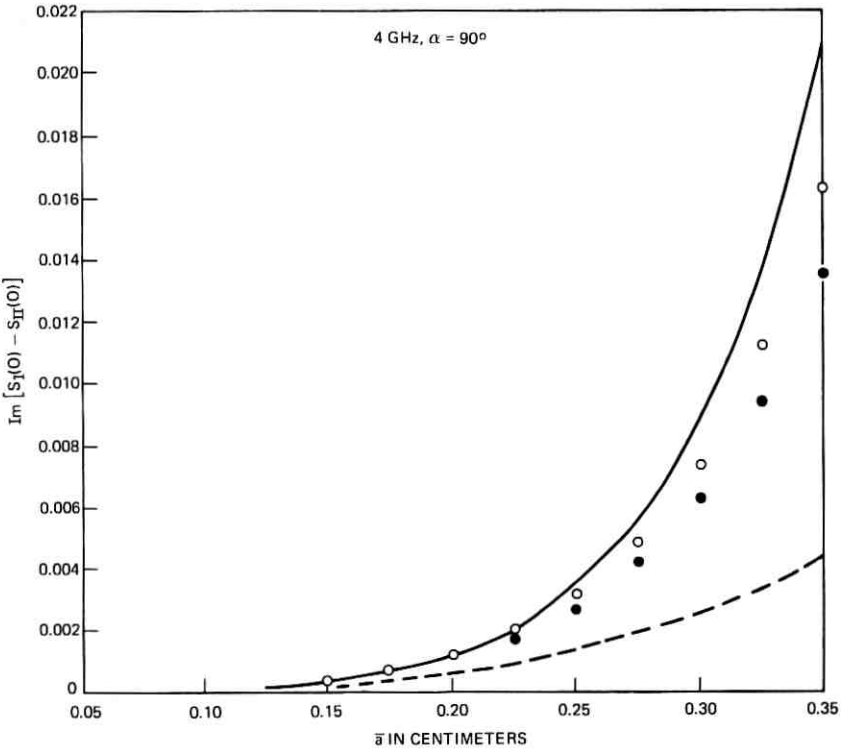

Fig. 19—Comparison of perturbation approximations to the least-squares-fitting value of $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ at 4 GHz with $\alpha = 90°$ as a function of drop size.

$S_{\mathrm{II}}(0)$, obtained by perturbing about the inscribed sphere, is in error by an order of magnitude for the largest drop size. It is seen that the best overall approximation is obtained by perturbing about the equivolumic sphere with perturbation parameter $\bar{\nu} = 2\bar{a}$, and in most cases there is a significant improvement over the approximation obtained by perturbing about the inscribed sphere as Oguchi[4] did. The second best overall approximation is obtained by perturbing about the equivolumic sphere with perturbation parameter $\nu = \bar{a}(2 - \bar{a})$, and is generally much better than the approximation obtained by perturbing about the inscribed sphere. The above ordering of the three sets of perturbation results is consistent with the order of the geometrical errors in the corresponding approximations in (63) and (65) to the oblate spheroid.

Although the comparison is not depicted for some of the smallest drop sizes, all three approximations are good for these, since the ec-

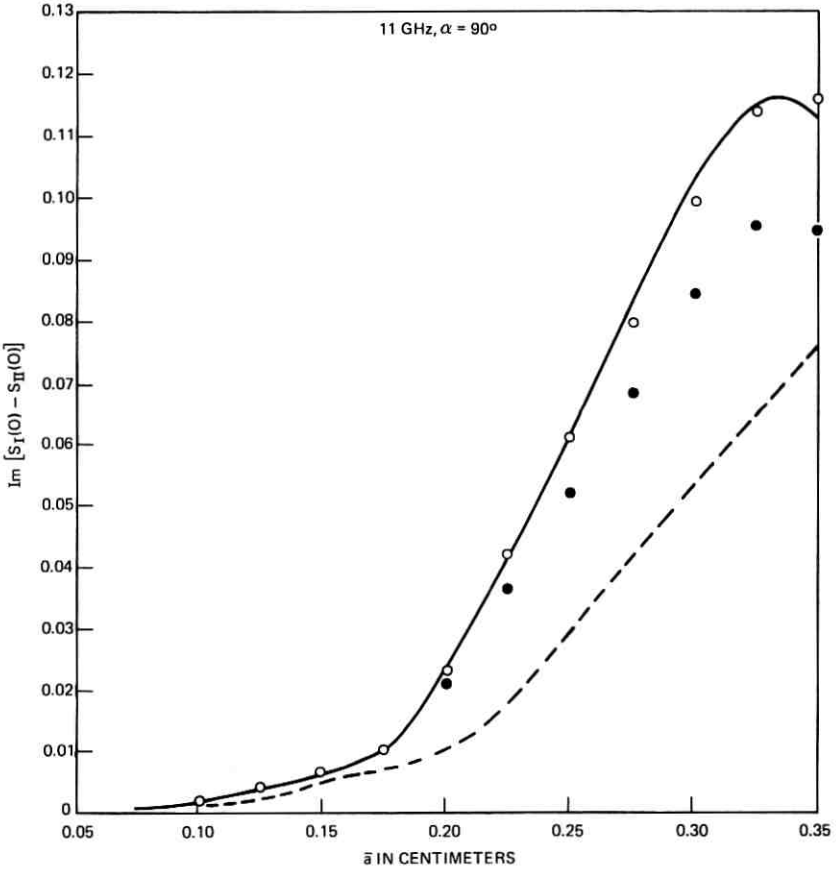

Fig. 20—Comparison of perturbation approximations to the least-squares-fitting value of $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ at 11 GHz with $\alpha = 90°$ as a function of drop size.

centricity is small. On the other hand, the approximations obtained by perturbing about the equivolumic sphere are remarkably good for the largest drop sizes, in view of the fact that neither the eccentricity nor the perturbation parameter is small. In particular, these approximations to the imaginary part of $S_{\mathrm{II}}(0)$, depicted in Figs. 10 to 14, are quite impressive. It is not too surprising that perturbing about the inscribed sphere leads to poor results for the larger drop sizes in the second polarization. The first-order approximations to the scattering cross sections Q_s^{I} and Q_s^{II} are very similar to those depicted in Figs. 3 to 9 for the real parts of $S_{\mathrm{I}}(0)$ and $S_{\mathrm{II}}(0)$, which are related to the total cross sections Q_t^{I} and Q_t^{II} by (39).

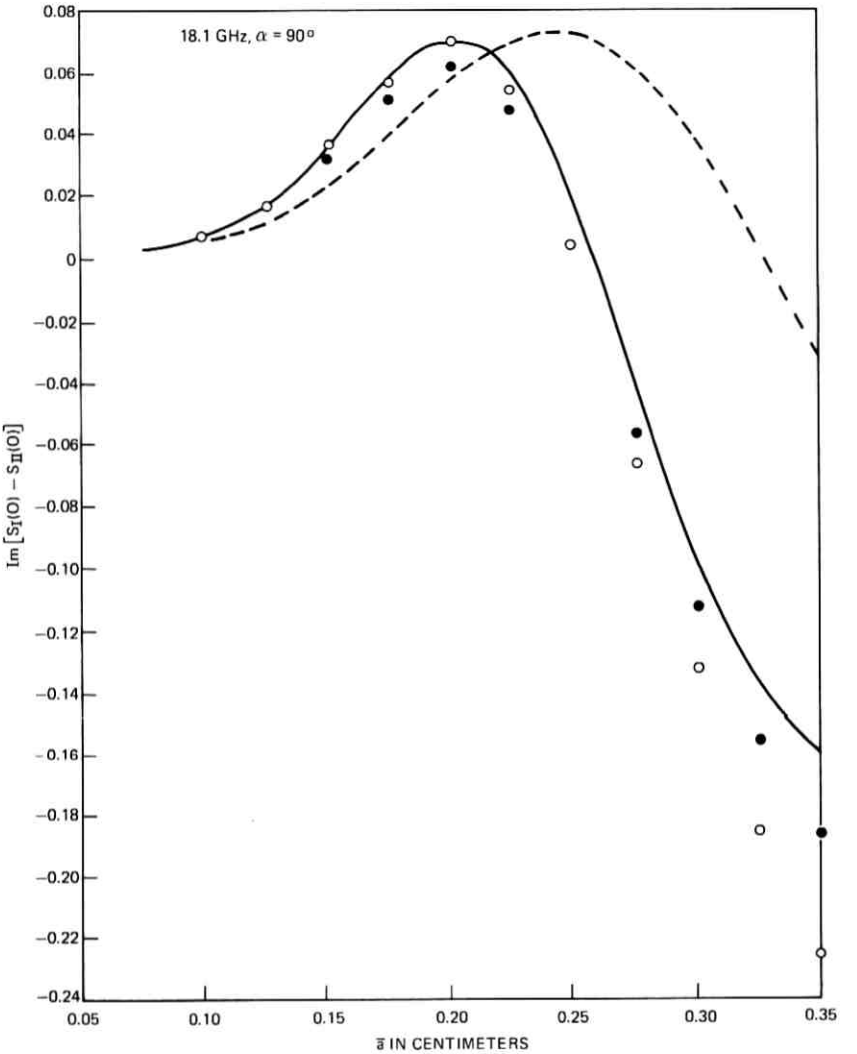

Fig. 21—Comparison of perturbation approximations to the least-squares-fitting value of $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ at 18.1 GHz with $\alpha = 90°$ as a function of drop size.

For purposes of comparison, the values of the forward scattering amplitude $S(0)$ and the total and scattering cross sections Q_t and Q_s for the equivolumic spherical drops are given in Tables XIV to XVII. These quantities do not depend on the polarization of the incident wave or on the angle of incidence α. As is seen from Tables VIII to XIII,

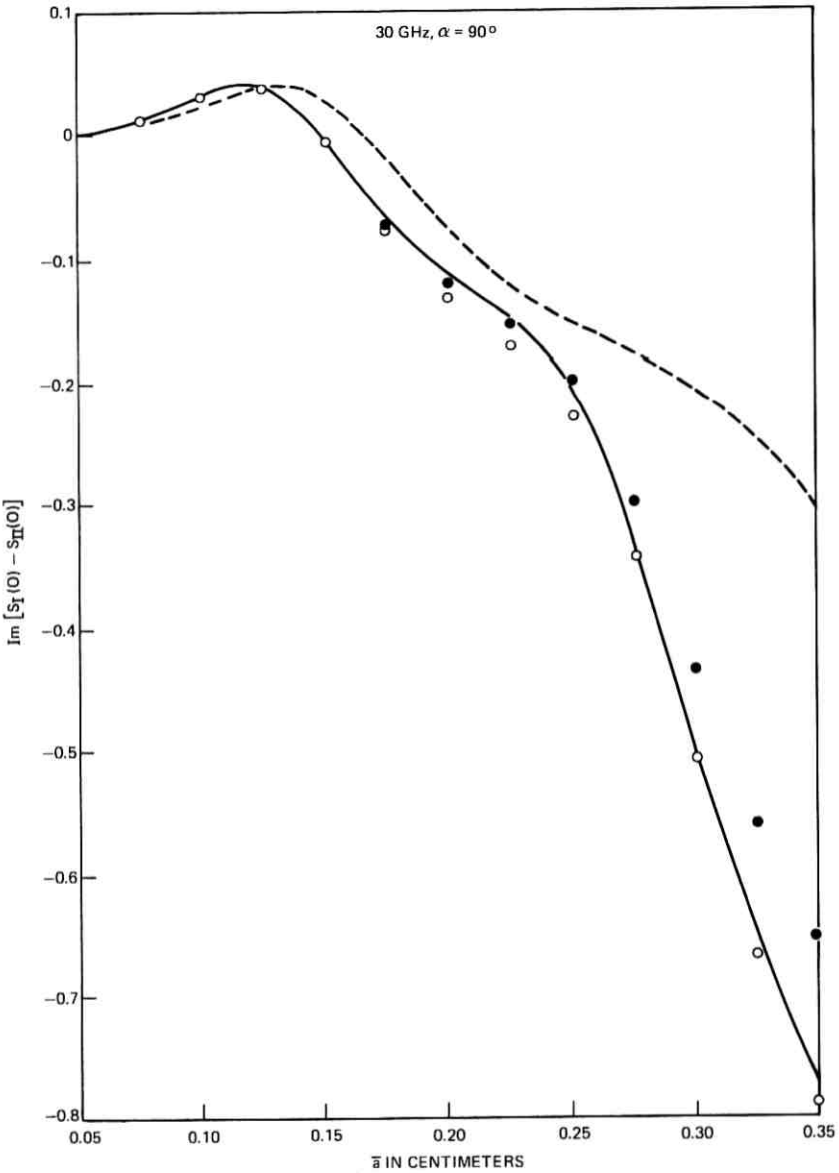

Fig. 22—Comparison of perturbation approximations to the least-squares-fitting value of $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ at 30 GHz with $\alpha = 90°$ as a function of drop size.

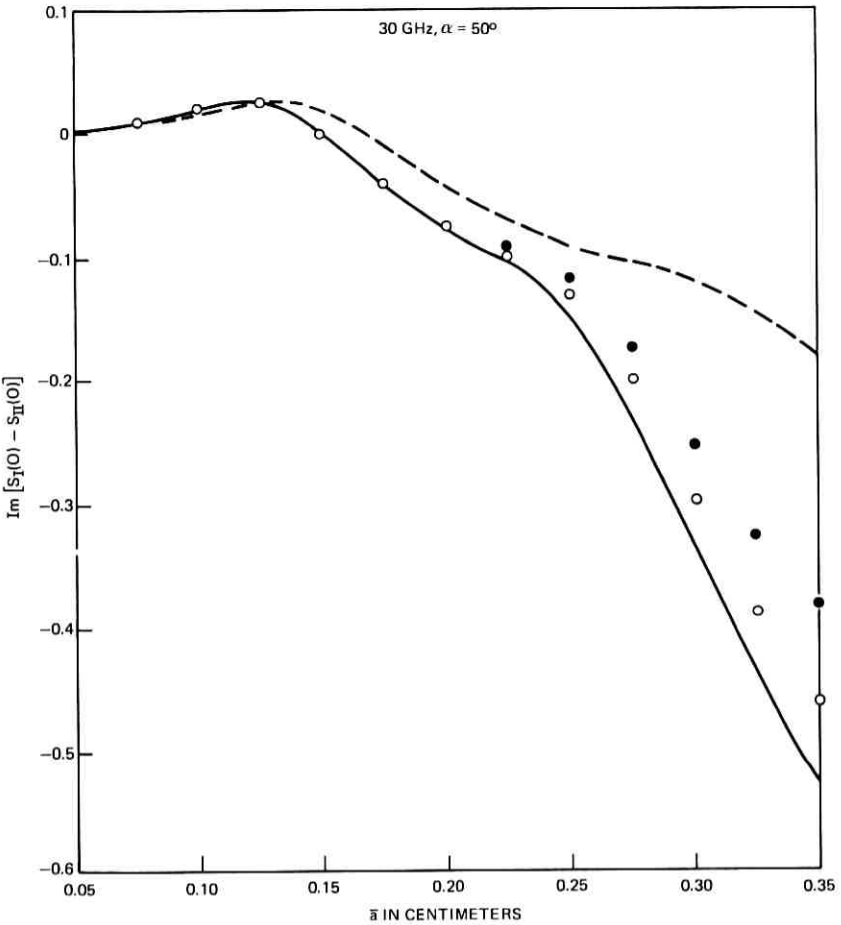

Fig. 23—Comparison of perturbation approximations to the least-squares-fitting value of $\mathrm{Im}[S_{\mathrm{I}}(0) - S_{\mathrm{II}}(0)]$ at 30 GHz with $\alpha = 50°$ as a function of drop size.

the value of Q_t lies between the corresponding values for the two polarizations for the oblate spheroidal drop of the same size and similarly for the value of Q_s. Although this happens to be true at 30 GHz for $\alpha = 90°$, 70°, and 50°, these relations should not be expected to hold for all values of α, since for $\alpha = 0°$ the cross sections are independent of the polarization because of the axial symmetry of the oblate spheroidal drop. We have verified that $Q_t \neq Q_t^{\mathrm{I}} = Q_t^{\mathrm{II}}$ and $Q_s \neq Q_s^{\mathrm{I}} = Q_s^{\mathrm{II}}$ for $\alpha = 0°$ at 11 GHz, for $\bar{a} = 0\,025$, 0.1, and 0.175.

The rain-induced differential attenuation and differential phase

Table XIV — Forward scattering amplitude and total and scattering cross sections for the equivolumic spherical drop at 4 GHz for different drop sizes

$\bar{a}$(cm)	$S(0)$	Q_t(cm)^2	Q_s(cm)^2
0.025	$7.1886 \times 10^{-8} - 8.8610 \times 10^{-6}i$	1.2871×10^{-6}	9.3508×10^{-10}
0.05	$6.3247 \times 10^{-7} - 7.1195 \times 10^{-5}i$	1.1324×10^{-5}	5.9937×10^{-8}
0.075	$2.4837 \times 10^{-6} - 2.4204 \times 10^{-4}i$	4.4470×10^{-5}	6.8450×10^{-7}
0.1	$7.1665 \times 10^{-6} - 5.7975 \times 10^{-4}i$	1.2832×10^{-4}	3.8606×10^{-6}
0.125	$1.7619 \times 10^{-5} - 1.1481 \times 10^{-3}i$	3.1547×10^{-4}	1.4804×10^{-5}
0.15	$3.9266 \times 10^{-5} - 2.0191 \times 10^{-3}i$	7.0305×10^{-4}	4.4509×10^{-5}
0.175	$8.1912 \times 10^{-5} - 3.2771 \times 10^{-3}i$	1.4666×10^{-3}	1.1325×10^{-4}
0.2	$1.6324 \times 10^{-4} - 5.0244 \times 10^{-3}i$	2.9228×10^{-3}	2.5537×10^{-4}
0.225	$3.1562 \times 10^{-4} - 7.3913 \times 10^{-3}i$	5.6511×10^{-3}	5.2608×10^{-4}
0.25	$6.0030 \times 10^{-4} - 1.0551 \times 10^{-2}i$	1.0748×10^{-2}	1.0123×10^{-3}
0.275	$1.1392 \times 10^{-3} - 1.4745 \times 10^{-2}i$	2.0397×10^{-2}	1.8527×10^{-3}
0.3	$2.1916 \times 10^{-3} - 2.0322 \times 10^{-2}i$	3.9240×10^{-2}	3.2838×10^{-3}
0.325	$4.3555 \times 10^{-3} - 2.7783 \times 10^{-2}i$	7.7985×10^{-2}	5.7649×10^{-3}
0.35	$9.1208 \times 10^{-3} - 3.7683 \times 10^{-2}i$	1.6331×10^{-1}	1.0372×10^{-2}

shift are obtained[1] by summing the real and imaginary parts of $S_{II}(0) - S_I(0)$ over the Laws and Parsons drop-size distribution.[7] In a recent short note,[9] the three first-order perturbation approximations have been compared to the least-squares fitting results for the differential attenuation and differential phase shift at several different rain rates. The same ordering of the overall closeness of the three approximations holds for these quantities. Since the perturbation results are obtained quite inexpensively whereas the least-squares-fitting procedure is very costly, approximations to the differential attenuation and differential phase shift at frequencies up to 100 GHz were obtained by per-

Table XV — Forward scattering amplitude and total and scattering cross sections for the equivolumic spherical drop at 11 GHz for different drop sizes

$\bar{a}$(cm)	$S(0)$	Q_t(cm)^2	Q_s(cm)^2
0.025	$4.9868 \times 10^{-6} - 1.8550 \times 10^{-4}i$	1.1804×10^{-5}	5.3599×10^{-8}
0.05	$6.2648 \times 10^{-5} - 1.5238 \times 10^{-3}i$	1.4829×10^{-4}	3.4733×10^{-6}
0.075	$3.8512 \times 10^{-4} - 5.3789 \times 10^{-3}i$	9.1163×10^{-4}	4.0552×10^{-5}
0.1	$1.7992 \times 10^{-3} - 1.3547 \times 10^{-2}i$	4.2588×10^{-3}	2.3883×10^{-4}
0.125	$7.1756 \times 10^{-3} - 2.7813 \times 10^{-2}i$	1.6985×10^{-2}	1.0004×10^{-3}
0.15	$2.2023 \times 10^{-2} - 4.4760 \times 10^{-2}i$	5.2132×10^{-2}	3.4809×10^{-3}
0.175	$3.9051 \times 10^{-2} - 5.5463 \times 10^{-2}i$	9.2438×10^{-2}	9.7067×10^{-3}
0.2	$5.0153 \times 10^{-2} - 7.3383 \times 10^{-2}i$	1.1872×10^{-1}	2.1692×10^{-2}
0.225	$6.8317 \times 10^{-2} - 1.0343 \times 10^{-1}i$	1.6171×10^{-1}	4.2294×10^{-2}
0.25	$9.6583 \times 10^{-2} - 1.3932 \times 10^{-1}i$	2.2862×10^{-1}	7.4229×10^{-2}
0.275	$1.3508 \times 10^{-1} - 1.8178 \times 10^{-1}i$	3.1976×10^{-1}	1.2397×10^{-1}
0.3	$1.8972 \times 10^{-1} - 2.2945 \times 10^{-1}i$	4.4910×10^{-1}	2.0165×10^{-1}
0.325	$2.6372 \times 10^{-1} - 2.7484 \times 10^{-1}i$	6.2427×10^{-1}	3.1470×10^{-1}
0.35	$3.5439 \times 10^{-1} - 3.1001 \times 10^{-1}i$	8.3887×10^{-1}	4.6135×10^{-1}

Table XVI — Forward scattering amplitude and total and scattering cross sections for the equivolumic spherical drop at 18.1 GHz for different drop sizes

$\bar{a}$(cm)	$S(0)$	Q_t(cm)^2	Q_s(cm)^2
0.025	$4.1444 \times 10^{-5} - 8.3170 \times 10^{-4}i$	3.6243×10^{-5}	3.9388×10^{-7}
0.05	$6.8431 \times 10^{-4} - 6.9938 \times 10^{-3}i$	5.9843×10^{-4}	2.6138×10^{-5}
0.075	$5.3548 \times 10^{-3} - 2.4905 \times 10^{-2}i$	4.6827×10^{-3}	3.2305×10^{-4}
0.1	$2.4004 \times 10^{-2} - 5.5611 \times 10^{-2}i$	2.0991×10^{-2}	2.0935×10^{-3}
0.125	$5.4019 \times 10^{-2} - 9.1780 \times 10^{-2}i$	4.7240×10^{-2}	8.8579×10^{-3}
0.15	$9.7706 \times 10^{-2} - 1.4719 \times 10^{-1}i$	8.5444×10^{-2}	2.6617×10^{-2}
0.175	$1.7343 \times 10^{-1} - 2.1684 \times 10^{-1}i$	1.5166×10^{-1}	6.2646×10^{-2}
0.2	$2.8868 \times 10^{-1} - 2.8757 \times 10^{-1}i$	2.5245×10^{-1}	1.2552×10^{-1}
0.225	$4.4398 \times 10^{-1} - 3.3858 \times 10^{-1}i$	3.8826×10^{-1}	2.1846×10^{-1}
0.25	$6.1859 \times 10^{-1} - 3.5216 \times 10^{-1}i$	5.4095×10^{-1}	3.2834×10^{-1}
0.275	$7.8546 \times 10^{-1} - 3.3543 \times 10^{-1}i$	6.8688×10^{-1}	4.3596×10^{-1}
0.3	$9.3560 \times 10^{-1} - 3.0993 \times 10^{-1}i$	8.1818×10^{-1}	5.3377×10^{-1}
0.325	$1.0751 - 2.9126 \times 10^{-1}i$	9.4019×10^{-1}	6.2372×10^{-1}
0.35	$1.2136 - 2.8673 \times 10^{-1}i$	1.0613	7.1027×10^{-1}

turbing about the equivolumic sphere. However, the results may be less reliable at the higher frequencies, particularly at the heavier rain rates.[18]

The difference $Q_t^{II} - Q_t^{I}$, which is related to the real part of $S_{II}(0) - S_I(0)$ by (39), is depicted in Figs. 15 to 18, and the imaginary part of $S_I(0) - S_{II}(0)$ is depicted in Figs. 19 to 23. We note that, although extra first-order correction terms arise in the expansions about the equivolumic sphere given in (65), they correspond to a constant change in the radius of the drop. Hence, the corresponding increments in the forward scattering amplitudes are the same for both polarizations, and

Table XVII — Forward scattering amplitude and total and scattering cross sections for the equivolumic spherical drop at 30 GHz for different drop sizes

$\bar{a}$(cm)	$S(0)$	Q_t(cm)^2	Q_s(cm)^2
0.025	$3.5549 \times 10^{-4} - 3.8212 \times 10^{-3}i$	1.1316×10^{-4}	2.9980×10^{-6}
0.05	$7.2950 \times 10^{-3} - 3.2466 \times 10^{-2}i$	2.3221×10^{-3}	2.1254×10^{-4}
0.075	$4.8577 \times 10^{-2} - 1.0204 \times 10^{-1}i$	1.5463×10^{-2}	2.8859×10^{-3}
0.1	$1.5024 \times 10^{-1} - 2.0855 \times 10^{-1}i$	4.7823×10^{-2}	1.7391×10^{-2}
0.125	$3.5560 \times 10^{-1} - 3.1825 \times 10^{-1}i$	1.1319×10^{-1}	5.6317×10^{-2}
0.15	$6.3545 \times 10^{-1} - 3.5742 \times 10^{-1}i$	2.0227×10^{-1}	1.1723×10^{-1}
0.175	$9.0831 \times 10^{-1} - 3.2571 \times 10^{-1}i$	2.8913×10^{-1}	1.8042×10^{-1}
0.2	$1.1444 - 2.9371 \times 10^{-1}i$	3.6426×10^{-1}	2.3403×10^{-1}
0.225	$1.3868 - 3.0778 \times 10^{-1}i$	4.4142×10^{-1}	2.8578×10^{-1}
0.25	$1.6859 - 3.5609 \times 10^{-1}i$	5.3662×10^{-1}	3.4813×10^{-1}
0.275	$2.0559 - 3.9425 \times 10^{-1}i$	6.5441×10^{-1}	4.2750×10^{-1}
0.3	$2.4657 - 3.9064 \times 10^{-1}i$	7.8486×10^{-1}	5.1929×10^{-1}
0.325	$2.8763 - 3.5872 \times 10^{-1}i$	9.1557×10^{-1}	6.1372×10^{-1}
0.35	$3.2831 - 3.3474 \times 10^{-1}i$	1.0450	7.0680×10^{-1}

therefore do not affect the difference $S_{II}(0) - S_I(0)$. Figures 15 to 23 show that the approximations to the differential quantities obtained by perturbing about the equivolumic sphere with perturbation parameter $\bar{\nu} = 2\bar{a}$ are overall remarkably close to the least-squares-fitting results and far better than the approximations obtained by perturbing about the inscribed sphere.

IX. ACKNOWLEDGMENTS

The authors are indebted to D. C. Hogg for bringing this problem to their attention, to T. S. Chu for some helpful discussions, to J. McKenna and N. L. Schryer for suggesting the matching and least-squares-fitting approaches and for several helpful discussions in relation to these, and to P. A. Businger, whose least-squares-fitting subroutine was incorporated into the main program. The authors are greatly indebted to Mary Ann Gatto, who took over the burdensome task of running the main program, set up the transfer of data onto a disc file, and made some other program modifications, and to Susan Hoffberg who wrote the programs for calculating the first-order perturbation approximations and carefully checked some special function routines against tables. The authors are grateful to T. Oguchi for his private communications and for making available at an early date a preprint of his detailed paper.[35] Finally, the authors are particularly indebted to T. S. Chu, D. C. Hogg, and J. McKenna for their continued encouragement throughout the lengthy course of this work.

APPENDIX A

We first derive the expansion of the incident plane wave in a Fourier series in the azimuthal angle φ, as given by (20). The unit vectors in Cartesian coordinates are given in terms of those in spherical coordinates by

$$\begin{aligned} \mathbf{i} &= \sin\theta\cos\varphi\mathbf{i}_1 + \cos\theta\cos\varphi\mathbf{i}_2 - \sin\varphi\mathbf{i}_3, \\ \mathbf{j} &= \sin\theta\sin\varphi\mathbf{i}_1 + \cos\theta\sin\varphi\mathbf{i}_2 + \cos\varphi\mathbf{i}_3, \end{aligned}$$

and

$$\mathbf{k} = \cos\theta\mathbf{i}_1 - \sin\theta\mathbf{i}_2. \tag{76}$$

Also, we have

$$x\sin\alpha + z\cos\alpha = r(\sin\alpha\sin\theta\cos\varphi + \cos\alpha\cos\theta). \tag{77}$$

But[36] for integer values of p,

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-ip\varphi}\exp(i\xi\cos\varphi)d\varphi = i^p J_p(\xi), \tag{78}$$

where J_p denotes the regular Bessel function (of the first kind) of order p.

It follows, from (5), (6), (20) to (22), and (76) to (78), using the recurrence relations for the Bessel functions,[37] that

$$\mathbf{f}_m(r,\theta) = i^m \exp(ik_0 r\cos\alpha\cos\theta)\Big[J_m(k_0 r\sin\alpha\sin\theta)\sin\alpha \times(\sin\theta\mathbf{i}_2 - \cos\theta\mathbf{i}_1) - iJ'_m(k_0 r\sin\alpha\sin\theta)\cos\alpha(\sin\theta\mathbf{i}_1 + \cos\theta\mathbf{i}_2) + \frac{mJ_m(k_0 r\sin\alpha\sin\theta)}{k_0 r\sin\alpha\sin\theta}\cos\alpha\mathbf{i}_3\Big] \quad (79)$$

and

$$\mathbf{g}_m(r,\theta) = -i^m \exp(ik_0 r\cos\alpha\cos\theta)\Big[\frac{mJ_m(k_0 r\sin\alpha\sin\theta)}{k_0 r\sin\alpha\sin\theta} \times(\sin\theta\mathbf{i}_1 + \cos\theta\mathbf{i}_2) + iJ'_m(k_0 r\sin\alpha\sin\theta)\mathbf{i}_3\Big], \quad (80)$$

where, as before, the prime denotes derivative with respect to the argument.

From (8) to (11), (13), (14), and (20), the boundary conditions (16) and (17), when multiplied by $e^{-im\varphi}$ and integrated with respect to φ from 0 to 2π, lead to the equations

$$e_{m3}(R,\theta) + \sum_{\substack{n\geq|m|\\ n\neq 0}} a_{mn}h_n^{(1)}(k_0R)\frac{dP_n^{|m|}(\cos\theta)}{d\theta} - \sum_{\substack{n\geq|m|\\ n\neq 0}} b_{mn}\Big[\frac{h_n^{(1)}(k_0R)}{k_0R} + h_n^{(1)'}(k_0R)\Big]\cdot\frac{im}{\sin\theta}P_n^{|m|}(\cos\theta) = \sum_{\substack{n\geq|m|\\ n\neq 0}} c_{mn}j_n(k_1R)\frac{dP_n^{|m|}(\cos\theta)}{d\theta} - \sum_{\substack{n\geq|m|\\ n\neq 0}} d_{mn}\Big[\frac{j_n(k_1R)}{k_1R} + j'_n(k_1R)\Big]\cdot\frac{im}{\sin\theta}P_n^{|m|}(\cos\theta) \quad (81)$$

and, using (4),

$$\frac{i\omega\mu_0}{k_0}h_{m3}(R,\theta) + \sum_{\substack{n\geq|m|\\ n\neq 0}} b_{mn}h_n^{(1)}(k_0R)\frac{dP_n^{|m|}(\cos\theta)}{d\theta} - \sum_{\substack{n\geq|m|\\ n\neq 0}} a_{mn}\Big[\frac{h_n^{(1)}(k_0R)}{k_0R} + h_n^{(1)'}(k_0R)\Big]\cdot\frac{im}{\sin\theta}P_n^{|m|}(\cos\theta) = N\sum_{\substack{n\geq|m|\\ n\neq 0}} d_{mn}j_n(k_1R)\frac{dP_n^{|m|}(\cos\theta)}{d\theta} - N\sum_{\substack{n\geq|m|\\ n\neq 0}} c_{mn}\Big[\frac{j_n(k_1R)}{k_1R} + j'_n(k_1R)\Big]\cdot\frac{im}{\sin\theta}P_n^{|m|}(\cos\theta). \quad (82)$$

We are using the notations $e_{mj} = \mathbf{e}_m\cdot\mathbf{i}_j$ and $h_{mj} = \mathbf{h}_m\cdot\mathbf{i}_j$.

Similarly, the boundary conditions (18) and (19) lead to the equations

$$
\begin{aligned}
& e_{m2}(R,\theta) + \frac{1}{R}\frac{dR}{d\theta}\, e_{m1}(R,\theta) - \sum_{\substack{n \geqq |m| \\ n \neq 0}} a_{mn} h_n^{(1)}(k_0 R) \cdot \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta) \\
& - \sum_{\substack{n \geqq |m| \\ n \neq 0}} b_{mn} \left\{ \left[\frac{h_n^{(1)}(k_0 R)}{k_0 R} + h_n^{(1)'}(k_0 R) \right] \frac{dP_n^{|m|}(\cos\theta)}{d\theta} \right. \\
& \left. + \frac{n(n+1)}{R} \frac{dR}{d\theta} \frac{h_n^{(1)}(k_0 R)}{k_0 R} P_n^{|m|}(\cos\theta) \right\} \\
& = - \sum_{\substack{n \geqq |m| \\ n \neq 0}} c_{mn} j_n(k_1 R) \cdot \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta) \\
& - \sum_{\substack{n \geqq |m| \\ n \neq 0}} d_{mn} \left\{ \left[\frac{j_n(k_1 R)}{k_1 R} + j_n'(k_1 R) \right] \frac{dP_n^{|m|}(\cos\theta)}{d\theta} \right. \\
& \left. + \frac{n(n+1)}{R} \frac{dR}{d\theta} \frac{j_n(k_1 R)}{k_1 R} P_n^{|m|}(\cos\theta) \right\} \qquad (83)
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{i\omega\mu_0}{k_0} \left[h_{m2}(R,\theta) + \frac{1}{R}\frac{dR}{d\theta}\, h_{m1}(R,\theta) \right] \\
& - \sum_{\substack{n \geqq |m| \\ n \neq 0}} b_{mn} h_n^{(1)}(k_0 R) \cdot \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta) \\
& - \sum_{\substack{n \geqq |m| \\ n \neq 0}} a_{mn} \left\{ \left[\frac{h_n^{(1)}(k_0 R)}{k_0 R} + h_n^{(1)'}(k_0 R) \right] \frac{dP_n^{|m|}(\cos\theta)}{d\theta} \right. \\
& \left. + \frac{n(n+1)}{R} \frac{dR}{d\theta} \frac{h_n^{(1)}(k_0 R)}{k_0 R} P_n^{|m|}(\cos\theta) \right\} \\
& = - N \sum_{\substack{n \geqq |m| \\ n \neq 0}} d_{mn} j_n(k_1 R) \cdot \frac{im}{\sin\theta} P_n^{|m|}(\cos\theta) \\
& - N \sum_{\substack{n \geqq |m| \\ n \neq 0}} c_{mn} \left\{ \left[\frac{j_n(k_1 R)}{k_1 R} + j_n'(k_1 R) \right] \frac{dP_n^{|m|}(\cos\theta)}{d\theta} \right. \\
& \left. + \frac{n(n+1)}{R} \frac{dR}{d\theta} \frac{j_n(k_1 R)}{k_1 R} P_n^{|m|}(\cos\theta) \right\}. \qquad (84)
\end{aligned}
$$

APPENDIX B

We consider here the case in which the raindrop is symmetrical about the plane $\theta = \pi/2$ so that

$$R(\pi - \theta) = R(\theta), \qquad 0 \leqq \theta \leqq \pi/2. \qquad (85)$$

Let $\hat{\alpha} = (\pi - \alpha)$. Then, from (21), (79), and (80), it follows that, cor-

responding to $\hat{\alpha}$,

$$\hat{e}^{\mathrm{I}}_{m3}(R, \theta) = -e^{\mathrm{I}}_{m3}(R, \pi - \theta), \tag{86}$$

$$\hat{h}^{\mathrm{I}}_{m3}(R, \theta) = h^{\mathrm{I}}_{m3}(R, \pi - \theta), \tag{87}$$

$$\left[\hat{e}^{\mathrm{I}}_{m2}(R, \theta) + \frac{1}{R}\frac{dR}{d\theta}\hat{e}_{m1}(R, \theta)\right] = \left[e^{\mathrm{I}}_{m2}(R, \pi - \theta) + \frac{1}{R}\frac{dR(\pi - \theta)}{d(\pi - \theta)}e^{\mathrm{I}}_{m1}(R, \pi - \theta)\right], \tag{88}$$

and

$$\left[\hat{h}^{\mathrm{I}}_{m2}(R, \theta) + \frac{1}{R}\frac{dR}{d\theta}\hat{h}^{\mathrm{I}}_{m1}(R, \theta)\right] = -\left[h^{\mathrm{I}}_{m2}(R, \pi - \theta) + \frac{1}{R}\frac{dR(\pi - \theta)}{d(\pi - \theta)}h^{\mathrm{I}}_{m1}(R, \pi - \theta)\right]. \tag{89}$$

But[14]

$$P^{|m|}_n(-\cos\theta) = (-1)^{n+|m|}P^{|m|}_n(\cos\theta). \tag{90}$$

It follows from (81) to (90) that

$$a^{\mathrm{I}}_{mn} + (-1)^{n+|m|+1}\hat{a}^{\mathrm{I}}_{mn} = 0, \qquad c^{\mathrm{I}}_{mn} + (-1)^{n+|m|+1}\hat{c}^{\mathrm{I}}_{mn} = 0 \tag{91}$$

and

$$b^{\mathrm{I}}_{mn} + (-1)^{n+|m|}\hat{b}^{\mathrm{I}}_{mn} = 0, \qquad d^{\mathrm{I}}_{mn} + (-1)^{n+|m|}\hat{d}^{\mathrm{I}}_{mn} = 0. \tag{92}$$

For $\alpha = \pi/2$, we have $\hat{\alpha} = \alpha$, and hence we obtain the relationships in (25).

Similarly, from (22), (79), and (80),

$$\hat{e}^{\mathrm{II}}_{m3}(R, \theta) = e^{\mathrm{II}}_{m3}(R, \pi - \theta), \tag{93}$$

$$\hat{h}^{\mathrm{II}}_{m3}(R, \theta) = -h^{\mathrm{II}}_{m3}(R, \pi - \theta), \tag{94}$$

$$\left[\hat{e}^{\mathrm{II}}_{m2}(R, \theta) + \frac{1}{R}\frac{dR}{d\theta}\hat{e}^{\mathrm{II}}_{m1}(R, \theta)\right] = -\left[e^{\mathrm{II}}_{m2}(R, \pi - \theta) + \frac{1}{R}\frac{dR(\pi - \theta)}{d(\pi - \theta)}e^{\mathrm{II}}_{m1}(R, \pi - \theta)\right], \tag{95}$$

and

$$\left[\hat{h}^{\mathrm{II}}_{m2}(R, \theta) + \frac{1}{R}\frac{dR}{d\theta}\hat{h}^{\mathrm{II}}_{m1}(R, \theta)\right] = \left[h^{\mathrm{II}}_{m2}(R, \pi - \theta) + \frac{1}{R}\frac{dR(\pi - \theta)}{d(\pi - \theta)}h^{\mathrm{II}}_{m1}(R, \pi - \theta)\right]. \tag{96}$$

It follows, from (81) to (85), (90), and (93) to (96), that

$$a_{mn}^{\mathrm{II}} + (-1)^{n+|m|}\hat{a}_{mn}^{\mathrm{II}} = 0, \qquad c_{mn}^{\mathrm{II}} + (-1)^{n+|m|}\hat{c}_{mn}^{\mathrm{II}} = 0 \quad (97)$$

and

$$b_{mn}^{\mathrm{II}} + (-1)^{n+|m|+1}\hat{b}_{mn}^{\mathrm{II}} = 0, \qquad d_{mn}^{\mathrm{II}} + (-1)^{n+|m|+1}\hat{d}_{mn}^{\mathrm{II}} = 0. \quad (98)$$

For $\alpha = \pi/2$ we obtain the relationships in (26).

For $\alpha \neq \pi/2$ we may consider the sum and the difference of the boundary conditions corresponding to α and to $\hat{\alpha} = (\pi - \alpha)$. Then the sums $(a_{mn} + \hat{a}_{mn})$, $(b_{mn} + \hat{b}_{mn})$, $(c_{mn} + \hat{c}_{mn})$, and $(d_{mn} + \hat{d}_{mn})$ and the differences $(a_{mn} - \hat{a}_{mn})$, $(b_{mn} - \hat{b}_{mn})$, $(c_{mn} - \hat{c}_{mn})$, and $(d_{mn} - \hat{d}_{mn})$ may be determined separately, and these sums and differences vanish for alternate values of n, depending on the polarization.

APPENDIX C

We first consider the calculation of the scattered energy W_s, which is defined by (35), by letting $r \to \infty$. From (29) it follows that

$$W_s = \lim_{r\to\infty} \left\{ \frac{k_0 r^2}{2\omega\mu_0} \int_0^{2\pi} \int_0^{\pi} (|E_2^s|^2 + |E_3^s|^2) \sin\theta \, d\theta \, d\varphi \right\}. \quad (99)$$

But from (28),

$$E_2^s \sim \frac{-e^{ik_0 r}}{k_0 r} \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geq |m| \\ n \neq 0}} (-i)^n \left[a_{mn} \cdot \frac{m}{\sin\theta} P_n^{|m|}(\cos\theta) + b_{mn} \frac{dP_n^{|m|}(\cos\theta)}{d\theta} \right] e^{im\varphi} \quad (100)$$

and

$$E_3^s \sim \frac{e^{ik_0 r}}{k_0 r} \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geq |m| \\ n \neq 0}} (-i)^{n+1} \left[a_{mn} \frac{dP_n^{|m|}(\cos\theta)}{d\theta} + b_{mn} \cdot \frac{m}{\sin\theta} P_n^{|m|}(\cos\theta) \right] e^{im\varphi}. \quad (101)$$

Substituting (100) and (101) into (99), the integration with respect to φ is straightforward. The integration with respect to θ readily follows with the help of the identities

$$m \int_0^{\pi} \left[P_l^{|m|}(\cos\theta) \frac{dP_n^{|m|}(\cos\theta)}{d\theta} + \frac{dP_l^{|m|}(\cos\theta)}{d\theta} P_n^{|m|}(\cos\theta) \right] d\theta = 0 \quad (102)$$

and[38]

$$\int_0^\pi \left[\frac{dP_l^{|m|}(\cos\theta)}{d\theta}\frac{dP_n^{|m|}(\cos\theta)}{d\theta} + \frac{m^2}{\sin^2\theta}P_l^{|m|}(\cos\theta)P_n^{|m|}(\cos\theta)\right]\sin\theta d\theta = \frac{2n(n+1)(n+|m|)!}{(2n+1)(n-|m|)!}\delta_{ln}, \quad (103)$$

where δ_{ln} denotes the Kronecker delta, i.e., $\delta_{ln} = 1$ for $l = n$, and 0 otherwise. Thus, the expression for W_s given in (36) is obtained.

We remark that there is no need to let $r \to \infty$ to obtain this expression for W_s. The same result follows from (35) by using the expressions (10) and (11) for the scattered field, wherein $\mathbf{M}_{mn}^{(3)}(k_0)$ and $\mathbf{N}_{mn}^{(3)}(k_0)$ are defined by (8) and (9), with $z_n(k_0r) = h_n^{(1)}(k_0r)$. The dependence on r is found to vanish, as is to be expected, in view of the Wronskian relationship[39]

$$j_n(k_0r)y_n'(k_0r) - y_n(k_0r)j_n'(k_0r) = \frac{1}{(k_0r)^2}. \quad (104)$$

We also remark that the expression in (36) holds quite generally, e.g., for scattering from nonaxisymmetric raindrops, since at this point we have made no use of the properties of the coefficients a_{mn} and b_{mn}.

We next consider the calculation of the total energy W_t, which is defined by (41). We begin by allowing for a general incident field, given by

$$\mathbf{E}^i = -\sum_{m=-\infty}^{\infty}\sum_{\substack{n\geq|m|\\ n\neq 0}}[A_{mn}\mathbf{M}_{mn}^{(1)}(k_0) + B_{mn}\mathbf{N}_{mn}^{(1)}(k_0)] \quad (105)$$

and

$$\mathbf{H}^i = \frac{ik_0}{\omega\mu_0}\sum_{m=-\infty}^{\infty}\sum_{\substack{n\geq|m|\\ n\neq 0}}[A_{mn}\mathbf{N}_{mn}^{(1)}(k_0) + B_{mn}\mathbf{M}_{mn}^{(1)}(k_0)], \quad (106)$$

where the superscript 1 indicates that $z_n(k_0r) = j_n(k_0r)$ in (8) and (9). The calculation of W_t is similar to that of W_s and it is found, after some reductions, that

$$W_t = \frac{-2\pi}{\omega\mu_0k_0}\operatorname{Re}\sum_{m=-\infty}^{\infty}\sum_{\substack{n\geq|m|\\ n\neq 0}}\frac{n(n+1)(n+|m|)!}{(2n+1)(n-|m|)!} \times (a_{mn}A_{mn}^* + b_{mn}B_{mn}^*). \quad (107)$$

We now consider the incident electric field given by (42), where $\mathbf{E}_{\mathrm{I}}^i$ and $\mathbf{E}_{\mathrm{II}}^i$ are given by (5) and (6), respectively. Then (43) holds and, from (10) and (11),

$$a_{mn} = a_{mn}^{\mathrm{I}} + a_{mn}^{\mathrm{II}}, \qquad b_{mn} = b_{mn}^{\mathrm{I}} + b_{mn}^{\mathrm{II}}. \quad (108)$$

From (5) and (6) and the expansions in (116) and (117), it follows from (105) to (107) that

$$W_t = \frac{2\pi}{\omega\mu_0 k_0} \operatorname{Re} \sum_{m=-\infty}^{\infty} \sum_{\substack{n \geqq |m| \\ n \neq 0}} (-i)^{n-1} \times \left\{ E_{\mathrm{I}}^* \left[a_{mn} \cdot \frac{m}{\sin \alpha} P_n^{|m|}(\cos \alpha) + b_{mn} \frac{dP_n^{|m|}(\cos \alpha)}{d\alpha} \right] + iE_{\mathrm{II}}^* \left[a_{mn} \frac{dP_n^{|m|}(\cos \alpha)}{d\alpha} + b_{mn} \cdot \frac{m}{\sin \alpha} P_n^{|m|}(\cos \alpha) \right] \right\}. \quad (109)$$

The relations (39) now follow from (33), (34), (40), and (108) by setting first $E_{\mathrm{II}} = 0$ and second $E_{\mathrm{I}} = 0$, in (109). We also note that, from (5), (6), (28), (30), (42), and (109),

$$W_t = \frac{2\pi}{\omega\mu_0 k_0} \operatorname{Re} \left[\lim_{r \to \infty} \left\{ \frac{-ik_0 r e^{-ik_0 r} (\mathbf{E}^i)^* \cdot \mathbf{E}^s \mid_{\theta=\alpha,\ \varphi=0}}{\exp[-ik_0(x \sin \alpha + z \cos \alpha)]} \right\} \right]. \quad (110)$$

We remark that both (109) and (110) hold, subject to (42), (43), and (108), for scattering from generally shaped raindrops.

Finally, we consider the particular case of an axisymmetric raindrop given by $r = R(\theta)$, so that (23) and (24) hold. Thus,

$$(a_{mn}^{\mathrm{I}} a_{mn}^{\mathrm{II}*} + b_{mn}^{\mathrm{I}} b_{mn}^{\mathrm{II}*}) = -(a_{-mn}^{\mathrm{I}} a_{-mn}^{\mathrm{II}*} + b_{-mn}^{\mathrm{I}} b_{-mn}^{\mathrm{II}*}). \quad (111)$$

Hence, from (36) and (108),

$$W_s = W_s^{\mathrm{I}} + W_s^{\mathrm{II}}. \quad (112)$$

Then, from (37), with

$$Q_s = \frac{2\omega\mu_0 W_s}{k_0(E_{\mathrm{I}} E_{\mathrm{I}}^* + E_{\mathrm{II}} E_{\mathrm{II}}^*)}, \quad (113)$$

we obtain (44). Also, from (23), (24), (108), and (109),

$$W_t = W_t^{\mathrm{I}} + W_t^{\mathrm{II}}. \quad (114)$$

Hence, from (40), with

$$Q_t = \frac{2\omega\mu_0 W_t}{k_0(E_{\mathrm{I}} E_{\mathrm{I}}^* + E_{\mathrm{II}} E_{\mathrm{II}}^*)}, \quad (115)$$

we obtain (45).

APPENDIX D

We first give the expansions for the incident wave in terms of spherical vector wave functions.[4,19] It is found that

$$(\cos\alpha\mathbf{i} - \sin\alpha\mathbf{k}) \exp[ik_0(x\sin\alpha + z\cos\alpha)]$$
$$= -\sum_{m=-\infty}^{\infty}\sum_{\substack{n=|m|\\ n\neq 0}}^{\infty} \frac{i^{n+1}(2n+1)(n-|m|)!}{n(n+1)(n+|m|)!}$$
$$\times\left[\frac{m}{\sin\alpha}P_n^{|m|}(\cos\alpha)\mathbf{M}_{mn}^{(1)}(k_0) + \frac{dP_n^{|m|}(\cos\alpha)}{d\alpha}\mathbf{N}_{mn}^{(1)}(k_0)\right] \quad (116)$$

and

$$\mathbf{j}\exp[ik_0(x\sin\alpha + z\cos\alpha)]$$
$$= -\sum_{m=-\infty}^{\infty}\sum_{\substack{n=|m|\\ n\neq 0}}^{\infty} \frac{i^{n}(2n+1)(n-|m|)!}{n(n+1)(n+|m|)!}$$
$$\times\left[\frac{dP_n^{|m|}(\cos\alpha)}{d\alpha}\mathbf{M}_{mn}^{(1)}(k_0) + \frac{m}{\sin\alpha}P_n^{|m|}(\cos\alpha)\mathbf{N}_{mn}^{(1)}(k_0)\right]. \quad (117)$$

Expressions for the quantities $\mathbf{e}_m(r,\theta)$ and $\mathbf{h}_m(r,\theta)$, defined in (20), then follow from (5), (6), (8), and (9). Thus, we may now consider the boundary conditions (81) to (84).

We first multiply (81) by $dP_l^{|m|}(\cos\theta)/d\theta\sin\theta$ and (83) by $im\,P_l^{|m|}(\cos\theta)$ and add, and then multiply (81) by $im\,P_l^{|m|}(\cos\theta)$ and (83) by $dP_l^{|m|}(\cos\theta)/d\theta\sin\theta$ and subtract, and integrate both these equations with respect to θ from 0 to π. In the zero-order approximation corresponding to $\nu = 0$ in (47), this leads, with the help of (102) and (103), to simultaneous linear equations for $a_{ml}^{(0)}$ and $c_{ml}^{(0)}$. Similarly, multiplying (82) by $dP_l^{|m|}(\cos\theta)/d\theta\sin\theta$ and (84) by $im\,P_l^{|m|}(\cos\theta)$ and adding, and multiplying (82) by $im\,P_l^{|m|}(\cos\theta)$ and (84) by $dP_l^{|m|}(\cos\theta)/d\theta\sin\theta$ and subtracting, and integrating both these equations with respect to θ from 0 to π, we obtain simultaneous linear equations for the zero-order coefficients $b_{ml}^{(0)}$ and $d_{ml}^{(0)}$. The solution of these two pairs of simultaneous equations leads to the relations (50) and (51), where the quantities α_{mn} and β_{mn} depend on the polarization, as given by (52) and (53). It remains to give the expressions for the quantities a_n, b_n, c_n, and d_n occurring in (50) and (51).

We define

$$F_n(\xi) = \left[\frac{h_n^{(1)}(\xi)}{\xi} + h_n^{(1)'}(\xi)\right], \qquad G_n(\xi) = \left[\frac{j_n(\xi)}{\xi} + j_n'(\xi)\right]. \quad (118)$$

Then, with $\rho = k_0 a$, it is found that

$$a_n = \frac{[j_n(\rho)NG_n(N\rho) - j_n(N\rho)G_n(\rho)]}{[h_n^{(1)}(\rho)NG_n(N\rho) - j_n(N\rho)F_n(\rho)]}, \quad (119)$$

$$b_n = \frac{[j_n(\rho)G_n(N\rho) - Nj_n(N\rho)G_n(\rho)]}{[h_n^{(1)}(\rho)G_n(N\rho) - Nj_n(N\rho)F_n(\rho)]}, \quad (120)$$

and

$$c_n = (i/\rho^2)[h_n^{(1)}(\rho)NG_n(N\rho) - j_n(N\rho)F_n(\rho)]^{-1}, \quad (121)$$

$$d_n = (i/\rho^2)[h_n^{(1)}(\rho)G_n(N\rho) - Nj_n(N\rho)F_n(\rho)]^{-1}. \quad (122)$$

In obtaining (121) and (122), we have made use of the Wronskian relationship[39]

$$j_n(\rho)h_n^{(1)\prime}(\rho) - h_n^{(1)}(\rho)j_n'(\rho) = (i/\rho^2). \quad (123)$$

Next, considering the first-order terms in ν in the integrated forms of the boundary conditions, and making use of (47) to (49), two pairs of simultaneous linear equations are obtained for $a_{mn}^{(1)}$ and $c_{mn}^{(1)}$ and for $b_{mn}^{(1)}$ and $d_{mn}^{(1)}$. These equations contain somewhat involved expressions, but after considerable reductions they lead to the expressions given in (54) to (56), subject to (57) to (62). In particular, use has been made of the differential equation satisfied by the spherical Bessel functions,[39]

$$\xi^2 z_l''(\xi) + 2\xi z_l'(\xi) + [\xi^2 - l(l+1)]z_l(\xi) = 0. \quad (124)$$

Moreover, from (118) to (124) it follows that

$$j_l(\rho) - a_l h_l^{(1)}(\rho) + c_l j_l(N\rho) = 0, \quad (125)$$

$$j_l'(\rho) - a_l h_l^{(1)\prime}(\rho) + Nc_l j_l'(N\rho) = 0, \quad (126)$$

$$G_l'(\rho) - a_l F_l'(\rho) + N^2 c_l G_l'(N\rho) = (1 - N^2)c_l j_l(N\rho), \quad (127)$$

$$Nj_l(\rho) - Nb_l h_l^{(1)}(\rho) + d_l j_l(N\rho) = (1 - N^2)d_l j_l(N\rho), \quad (128)$$

$$j_l'(\rho) - b_l h_l^{(1)\prime}(\rho) + N^2 d_l j_l'(N\rho) = (N^2 - 1)d_l G_l(N\rho), \quad (129)$$

and

$$N\rho^2[G_l'(\rho) - b_l F_l'(\rho) + Nd_l G_l'(N\rho)] = l(l+1)(1 - N^2)d_l j_l(N\rho). \quad (130)$$

We have also used the fact that

$$\int_0^\pi P_l^{|m|}(\cos\theta)\frac{dP_n^{|m|}(\cos\theta)}{d\theta}\sin\theta\frac{d\sigma_1}{d\theta}d\theta = \frac{n(n+1)}{l(l+1)}\mathfrak{Re}_{nl}^m - \mathfrak{g}_{nl}^m, \quad (131)$$

where $\mathfrak{Re}_{nl}^m$ and $\mathfrak{g}_{nl}^m$ are given by (60) and (61). This result follows directly by integration by parts and use of the differential equation satisfied by the associated Legendre functions,[14]

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dP_n^{|m|}(\cos\theta)}{d\theta}\right] + \left[n(n+1) - \frac{m^2}{\sin^2\theta}\right]P_n^{|m|}(\cos\theta) = 0. \quad (132)$$

APPENDIX E

We outline here the calculation of the integrals in (60) to (62) in the case

$$\sigma_1(\theta) = \tfrac{1}{2}\sin^2\theta. \tag{133}$$

It is assumed that $l \geqq |m|$, $n \geqq |m|$, $l \neq 0$ and $n \neq 0$. Integration by parts of the expression in (62) leads to

$$\begin{aligned} \mathcal{J}_{nl}^{m} &= -m\int_0^{\pi} P_l^{|m|}(\cos\theta)P_n^{|m|}(\cos\theta)\sin\theta\cos\theta d\theta \\ &= -m\int_{-1}^{1} xP_l^{|m|}(x)P_n^{|m|}(x)dx. \end{aligned} \tag{134}$$

But,[14] with $P_{|m|-1}^{|m|}(x) \equiv 0$,

$$(2l+1)xP_l^{|m|}(x) = (l-|m|+1)P_{l+1}^{|m|}(x) + (l+|m|)P_{l-1}^{|m|}(x). \tag{135}$$

Substituting (134) into (135), and using the relationship[14]

$$\int_{-1}^{1} P_l^{|m|}(x)P_n^{|m|}(x)dx = \frac{2(n+|m|)!}{(2n+1)(n-|m|)!}\,\delta_{ln}, \tag{136}$$

it follows from (62) that

$$J_{nl}^{m} = \frac{-m}{n(n+1)}\left[\frac{(n-|m|)}{(2n-1)}\,\delta_{l,n-1} + \frac{(n+|m|+1)}{(2n+3)}\,\delta_{l,n+1}\right]. \tag{137}$$

Next, from (60) and (133),

$$\mathcal{K}_{nl}^{m} = \frac{1}{2}l(l+1)\int_{-1}^{1}(1-x^2)P_l^{|m|}(x)P_n^{|m|}(x)dx. \tag{138}$$

The integral in (138) may be evaluated by using (136) and the recurrence relation (135), with l replaced by n also. Then, from (60), it is found that

$$\begin{aligned} H_{nl}^{m} = {} & \frac{n(n+1)(m^2+n^2+n-1)}{(2n-1)(2n+3)}\,\delta_{ln} \\ & - \frac{(n+2)(n+3)(n+|m|+1)(n+|m|+2)}{2(2n+3)(2n+5)}\,\delta_{l,n+2} \\ & - \frac{(n-2)(n-1)(n-|m|)(n-|m|-1)}{2(2n-3)(2n-1)}\,\delta_{l,n-2}. \end{aligned} \tag{139}$$

Finally, from (61) and (133),

$$g_{nl}^{m} = \frac{1}{2}\int_{-1}^{1}\left[(1-x^2)^2\frac{dP_l^{|m|}(x)}{dx}\frac{dP_n^{|m|}(x)}{dx} + m^2P_l^{|m|}(x)P_n^{|m|}(x)\right]dx. \tag{140}$$

The integral in (140) may be evaluated by using (136) and the relation[14]

$$(2k+1)(1-x^2)\frac{dP_k^{|m|}(x)}{dx} = (k+1)(k+|m|)P_{k-1}^{|m|}(x) - k(k-|m|+1)P_{k+1}^{|m|}(x). \quad (141)$$

Then from (61) it is found that

$$I_{nl}^m = \frac{1}{2}\left\{\frac{m^2}{n(n+1)} + \frac{n[(n+1)^2-m^2]}{(n+1)(2n+1)(2n+3)} + \frac{(n+1)(n^2-m^2)}{n(4n^2-1)}\right\}\delta_{ln} - \frac{(n+3)(n+|m|+1)(n+|m|+2)}{2(n+1)(2n+3)(2n+5)} \times \delta_{l,n+2} - \frac{(n-2)(n-|m|)(n-|m|-1)}{2n(2n-3)(2n-1)}\delta_{l,n-2}. \quad (142)$$

We note that the above results are consistent with the expressions given by Oguchi,[4] without derivation, for the integrals in (131), (61), and (62), subject to (133).

REFERENCES

1. J. A. Morrison, M.-J. Cross, and T. S. Chu, "Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 4 (April 1973), pp. 599–604.
2. H. R. Pruppacher and R. L. Pitter, "A Semi-Empirical Determination of the Shape of a Cloud and Rain Drops," J. Atmosph. Sci., *28*, No. 1 (January 1971), pp. 86–94.
3. H. C. Van de Hulst, *Light Scattering by Small Particles*, New York: Wiley, 1957, p. 34.
4. T. Oguchi, "Attenuation of Electromagnetic Wave Due to Rain with Distorted Raindrops," J. Radio Res. Labs. (Tokyo), *7*, No. 33 (September 1960), pp. 467–485.
5. T. Oguchi, "Attenuation and Phase Rotation of Radio Waves Due to Rain: Calculations at 19.3 and 34.8 GHz," Radio Sci., *8*, No. 1 (January 1973), pp. 31–38.
6. J. A. Stratton, *Electromagnetic Theory*, New York: McGraw-Hill, 1941, pp. 563–565.
7. D. E. Setzer, "Computed Transmission Through Rain at Microwave and Visible Frequencies," B.S.T.J., *49*, No. 8 (October 1970), pp. 1873–1892.
8. P. S. Ray, "Broadband Complex Refractive Indices of Ice and Water," Appl. Opt., *11*, No. 8 (August 1972), pp. 1836–1844.
9. J. A. Morrison and T. S. Chu, "Perturbation Calculations of Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 10 (December 1973), pp. 1907–1913.
10. Ref. 6, p. 137.
11. Ref. 6, p. 37.
12. Ref. 6, pp. 414–416.
13. Ref. 6, pp. 404–406.
14. W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Functions of Mathematical Physics*, New York: Chelsea, 1954, pp. 53–54.
15. Ref. 6, pp. 568–569.
16. Ref. 3, p. 13.
17. Ref. 3, p. 31.

18. T. Oguchi, "Attenuation of Electromagnetic Wave Due to Rain with Distorted Raindrops (Part II)," J. Radio Res. Labs. (Tokyo), *11*, No. 53 (January 1964), pp. 19–44.
19. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, New York: McGraw-Hill, 1953, p. 1866.
20. C. R. Mullin, R. Sandburg, and C. O. Velline, "A Numerical Technique for the Determination of Scattering Cross Sections of Infinite Cylinders of Arbitrary Geometrical Cross Section," IEEE Trans. Ant. and Prop., *AP-13*, No. 1 (January 1965), pp. 141–149.
21. P. Businger and G. H. Golub, "Linear Least Squares Solutions by Householder Transformations," Numer. Math., *7*, No. 4 (September 1965), pp. 269–276.
22. E. Sonnenblick, "A Program to Compute Bessel Functions J and Y of Complex Argument, Integer Order," unpublished work.
23. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Washington: National Bureau of Standards, 1964, pp. 390–407.
24. D. S. Drumheller, "A New Method for Generating Spherical Bessel Functions," unpublished work.
25. Ref. 23, p. 439.
26. Ref. 23, pp. 457–466.
27. U. S. Math Tables Project, *Tables of Spherical Bessel Functions*, Vols. I and II, New York: Columbia, 1947.
28. A. E. Kaplan, "Numerical Generation of Spherical Bessel Functions of Real and Complex Arguments," unpublished work.
29. Ref. 23, pp. 469–473.
30. L. Robin, *Fonctions Sphériques de Legendre et Fonctions Sphéroidales*, Tome 1, Paris: Gauthier-Villars, 1957, pp. 74–75, 82–83.
31. Ref. 6, p. 401.
32. S. L. Belousov, *Tables of Normalized Associated Legendre Functions*, New York: Macmillan, 1962.
33. U. S. Math Tables Project, *Tables of Associated Legendre Functions*, New York: Columbia, 1945.
34. T. Oguchi, private communication.
35. T. Oguchi, "Scattering Properties of Oblate Raindrops and Cross Polarization of Radio Waves Due to Rain: Calculations at 19.3 and 34.8 GHz," J. Radio Res. Labs. (Tokyo), *20*, No. 102 (1973), pp. 79–118.
36. Ref. 23, p. 360.
37. Ref. 23, p. 361.
38. Ref. 6, p. 417.
39. Ref. 23, p. 437.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

Low-Loss Single-Material Fibers Made From Pure Fused Silica

By P. KAISER and H. W. ASTLE

(Manuscript received January 22, 1974)

Low-loss single- and multimode optical fibers were fabricated solely from pure fused silica. Their spectral losses corresponded closely to those of unclad fibers drawn from the same material, provided the cores of the single-material fiber preform were redrawn under pure conditions. The lowest steady-state loss of about 3 dB/km at a wavelength of 1.1 μm was obtained with a fiber 130 meters long that had a Spectrosil WF core. Experimental numerical apertures agreed excellently with theoretical predictions.

I. INTRODUCTION

Recently, we introduced an optical fiber that utilizes only a single, low-loss material in a unique structural form.[1] In the single-material fiber, the light is guided in a core of arbitrary shape which is supported by spoke-like membranes within a protective tubing. The guided modes have exponentially decaying fields in the supporting slabs, so that for proper design no power is lost to the surrounding tube. For a slab of arbitrary thickness, single- and multimode operation can be obtained by choosing the proper size of the central core region. In practice, though, this thickness cannot be too large if the field amplitude is to be sufficiently small at the end of the slabs to result in a fiber of reasonable size.

Theoretical analysis of the single-material fiber has been carried out and is presented in Refs. 2 and 3. In this paper, we concentrate on the experimental evaluation of the transmission characteristic and bring only a summary of those theoretical results that help us to understand the experimental data.

Cross-sectional views of typical single- and multimode fibers are shown in Fig. 1. For a unitary aspect ratio, $h = w$, the rectangular

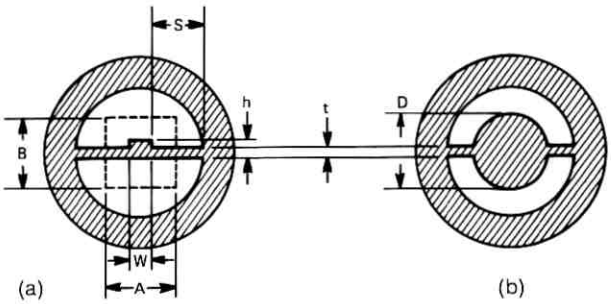

Fig. 1—Cross-sectional views of rectangular and cylindrical core, single-mode ($w \times h$) and multimode ($A \times B$; D) single-material fibers.

core of Fig. 1a has single-mode propagation if the condition

$$h \leqq 2t \tag{1}$$

is satisfied.[2] This result is based on the assumption that the energy is primarily concentrated in the dielectric, and that wavelength λ is small compared to the slab thickness t:

$$\lambda \ll t. \tag{2}$$

For a multimode guide, the wave propagation effects of the slab support can be represented by a uniform-index side support having the same height as the core and an equivalent index $n_e = n(1 - \Delta)$, where n is the refractive index of the core and Δ is computed from

$$\Delta = \frac{1}{8}\left(\frac{\lambda}{tn}\right)^2. \tag{3}$$

For a given n and λ, the relative refractive-index difference Δ depends only on slab thickness. From Δ, we obtain the numerical aperture (NA) and the modal dispersion, τ,

$$\text{NA} = n\sqrt{2\Delta} = \frac{\lambda}{2t} \tag{4}$$

and

$$\tau = \frac{L}{c} n\Delta, \tag{5}$$

where L is the length of the fiber, and c is the velocity of light in free space. The number of modes N for a core of diameter D (Fig. 1) is approximately given by

$$N \cong \frac{\pi^2}{8}\left(\frac{D}{t}\right)^2. \tag{6}$$

For all guided modes, the field decays exponentially along the slab. The field penetration is the largest for the highest-order mode and decays by $1/e$ in a length l, where

$$l \cong \frac{t}{\pi} \tag{7}$$

provided that h and w are large compared to t.

II. FABRICATION OF SINGLE-MATERIAL FIBERS

The preforms from which single-material fibers were drawn typically consisted of a core rod, a thin, polished plate, and a surrounding thick-walled cladding tube* (Fig. 2). Whereas high-grade synthetic silicas were used for core and plate, commercial-grade fused quartz was satisfactory for the cladding tube. The core rods were drawn from about 7-mm-diameter drawn or polished rods just before assembling the single-material fiber preform. To avoid contaminating these rods, a CO_2 laser or an oxy-hydrogen torch of high purity had to be employed for the redraw operation. Copper, with an associated broad absorption band centered between 0.8 and 0.9 μm, was considered the predominant contaminant in less pure systems.

The core rods were either centered in the cladding tube by means of a capillary tube, as indicated in Fig. 2, or they were attached to the top end of the plates with high-temperature cement. Support plates up to 15 cm long were cut from about 0.1-mm-thick polished plates to fit into the approximately 6.5-mm interior diameter of the cladding tubes (10 mm o.d.). The plates and the inner surface of the cladding tubes were cleaned with acetone and hydrofluoric acid and subsequently rinsed with deionized water. The assembled preform was lowered through a moderately hot oxy-hydrogen torch while helium was blown through the tube to carry away residual contaminants evaporating from the surfaces of the preform elements. The same gas was also used as protective atmosphere during the drawing operation. The fibers were drawn with an oxy-hydrogen ring-burner with an approximate draw-down ratio of 100 to 1. The preform geometry was essentially maintained in the drawn fiber, provided we used a proper drawing temperature. Temperatures that were too high caused the tube to collapse or resulted in slabs that were too short to enable low-loss guidance. Temperatures too low, on the other hand, resulted in

* For simplicity, the outer supporting cylinder is referred to as the cladding tube in the remainder of the text; its different meaning from the cladding of the conventional fiber should be kept in mind.

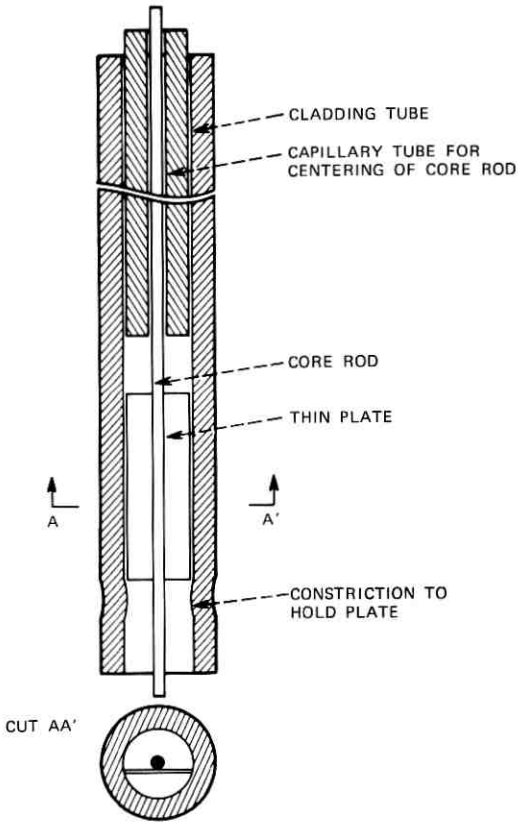

Fig. 2—Preform for single-material fibers.

brittle fibers. Since the core rod is thermally insulated from the heated cladding tube, the maximum diameter of the core rods that could be drawn without difficulty was about 1.5 mm. To avoid collapse of the cladding tube because of excessive heating, a wall thickness of about 1.7 mm was selected. This heavy wall thickness also helped to prevent the tube from being excessively deformed by the surface tension of the slab during the drawing process. Nevertheless, single-mode fibers typically have an elliptical cross section because of this force.

III. SINGLE-MODE, SINGLE-MATERIAL FIBERS

The single-mode fiber shown in Fig. 3 was drawn from an unclad fiber approximately 0.2 mm in diameter and a support plate 0.18 mm thick. The fiber had a slab thickness of 4 μm, a total core height h of

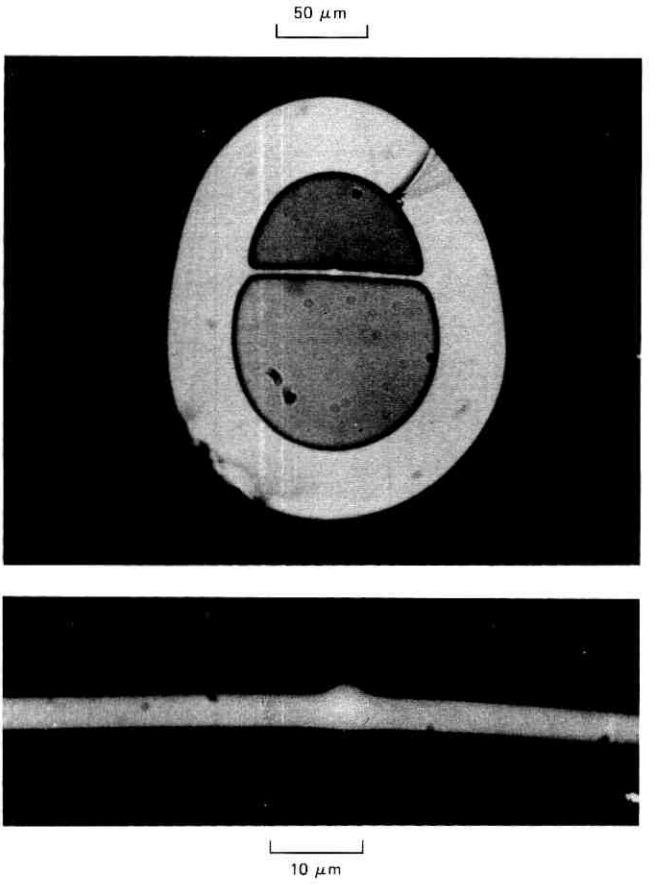

Fig. 3—Cross-sectional picture of single-mode, single-material fiber.

6 μm, and an approximate width w of 5 μm. According to eq. (1), single-mode operation was to be expected. While excited with a HeNe laser at 0.6328 μm, the intensity distribution of the guided wave was measured by projecting the field distribution of the fiber end with a 40X microscope objective lens to a distant target with a pinhole recording system (Fig. 4). For larger distances from the core region, the intensity in the direction of the slab $(y - y')$ decreases exponentially with a $1/e$ length of 2.3 μm. As shown in Fig. 4, the intensity distribution in the direction perpendicular to the slab could be well approximated by a $\cos^2\left(\frac{\pi}{2}\frac{x}{3}\right)$ distribution (x in μm). The far-field patterns

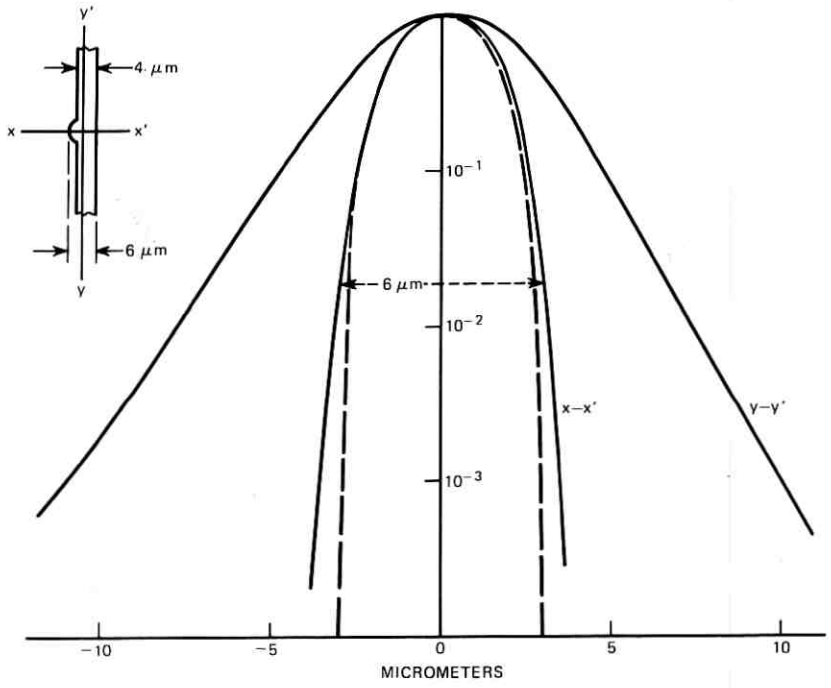

Fig. 4—Near-field intensity of the single-mode, single-material fiber of Fig. 3.

corresponding to this near-field distribution are shown in Fig. 5. These patterns were taken at a distance of 3.4 cm from the fiber end face with a 0.25-mm-diameter pinhole detector. Whereas the broad, exponentially decaying slab field results in a narrow, far-field pattern, the narrow $\cos^2$ distribution of the perpendicular plane results in a correspondingly broader distribution that has a zero at an angle θ_0:

$$\sin \theta_0 = \frac{3}{2} \frac{\lambda}{h_{\text{eff}}} \tag{8}$$

with h_{eff} being the effective height of the near-field distribution. With an experimental θ_0 of 11.7 degrees, h_{eff} is computed to be 4.7 μm, which emphasizes that a substantial part of the energy is propagating in the 4-μm-thick slab.

In addition to this usefully guided field in the core region, we observe other types of modes via their radiation patterns. These modes are more noticeable when the fiber is shorter (less than about 2 m) and straighter. Slab modes typically had a single intensity maximum

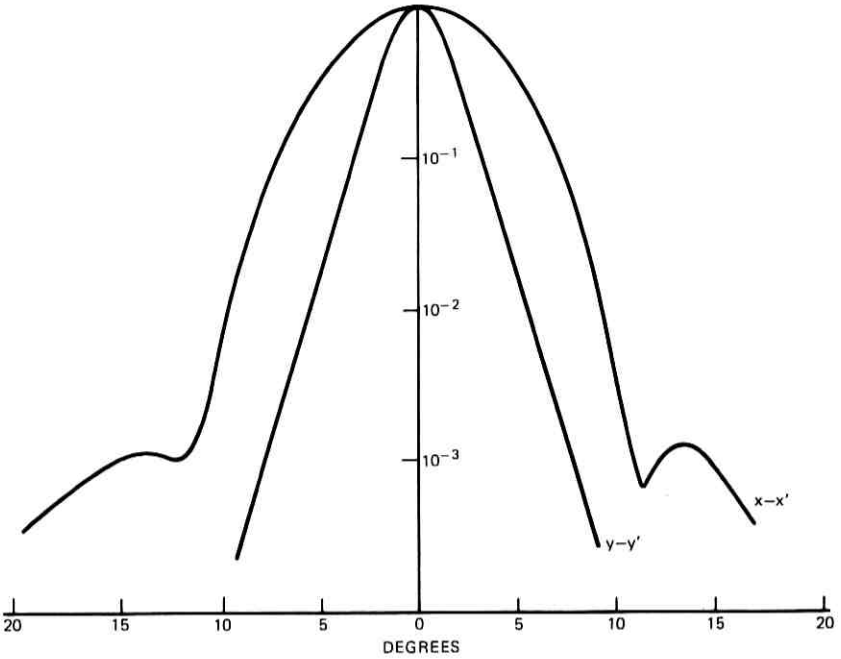

Fig. 5—Far-field intensity of the single-mode, single-material fiber of Figs. 3 and 4.

perpendicular to the slab, and several maxima and minima in the direction of the slab. For guide lengths in the order of a few centimeters, the slab modes are seen to extend to the cladding tube. The first higher mode in the direction perpendicular to the slab could be observed for slab thicknesses in the order of 5 μm. Besides the rapidly decaying slab modes, we also observed hollow-dielectric waveguide[4] and cladding modes. Hollow waveguide modes propagating in the voids of the fiber were only excited with low efficiency, and they rapidly decayed within a distance of a few decimeters. For accurate loss measurement, it is important to leave a sufficient length of fiber at the launching end (in excess of 1 m, for example) so that the slab and hollow waveguide modes are sufficiently attenuated. Cladding modes can easily be extracted with a matching liquid.

The single-material fiber losses were measured by breaking off known lengths of fiber and measuring the difference between the power levels. The fibers were broken by scoring them with a diamond while under tension. The fiber end could not be index matched with a liquid to the

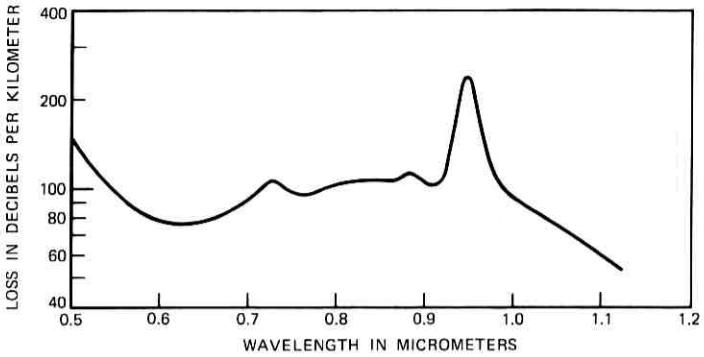

Fig. 6—Spectral losses of a single-mode, single-material fiber made from a Spectrosil WF unclad fiber on a Suprasil 2 plate.

detector surface due to the capillary action exerted by the hollow parts of the fiber. Using a measuring apparatus described elsewhere,[5] the spectral transmission losses of the single-material fibers were measured between 0.5 and 1.15 μm. Minimum losses achieved with a single-mode, single-material fiber amounted to about 50 dB/km at 1.06 μm (Fig. 6). Despite containing a Spectrosil WF core on a Suprasil 2 plate, the losses of this fiber were comparatively high since the core rod was redrawn with a low-purity oxy-hydrogen flame. Furthermore, since for

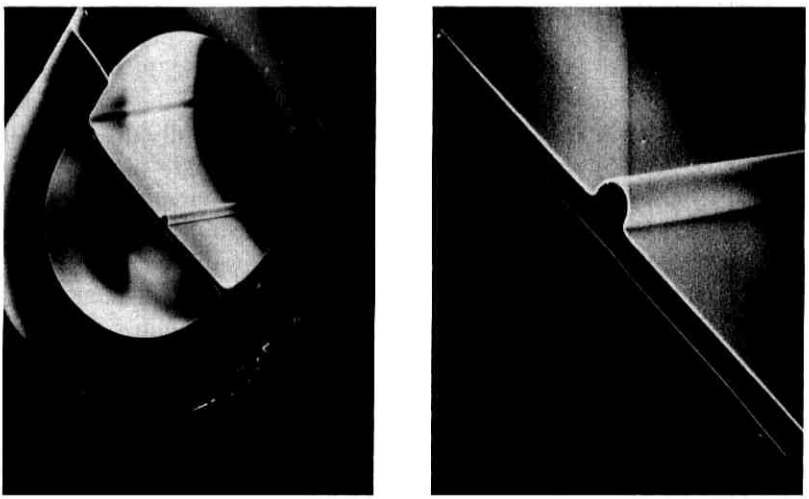

Fig. 7—Electron-microscope pictures of the single-mode, single-material fiber.

single-mode fibers a substantial portion of the energy propagates in the slab, contamination introduced during the polishing process may have contributed to the high losses. Electron-microscope pictures of a single-mode fiber are shown in Fig. 7.

IV. MULTIMODE, SINGLE-MATERIAL FIBERS

Cross-sectional views of multimode, single-material fibers are shown in Figs. 8 and 9. The diameter of the cores typically varied between 20 and 30 μm, depending on the size of the core rod and the draw-down ratio employed. A representative slab thickness was 2 to 3 μm, but fibers with slabs less than 1 μm thick and up to 5 μm thick have also been drawn (Fig. 8). To avoid losses via the exponential tail in the slab, the required minimum length-to-thickness ratio needed to be larger than about a factor of 7. Electron-microscope pictures of the interior structure of a multimode, single-material fiber (Fig. 10) demonstrated the intimate fusion of the preform parts.

The NAs of the multimode, single-material fibers were typically determined at 0.6328 μm from the diameters of the radiation patterns obtained from fibers that were a few meters long. In Fig. 11, the NAs of numerous single-material fibers that vary between 0.07 and 0.32 are compared with theoretical predictions [eq. (4)] and excellent agreement is realized. Diffraction effects made a determination of the NA difficult only for thicker slabs and resulting NAs below about 0.07. It

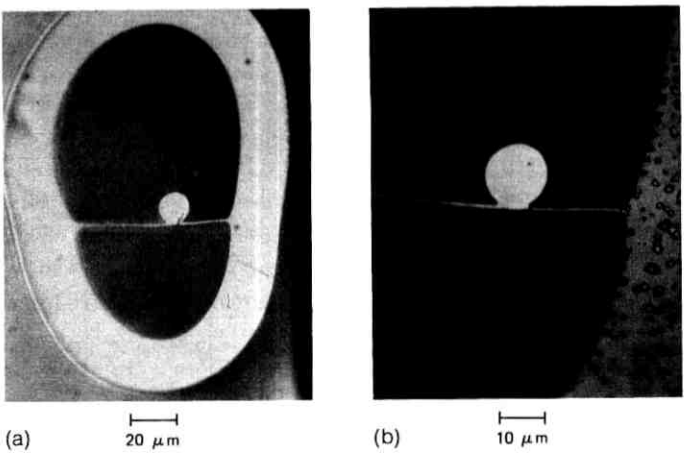

Fig. 8—(a) Cross section of a large-numerical-aperture, multimode fiber with 1-μm-thick supporting slab. ($NA_{0.6328\ \mu m}$ = 0.32.) (b) Magnified core region.

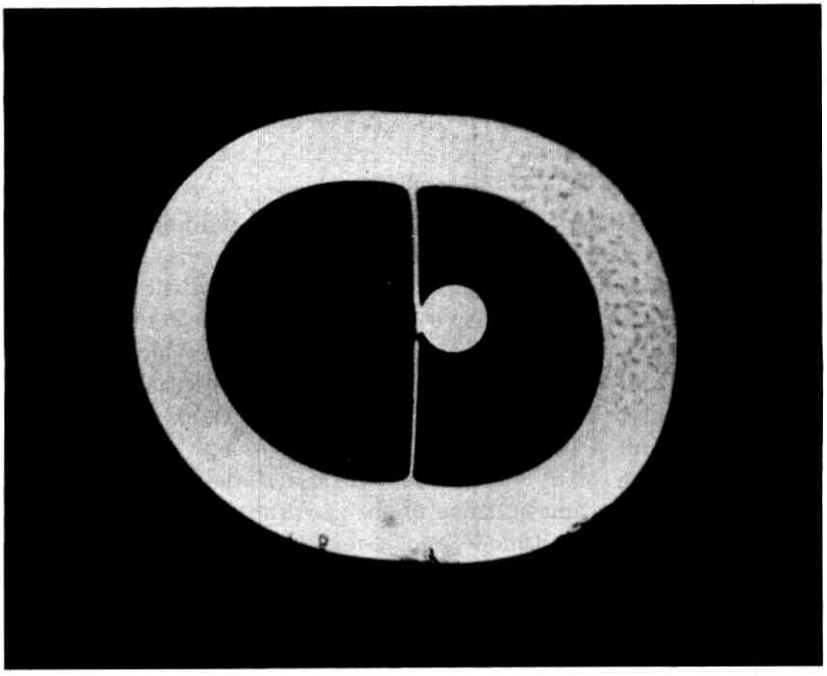

Fig. 9—Cross section of multimode fiber SMF 56 with a Spectrosil WF core on a Suprasil W1 slab; core diameter $\cong 25\ \mu m$, slab thickness $\cong 2.2\ \mu m$, $NA_{0.6328\ \mu m} \cong 0.14$.

is noteworthy that the good agreement between theory and experiment enabled us to determine the slab thickness through NA measurements derived from the transmitted radiation pattern. The linear increase of the NA with wavelength also agrees with theory, as illustrated in Fig. 12. Here, a fiber whose NA corresponding to a 2-μm-thick slab was 0.16 at 0.6328 μm, was illuminated by 10-nm bands filtered from a xenon arc lamp,[5] and the NA was calculated from the 1/100-power-points of the transmitted radiation pattern. Lack of sensitivity limited the data acquisition to the intermediate wavelength region shown, but HeNe-laser measurements at 1.15 μm confirmed a linear dependence in the whole wavelength range investigated.

The transmission losses of single-material fibers were expected to be identical or lower than those of unclad fibers drawn from the same material. Provided that the core rod was drawn in a pure oxy-hydrogen flame, as noted earlier, close agreement was indeed realized (see Fig. 13). Total losses of 10.6 and 10 dB/km were obtained at 0.8 and

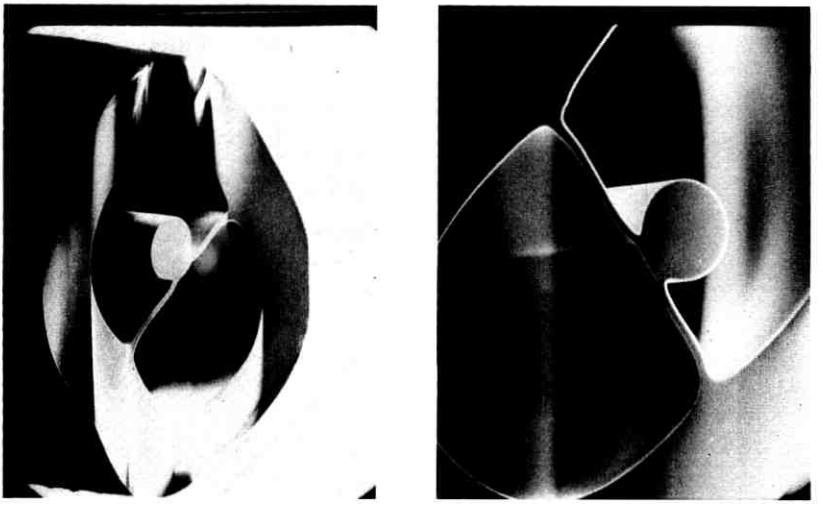

Fig. 10—Electron microscope pictures of a multimode, single-material fiber.

1.06 μm, respectively, with a fiber 210 m long that had a Suprasil 2 core (Fig. 13). The unclad-fiber losses at these wavelengths amounted to 7 and 10.5 dB/km. The various absorption peaks visible in the loss spectrum of Suprasil 2 are, as in Suprasil 1, due to its high OH content of 1200 ppm.[6] In contrast, only a weak OH band of 3 dB/km appeared in the loss spectrum of a multimode, single-material fiber (SMF 56) with a Spectrosil WF core on a Suprasil W1 slab (Figs. 9 and 14). The approximate steady-state losses of this 130-m-long fiber remained below the 7.5 dB/km level from 0.75 μm to the end of the spectral range investigated, and amounted to 6, 4.5, and 3 dB/km at 0.8, 0.9, and 1.1 μm, respectively.

Approximate steady-state losses were obtained by launching beams with different NAs into the fiber and by measuring the far- and near-end radiation patterns as a function of the launch NA, in addition to the wavelength-dependent transmission losses.[7] The losses associated with that NA for which the radiation pattern changed least from one end to the other can be considered as the steady-state losses. The radiation patterns were measured at a wavelength of 0.88 μm, which coincides with a resonance peak of the xenon arc lamp. At this wavelength, the 1/100-power-point-equivalent NA was 0.2, which agrees with the theoretical NA and corresponds to that of a 2.2-μm-thick slab. Whereas the NA of single-material fibers increased with wavelength, the NA

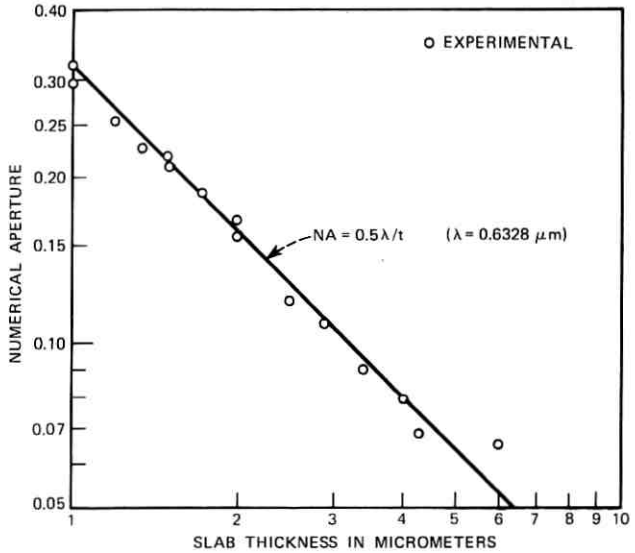

Fig. 11—Numerical aperture as function of slab thickness for multimode, single-material fibers.

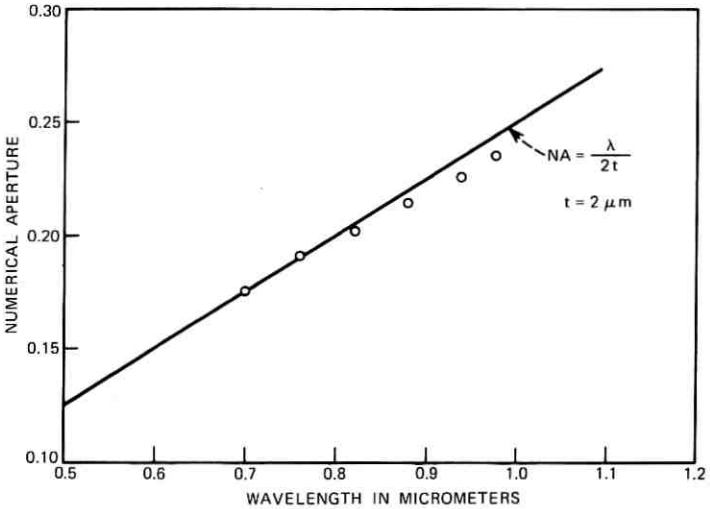

Fig. 12—Numerical aperture as function of wavelength for a multimode, single-material fiber.

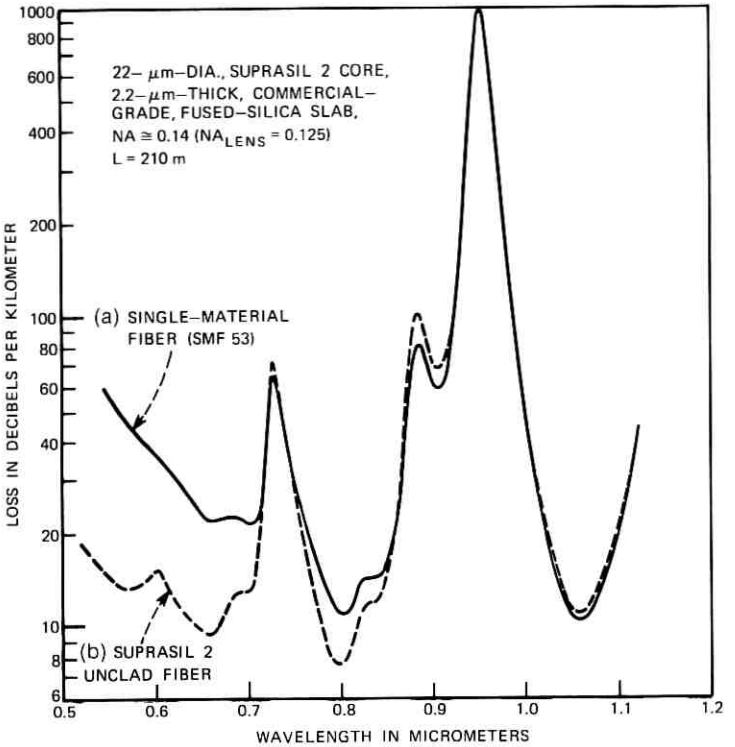

Fig. 13—Spectral losses of a Suprasil 2-cored, single-material fiber compared with unclad fiber losses. (a) Single-material fiber with 22-μm-diameter core and 2.2-μm-slab thickness, L = 210 m. (b) Suprasil 2 unclad fiber with approximately 0.2 mm dia; L ≅ 60 m.

of the launching beam was kept constant during a wavelength scan. Hence, the data obtained are increasingly too high at shorter wavelengths and too low at longer wavelengths relative to the steady-state losses at those wavelengths. We can get an estimate of the possible error by injecting a beam with a small NA into the fiber. The resulting losses are shown as curve b in Fig. 14. Even lower losses were achieved when we cooled the aluminum drum on which the fiber was wound with dry ice to reduce stress-induced losses (Fig. 14, curve c).[8] With the new minimum losses at 0.8, 0.9, and 1.1 μm amounting to 4, 3.4, and 2.6 dB/km, respectively, this curve represents the lowest loss spectrum obtained for pure fused silica, and corresponds closely to the spectra of the lowest-loss Corning fibers.[9]

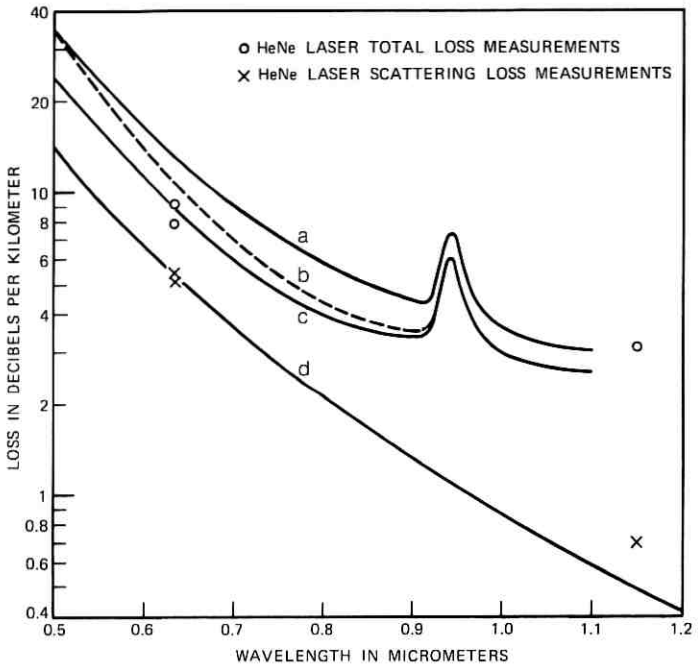

Fig. 14—Spectral losses of multimode fiber SMF 56 with Spectrosil WF core on a Suprasil W1 slab; L = 130 m. (a) Approximate steady-state losses. (b) Losses for small-angle excitation. (c) Same as (b) but after stress was relieved by cooling the drum. (d) Rayleigh scattering losses of bulk fused silica.

Total losses of 9 and 3.1 dB/km measured with HeNe lasers at 0.6328 and 1.15 μm agree well with the incoherent-source losses. A 7.8-dB/km loss was obtained at 0.6328 μm for low-order-mode laser excitation. Since scattering losses for this excitation condition amounted to 5.1 dB/km, the approximate absorption losses at 0.6328 μm were 2.7 dB/km. With the scattering losses at 1.15 μm amounting to 0.7 dB/km, the absorption losses there were 2.4 dB/km. The absorption losses of a different Spectrosil WF bulk sample were measured by Rich[10] to be less than 1.6 dB/km at 1.06 μm.

The scattering losses were measured with a 4-cm-long integrating cell built with silicon photo detectors.[11] Highly reproducible data were obtained when cladding mode strippers were provided on both sides of the cell, and when small droplets of matching liquid were deposited on the 4-cm-fiber section in the cell to reradiate the otherwise captured cladding power. Scattering losses of 5.5 dB/km were measured at 0.6328 μm when the fiber was filled with a mode spectrum correspond-

ing to the fiber NA of 0.14. This loss value is lower than that reported for a recent low-loss Corning fiber,[9] and agrees well with the bulk scattering losses of 5 to 6 dB/km measured for numerous fused silica samples by Tynes.[12] Furthermore, it corresponds closely to a value of 5.4 dB/km that was computed from Rich and Pinnow's 12.4 dB/km loss measured at 0.5145 μm, using a λ^{-4} Rayleigh-scattering dependence (curve d in Fig. 14).[13] It is noteworthy that particularly at shorter wavelengths SMF 56 has approximately the λ^{-4} dependence of Rayleigh-scattering losses.

Whereas the spectral losses of SMF 56 agree closely with the unclad fiber losses of a different batch of Spectrosil WF[14] (Fig. 15, curve b), the unclad fiber losses of the same raw material from which the core rod of SMF 56 was prepared were relatively high (Fig. 15, curve a). This is attributed to imperfections in the preform rod and accidental contamination of the unclad fiber surface. In contrast to an about 11-dB/km OH peak in the unclad fiber loss curve, the single-material-fiber peak at 0.95 μm was a remarkably low 3 dB/km in spite of the

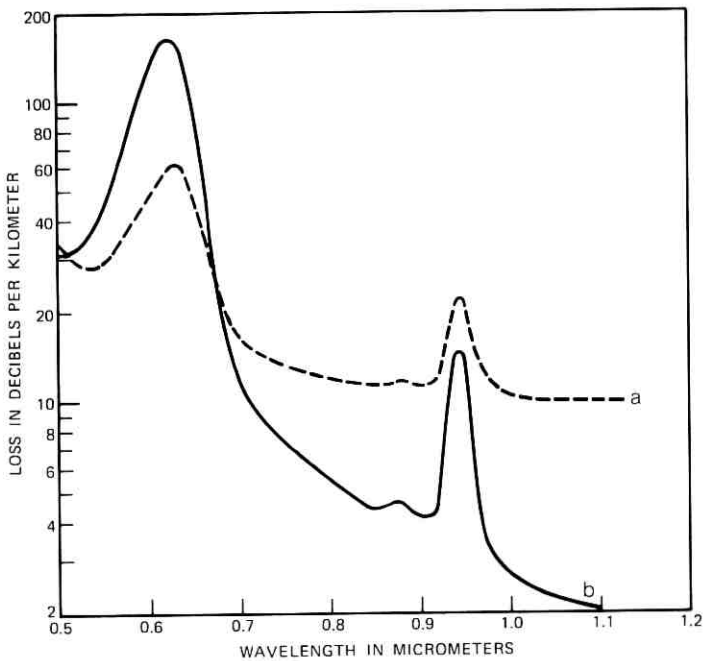

Fig. 15—Loss spectrum of unclad Spectrosil WF fiber. (a) Made from same bulk material as the core of SMF 56. (b) Previously measured sample.

fact that an oxy-hydrogen torch was used for the redraw operation. The difference must be due to the fact that the core rod was only drawn to a diameter of approximately 1.5 mm (compared to the 0.2-mm diameter of the unclad fiber), after which it was surrounded with an inert atmosphere while it was drawn into the single-material fiber. We conclude, therefore, that most of the water in the unclad fiber, and possibly still in the single-material fiber, was introduced in the drawing process, and that the OH content of the raw material could be as low as 1 ppm.[6]

Aside from the water peak at 0.95 μm, the losses of SMF 56 monotonically decreased with wavelength. In contrast, the loss spectra of other single-material and unclad fibers drawn from silica with low OH content exhibit a strong loss band at 0.63 μm.[6,14,15] Similarly, the 0.63-μm band exists in the unclad fiber drawn from the same raw material (as shown in Fig. 15, curve a), although to a much lesser degree than in other samples evaluated previously. The reason for this is unknown. As shown in Ref. 15, the intensity of the 0.63-μm band depends, among other factors, on the drawing conditions, and it is typically less pronounced in single-material fibers than in related unclad fibers. Also, spontaneous annealing of this band has been observed.

The loss spectrum of a single-material fiber whose Suprasil W1 core was drawn using standard gases and brass fittings, is shown for comparison in Fig. 16. The resulting broad loss band centered between 0.8 and 0.9 μm is believed to be caused by contamination with copper.[16]

Preliminary dispersion measurements performed by Cohen[17] with single-material fibers up to 210 m long indicate a weak coupling between modes. The maximum pulse dispersion followed closely the theoretical prediction expressed by eq. (5).

Instead of using the rod-plate technique for the preform preparation, we can achieve longer preform lengths by using thin-walled tubes as supports. A single-mode, single-material fiber created at the intersection of two tubes is shown in Fig. 17.[18] Using three such tubes results in three junctions and associated single-mode guides within the same cladding tube (Fig. 18). Single- and multicore multimode guides using this approach can also be envisioned by supporting one or more core rods by a suitable number of tubes.[19]

V. SUMMARY

Low-loss, single- and multimode optical fibers were fabricated solely from pure fused silicas. The single-material fibers consisted of small-diameter rods supported on thin plates in the center of larger-diameter

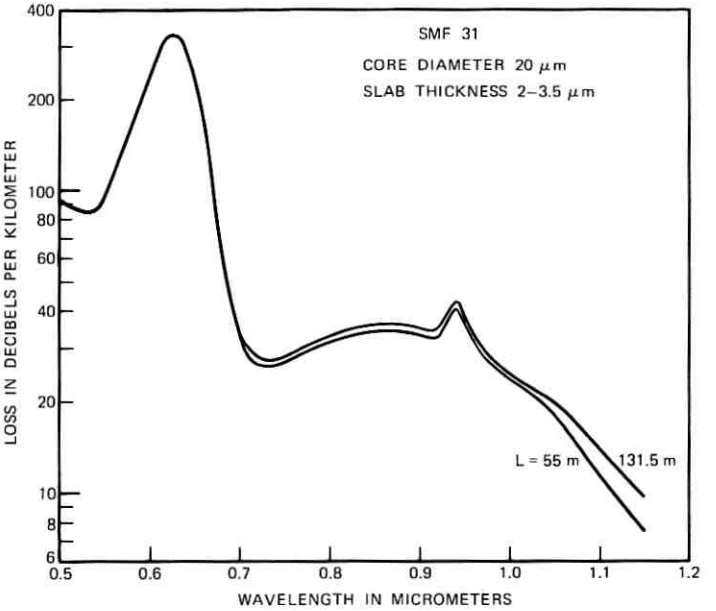

Fig. 16—Spectral losses of a Suprasil W1-cored, single-material fiber whose core was contaminated by an impure flame during the redraw operation.

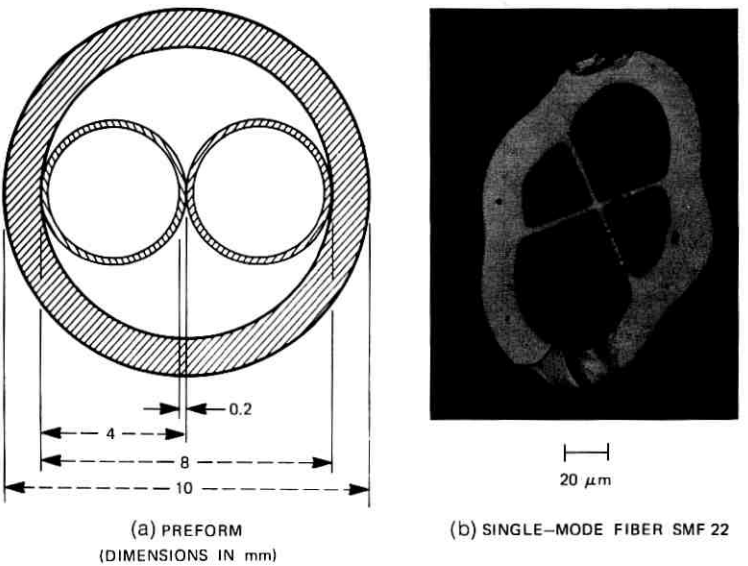

Fig. 17—Single-mode, single-material fiber made by fusing two thin-walled tubes.

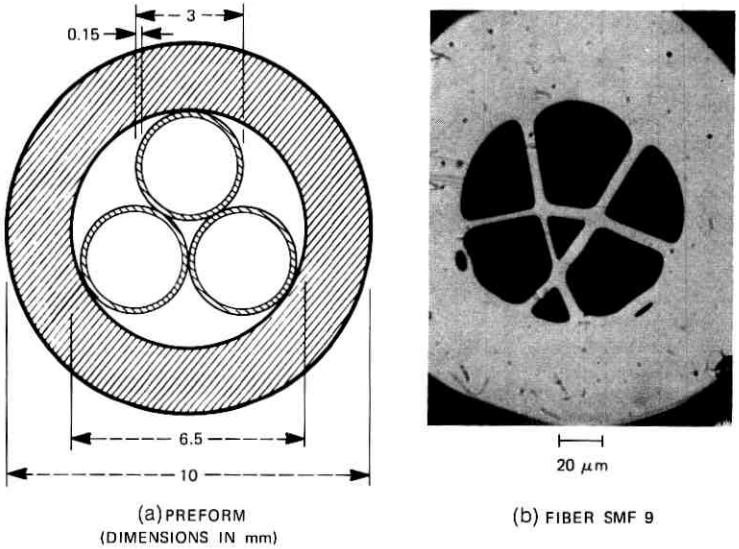

Fig. 18—Multiple-core, single-mode, single-material fiber made by fusing three thin-walled tubes.

protective tubings. The loss spectra of these fibers approached those of unclad fibers drawn from the same raw material. Specifically, steady-state losses of a 130-meter-long multimode fiber with Spectrosil WF core were approximately 6, 4.5, and 3 dB/km at wavelengths of 0.8, 0.9 and 1.1 μm, respectively, with even lower losses existing for low-order mode excitation. Aside from a small 3-dB/km OH band at 0.95 μm, the losses monotonically decreased throughout the 0.5- to 1.15-μm wavelength range investigated. Scattering losses, measured with HeNe lasers at 0.6328 and 1.15 μm, amounted to 5.5 and 0.7 dB/km, respectively.

For supporting plate thicknesses varying between 1 and 4 μm, the experimental NA of multimode fibers changed between 0.32 and 0.08 μm (at 0.6328 μm), which agrees excellently with theoretical predictions. Similarly, the predicted linear increase of the NA with wavelength was confirmed experimentally.

VI. ACKNOWLEDGMENTS

Valuable discussions with and recommendations by E. A. J. Marcatili and S. E. Miller are gratefully acknowledged. We thank A. D. Pearson, W. C. French, and D. D. Bacon for generously making avail-

able their clean-room facilities. R. E. Jaeger's CO_2-laser-drawing assistance, and R. D. Standley's electron-microscope pictures are very much appreciated. S. Gottfried contributed to this paper with some loss measurements of SMF 56.

REFERENCES

1. P. Kaiser, E. A. J. Marcatili, and S. E. Miller, "A New Optical Fiber," B.S.T.J., *52*, No. 2 (February 1973), pp. 265–269.
2. E. A. J. Marcatili, "Slab-Coupled Waveguides," B.S.T.J., *53*, No. 4 (April 1974), pp. 645–674.
3. S. E. Miller, unpublished work.
4. E. A. J. Marcatili and R. A. Schmeltzer, "Hollow Metallic and Dielectric Waveguides for Long Distance Optical Transmission and Lasers," B.S.T.J., *43*, No. 4 (July 1964), pp. 1783–1809.
5. P. Kaiser and H. W. Astle, "Measurement of Spectral Total and Scattering Losses in Unclad Optical Fibers," J. Opt. Soc. Am., *64*, No. 4 (April 1974), pp. 469–474.
6. P. Kaiser et al., "Spectral Losses of Unclad Vitreous Silica and Soda-Lime-Silicate Fibers," J. Opt. Soc. Am., *63*, No. 9 (September 1973), pp. 1141–1148.
7. P. Kaiser, unpublished work.
8. W. B. Gandrud, unpublished work.
9. D. B. Keck, R. D. Maurer, and P. C. Schultz, "On the Ultimate Lower Limit of Attenuation in Glass Optical Waveguides," Appl. Phys. Lett., *22*, No. 7 (April 1973), pp. 307–309. Note also announcement by P. C. Schultz on fabrication of a 2-dB/km fiber at Annual Meeting of American Ceramic Society, Cincinnati, Ohio, April 30–May 2, 1973.
10. T. C. Rich, unpublished work.
11. J. Stone, "Measurement of Rayleigh Scattering in Liquids Using Optical Fibers," Appl. Opt., *12*, No. 8 (August 1973), pp. 1824–1827.
12. A. R. Tynes, unpublished work.
13. D. A. Pinnow, T. C. Rich, F. W. Ostermayer, Jr., and M. DiDomenico, Jr., "Fundamental Optical Attenuation Limits in the Liquid and Glassy State with Application to Fiber Optical Waveguide Materials," Appl. Phys. Lett., *22*, No. 10 (May 1973), pp. 527–529.
14. P. Kaiser, "Spectral Losses of Unclad Fibers Made From High-Grade Vitreous Silica," Appl. Phys. Lett., *23*, No. 1 (July 1973), pp. 45–46.
15. P. Kaiser, "Drawing-Induced Coloration in Vitreous Silica Fibers," J. Opt. Soc. Am., *64*, No. 4 (April 1974), pp. 475–481.
16. H. L. Smith and A. J. Cohen, "Absorption Spectra of Cations in Alkali-Silicate Glasses of High Ultra-Violet Transmission," Phys. Chem. Glasses, *4*, No. 5, (October 1963), pp. 173–187.
17. L. G. Cohen, unpublished work.
18. E. A. J. Marcatili, patent filed.
19. P. Kaiser, patent filed.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

Treatment of Microscopic Fluctuations in Noise Theory

By K. K. THORNBER

(Manuscript received January 8, 1974)

A new method is introduced and used to calculate the statistics of the microscopic fluctuations of charge carriers in devices. By expressing the fluctuations of the carriers in terms of elementary transfer fluctuations, we are able to separate the induced fluctuations from the spontaneous fluctuations experienced by such carriers. This enables us to treat correlation effects in the dynamical portion of the problem and reserve for the statistical portion only well-defined, uncorrelated random forces whose statistics are readily calculated. The method includes all important correlation effects as well as multiple-decay-time relaxation effects and, thus, it fills a gap in the Langevin method as well as in the impedance-field method of calculating noise in devices. The method is suitable for treating nonstationary as well as stationary noise, and in some cases can be used directly on macroscopic problems. We also present a derivation of a recently introduced expression for diffusion noise of carriers whose mobility is a nonlinear function of applied electric field. This microscopic approach may further illustrate the origin, nature, and treatment of fluctuations in devices.

I. INTRODUCTION

In this article we describe a very simple means of treating microscopic fluctuations in noise theory. The method is simple in the sense that it focuses attention directly on the heart of the matter, the elementary processes which give rise to device noise. It is also simple in the sense that no sophistication in probability theory is used beyond an understanding of simple shot noise. Nonetheless, the method is rigorous under the rather mild constraint that the fluctuations are sufficiently small that the equations governing the noise are linear. The method has the added advantage that it can be used nearly as easily for nonstationary noise as for stationary noise. We make no

claim that this method is an advance in the philosophy of noise; we do claim, however, that it is adequate for solving many noise problems of practical interest.

There are two equivalent[1,2] methods of calculating device noise, the Langevin[3] method (LM) and the impedance-field[4] method (IFM). Both methods are characterized by an inherent simplification: namely, the separation of the task of calculating the spontaneous fluctuations of the current carriers in each elemental region of the device, and the task of calculating the observable response to these fluctuations at the external contacts of the device. The former task, the treatment of the microscopic fluctuations, is simplified because in dealing with the source of the fluctuations one can focus attention on the statistics of the microscopic variations inherent to the local physical conditions, which in turn are determined by the (noiseless) state of the device during operation. As a result, both the LM and the IFM are primarily concerned with the latter task, the coupling of the microscopic fluctuations to the macroscopic, observable voltages and currents. This task is also well-defined because the influence of the carriers in one region of the device on the carriers in another region, and on the contacts, has often been studied in detail in attempts to understand the dc and ac operation of the device. Thus, it is important to complement the LM or the IFM with the microscopic method described below. When this is done, it can be claimed that in most cases, if one understands the device sufficiently to calculate its noiseless operation, one can calculate the device noise as well. This should be of assistance to those tackling the noise in new and/or unfamiliar devices from scratch.

For example, if a device is sufficiently well-understood to be characterizable by an equivalent circuit, one can often introduce equivalent, random voltage and current sources to simulate the noise in the device.[5–8] Using such sources, a circuit designer with little interest in noise theory can readily calculate the size of the noise in the circuit employing the devices of interest. In a similar way, a person working with individual devices may find it convenient to have a simple scheme to translate the physical processes with which he is familiar into noise sources and to quickly evaluate their effect on the performance of the device of interest.[9–10] It is hoped that the method presented here will be used in such situations.

It is important to realize that charge carriers in any device fluctuate in response to random forces exerted on them.[3] The response to a specific impulse continues in general long after the impulse causing it

has ceased, and, in the interim, subsequent impulses will further alter the induced fluctuation. In addition, a fluctuation in the distribution of one type of carrier can induce fluctuations in the distributions of other types of carriers. Thus, whereas the statistics of the spontaneous, random forces may be quite simple, those of the carriers can be somewhat complicated owing to the correlation between the various induced fluctuations. The key to the simplicity of the method presented here results from the separation of the correlation effects from the statistical problem. These correlation effects are *not* neglected. Rather, they are included in the dynamical portion instead of in the statistical portion of the treatment. As it turns out, this results in the primary simplification achieved with our method.

In what follows, we shall use several examples to introduce and elaborate our microscopic approach. The first example, the decay of charge stored on a leaky capacitor, will motivate the method and illustrate how this technique can be used to treat certain macroscopic problems as well as microscopic ones. The second example concerns transfers between a two-level system. Here correlations are of primary importance, and our method is seen to treat these adequately yet simply. The third example illustrates how velocity fluctuations can be decomposed into transfer fluctuations, which are much easier to treat. Complicated scattering mechanisms including multiple decay times, which are not normally covered in the usual Langevin[3] or impedance-field[4] methods, can be handled with relative ease. The fourth example considers recombination to illustrate how correlation effects can be treated efficiently and effectively. By working in the time domain, we can see how the method works for nonstationary[9] as well as for stationary statistics. The statistics are treated in detail in the appendices. In particular, a recently used expression[11] for the diffusion noise of carriers having a nonlinear mobility is derived.

It is hoped that our discussion will provide, for the nonspecialist, further insight into the physical nature and mathematical representation of noise in general.

II. SOME SIMPLE EXAMPLES

Since the primary purpose of this paper is to elucidate a general, practical approach to solving the microscopic portion of noise problems, it seems best at first to describe this approach in terms of simple examples. Although this method is best suited to treat noise at the microscopic level, for purposes of illustration we shall commence with a macroscopic example.

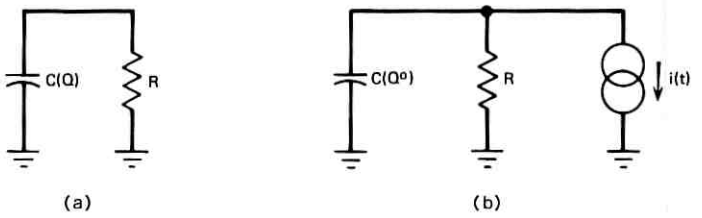

Fig. 1—(a) *RC* circuit in which charge initially placed on the capacitor decays to ground through the resistor. (b) Noise equivalent circuit of (a).

Consider the *RC* circuit shown in Fig. 1a in which we assume that capacitance C is a function $C(Q)$ of the charge Q, which it stores. Let us assume that initially ($t = 0$) a charge of size Q_o^o is stored on the capacitor. Subsequently ($t > 0$) the charge will decay away through the resistor R. In the absence of noise, the charge as a function of time $Q^o(t)$ satisfies the equation

$$\frac{dQ^o(t)}{dt} = -\frac{Q^o(t)}{RC[Q^o(t)]}, \tag{1}$$

subject to the boundary condition that $Q^o(0) = Q_o^o$.

We know, however, that the thermal (Brownian) motion of the electrons in the resistor gives rise to a noise current. Thus, rather than the uniform charge decay predicted by (1) for noiseless conditions, the charge decay is in fact somewhat random. Moreover, the noise is not solely governed by the resistor. If, during a time interval, the noise current in the resistor is such as to draw too many charges from the capacitor, then the voltage on the capacitor will decrease and the subsequent current will be reduced. This is just the response of the circuit to the noise generated by the resistor. In such a problem, we would ordinarily determine the statistics of the noise current $i(t)$ associated with the resistance itself, and then determine the fluctuations in the charge $Q(t)$ on the capacitor in terms of $i(t)$. We would be implicitly assuming, and rightly so, that only the resistor, and not the circuit exterior to the resistor, determines the statistics of $i(t)$, the thermal noise current generated by the resistor. We would then relate $Q(t)$ to $i(t)$ using an equation of the form

$$\frac{dQ}{dt} = -I(Q) - i(t), \qquad I(Q) \equiv \frac{Q}{RC(Q)}. \tag{2}$$

(The noise equivalent circuit is shown in Fig. 1b.) We would then write $Q(t)$ as a sum of its stochastic average $Q^o(t)$ and a fluctuation

$\delta Q(t) \equiv Q(t) - Q^o(t)$; assume that $\langle \delta Q(t)^2 \rangle^{\frac{1}{2}} \ll \Delta Q$, where

$$\Delta Q = (\partial I/\partial Q)/(\partial^2 I/\partial Q^2);$$

and expand eq. (2) to first order in $\delta Q(t)$, obtaining a nonlinear equation [eq. (1)] for the $Q^o(t)$ and a linear equation,

$$\frac{d\delta Q}{dt} = -\frac{\delta Q}{\tau(t)} - i(t), \tag{3}$$

for δQ in terms of $i(t)$. In eq. (3) $\tau(t)$ is defined by

$$\frac{1}{\tau(t)} \equiv \frac{d}{dQ}\left(\frac{Q}{RC(Q)}\right)\Bigg|_{Q=Q^o(t)}. \tag{4}$$

Solving (3) for $\delta Q(t)$ in terms of $i(t)$ permits the statistics of $\delta Q(t)$ to be obtained from those of $i(t)$, which for now we assume we know. In this way, we can determine the noise associated with the variable of interest $Q(t)$ from a knowledge of the noise associated with a more simply characterized noise variable $i(t)$.

The foregoing procedure has one drawback: one must be extremely careful, in general, to properly introduce such noise terms as $i(t)$ into otherwise noiseless dynamical relations. As we shall see in the examples considered below, there is a one-to-one correspondence between the transfer processes characterizing the problem of interest and the noise terms which one introduces. Thus, if one writes down several coupled-rate equations, each involving several transfer processes, and introduces but one noise term per equation, one finds, in general, that these noise terms are not simple, being correlated statistically to one another. Since, for simplicity, one would desire noise terms to be uncorrelated, some care must be used in including them in the rate equations. We now outline a procedure in the context of the above example that can be used in more complicated problems to insure that such noise terms are included properly.

The decay of charge from capacitor to ground is accomplished by transfer of individual electrons. Let t_i be the time at which the ith electron leaves the capacitor. Then we may write the following kinematic relation for the decay of the charge on the capacitor,

$$\frac{dQ(t)}{dt} = -q \sum_i \delta(t - t_i), \tag{5}$$

where q is the size of an elementary charge. If we consider an ensemble of RC circuits of the type shown in Fig. 1a, in which an initial charge of Q_o^o is decaying, each decay will be characterized by a different set of

times $\{t_i\}$ at which elementary charges leave the capacitors. Each t_i, therefore, is a random variable whose probability distribution is, in general, dependent upon all t_j for the preceding events ($t_j < t_i$). A completely rigorous derivation of the noise would, therefore, involve calculating the probability of each such sequence of times $\{t_i\}$, including all correlation effects, and then using these probabilities to ascertain the statistics of the noise. Were it not for these correlation effects, the t_i would be independent, and the statistical problem would be greatly simplified.

We shall now recast eq. (5) into a form that greatly simplifies the correlation problem by decomposing the current into a spontaneous portion $d(t)$ and an induced portion $R(t)$. The spontaneous portion is governed by the sources of the noise, and the induced portion is controlled by the instantaneous state of the device, in this case the stored charge $Q(t)$. Returning to the charge decay problem, we rewrite eq. (5) in the following form:

$$\frac{dQ}{dt} = -qR(t) - qd(t), \tag{6}$$

where

$$d(t) \equiv \sum_i \delta(t - t_i) - R(t), \tag{7}$$

and where $R(t)$ is the "dynamical" rate of charge loss. By "dynamical" rate we mean that $R(t)$ is, in general, a function of (that is, is determined by) the dynamical variables of the problem. In this case, we have

$$qR(t) = \frac{Q(t)}{RC[Q(t)]}. \tag{8}$$

If there is a fluctuation δQ in Q, then a fluctuation $\delta R(t)$ occurs in $R(t)$ also, which in this case is given to first order in $\delta Q(t)$:

$$q\delta R(t) = \delta Q(t)/\tau(t), \tag{9}$$

where τ is defined in eq. (4). Returning to eq. (6), if we write $Q = Q^o + \delta Q$ as we did in passing from eq. (2) to eq. (3), then we obtain an equation for the noise δQ [which corresponds to eq. (3) above],

$$\frac{d\delta Q}{dt} = -\frac{\delta Q}{\tau(t)} - qd(t). \tag{10}$$

In this case, $qd(t)$ corresponds to the noise current $i(t)$ generated by the resistor, which, as discussed above, is independent of the δQ associated with the capacitor. Thus, $d(t)$ serves as a statistical driving term. We

may calculate $\delta Q(t)$ in terms of $d(t)$ and from the statistics of $d(t)$ determine those of $\delta Q(t)$.

Let us return to eq. (7) for a moment to point out why $d(t)$ is to be regarded as the spontaneous portion of the current. The $-e \sum_i \delta(t - t_i)$ term is, of course, the entire current associated with the charge decay. The dynamical rate term $R(t)$, however, is a function only of the dynamical variables of the problem and does not contain noise sources. In our example, $R(t)$ involves only the charge $Q(t) = Q^o(t) + \delta Q(t)$, where $\delta Q(t)$ is the charge fluctuation induced by the noise sources acting on the device. $R(t)$ does not involve the noise sources themselves. Thus, if $d(t) = 0$, δQ, the response, vanishes. Hence, inasmuch as $d(t)$ is the difference between the total current and the noiseless-plus-induced portion of the current, it follows that $d(t)$ can contain only the spontaneous portion of the current. The advantage of starting with eq. (5) and proceeding as we did to eq. (10), rather than attempting to write eq. (2) [or eq. (3)] *a priori*, will become evident when more than one process is involved in the problem. By introducing a $d_i(t)$ and $R_i(t)$ for each process i, and noting that each $d_i(t)$ involves only spontaneous fluctuations and, hence, must be independent of all the other d_j, $j \neq i$, we can readily express the fluctuations of interest in terms of the independent statistical driving terms d_i.

If we solve eq. (10) for $\delta Q(t)$, we find that

$$\delta Q(t) = \int_{-\infty}^{t} dt' \exp\left[- \int_{t'}^{t} dt''/\tau(t'') \right] [-qd(t')]. \tag{11}$$

Often one is most interested in the mean-square fluctuation $\langle \delta Q^2 \rangle$ for some time t. From (11) this is given by

$$\langle \delta Q^2(t) \rangle = q^2 \int_{-\infty}^{t} dt_1' \int_{-\infty}^{t} dt_2' \exp\left[- \int_{t_1'}^{t} dt''/\tau(t'') \right] \times \exp\left[- \int_{t_2'}^{t} dt''/\tau(t'') \right] \langle d(t_1')d(t_2') \rangle. \tag{12}$$

If we know the statistics of $d(t)$ (in this case, those of $i(t)/q$ associated with the resistor), we can calculate $\langle d(t_1)d(t_2) \rangle$, and hence $\langle \delta Q^2(t) \rangle$. Note that, in general, we must also know the noiseless solution $Q^o(t)$ [eq. (1), and also see eq. (4)]. For simplicity here, let us assume that $i(t)$ is pure thermal noise so that

$$q^2 \langle d(t_1)d(t_2) \rangle = 2kTG\delta(t_1 - t_2), \qquad G = 1/R.$$

It follows then that

$$\langle \delta Q^2(t) \rangle = kTC, \tag{13}$$

the usual result. [Note: $\langle \delta Q(t_1)\delta Q(t_2)\rangle = kTC \exp(-|t_1 - t_2|/\tau)$. This illustrates how eq. (10) maintains the correlation between fluctuations at two different times, while the fluctuations themselves are driven by a source without correlation, $\delta(t - t')$.] When we turn to purely microscopic processes, we shall find the statistics of $d(t)$ are governed by the rate function $R(t)$.

In what follows, the very important distinction between spontaneous and induced fluctuations will be used repeatedly. The advantage of this macroscopic example is that the separation between the two may be clearly visualized. From the point of view of the RC circuit, fluctuations in the current generator are spontaneous and induce fluctuations in the charge decay, δQ. In addition, the rate term $R(t)$ is a function only of Q and clearly includes only the induced and *not* the spontaneous fluctuations. The reader may find it helpful in subsequent examples to refer back to this simple model to clarify the somewhat more subtle distinctions between induced and spontaneous fluctuations at the microscopic level.

We now consider another simple example, this time a truly microscopic one. Let us consider two states a and b containing $n_a(t)$ and $n_b(t)$ charges, respectively. [These states, for example, might be two regions of phase space $(d\mathbf{x}d\mathbf{v})$:* one region for $(\mathbf{x}_a, \mathbf{v}_a)$ and one for $(\mathbf{x}_b, \mathbf{v}_b)$, or a might be trapped electrons and b free electrons.] Let us assume that charges are flowing to b from a at a rate $R_{ba} = R_{ba}(n_a, n_b)$ and to a from b at a rate $R_{ab} = R_{ab}(n_b, n_a)$. If charges leave a and enter b at times t_{bai} and leave b and enter a at times t_{abi}, then by analogy with (5) we write

$$\dot{n}_a = -\sum_i \delta(t - t_{bai}) + \sum_i \delta(t - t_{abi}) \tag{14a}$$

and

$$\dot{n}_b = -\sum_i \delta(t - t_{abi}) + \sum_i \delta(t - t_{bai}). \tag{14b}$$

Following the previous example, we rewrite (14) in the form

$$\dot{n}_a = -R_{ba}(t) + R_{ab}(t) - d_{ba}(t) + d_{ab}(t) \tag{15a}$$

and

$$\dot{n}_b = -R_{ab}(t) + R_{ba}(t) - d_{ab}(t) + d_{ba}(t), \tag{15b}$$

where, of course,

$$d_{ab}(t) \equiv \sum_i \delta(t - t_{abi}) - R_{ab}(t) \tag{16a}$$

* Note that boldface capital letters denote matrices; boldface lower-case letters denote vectors.

and

$$d_{ba}(t) \equiv \sum_i \delta(t - t_{bai}) - R_{ba}(t). \tag{16b}$$

If we write $n_a = n_a^o + \delta n_a$, $n_b = n_b^o + \delta n_b$, insert into (15), and expand, we obtain the following equations for the noiseless quantities n_a^o, n_b^o:

$$\dot{n}_a^o = -R_{ba}^o(t) + R_{ab}^o(t) \tag{17a}$$

and

$$\dot{n}_b^o = -R_{ab}^o(t) + R_{ba}^o(t), \tag{17b}$$

where

$$R_{ba}^o \equiv R_{ba}(n_a^o(t), n_b^o(t)), \qquad R_{ab}^o \equiv R_{ab}(n_b^o(t), n_a^o(t)).$$

For the noise δn_a, δn_b, we obtain the linear relations

$$\delta\dot{n}_a = -\left(\frac{\delta R_{ba}}{\delta n_a} - \frac{\delta R_{ab}}{\delta n_a}\right)\delta n_a - \left(\frac{\delta R_{ba}}{\delta n_b} - \frac{\delta R_{ab}}{\delta n_b}\right)\delta n_b - d_{ba}(t) + d_{ab}(t) \tag{18a}$$

and

$$\delta\dot{n}_b = -\left(\frac{\delta R_{ab}}{\delta n_a} - \frac{\delta R_{ba}}{\delta n_a}\right)\delta n_a - \left(\frac{\delta R_{ab}}{\delta n_b} - \frac{\delta R_{ba}}{\delta n_a}\right)\delta n_b - d_{ab}(t) + d_{ba}(t), \tag{18b}$$

from which δn_a and δn_b can be determined in terms of d_{ab} and d_{ba}. (The linear operators of the form $\delta R/\delta n$ are evaluated at their noiseless values.) Since there are only two states, and since $n_a + n_b =$ constant ($\dot{n}_a + \dot{n}_b = 0$), it comes as no surprise that $\delta n_a(t) = -\delta n_b(t)$; one state's loss is the other's gain. Nonetheless, the source or driving terms, d_{ab} and d_{ba}, are independent, and the correlations between δn_a and δn_b are included in (18) through the presence of the dynamical terms. Since this is an important point, we shall discuss it more fully below.

If we ignore for the moment the random nature of the flow of charges from a to b (and from b to a), then eq. (17) tells us that we have a "smooth" continuous flow of charges from a to b at a rate of $R_{ba}^o(t)$ and from b to a at a rate of $R_{ab}^o(t)$. Noise enters the problem when we note (as we have above) that charges are actually transferred at times $\{t_{bai}\}$ and $\{t_{abi}\}$, the (ensemble) average rate of occurrence of these times being $R_{ba}^o(t)$ and $R_{ab}^o(t)$. Since $R_{ba}^o(t)$ and $R_{ab}^o(t)$ depend only on the steady-state solution (n_a^o, n_b^o) to eq. (17), these quantities are not affected by the details of a particular set of fluctuations. This means that if R_{ba}^o and R_{ab}^o govern the statistics of the transfers from a to b and b to a, then the individual transfer times $\{t_{bai}^o\}$, $\{t_{abi}^o\}$ associated with

R^o_{ba} and R^o_{ab} must be statistically independent. [If they were correlated, then R^o_{ba} and R^o_{ab} would reflect this correlation much as do R_{ba} and R_{ab} in eq. (15).] In discussing the discharge of the capacitor, we pointed out how $d(t)$ acted as a noise source containing only the spontaneous fluctuations. Above, we have noted that these fluctuations are governed by the average rates R^o. This suggests that to calculate the statistics of d_{ab} and d_{ba}, which according to (18) are needed to calculate the statistics of δn_a and δn_b, we may write the $d(t)$'s in the following form:

$$d_{ab}(t) = \sum_i \delta(t - t^o_{abi}) - R^o_{ab}(t) \tag{19a}$$

and

$$d_{ba}(t) = \sum_i \delta(t - t^o_{bai}) - R^o_{ba}(t). \tag{19b}$$

We stress at this point that for a specific event ($\{t_{abi}, t_{bai}\}$) we need not demand that the right-hand sides of (16) and (19) be equal. This is because we eventually average the dependence of n_a and n_b on d_{ab} and d_{ba} over all events in calculating correlation functions and spectral densities. All that is necessary is that the statistical properties of the two forms of d_{ab} and d_{ba} be the same, at least to lowest order in the size of the fluctuations. We shall see below that this is indeed the case. We are able to calculate simply the statistics of d_{ab} and d_{ba} because in writing (19) we have cast these random variables into a form in which the distinction between spontaneous and induced fluctuation no longer enters. We can do this because d involves only spontaneous (hence independent) fluctuations and, therefore, can be written in terms of independent events.

With independent transfer times, it is straightforward to calculate the statistical distributions of d_{ab} and d_{ba} as given in (19). This calculation is carried out in Appendix A. The autocorrelation functions for d_{ab} and d_{ba}, which are the spontaneous-fluctuation, noise-source terms, are given by

$$\langle d_{ab}(t_1) d_{ab}(t_2) \rangle = R^o_{ab}(t_1)\delta(t_1 - t_2) \tag{20a}$$

and

$$\langle d_{ba}(t_1) d_{ba}(t_2) \rangle = R^o_{ba}(t_1)\delta(t_1 - t_2), \tag{20b}$$

as expected for pure shot noise associated with independent events. In addition, if R^o_{ab} and R^o_{ba} are independent of time (so that the noise is stationary), then the spectral densities of d_{ab} and d_{ba} are given by

$$S^d_{ab}(f) = 2R^o_{ab} \tag{20c}$$

and

$$S^d_{ba}(f) = 2R^o_{ba}. \tag{20d}$$

The form given d_{ab} and d_{ba} in eq. (19), or rather their statistical form, eq. (20), is the basic assumption that enters our approach. That it is valid so long as the fluctuations δR in the rates $R = R^o + \delta R$ satisfy

$$\langle \delta R^2 \rangle^{\frac{1}{2}} \ll R^o \tag{21}$$

may be motivated by the following. If a fluctuation δn in n leads to a fluctuation δR in R, then the statistics of the immediately following t_i will be governed by the altered rate. This suggests as a first-order iteration in the fluctuation that

$$\langle d(t_1)d(t_2)\rangle \approx R(t_1)\delta(t_1 - t_2). \tag{22}$$

However, $R = R^o + \delta R$; thus, the error in (20) is of the order of $\delta R/R^o$, which we assume to be small. A similar argument applies to residual correlation effects among the d's. Thus, so long as (21) is satisfied, we expect (20) to be valid to order $\delta R/R^o$. Since we are discussing a linear theory, this is good enough for our purposes.

Returning to (18), assuming stationary noise so that we may use Fourier-transform techniques, and representing that the linear operators are simple decay rates for purposes of illustration, we may write

$$i\omega\delta n_a = -\delta n_a/\tau_a + \delta n_b/\tau_b - d_{ba} + d_{ab} \tag{23a}$$

and

$$i\omega\delta n_b = +\delta n_a/\tau_a - \delta n_b/\tau_b - d_{ab} + d_{ba}, \tag{23b}$$

which become, upon solving for $\delta n_a(\omega)$ and $\delta n_b(\omega)$ in terms of $d_{ba}(\omega)$ and $d_{ab}(\omega)$,

$$\delta n_a = -\delta n_b = (d_{ab} - d_{ba})(i\omega + 1/\tau)^{-1}, \tag{23c}$$

where

$$1/\tau = 1/\tau_a + 1/\tau_b. \tag{23d}$$

Since the d_{ab} and d_{ba} are independent, for spectral densities we obtain

$$S^{\delta n}_a(f) = S^{\delta n}_b(f) = |i\omega + 1/\tau|^{-2}[S^d_{ab}(f) + S^d_{ba}(f)] \tag{24a}$$

$$= |i\omega + 1/\tau|^{-2}(2R^o_{ab} + 2R^o_{ba}) \tag{24b}$$

$$= -S^{\delta n}_{ab}(f) = -S^{\delta n}_{ba}(f), \tag{24c}$$

where the expressions in (24c) are cross-spectral densities ($f = \omega/2\pi$). In such a problem as this with transfers permitted only between two states, the correlation between δn_a and δn_b is maximal. (Introducing

more states will lessen the correlation between any two.) Nonetheless, the important features of the correlation are clearly obtained in the result (24b): the small-signal decay rate $(1/\tau)$ is the sum of the separate decay rates, and the spectral density is proportional to a full shot-noise term derived from the sum of the noiseless rates $(R^o_{ab} + R^o_{ba})$. This example is admittedly very trivial and could have been done more easily by other means. In the following sections, the advantages of the present method will become evident. Note, however, that there was no question as to how to include the driving terms when writing down the coupled noise equations (18) and no question regarding the independence of these driving terms when it came to calculating spectral densities in (24).

Let us pause before going on to summarize the logic we have used in arriving at our method. First we recognized that noise arises owing to the randomness in the times at which charge carriers change their state. Then for each transfer process, we separated the expression for the particle current flowing between any two states into a dynamical-rate term $R(t)$ that includes all induced fluctuations, and a driving term $d(t)$ that includes only spontaneous (and therefore independent) fluctuations. In this way, we are able to shift the statistical correlations, so to speak, out of the statistical term and into the dynamical term. To calculate the statistics of our driving term, we noted that since the fluctuations were characterized by a mean rate of $R^o(t)$, which depended only on the noiseless solution, we could rewrite $d(t)$ in a natural way in terms of independent events. [These latter events cannot be coupled since their statistics are governed only by $R^o(t)$ and not by preceding events.] Although this step represented an approximation, we argued that it should be all right for small fluctuations. With the new form for $d(t)$, its statistics were readily calculated. Finally, knowing the statistics of d and the linear relation between δn and d, the statistics of δn including all correlation effects of interest, which we desire, are straightforward to obtain. In the next two sections, we shall carry out the above procedure for two examples from start to finish.

In passing, we make reference to Lax's discussion[12] in which he showed the basic limitation of any source theory of noise. The heart of Lax's argument is that one cannot expect to model on the time scale of the duration of a spontaneous fluctuation (during which time the system appears to be reversible) by using the essentially irreversible dynamics embodied in eq. (18). As for time scales on the order of the response time of the carriers to spontaneous fluctuations, however, all is well. Thus, as Langevin probably recognized long ago, the price paid

for using source terms is small. The whole trick is to include them unambiguously.

III. NOISE IN CHARGE TRANSFER

The purpose of this section is to stress the insight one can gain by starting with the most elementary transfer processes taking place in a device. In this case, we avoid making assumptions about the statistics of the velocity fluctuations of carriers by calculating them from occupancy fluctuations, which can be understood much more simply using our method. For simplicity, we shall ignore correlation effects, which in fact are shown to be negligible for the problem we discuss. In the next section, which is on recombination, we shall stress correlation effects and how they can be dealt with using the microscopic method.

One of the most basic approaches to the problem of the storage and transport of carriers in a device is first to assign a density function to each carrier state. For example, $n_a(\mathbf{x}, \mathbf{v}, t)d\mathbf{x}d\mathbf{v}$ represents the number of carriers of type a at $\mathbf{x}$ with velocity $\mathbf{v}$ in the element $d\mathbf{x}d\mathbf{v}$ of phase space. Similarly, $n_b(\mathbf{x}, t)d\mathbf{x}$ represents the number of carriers of type b trapped at $\mathbf{x}$ in volume $d\mathbf{x}$. In this example, let us ignore trapping and recombination effects and focus attention on the transport via scattering of a single type of carrier. This will lead to a general expression for diffusion noise which we can compare with the results obtained by Langevin and by Shockley et al. using the Langevin method and the impedance-field methods, respectively.

We proceed as follows. If at times t_{ijl} a particle is scattered from state j to state i, then we have (by analogy with our treatments above) that

$$\frac{dn_i}{dt} = -\sum_{jl} \delta(t - t_{jil}) + \sum_{jl} \delta(t - t_{ijl}) \tag{25}$$

and

$$\frac{dn_i}{dt} = -\sum_j R_{ji}(t) + \sum_j R_{ij}(t) - \sum_j d_{ji}(t) + \sum_j d_{ij}(t). \tag{26}$$

The subscripts i, j each designate a particular region of phase space $d\mathbf{x}d\mathbf{v}$ at $(\mathbf{x}, \mathbf{v})$. (We do this to simplify the notation.) In (26), $d_{ij}(t)$ is defined by

$$d_{ij}(t) = \sum_l \delta(t - t_{ijl}) - R_{ij}(t). \tag{27}$$

It is important to note that we introduce a source term for *each* elementary process, that is, for each transfer process indexed by the ordered pair (i, j).

As before, we let $n_i \equiv n_i^o + \delta n_i$, insert into (26), and expand to lowest order in δn_i. The noiseless equation that results in

$$\frac{dn_i^o}{dt} = -\sum_j R_{ji}^o(t) + \sum_j R_{ij}^o(t) \tag{28}$$

is just the Boltzmann equation, R_{ij}^o being a function of n_i^o and n_j^o. Using $\mathbf{x}$, $\mathbf{v}$ notation, (28) becomes

$$\begin{aligned}\frac{dn(\mathbf{x}, \mathbf{v}, t)}{dt} &= \frac{\partial n(\mathbf{x}, \mathbf{v}, t)}{\partial t} + \mathbf{v} \cdot \frac{\partial n(\mathbf{x}, \mathbf{v}, t)}{\partial \mathbf{x}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial n(\mathbf{x}, \mathbf{v}, t)}{\partial \mathbf{v}} \\ &= -\int d\mathbf{x}'\, d\mathbf{v}' (R(\mathbf{x}, \mathbf{v}; \mathbf{x}', \mathbf{v}') - R(\mathbf{x}', \mathbf{v}'; \mathbf{x}, \mathbf{v})),\end{aligned}$$

the more usual form of the Boltzmann equation, where $R(1, 2)$ is the average rate at which particles are scattered from 2 to 1. In general, $R(1, 2)$ is a function of n_1 and n_2. Usually $R(1, 2)$ is taken to be proportional to n_2, however. To determine such noiseless quantities as mobility, etc., (28) must be solved for the noiseless $n^o(\mathbf{x}, \mathbf{v}, t)$.

The equation for the noise $\delta n(\mathbf{x}, \mathbf{v}, t) = \delta n_i$ obtained from (26) is

$$\begin{aligned}\delta \dot{n}_i = &-\sum_{j,k} \frac{\delta R_{ji}}{\delta n_k} \delta n_k + \sum_{j,k} \frac{\delta R_{ij}}{\delta n_k} \delta n_k \\ &- \sum_j d_{ji}(t) + \sum_j d_{ij}(t).\end{aligned} \tag{29}$$

In general, this is a coupled (linear, integral-differential) equation among the δn_i. [As before, the linear operator $\delta R(n)/\delta n$ is evaluated at its noiseless value by inserting the noiseless solutions n_i^o of (28).] For our purposes here, we simplify (29) by assuming (*i*) that R_{ji} is a function of (and not an operator on) the n_k, and (*ii*) that in (29) we can ignore δn_k where $k \neq i$. Assumption (*i*) is in fact the usual situation one has with the Boltzmann equation. Assumption (*ii*) can be made plausible in the following manner. By ignoring in (29) δn_k, $k \neq i$, we are ignoring correlations among the fluctuations δn_i and δn_k. Let us suppose that the scattering rates between i and k are equal for all $N(N + 1)$ (i, k) pairs. Then a fluctuation δn_i will lead to a δn_k on the order of $\delta n_i/N$. The effect of δn_k back on state i will be of the order $\delta n_k/N$. Thus, summing over the N states $k \neq i$, the correlated contribution to δn_i which we lose by ignoring the δn_k in (29) is of the order of $\delta n_i/N$ $[= N \cdot (\delta n_i/N)/N]$, which for large N is entirely negligible. While we have assumed that the transfer rates between all i and k states are equal, a similar argument applies so long as the scattering to a few states is not favored. In this case, the δn_k dependence of these

few states must be included in (29). (We have already considered the two-state problem where correlation is largest. In our next example, we shall consider recombination in which correlation among several states is important.) Finally, we should note that our approximation is *not* equivalent to a relaxation-time approximation, which might be used to simplify (28).

Returning to our example, and based upon the reasoning given above, we approximate (29) by

$$\begin{aligned}\delta\dot{n}_i = &- \sum_j \frac{\delta R_{ji}}{\delta n_i}\,\delta n_i + \sum_j \frac{\delta R_{ij}}{\delta n_i}\,\delta n_i \\ &- \sum_j d_{ji}(t) + \sum_j d_{ij}(t)\end{aligned} \tag{30a}$$

and

$$\delta\dot{n}_i = -\delta n_i/\tau_i(t) - d_i^o(t) + d_i^e(t), \tag{30b}$$

where

$$1/\tau_i(t) \equiv \sum_j \left(\frac{\delta R_{ji}}{\delta n_i} - \frac{\delta R_{ij}}{\delta n_i}\right), \tag{30c}$$

$$d_i^o(t) \equiv \sum_{j,l} \delta(t - t_{jil}) - \sum_j R_{ji}(t), \tag{31a}$$

and

$$d_i^e(t) \equiv \sum_{j,l} \delta(t - t_{ijl}) - \sum_j R_{ij}(t). \tag{31b}$$

The superscripts on d_i, "o" and "e," designate "out" and "in" scattering, respectively. We are able to lump the driving terms together in this way because (30a, b) involve only δn_i and because the residual correlation between d_i^o and d_i^e is of order $1/N$.

To proceed further, we approximate (31a, b) in the manner discussed above [see (19)] and determine the statistics of δn_i. Using the general results derived in Appendix A for d and then in Appendix B for $2n$ as a function of d, we find for the correlation function of δn_i, δn_j the expression

$$\langle \delta n_i(t_1)\delta n_j(t_2)\rangle = \delta_{ij}[R_i^o(\text{in}) + R_i^o(\text{out})](\tau_i/2)e^{-|t_1-t_2|/\tau_i}, \tag{32}$$

where we have assumed that the noiseless rates $R_i^o(\text{in}) = \sum_j R_{ij}^o$ and $R_j^o(\text{out}) = \sum_j R_{ji}^o$ [and therefore τ_i defined in (30)] are independent of time. One further simplification is often justified. Under the stationary conditions taken here to arrive at (32), $R_i^o(\text{in}) = R_i^o(\text{out})$. Furthermore, if R_{ji} depends only on n_i and is directly proportional thereto, it follows

that

$$R_i^o(\text{out})\tau_i = n_i^o. \tag{33}$$

Thus, we find the expected result that

$$\langle \delta n_i(t_1)\delta n_j(t_2)\rangle = \delta_{ij} n_i^o e^{-|t_1-t_2|/\tau_i}. \tag{34}$$

We shall now use this result to determine the statistics of the velocity fluctuations of the individual carriers. It is velocity fluctuations rather than occupancy fluctuations that are usually taken as basic. We shall see, however, that occupancy fluctuations are the more fundamental of the two.

Diffusion noise arises from the velocity fluctuations that a charge carrier undergoes owing to its interaction with the material in which it is confined as well as with the other carriers. Since the current in a device is made up of the linear superposition of the currents carried by each carrier, and since each contribution is proportional to the velocity of the individual carrier, a calculation of the current correlation function (to determine device noise) requires knowledge of the velocity autocorrelation function for each carrier. This latter quantity we can calculate at once from (34). We proceed in the following manner.

By definition, the velocity of a carrier is given by

$$\mathbf{v}(\mathbf{x}, t) = \frac{\int d^3\mathbf{v}'\mathbf{v}' n(\mathbf{x}, \mathbf{v}', t)}{\int d^3\mathbf{v}' n(\mathbf{x}, \mathbf{v}', t)}. \tag{35}$$

If we recall that $n = n^o + \delta n$, we can write (35) in the form $\mathbf{v} = \mathbf{v}^o + \delta\mathbf{v}$ to determine that

$$\mathbf{v}^o(x_{n,t}) = \frac{\int d^3\mathbf{v}'\mathbf{v}' n^o(\mathbf{x}, \mathbf{v}', t)}{\int d^3\mathbf{v}' n^o(\mathbf{x}, \mathbf{v}', t)} \tag{36}$$

and that

$$\delta\mathbf{v}(x, t) = \frac{\int d^3\mathbf{v}'\mathbf{v}'\delta n(\mathbf{x}, \mathbf{v}', t)}{\int d^3\mathbf{v}' n^o(\mathbf{x}, \mathbf{v}', t)}, \tag{37}$$

where we have assumed that the total number of carriers within a volume element $d\mathbf{x}$ remains at the noiseless value. We make this assumption only because we are interested here in diffusion noise only

and not noise from fluctuations in particle density. It follows at once from (34) (assuming stationarity) and (37) that the velocity correlation function of interest is given by

$$\langle \delta v_i(\mathbf{x}_1, t)\delta v_j(\mathbf{x}_2, t_2)\rangle = \delta(\mathbf{x}_1 - \mathbf{x}_2)\frac{\int d^3\mathbf{v}' v_i' v_j' n^o(\mathbf{x}_1, \mathbf{v}')}{n^o(\mathbf{x}_1)\int d^3\mathbf{v}' n^o(\mathbf{x}_1, \mathbf{v}')} e^{-|t_1-t_2|/\tau(\mathbf{x}_1,\mathbf{v}')}. \quad (38)$$

In (38), i and j now correspond to the components of the velocity $\mathbf{v} = (v_x, v_y, v_z)$.

When one is dealing with stationary noise, it is more convenient to work with spectral densities than autocorrelation functions. Using the standard definitions of the spectral density function,[13] one finds using (34) that

$$\begin{aligned} S_{\delta n}(\mathbf{x}_1, \mathbf{v}_1, \mathbf{x}_2, \mathbf{v}_2, f) &= \delta(\mathbf{x}_1 - \mathbf{x}_2)\delta(\mathbf{v}_1 - \mathbf{v}_2)4\int_0^\infty e^{-t/\tau(\mathbf{x}_1,\mathbf{v}_1)} n^o(\mathbf{x}_1, \mathbf{v}_1)\cos \omega t dt \\ &= \delta(\mathbf{x}_1 - \mathbf{x}_2)\delta(\mathbf{v}_1 - \mathbf{v}_2)\frac{4n^o(\mathbf{x}_1, \mathbf{v}_1)\tau(\mathbf{x}_1, \mathbf{v}_1)}{1 + \omega^2\tau^2(\mathbf{x}_1, \mathbf{v}_1)}. \end{aligned} \quad (39)$$

It follows that the spectral density of the velocity fluctuations is given by

$$S_{\delta\mathbf{v},ij}(\mathbf{x}_1, \mathbf{x}_2, f) = \delta(\mathbf{x}_1 - \mathbf{x}_2)\int d\mathbf{v}\frac{4n^o(\mathbf{x}_1, \mathbf{v})\tau(\mathbf{x}_1, \mathbf{v})}{1 + \omega^2\tau^2(\mathbf{x}_1, \mathbf{v})}\frac{v_i v_j}{n^o(\mathbf{x}_1)^2}, \quad (40)$$

an expression which could have been obtained directly from (38).

The spectral density given in (40) has several interesting features that we shall touch on briefly. Ordinarily for the frequencies of interest in devices, $\omega^2\tau^2 \ll 1$, owing to the very rapid carrier-scattering rates. In this case (40) reduces to[4,14]

$$S_{\delta\mathbf{v},ij} = \delta(\mathbf{x}_1 - \mathbf{x}_2)4D_{ij}/n^o(\mathbf{x}_1), \quad (41)$$

where D_{ij} is the ij component of the diffusion tensor defined by

$$D_{ij} \equiv \frac{1}{2}\frac{d}{dt}\langle \delta\mathbf{x}_i(t_1)\delta\mathbf{x}_j(t_1 + t)\rangle. \quad (42)$$

This expression (42) can be derived from (38) by noting that

$$\delta x(t) = \int_{-\infty}^{t} dt'\delta v(t').$$

The diagonal components of D, $D_{ij} \equiv D_i$ satisfy

$$D_i = \frac{\int d\mathbf{v} v_i^2 n^o(\mathbf{x}, \mathbf{v}) \tau(\mathbf{x}, \mathbf{v})}{\int d\mathbf{v} v_i^2 n^o(\mathbf{x}, \mathbf{v})} \int d\mathbf{v} v_i^2 n^o(\mathbf{x}, \mathbf{v}) / n^o(\mathbf{x}). \tag{43}$$

In this form, we recognize that the first factor is $m\mu_i(\mathbf{x})/q$, while for a thermal distribution, the second factor is kT/m. Here, $\mu_i(\mathbf{x})$ is the carrier mobility in the ith direction at $\mathbf{x}$ derived using a standard Boltzmann equation approach. Inserting into (41), we find the usual result for thermal noise, namely that

$$S_{\delta v}(\mathbf{x}_1, \mathbf{x}_2, f) = \delta(\mathbf{x}_1 - \mathbf{x}_2) 4kT\mu(\mathbf{x}) / (qn^o(\mathbf{x}_1)). \tag{44a}$$

In Appendix C we generalize (44) to the case of field-dependent mobilities $\mu(E)$. The result when an effective temperature T can be defined is that

$$S_{\delta v}(\mathbf{x}_1, \mathbf{x}_2, f) = \delta(\mathbf{x}_1 - \mathbf{x}_2) 4 \left(\frac{kT}{q} \right) \frac{d[\mu(E)E]}{dE} \Big/ n^o(\mathbf{x}_1). \tag{44b}$$

We are now in a position to compare the results obtained here with those obtained by Langevin[3] and by Shockley et al.[4] using additional assumptions. For simplicity, we shall work in one dimension and ignore spatial variations. In the Langevin[3] method, one begins by decomposing the force acting on each carrier into two parts, a damping force proportional to the velocity and a stochastic force of zero mean. The purpose of the latter is, of course, to produce the random fluctuations in the velocity. Thus, one writes

$$\delta\dot{v} = -\delta v/\tau + h(t), \tag{45}$$

where $\tau = m\mu/q$. One then calculates the spectral density of δv in terms of that of h, obtaining

$$S_{\delta v}(f) = \frac{S_h(f)\tau^2}{1 + \omega^2\tau^2}. \tag{46}$$

If one assumes that $S_h(\omega) = S_h(0)$ [white noise corresponding to totally uncorrelated $h(t)$], then since

$$\frac{kT}{m} = \langle \delta v^2 \rangle = \int_0^\infty df S_{\delta v}(f) = S_h(0)\tau/4, \tag{47}$$

one finds that $S_h(\omega) = 4kT/m\tau$. It follows then from (46) that

$$S_{\delta v}(f) = \frac{4kT\tau/m}{1 + \omega^2\tau^2} = \frac{4kT\mu/e}{1 + \omega^2\tau^2} \tag{48a}$$

$$= \frac{4\int dv n^o(v)v^2\tau}{1 + \omega^2\tau^2}. \tag{48b}$$

This result agrees with (44) for $\omega\tau \ll 1$. If, however, we are interested in ω, for which $\omega\tau \approx 1$, then this result differs markedly from (40), which we write as

$$S_{\delta v}(f) = 4\int dv \frac{n^o(v)v^2\tau(v)}{1 + \omega^2\tau^2(v)}, \tag{49}$$

unless $\tau(v) = \tau$, a constant. This, however, is seldom the case in realistic scattering problems. We may rescue the Langevin approach if, in place of (45), we pass to the frequency domain to write

$$i\omega\delta v = -\delta v/\tau_\omega + h_\omega, \tag{50}$$

in which the effective damping $1/\tau_\omega$ is frequency dependent; and now

$$S_{\delta v}(f) = \frac{S_h(f)\tau_\omega^2}{1 + \omega^2\tau_\omega^2}. \tag{51}$$

Again, we choose $S_h(f) = S_h(0)$ and note that as $f \to 0$ we must obtain (48). Thus, $S_h(f) = 4kT/m\tau$, where $\tau = \tau(\omega = 0)$. Inserting into (51) we obtain

$$S_{\delta v}(f) = \frac{4kT/m\tau}{\omega^2 + 1/\tau_\omega^2} = \frac{4\int dv n^o(v)v^2/\tau}{\omega^2 + 1/\tau_\omega^2}. \tag{52}$$

If (52) is equated to (49), one can solve for $1/\tau_\omega$, that is, for the appropriate damping term to be used in the Langevin equation. This illustrates a major defect of the Langevin[3] approach, quite apart from inserting h_ω in a rather arbitrary manner. If the appropriate τ_ω is not known *a priori*, one must return to a more fundamental approach such as that given above. This is important, for if $\dot{x}_\omega = \mu_\omega E_\omega$, then

$$\mu_\omega = (e/m)(i\omega + 1/\tau_\omega')^{-1} = (e/kT)\int dv v^2 n(v)[i\omega + 1/\tau(v)]^{-1}.$$

Setting $\tau_\omega = \tau_\omega'$ does *not* make (52) equal (49). Thus, the original Langevin idea is not internally consistent in general.

In developing the impedance-field[4] method, Shockley et al. focus attention on the trajectory of a given carrier and on the carrier's deviations from its noiseless trajectory. Nonetheless, when it comes to writing an expression for the autocorrelation function of the velocity, they must postulate an expression equivalent to (38). [See eq. (48) of Ref. (4).] In more complicated problems, it seems best to have a fundamental approach from which such microscopic correlation functions can be derived. For example, were it necessary to include correlation between certain pairs of states, several characteristic decay times would be present in the correlation expression. These can be included in a natural way if we derive $\delta \mathbf{v}$ from eq. (37), as we have done here. Using less fundamental approaches, one must often rely on analogy and intuition.

IV. RECOMBINATION-GENERATION NOISE

Up to this point, we have been concerned with deriving the statistical distributions of microscopic variables. From a knowledge of the dependence of the macroscopic quantities of interest on these microscopic variables, the statistical distributions of the former can be calculated from those of the latter. In some cases, e.g., thermal noise in resistors, the macroscopic current or voltage spectral densities can be obtained very simply using thermodynamic arguments without recourse to microscopic methods. We now turn to an example, that of recombination noise, for which the power of treating correlations among fluctuations using the microscopic method developed in this paper becomes apparent.

Let us consider the following model of recombination. Let n be the number of mobile electrons, p the number of holes, n_{to} the number of unoccupied traps, n_{tp} the number of traps containing a trapped hole, and n_{tn} the number of traps containing trapped electron, all within a unit volume. Recombination occurs when a trap containing an electron captures a hole or when a trap containing a hole captures an electron. In Fig. 2, we present schematically the eight different processes which enter into this recombination-generation model. In Fig. 2, as in the equations for the dynamics of this system, the r's represent trapping rates, e.g., r_{ton} the rate at which electrons (n) are trapped by empty traps (t_o) per electron per trap, and the g's represent release rates from the traps per trap. As we shall see, the key to determining the noise correctly within this model is to introduce source terms for each of these eight processes.

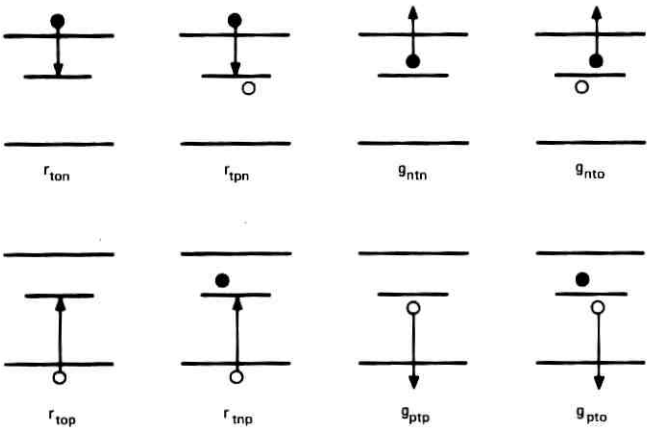

Fig. 2—The eight trapping processes contributing to recombination.

Even before we consider this problem quantitatively, we realize that correlations among fluctuations in the various particle densities will be complicated. For example, a positive fluctuation δn in the number of mobile electrons will lead to positive fluctuations in the recombination directly via increased trapping in the t_p traps and indirectly via increased trapping in the t_o traps, and the increased number of t_o traps. Using our microscopic approach we can deal with these correlations in a routine manner.

The step-by-step application of our method to this problem is carried out in some detail in Appendix D. If we make use of the definitions $\delta n \equiv \delta n_1$, $\delta p \equiv \delta n_2$, $\delta n_{tn} = \delta n_3$, $\delta n_{tp} = \delta n_4$, $\delta \mathbf{n} = (\delta n_1, \delta n_2, \delta n_3, \delta n_4)$, then the cross-spectral density of δn_i, δn_j is given by

$$S_{ij}^{\delta \mathbf{n}}(f) = \sum_{kl} M_{ik}(\omega) S_{kl}^{\mathbf{d}}(f) M_{lj}^{*}(\omega), \tag{53}$$

according to (124a). In (53), $\mathbf{S}^{\mathbf{d}}$ is the (4×4) cross-spectral density of the driving sources given by (124b) and $\mathbf{M}$ is the (4×4) matrix expressing the induced fluctuations $\delta \mathbf{n}$ in terms of the spontaneous ones $\mathbf{d}$:

$$\delta \mathbf{n}(\omega) = \mathbf{M}(\omega) \mathbf{d}(\omega), \tag{54}$$

according to (123). [Note from Appendix D that the vector $\mathbf{d} \equiv (d_1, d_2, d_3, d_4)$ consists of four linear combinations of the eight fundamental noise driving terms that arise using this method. That these terms are mutually correlated can be seen from (124b).] Thus, from (53) we can calculate the spectral densities of interest. These include S_{11} for the

spectral density of δn, S_{22} for that of δp, and S_{12} and S_{21} for the cross-spectral densities between δn and δp, all of which contribute to the current fluctuations observed at the contacts of the device. Even more components of $\mathbf{S}^{\delta n}$ are needed to calculate fluctuations in recombination radiation, if such results are desired.

We should stress that (53) contains all correlation effects (through $\mathbf{S}^d$) and all relaxation effects (through $\mathbf{M}$ and $\mathbf{M}^*$) contained in this problem. If certain trapping or release rates are small relative to others, the contribution of such processes to $\mathbf{S}^d$ can be neglected; if $\mathbf{M}$ can be characterized by a single time constant τ, $\mathbf{MM}^*$ becomes proportional to $(\omega^2 + 1/\tau^2)^{-1}$, as is usually assumed.[8] The point to be made is that with *no* additional effort of a statistical nature we can solve device-noise problems which contain rather complicated statistical correlations. The key is to put the correlations in the dynamics of the problem of interest, thereby keeping the statistics simple.

Finally, the reader is cautioned against making a normal mode analysis of the linearized noise equations and then introducing a noise-source term for each normal mode, since such terms will not be statistically independent in general. We note in concluding that, in obtaining (53), nowhere did we have to make use of normal-mode analysis.

V. CONCLUSION

In this paper we have discussed a straightforward, microscopic approach which can be used to calculate the statistical fluctuations accompanying any transfer process. Since nearly all charge-carrier velocity fluctuations can be characterized in terms of transfer processes, we, therefore, are able to calculate the statistics of the velocity fluctuations from those of the transfer fluctuations. Knowledge of the velocity fluctuations is often all the microscopic information that is needed to insert into the Langevin method or the impedance-field method to calculate device noise. (These methods explain in great detail how to convert velocity fluctuations into observed current and voltage fluctuations.) In so doing, we are able to insure that all important correlation effects and relaxation effects are included in the results. We further simplify the statistical portion of the calculation by separating the spontaneous from the induced fluctuations, and expressing the statistics of the latter in terms of the former. In this way, we have to calculate the statistics of only the uncorrelated, spontaneous fluctuations from probability theory, while the more complicated statistics of the induced fluctuations can be obtained directly from those of the spontaneous

type. The correlation effects are thus shifted from the statistical portion of the problem to the dynamical portion. We have included (in Appendix C) a derivation of an important result for the diffusion noise of "hot" charge carriers, that is, for charge carriers whose mobilities are field dependent. As the method developed here is readily generalized to nonstationary noise, our results will be used extensively in treating noise in charge-transfer devices.

VI. ACKNOWLEDGMENT

It is a pleasure to thank J. R. Brews for helpful discussions.

APPENDIX A

In this appendix, we consider in detail a single elementary random process and calculate its statistical distribution. As we have shown in the text, many complicated random processes can be decomposed into simpler processes of the type considered here. Once the statistical distribution of these simpler processes is understood, one can determine the distribution of the more complicated process of interest. In the text, we have focused attention on autocorrelation functions, and on how to distinguish the autocorrelation function of a complicated process from those of the simpler processes of which it is composed. In Appendix B, we show how to obtain distribution functionals of complicated processes from those of simpler ones.

The random variable whose statistics we seek is

$$d(t) = \sum_i \delta(t - t_i) - R^o(t), \tag{55}$$

where $R^o(t)$ (a specific function of time) is the mean rate of occurrence of the t_i at time t. In other words, the t_i are independent random variables, each specifying the time at which an independent random event occurs. To be a meaningful construct, it is necessary that $R^o(t)$ satisfy

$$R^o(t) \gg \frac{dR^o(t)}{dt} \frac{1}{R^o(t)}, \tag{56}$$

that is, that the characteristic time in which $R^o(t)$ changes is much longer than the average time necessary for an event to occur. If (56) is satisfied, many events will almost surely occur in time intervals during which $R^o(t)$ changes only by a small amount. During these intervals, the statistics of the t_i will be Poisson with mean rate $R^o(t)$.

In probability theory, one often works with distribution functions. Thus, if a random variable x has a probability of $p_x(x_1)dx_1$ at x_1, that

x is in dx_1, then the distribution function $q_p(k)$ of $p_x(x)$ is defined by

$$q_p(k) = \int_{-\infty}^{\infty} dx e^{ikx} p_x(x). \tag{57}$$

Since

$$p_x(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{-ikx} q_p(x), \tag{58}$$

it is clear that $q_p(k)$ contains as much information about x as does $p_x(x)$. What is important is that it contains this information in a more convenient form, since, for example, all moments of x can be obtained at once from $q_p(k)$ by suitable differentiations with respect to k.

Here, we shall calculate the distribution functional[15] $Q_d(k(t))$ of $d(t)$ defined by

$$Q_d[k(t)] = \left\langle \exp\left(i \int_{t_1}^{t_2} dt k(t) d(t) \right) \right\rangle, \tag{59}$$

where the brackets denote averaging over the probability distribution of $d(t)$. Rather than use the form of $d(t)$ given in (55), let us use instead the form

$$d(t) = \sum_i f(t - t_i) - R^o(t), \tag{60}$$

where $f(t)$ is a function of t satisfying

$$\int_{-\infty}^{\infty} dt f(t) = 1. \tag{61}$$

Equation (59) then becomes

$$Q_d[k(t)] = \left\langle \exp\left[i \int_{t_1}^{t_2} dt k(t) \left(\sum_i f(t - t_i) - R^o(t) \right) \right] \right\rangle. \tag{62}$$

Let us now evaluate this average.

To determine $Q_d[k(t)]$ we note the following.

(*i*) $R^o(t)$ is a specific function of time and, hence,

$$\exp\left(-i \int_{t_1}^{t_2} dt k(t) R^o(t) \right)$$

can be factored out of the brackets.

(*ii*) The t_i are independent and, hence, the average of the exponential in $\sum_i f(t - t_i)$ can be factored into products of averages of an exponential in each $f(t - t_i)$, each of the n averages being equal to all the others.

(*iii*) The probability that n events occur in the interval $t_1 < t < t_2$ is Poisson.

From these three considerations, it follows that

$$Q_d[k(t)] = \exp\left(-i\int_{t_1}^{t_2} dtk(t)R^o(t)\right) \times \sum_{n=0}^{\infty} P_n\left[\left\langle\exp\left(i\int_{t_1}^{t_2} dtk(t)f(t-t_i)\right)\right\rangle\right]^n, \quad (63)$$

where P_n is given by

$$P_n = e^{-\Lambda}\Lambda^n/n! \quad (64)$$

and Λ is given by

$$\Lambda = \int_{t_1}^{t_2} R^o(t)dt. \quad (65)$$

To evaluate the average in (63), we note that the probability that t_i occurs in dt' at t' is $R^o(t')dt'/\Lambda$. Thus,

$$\left\langle\exp\left(i\int_{t_1}^{t_2} dtk(t)f(t-t_i)\right)\right\rangle = \frac{1}{\Lambda}\int_{t_1}^{t_2} dt_i \exp\left(i\int_{t_1}^{t_2} dtk(t)f(t-t_i)\right)R^o(t_i). \quad (66)$$

Inserting this into (63) and carrying out the sum on n, one obtains finally for $Q_d[k(t)]$

$$Q_d[k(t)] = \exp\left\{\int_{t_1}^{t_2} dt'R^o(t')\left[\exp\left(i\int_{t_1}^{t_2} dtk(t)f(t-t')\right)-1\right]\right\} \times \exp\left(-i\int_{t_1}^{t_2} k(t)R^o(t)dt\right). \quad (67)$$

We note from (59) that if $k(t) = 0$, then Q_d should equal 1. Since (67) satisfies this condition, we are assured that $Q_d[k(t)]$ is properly normalized. As we shall see in Appendix B, from (67) we can calculate the distribution functional of nearly any function of $d(t)$.

It often happens that the particular process of interest involves numerous events, each of which contributes only a very small portion to the total fluctuation. In this case, we can expand the exponential in $f(t-t')$ to quadratic order, obtaining for $Q_d[k(t)]$

$$Q_d[k(t)] = \exp\left[-\frac{1}{2}\int_{t_1}^{t_2} dt'R^o(t')\left(\int_{t_1}^{t_2} dtf(t-t')k(t)\right)^2\right], \quad (68)$$

where we have assumed that $f(t)$ is of sufficiently short duration that

$$\int_{t_1}^{t_2} dt'R^o(t')f(t-t') = R^o(t). \quad (69)$$

Recalling finally that $f(t - t') = \delta(t - t')$ for the $d(t)$ of (55), it follows that

$$Q_d[k(t)] = \exp\left(-\frac{1}{2}\int_{t_1}^{t_2} dtk(t)k(t)R^o(t)\right). \tag{70}$$

In the text, we made repeated use of the autocorrelation function of d, $\langle d(t_1)d(t_2)\rangle$. This we may calculate from (67) with $f(t) = \delta(t)$, or from (70), the results being the same. To do this, we note that from (59)

$$\langle d(t_1)d(t_2)\rangle = -\frac{\partial}{\partial\gamma_1}\bigg|_{\gamma_1=0}\frac{\partial}{\partial\gamma_2}\bigg|_{\gamma_2=0} Q_d[\gamma_1\delta(t - t_1) + \gamma_2\delta(t - t_2)]. \tag{71}$$

Thus, from (70) we readily find that

$$\langle d(t_1)d(t_2)\rangle = R^o(t_1)\delta(t_1 - t_2). \tag{72}$$

Although (67) is our most general expression for the statistics of $d(t)$, (72) is usually all that is needed in noise calculations for devices. In simplest terms, it is the autocorrelation of the (shot) noise associated with the process of which $d(t)$ represents the source term and $R^o(t)$ the average rate at time t. It is the building block from which most device noise can be constructed. For stationary processes, $R^o(t_1) = R^o$, a constant depending on the noiseless solution. The spectral density of d under such conditions is a useful concept and is given by

$$S_d(f) = 2R^o. \tag{73}$$

APPENDIX B

In this appendix, we show how to obtain the distribution functional of a statistical process composed of a number of simple independent processes of the type discussed in Appendix A. Suppose that the process of interest, $\delta n(t) \equiv n(t) - n^o(t)$, is a linear functional of a number of $d(t)$, $\{d_i(t), 1 \leqq i \leqq m\}$. Then

$$\delta n(t) = \mathfrak{F}_n[d_1(t), \cdots, d_m(t), t]. \tag{74}$$

Since δn is linear in the d_i, we can write $\mathfrak{F}_n$ in the form

$$\mathfrak{F}_n = \sum_{i=1}^{m} \mathfrak{F}_{ni}[d_i(t), t]. \tag{75}$$

The distribution functional $Q_n[k(t)]$ of $\delta n(t)$ is then given by

$$Q_n[k(t)] = \left\langle \exp\left(i \int_{t_1}^{t_2} k(t)\mathfrak{F}_n dt \right) \right\rangle \tag{76}$$

$$= \prod_{i=1}^{m} \left\langle \exp\left(i \int_{t_1}^{t_2} k(t)\mathfrak{F}_{ni}[d_i(t), t]dt \right) \right\rangle \tag{77}$$

$$\equiv \prod_{i=1}^{m} Q_{ni}[k(t)], \tag{78}$$

where (77) follows from the independence of the d_i.

A typical functional dependence of $\delta n(t)$ on $d_i(t)$ is

$$\delta n(t) = \sum_{i=1}^{m} a_i \int_{t_1}^{t} dt' \exp\left(- \int_{t'}^{t} dt''/\tau_i(t'') \right) d_i(t'). \tag{79}$$

It follows from (78) that $Q_{ni}[k(t)]$ is given by

$$Q_{ni}[k(t)]$$

$$= \left\langle \exp\left[i \int_{t_1}^{t_2} dt k(t) \int_{t_1}^{t} dt' a_i d_i(t') \exp\left(- \int_{t'}^{t} dt''/\tau_i(t'') \right) \right] \right\rangle \tag{80}$$

$$= \left\langle \exp\left[i \int_{t_1}^{t_2} dt a_i d_i(t) \int_{t}^{t_2} dt' k(t') \exp\left(- \int_{t}^{t'} dt''/\tau_i(t'') \right) \right] \right\rangle \tag{81}$$

$$= Q_{di}\left[a_i \int_{t}^{t_2} dt' k(t') \exp\left(- \int_{t}^{t'} dt''/\tau_i(t'') \right] \right). \tag{82}$$

If we use (70) for Q_{di}, then we find that

$$Q_{ni}[k(t)] = \exp\left[- \frac{1}{2} \int_{t_1}^{t_2} dt \int_{t_1}^{t_2} dt' k(t)k(t')F_i(t, t') \right], \tag{83}$$

where

$$F_i(t, t') = a_i^2 \exp\left[- \int_{\tau_{\min}}^{\tau_{\max}} dt''/\tau_i(t'') \right] \times \int_{t_1}^{\tau_{\min}} d\tau R_i^o(\tau) \exp\left[-2 \int_{\tau}^{\tau_{\min}} dt''/\tau_i(t'') \right] \tag{84}$$

and where $\tau_{\max} = \max(t, t')$, $\tau_{\min} = \min(t, t')$. Inserting (83) into (78) determines Q_n as desired.

From Q_n we can determine the autocorrelation function of $n(t)$. First, we note from (74) and (76) that

$$\langle \delta n(t_1)\delta n(t_2) \rangle = - \frac{\partial}{\partial \gamma_1}\bigg|_{\gamma_1=0} \frac{\partial}{\partial \gamma_2}\bigg|_{\gamma_2=0} Q[\gamma_1\delta(t - t_1) + \gamma_2\delta(t - t_2)]. \tag{85}$$

Using (78) and (83) this becomes

$$\langle \delta n(t_1)\delta n(t_2)\rangle = \sum_i F_i(t_1, t_2), \tag{86}$$

where F_i is given in (84).

If τ_i and R_i^o are independent of t, an interesting result is obtained. By performing the integrations in (84), it follows at once that

$$F_i(t_1, t_2)/a_i^2 = (R_i^o\tau_i/2) \exp(-|t_2 - t_1|/\tau_i), \tag{87}$$

where the correlation of the fluctuations contributing to δn is readily apparent. Expression (87) occurs often in the theory of stationary processes. [Note in (87) that in the limit $\tau_i \to 0$, $F_i''(t_1, t_2)/a_i^2$ approach $R_i^o\delta(t_2 - t_1)$ as expected from (72).]

APPENDIX C

Fluctuation-Dissipation Theorem

Relations between different physical phenomena have always attracted considerable interest, and justifiably so. No exception is the Einstein relation between the mobility μ and the diffusion coefficient D:

$$\mu = \frac{qD}{kT}. \tag{88}$$

In this expression, μ relates the carrier's velocity v to the electric field E acting on the carrier according to

$$v = \mu E, \tag{89}$$

and D relates the mean-square distance which the carrier diffuses in equilibrium to the time t in which it has been diffusing according to

$$\langle (x_t - x_o)^2\rangle = 2Dt. \tag{90}$$

Although usually not emphasized, all derivations of (88) are based on small fluctuations from equilibrium (in the absence of driving forces). In the presence of driving forces, one must be careful to use the appropriate small-signal quantities when relating diffusion to transport. One of the purposes of this appendix is to explain how this can be done in some cases.

It is very important at the outset to understand the physical origin of the Einstein relation, as well as that of the more general fluctuation-dissipation theorem[16–18] (FD theorem). This is especially important for noise theory since it often enables one to express fluctuation properties in terms of transport properties. This is a valuable aid since transport properties have already been carefully studied in attempts to

understand the noiseless behavior of the device. In addition, an understanding of how the microscopic noiseless motion of a single carrier contributes to the noiseless device current can be carried over at once to calculating how the microscopic fluctuations contribute to the device noise. Thus, we ask the question, Why are fluctuations and dissipation so closely related?

What is the physical origin of the fluctuations undergone by charge carriers? The answer clearly lies in the scattering of all sorts that such carriers experience within the material. What is the physical origin of the damping force experienced by charge carriers? The answer clearly is the same. This connection may be phrased in several ways. A fluctuation corresponds to the response of the carrier to a random force. The velocity-field dependence is a similar relation of response to applied force. Alternatively, under steady-state conditions, the average gain in the energy of a carrier from the material due to fluctuations must be dynamically balanced by the loss of energy due to damping. The gain and loss are, therefore, closely linked. Still another way of seeing the connection is to note that a fluctuation is a departure from equilibrium, such as when a small, disturbing probe force is applied to the system. However it is viewed, a close connection between fluctuation and dissipation must clearly exist, which we shall derive below.

In the impedance-field[4] method, we arrive at the following equation for the voltage spectral density of the device in terms of the elementary thermal-velocity fluctuations of the carriers:

$$S(\delta V_N, f) = \sum_{i=1}^{3} \int d\mathbf{x} \, |\nabla_i Z_{N\mathbf{x}}|^2 4q^2 n(\mathbf{x}) D_{\mathbf{x}}(\delta v_i, f), \tag{91}$$

where q is the elementary carrier charge, $n(\mathbf{x})$ is the carrier density, $Z_{N\mathbf{x}}$ is the impedance field between the contact at N and the field point at $\mathbf{x}$, and

$$S_v(\mathbf{x}, f) = 4D_{\mathbf{x}}(\delta v_i, f) = 4\mathrm{Re} \int_0^\infty e^{i\omega t} \langle \delta v_i(t) \delta v_i(0) \rangle dt \tag{92}$$

is the velocity spectral density in the ith ($=\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$) direction. [See eqs. (54), (56), (27), (28), and (35) of Ref. 4.] The quantity $D_{\mathbf{x}}(\delta v_i, f)$ is also referred to as the diffusion of δv_i at frequency f in the region of $\mathbf{x}$ for a single carrier. In discussing the impedance-field method for the case of $v = \mu E$, where μ is a constant, it is found that

$$D_{\mathbf{x}}(\delta v_x, f) = \left(\frac{kT}{q} \right) \mu \tag{93}$$

for f much less than the scattering frequency. In what follows, we use the FD theorem to prove that if $\mu = \mu(E) = v(E)/E$ and if an effective temperature T can be defined, then[11]

$$D_x(\delta v_x, f) = \left(\frac{kT}{q}\right)\frac{d[\mu(E)E]}{dE}. \tag{94}$$

[Intuitively, (94) follows from (93) if we recall that fluctuations in velocity result from spontaneous fluctuations in the electric field, and if $v = v(E)$, then $\delta v = [dv(E)/dE]\cdot\delta E = (d[\mu(E)E]/dE)\cdot\delta E.$] In the Langevin[3] method, diffusion noise enters through a spectral density of the form

$$S_h(\mathbf{x}, \mathbf{x}', f) = 4q^2 n(\mathbf{x}) D_n \delta(\mathbf{x} - \mathbf{x}'), \tag{95}$$

where again D_n is the diffusion constant, as defined in (42) for frequencies much less than the scattering frequency. Noting how (42) is related to $\langle \delta v(t_1)\delta v(t_2)\rangle$, we note that, once we have shown that (94) is true, it will follow at once that

$$S_h(\mathbf{x}, \mathbf{x}', f) = 4q^2 n(\mathbf{x})\frac{kT}{q}\frac{d(\mu E)}{dE}\delta(\mathbf{x} - \mathbf{x}'). \tag{96}$$

The first quantum-mechanical derivation of the FD theorem is usually attributed to Callen and Welton.[16] This subject has subsequently been treated in greater detail in Refs. 19 and 20. For the sake of completeness, we shall rederive the FD theorem here primarily for the purpose of calling attention to its application to cases of steady-state but nonequilibrium conditions. We conclude this appendix with the derivation of more general relations for fluctuations and dissipation valid for any energy distribution (especially nonthermal) of the states of the system. These relations show how closely the two are related even under nonthermal conditions.

Let us suppose that we have some system which can be described by a Hamiltonian H^s, which includes applied electric and magnetic fields giving rise to currents, etc. Let $\hat{r}_i$ be an operator whose expectation value we seek as a function of time in response to a unit impulse in a probe force f_j, which enters the total Hamiltonian H according to

$$H = H^s - f_j\hat{r}_j, \tag{97}$$

where $\hat{r}_j$ is another operator. Then the FD theorem states that

$$\text{Re}\,[K_{ij}(\omega)] = \text{Im}\,[X_{ij}(\omega)]\frac{\hbar}{2}\frac{1 + \exp(-\beta\hbar\omega)}{1 - \exp(-\beta\hbar\omega)}, \tag{98a}$$

where $\beta = 1/kT$, T is the effective temperature of the system, if such can be defined,

$$K_{ij}(\omega) \equiv \int_0^\infty dt k_{ij}(t) e^{i\omega t}, \tag{98b}$$

$$k_{ij}(t) \equiv \tfrac{1}{2}\langle[\hat{r}_i(t), \hat{r}_j(0)]_+\rangle = \tfrac{1}{2}\langle \hat{r}_i(t)\hat{r}_j(0) + \hat{r}_j(0)\hat{r}_i(t)\rangle, \tag{98c}$$

$$X_{ij}(\omega) \equiv \int_{-\infty}^\infty dt x_{ij}(t) e^{i\omega t}, \tag{98d}$$

and $x_{ij}(t)$ is the impulse response of $\hat{r}_i$ to the unit impulse f_j in (97). From (98b, c) it is clear that 4 Re $[K_{ij}(\omega)] = S_{ij}(f)$, the cross-spectral density between r_i and r_j, whereas from (98d) it follows that Im $[X_{ij}(\omega)]$ is the absorption through r_i of the energy put into the system via the coupling (interaction energy) $-f_j r_j$.

Returning for the moment to (92), we note that if we let $\hat{r}_i = \hat{r}_j = \delta v_i = \delta\dot{x}$ (for $i = x$), in the classical limit ($\hbar \to 0$), we obtain from (98a)

$$S_v(\mathbf{x}, f) = 4 \text{ Im } [X_{\dot{x}\dot{x}}(\omega)]kT/\omega. \tag{99}$$

After we derive (98a), we shall show that for ω much less than the scattering frequency

$$X_{\dot{x}\dot{x}}(\omega) = i\omega\mu_{ac}/q, \tag{100}$$

where $\mu_{ac} = d[\mu(E)E]/dE$. The important result (94) then follows at once.

We shall derive (98a) by evaluating $K_{ij}(\omega)$ and $X_{ij}(\omega)$ and comparing the results. By definition

$$x_{ij}(t) = \frac{i}{\hbar}\langle \hat{r}_i(t)\hat{r}_j(0) - \hat{r}_j(0)\hat{r}_i(t)\rangle, \tag{101}$$

where we assume the impulse occurs at $t = 0$. Clearly $x_{ij}(t) = 0, t < 0$. Thus,

$$X_{ij}(\omega) = \frac{i}{\hbar}\int_0^\infty dt e^{i\omega t} \sum_{lm} \langle l|\hat{r}_i|m\rangle\langle m|\hat{r}_j|l\rangle(e^{-\beta E_l} - e^{-\beta E_m}) \times e^{-i/\hbar(E_m - E_l)t}/\sum_l e^{-\beta E_l}, \tag{102}$$

which follows if we assume that the system described by H^s can be characterized by a temperature T. This insures that the probability that the system is in eigenstate ψ_n with eigenenergy E_n is exp $(-\beta E_n)/N$, where N is the normalization factor used in (102).

Performing the integration on t, we obtain

$$X_{ij}(\omega) = \sum_{lm} (e^{-\beta E_l} - e^{-\beta E_m})\langle l|\hat{r}_i|m\rangle\langle m|\hat{r}_j|l\rangle \times (-1)[\hbar\omega - (E_m - E_l) + i\epsilon]^{-1}/N, \quad (103)$$

and, finally, taking the imaginary part we find that

$$\text{Im}\,[X_{ij}(\omega)] = \pi \sum_{lm} (e^{-\beta E_l} - e^{-\beta E_m})\langle l|\hat{r}_i|m\rangle\langle m|\hat{r}_j|l\rangle\delta[\hbar\omega - (E_m - E_l)]/N. \quad (104)$$

If now we compare the definition of $k_{ij}(t)$ (98c) with (101), it follows at once from making the appropriate changes in (103) that

$$K_{ij}(\omega) = \frac{\hbar}{2i} \sum_{lm} (e^{-\beta E_l} + e^{-\beta E_m})\langle l|\hat{r}_i|m\rangle\langle m|\hat{r}_j|l\rangle \times (-1)[\hbar\omega - (E_m - E_l) + i\epsilon]^{-1}/N. \quad (105)$$

And upon taking the real part of (105),

$$\text{Re}\,[K_{ij}(\omega)] = \pi \frac{\hbar}{2} \sum_{lm} (e^{-\beta E_l} + e^{-\beta E_m}) \times \langle l|\hat{r}_i|m\rangle\langle m|\hat{r}_j|l\rangle\delta[\hbar\omega - (E_m - E_l)]/N \quad (106)$$

is obtained. In (104) and (106), we may replace E_m in the exponent by $(E_l + \hbar\omega)$ owing to the presence of the delta function. This then permits the factor $[1 - \exp(-\beta\hbar\omega)]$ to be pulled out of the sum in (104) and $[1 + \exp(-\beta\hbar\omega)]$ to be pulled out in (106). The FD theorem (98a) follows at once.

To derive (100), we must look more closely at the exact response x_{ij}^e of $\hat{r}_i$ to a general time dependent f_j in $H = H_s - f_j\hat{r}_j$. If H is the Hamiltonian, then

$$x_{ij}^e(t) = \langle \hat{r}_i \rangle_t = \text{Tr}\,[\hat{r}_i\rho(t)]/\text{Tr}\,[\rho(t)], \quad (107)$$

where $\rho(t)$ is the density matrix of the system at time t. The density matrix $\rho(t)$ at t can be obtained from that at t_1 according to

$$\rho(t) = P(t, t_1)\rho(t_1)P^\dagger(t, t_1), \quad (108)$$

where the propagator $P(t, t_1)$ is defined by

$$P(t, t_1) = \exp\left(-\frac{i}{\hbar}\int_{t_1}^{t} d\tau H_\tau\right), \quad (109)$$

and where we are using the Feynman[21] ordered-operator notation. [If

t_1 is prior to the turning on of f and the system can be described by an effective temperature T, then $\rho(t_1) = \exp(-\beta H^s)$.]

The specific problem in which we are most interested is the case where $\hat{r}_i$, $\hat{r}_j$ represent velocity fluctuations from the expectation value of the velocity $\mathbf{v}_o$ of a carrier under the influence of an electric field of arbitrary strength contained in H^s. Let us place ourselves in a reference frame drifting with velocity $\mathbf{v}_o$. In such a frame, we may calculate either the velocity fluctuations $k_{\dot{x}\dot{x}}(t)$ from the noiseless motion or the response $\dot{x}_{ij}(t)$ to a probe force f_j. If we were applying a small, ac electric field $e(t)$ to the carrier, then $H = H^s - qe(t)\hat{x}$, where $\hat{x}$ is the position operator corresponding to the x coordinate (in the drifting frame). Also, we know that the velocity response to such a field is given by

$$\delta v(t) = [dv(E)/dE]e(t) \tag{110}$$

or, taking Fourier transforms,

$$\delta v(\omega) = \{d[\mu(E)E]/dE\}e(\omega), \tag{111}$$

since $v(E) = \mu(E)E$, μ being the mobility. However, $X_{\dot{x}\dot{x}}(\omega)$ arises from the velocity response to $H = H^s - f(t)\hat{\dot{x}}$. We can relate $f(t)$ to $e(t)$ if we note that from (109), $f(t)$ enters in the form

$$\exp\left(-\frac{i}{\hbar}\int_{t_1}^{t} d\tau f(\tau)\dot{x}\right) = \exp\left(-\frac{i}{\hbar}\int_{t_1}^{t} d\tau(-)\dot{f}(\tau)x\right), \tag{112}$$

where we can take $f(t) = 0$ for simplicity. Thus, $qe(t) = -\dot{f}(t)$, or $qe(\omega) = i\omega f(\omega)$. Also, using (111), we obtain

$$\begin{aligned} X_{\dot{x}\dot{x}}(\omega) &\equiv \frac{\delta v(\omega)}{f(\omega)} = \frac{\delta v(\omega)}{e(\omega)}\frac{i\omega}{q} = \frac{d[\mu(E)E]}{dE}\frac{i\omega}{q} \\ &= i\omega\mu_{ac}/q, \end{aligned} \tag{113}$$

which proves (86). Our principal result (94) is therefore demonstrated.

We promised that we would conclude with more general relations for fluctuations and dissipation valid for any probability distribution of the eigenstates of a system according to eigenenergies. We proceed as follows.

Let the (normalized) probability that the system be in a state of eigenenergy E be $f(E)$. Choose the scale of energy such that the lowest eigenenergy is zero (0). Take the Laplace transform of $f(E)$ to obtain

$$F(s) = \int_0^\infty dE f(E)e^{-sE} \tag{114a}$$

and, of course,

$$f(E) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} ds e^{sE} F(s). \tag{114b}$$

[For a thermal distribution, $f(E) = \exp(-\beta E)/N$ and $F(s) = (s + \beta)^{-1}/N$.] It follows from (104) and (106) that

$$\frac{2}{\hbar} \operatorname{Re}[K_{ij}(\omega)] = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} ds G(s, \omega)(1 + e^{s\hbar\omega}) \tag{115a}$$

and

$$\operatorname{Im}[X_{ij}(\omega)] = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} ds G(s, \omega)(1 - e^{s\hbar\omega}), \tag{115b}$$

where

$$G(s, \omega) \equiv \pi F(s) \sum_{lm} e^{sE_l} \langle l | \hat{r}_i | m \rangle \langle m | \hat{r}_j | l \rangle \delta[\hbar\omega - (E_m - E_l)]. \tag{115c}$$

The difference between (115a) and (115b) consists of only one sign: plus in (115a), minus in (115b). For the special case of a thermal distribution, this sign difference yields the ratio factor that appears in (98a). In general, while such a simple relation between K_{ij} and X_{ij} is no longer valid, it is clear from (115) that fluctuations and dissipation have a common origin. These are the more general relations we promised.

APPENDIX D

The purpose of this appendix is merely to carry out the routine mathematical steps necessary to arrive at the spectral densities of recombination-generation noise. These results are referred to in Section IV.

We proceed as follows. We can abbreviate our exposition since we have been through the necessary steps several times in the previous sections. The steps are as follows:

(*i*) Write the microscopic kinematic equations:

$$\frac{dn}{dt} = -\sum_m \delta[t - t_{ton}(m)] - \sum_m \delta[t - t_{tpn}(m)] + \sum_m \delta[t - t_{ntn}(m)] + \sum_m \delta[t - t_{nto}(m)], \tag{116a}$$

$$\frac{dp}{dt} = -\sum_m \delta[t - t_{top}(m)] - \sum_m \delta[t - t_{tnp}(m)] + \sum_m \delta[t - t_{ptp}(m)] + \sum_m \delta[t - t_{pto}(m)], \tag{116b}$$

$$\frac{dn_{tn}}{dt} = + \sum_m \delta[t - t_{ton}(m)] - \sum_m \delta[t - t_{ntn}(m)] - \sum_m \delta[t - t_{tnp}(m)] + \sum_m \delta[t - t_{pto}(m)], \quad (116c)$$

$$\frac{dn_{tp}}{dt} = - \sum_m \delta[t - t_{tpn}(m)] + \sum_m \delta[t - t_{nto}(m)] + \sum_m \delta[t - t_{top}(m)] - \sum_m \delta[t - t_{ptp}(m)], \quad (116d)$$

$$\frac{dn_{to}}{dt} = - \frac{dn_{tn}}{dt} - \frac{dn_{tp}}{dt}. \quad (116e)$$

(*ii*) Rewrite in terms of sources and responses:

$$\frac{dn}{dt} = -r_{ton}n_{to}n - r_{tpn}n_{tp}n + g_{ntn}n_{tn} + g_{nto}n_{to} - d_{ton} - d_{tpn} + d_{ntn} + d_{nto}, \quad (117a)$$

$$\frac{dp}{dt} = -r_{top}n_{to}p - r_{tnp}n_{tn}p + g_{ptp}n_{tp} + g_{pto}n_{to} - d_{top} - d_{tnp} + d_{ptp} + d_{pto}, \quad (117b)$$

$$\frac{dn_{tn}}{dt} = r_{ton}n_{to}n - g_{ntn}n_{tn} - r_{tnp}n_{tn}p + g_{pto}n_{to} + d_{ton} - d_{ntn} - d_{tnp} + d_{pto}, \quad (117c)$$

$$\frac{dn_{tp}}{dt} = - r_{tpn}n_{tp}n + g_{nto}n_{to} + r_{top}n_{to}p - g_{ptp}n_{tp} - d_{tpn} + d_{nto} + d_{top} - d_{ptp}, \quad (117d)$$

$$\frac{dn_{to}}{dt} = - \frac{dn_{tn}}{dt} - \frac{dn_{tp}}{dt}, \quad (117e)$$

where a typical d function is defined as

$$d_{ton} \equiv \sum_m \delta(t - t_{ton}) - r_{ton}n_{to}n \quad (118a)$$

and is equivalent for small fluctuations to

$$d_{ton} = \sum_m \delta(t - t^o_{ton}) - r^o_{ton}n^o_{to}n^o. \quad (118b)$$

(*iii*) Write each variable as the sum of a noiseless contribution and a noise contribution, and linearize (117) to obtain the following nonlinear equations for the noiseless solution:

$$\dot{n}^o = -r_{ton}n^o_{to}n^o - r_{tpn}n^o_{tp}n^o + g_{ntn}n^o_{tn} + g_{nto}n^o_{to}, \quad (119a)$$

$$\dot{p}^o = -r_{top}n^o_{to}p^o - r_{tnp}n^o_{tn}p^o + g_{ptp}n^o_{tp} + g_{pto}n^o_{to}, \quad (119b)$$

$$\dot{n}^o_{tn} = r_{ton}n^o_{to}n^o - g_{ntn}n^o_{tn} - r_{tnp}n^o_{tn}p^o + g_{pto}n^o_{to}, \tag{119c}$$

$$\dot{n}^o_{tp} = -r_{tpn}n^o_{tp}n^o + g_{nto}n^o_{to} + r_{top}n^o_{to}p^o - g_{ptp}n^o_{tp}, \tag{119d}$$

$$\dot{n}^o_{to} = -\dot{n}^o_{tn} - \dot{n}^o_{tp}, \tag{119e}$$

and to obtain the following linear equations for the noise solution:

$$\begin{aligned}\delta\dot{n} = {} & -r_{ton}(n^o_{to}\delta n + \delta n_{to}n^o) - r_{tpn}(n^o_{tp}\delta n + \delta n_{tp}n^o) \\ & + g_{ntn}\delta n_{tn} + g_{nto}\delta n_{to} - d_{ton} - d_{tpn} + d_{ntn} + d_{nto},\end{aligned} \tag{120a}$$

$$\begin{aligned}\delta\dot{p} = {} & -r_{top}(n^o_{to}\delta p + \delta n_{to}p^o) - r_{tnp}(n^o_{tn}\delta p + \delta n_{tn}p^o) \\ & + g_{ptp}\delta n_{tp} + g_{pto}\delta n_{to} - d_{top} - d_{tnp} + d_{ptp} + d_{pto},\end{aligned} \tag{120b}$$

$$\begin{aligned}\delta\dot{n}_{tn} = {} & r_{ton}(n^o_{to}\delta n + \delta n_{to}n^o) - g_{ntn}\delta n_{tn} - r_{tnp}(n^o_{tn}\delta p + \delta n_{tn}p^o) \\ & + g_{pto}\delta n_{to} + d_{ton} - d_{ntn} - d_{tnp} + d_{pto},\end{aligned} \tag{120c}$$

$$\begin{aligned}\delta\dot{n}_{tp} = {} & -r_{tpn}(n^o_{tp}\delta n + \delta n_{tp}n^o) + g_{nto}\delta n_{to} + r_{top}(n^o_{to}\delta p + \delta n_{to}p^o) \\ & - g_{ptp}\delta n_{tp} - d_{tpn} + d_{nto} + d_{top} - d_{ptp},\end{aligned} \tag{120d}$$

$$\delta n_{to} = -\delta n_{tn} - \delta n_{tp}. \tag{120e}$$

(*iv*) Finally, solve (120) for δn, δp, δn_{tn}, δn_{tp}, and δn_{to} in terms of d_{ton}, $\cdots$. This latter step involves solving five equations for five unknowns. If we are interested in the stationary solution to (119) ($\dot{n}^o = \dot{p}^o = \dot{n}^o_{tn} = \dot{n}^o_{tp} = 0$), then the coefficients of the δn, δp, $\cdots$ (120) are independent of time and a Fourier analysis of (120) is most expedient. Using (120e) to eliminate δn_{to}, one obtains the following set of equations:

$$i\omega\delta n = -R_n\delta n + R_{ntn}\delta n_{tn} - R_{ntp}\delta n_{tp} + d_1, \tag{121a}$$

$$i\omega\delta p = -R_p\delta p - R_{ptn}\delta n_{tn} + R_{ptp}\delta n_{tp} + d_2, \tag{121b}$$

$$i\omega\delta n_{tn} = -R_{tn}\delta n_{tn} + R_{tnn}\delta n - R_{tnp}\delta p - R_{tntp}\delta n_{tp} + d_3, \tag{121c}$$

$$i\omega\delta n_{tp} = -R_{tp}\delta n_{tp} - R_{tpn}\delta n + R_{tpp}\delta p - R_{tptn}\delta n_{tn} + d_4, \tag{121d}$$

where

$$d_1 \equiv -d_{ton} - d_{tpn} + d_{ntn} + d_{nto}, \tag{122a}$$

$$d_2 \equiv -d_{top} - d_{tnp} + d_{ptp} + d_{pto}, \tag{122b}$$

$$d_3 \equiv d_{ton} - d_{ntn} - d_{tnp} + d_{pto}, \tag{122c}$$

and

$$d_4 \equiv -d_{tnp} + d_{nto} + d_{top} - d_{ptp}. \tag{122d}$$

By combining the coefficients of the δn, δp, $\cdots$ in (120), the reader can determine the R_n, R_p, $\cdots$ rates in (121). For example, $R_n = r_{ton}n^o_{to} + r_{tpn}n^o_{tp}$, etc. The important point to be made here is that the d_1, d_2, d_3, d_4 are correlated, that is, they are not mutually independent. Thus,

one cannot *a priori* introduce independent d_1, d_2, d_3, d_4 into equations of the form of (121) at this stage of the calculation. Using our microscopic approach, however, these source terms and their correlations arise in a well-defined manner.

If we let $\delta n \equiv \delta n_1$, $\delta p \equiv \delta n_2$, $\delta n_{tn} \equiv \delta n_3$, $\delta n_{tp} = \delta n_4$, $\boldsymbol{\delta n} \equiv (\delta n_1, \delta n_2, \delta n_3, \delta n_4)$, $\mathbf{d} \equiv (d_1, d_2, d_3, d_4)$, then (121) can be written in the form

$$i\omega \mathbf{I}\boldsymbol{\delta n} = \mathbf{R}\boldsymbol{\delta n} + \mathbf{d} \tag{123a}$$

and solved for $\boldsymbol{\delta n}$ at once:

$$\boldsymbol{\delta n}(\omega) = \mathbf{M}(\omega)\mathbf{d}(\omega), \tag{123b}$$

where

$$\mathbf{M}(\omega) \equiv (i\omega\mathbf{I} - \mathbf{R})^{-1}. \tag{123c}$$

The cross-spectral density of δn_i, δn_j is then given by ($\omega = 2\pi f$):

$$S_{ij}^{\boldsymbol{\delta n}}(f) = \sum_{kl} M_{ik}(\omega) S_{kl}^{\mathbf{d}}(f) M_{lj}^{*}(\omega), \tag{124a}$$

where $S_{kl}^{\mathbf{d}}(f)$ is the cross-spectral density of d_k, d_l:

$$\mathbf{S}^{d} = \begin{bmatrix} S_{ton}^{d}+S_{tpn}^{d}+S_{ntn}^{d}+S_{nto}^{d} & 0 & -S_{ton}^{d}-S_{ntn}^{d} & S_{tpn}^{d}+S_{nto}^{d} \\ 0 & S_{top}^{d}+S_{tnp}^{d}+S_{ptp}^{d}+S_{pto}^{d} & S_{tnp}^{d}+S_{pto}^{d} & -S_{top}^{d}-S_{ptp}^{d} \\ S_{tpn}^{d}+S_{nto}^{d} & -S_{top}^{d}-S_{ptp}^{d} & S_{ton}^{d}+S_{ntn}^{d}+S_{tnp}^{d}+S_{pto}^{d} & 0 \\ -S_{ton}^{d}-S_{ntn}^{d} & S_{tnp}^{d}+S_{pto}^{d} & 0 & S_{tpr}^{d}+S_{nto}^{d}+S_{top}^{d}+S_{ptp}^{d} \end{bmatrix}, \tag{124b}$$

and

$$\begin{aligned} &S_{ton}^{d} = 2r_{ton}n^{o},\ S_{top}^{d} = 2r_{top}p^{o},\ S_{tpn}^{d} = 2r_{tpn}n^{o}, \\ &S_{tnp}^{d} = 2r_{tnp}p^{o},\ S_{ntn}^{d} = 2g_{ntn},\ S_{ptp}^{d} = 2g_{ptp},\ S_{nto}^{d} = 2g_{nto}, \\ &S_{pto}^{d} = 2g_{pto} \end{aligned}$$

from (73). Using (123b, c) and (124a, b), the cross-spectral density of δn_i, δn_j including *all* correlation effects can be readily calculated. Of greatest interest are the spectral densities of δn and δp given by $S_{21}^{\boldsymbol{\delta n}}(f)$ $S_{22}^{\boldsymbol{\delta n}}(f)$, *and* of δn, δp given by $S_{12}^{\boldsymbol{\delta n}}(f) = S_{21}^{\boldsymbol{\delta n}}(f)$. These are the terms involving mobile carriers, which contribute to the current fluctuations seen at the contacts of the device. It should be clear that trapping can be treated by using an approach similar to the above.

There is little point in carrying out the algebra implicit in (124a). Evaluating the inverse of the 4×4 matrix in (123c) and carrying out the sum in (124a) can be done readily, although it is somewhat tedious.

What we have accomplished is to express the cross-spectral densities of the electrons, holes, and various trapping states in terms of known, easily calculated, cross-spectral densities of the source-driving terms. And in so doing we have included all important correlation effects in a very natural way.

REFERENCES

1. T. C. McGill, M.-A. Nicolet, and K. K. Thornber, "Equivalence of the Langevin Method and the Impedance-Field Method of Calculating Noise in Devices," Solid-State Elec. *17*, No. 1 (January 1974), pp. 107–108.
2. K. K. Thornber, T. C. McGill, and M.-A. Nicolet, "Structure of the Langevin and Impedance-Field Methods of Calculating Noise in Devices," Solid-State Elec., in press.
3. P. Langevin, "On the Theory of Brownian Motion," Compt. Rend., *146* (March 9, 1908), pp. 503–533.
4. W. Shockley, J. A. Copeland, and R. P. James, "The Impedance-Field Method of Noise Calculations in Active Semiconductor Devices," P. O. Löwdin, ed., *Quantum Theory of Atoms, Molecules, and the Solid State*, New York: Academic Press, 1966, pp. 537–563.
5. A. van der Ziel, *Noise*, Englewood Cliffs, N. J.: Prentice-Hall, 1956.
6. A. van der Ziel, *Fluctuation Phenomena in Semiconductors*, New York: Academic Press, 1959.
7. A. van der Ziel, *Noise: Sources, Characterization, Measurement*, Englewood Cliffs, N. J.: Prentice-Hall, 1970.
8. A. van der Ziel, "Noise in Solid-State Devices and Lasers," Proc. IEEE, *58*, No. 8 (August 1970), pp. 1178–1206.
9. K. K. Thornber, "Noise Suppression in Charge Transfer Devices," Proc. IEEE, *60*, No. 9 (September 1972), pp. 1113–1114.
10. K. K. Thornber and M. F. Tompsett, "Spectral Density of Noise Generated in Charge Transfer Devices," IEEE Trans. Elec. Devices, *ED-20*, No. 4 (April 1973), p. 456.
11. K. K. Thornber, "Some Consequences of Spatial Correlation on Noise Calculations," Solid-State Elec., *17*, No. 1 (January 1974), pp. 95–97.
12. M. Lax, "Fluctuations from the Nonequilibrium Steady-State," Rev. Mod. Phys., *32*, No. 1 (January 1960), pp. 25–64. (See especially Section 6 and Fig. 1.)
13. Ref. 7, p. 10.
14. K. M. van Vliet, "Noise Sources in Transport Equations Associated with Ambipolar Diffusion and Shockley-Read Recombination," Solid-State Elec., *13*, No. 5 (May 1970), pp. 649–657.
15. R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill, 1965, Ch. 12. These authors assume stationary noise, whereas our results are valid for nonstationary noise as well.
16. H. B. Callen and T. A. Welton, "Irreversibility and Generalized Noise," Phys. Rev., *83*, No. 1 (July 1, 1951), pp. 34–40.
17. H. B. Callen, "Path Distribution for Irreversible Processes," Phys. Rev., *111*, No. 2 (July 15, 1958), pp. 367–372.
18. W. Bernard and H. B. Callen, "Irreversible Thermodynamics of Nonlinear Processes and Noise in Driven Systems," Rev. Mod. Phys., *31*, No. 4 (October 1959), pp. 1017–1044.
19. H. B. Callen, "Fluctuation-Dissipation Theorem and Irreversible Thermodynamics," D. ter Haar, ed., *Fluctuation, Relaxation, and Resonance in Magnetic Systems*, New York: Plenum Press, 1962.
20. L. D. Landau and E. M. Lifshitz, *Statistical Physics*, London: Pergamon Press, 1958, Sections 123-4.
21. R. P. Feynman, "An Operator Calculus Having Applications in Quantum Electrodynamics," Phys. Rev., *84*, No. 1 (October 1, 1951), pp. 108–128.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

Bent Optical Waveguide With Lossy Jacket

By D. MARCUSE

(Manuscript received December 20, 1973)

The influence of a lossy jacket on the curvature losses of a bent optical waveguide is studied for the special case of the TE modes of a slab waveguide. This paper presents an approximate theory of curvature losses of the TE modes of dielectric slabs that can be used to obtain numerical answers with the help of a computer. We conclude that the presence of a jacket can increase the curvature losses very substantially. A jacket whose refractive index is larger than that of the waveguide cladding is most effective in increasing cladding losses. It is advisable to keep a jacket at a safe distance from the waveguide core.

I. INTRODUCTION

To avoid crosstalk between adjacent fibers in a cable and also to suppress unwanted cladding modes, optical fibers for communication purposes need lossy jackets.[1] Each fiber thus consists of a core of refractive index n_1 and a cladding with index n_2. Since core and cladding are made of low-loss materials, we consider n_1 and n_2 real constants. The refractive index of the lossy jacket is considered complex:

$$n_3 = n_{3r} - in_{3i}. \tag{1}$$

The negative sign is necessary since we use the time dependence

$$e^{i\omega t} \tag{2}$$

for the optical waves.

The guided-mode fields decrease in intensity exponentially with increasing distance from the fiber core. At the boundary between the cladding and the lossy jacket, the intensity of the modes should decrease to insignificant values. If the cladding is too thin, so that the modes arrive at the cladding-jacket boundary with appreciable field intensities, considerable amounts of power would be dissipated in the lossy jacket, resulting in intolerably high waveguide losses. The de-

signer must provide for a cladding of sufficient thickness to keep the fiber losses low.

So far, we have considered a fiber that is perfectly straight. However, an advantage of optical fiber systems is that light transmission is maintained as the fiber is curved. Since curvature of the fiber axis distorts the shape of the guided modes,[2] it is necessary to study the effect of the lossy jacket in the presence of fiber curvature. A curved fiber radiates a certain amount of power even if its cladding extends infinitely far from the core.[2,3] The amount of radiated power is modified by the presence of the lossy jacket.

It is the purpose of this paper to investigate the influence of the lossy jacket on the curvature losses of optical waveguides. Because of the complexity of the problem, we use the TE modes of the symmetric slab waveguide as a model.

The following sections are devoted to the derivation of the theory. Readers not interested in the theoretical details are advised to turn to Section VI, on numerical examples.

II. OUTLINE OF THE METHOD OF SOLUTION

A curved slab waveguide with lossy jacket is schematically shown in Fig. 1. The core with refractive index n_1 has the full width $2d$. The center line of the core is curved with radius of curvature R. The cladding with index n_2 has the thickness $D - d$. The refractive index of the jacket is assumed to be a complex quantity. A straightforward solution of this problem would involve writing down solutions of Maxwell's equations in the five different regions of the structure. These solutions can be expressed in terms of cylinder functions. The waveguide modes are obtained by joining the solutions in the different regions with the help of boundary conditions. This straightforward procedure is not practical for the determination of the fiber losses. To understand the difficulty, we must consider that the cylinder functions, expressing the solutions of Maxwell's equations, have very large order numbers and arguments that are of the same order of magnitude as the order numbers. The problem consists in finding the order number as a solution of an eigenvalue problem. Since we expect to compute the waveguide losses, the order of the Bessel functions must be a complex quantity. We are thus faced with solving a determinantal equation whose elements are cylinder functions of very large complex order. Cylinder functions of this type cannot be computed with the help of power series expansions. The functions must be obtained from approximate asymptotic expressions. The solution of the complex

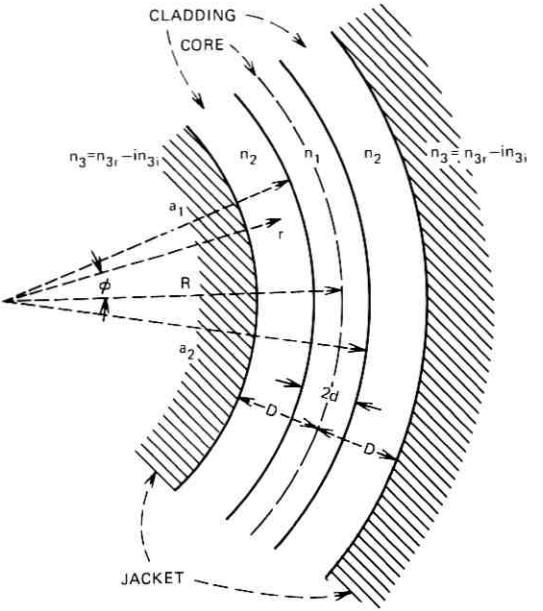

Fig. 1—Schematic of the bent slab waveguide with lossy jacket. The z-axis is directed normal to the plane of the figure.

transcendental eigenvalue equation thus not only is a difficult numerical task, but also may be expected to yield poor accuracy since we expect the imaginary part of the eigenvalue (the order number of the cylinder functions) to be small so that it could be obtained with high accuracy only if the functions themselves are known to high precision.

Since the straightforward approach seems to present an almost insurmountable obstacle, we use a different approach. Instead of solving the problem sketched in Fig. 1, we begin by solving the simpler problem that results if we let $D \rightarrow \infty$. The exact solution of the bent slab with infinite cladding thickness still results in a complex eigenvalue, since radiation losses occur. However, we are not interested in computing the radiation losses at this point and modify the eigenvalue equation so that its imaginary part is neglected. We are now left with a relatively simple eigenvalue problem. It is still necessary to compute cylinder functions of large order and argument. But since only a real eigenvalue is computed with the help of real cylinder functions, the usual asymptotic approximations of the cylinder functions can be used.

The next step of our approximate procedure consists in determining the reflection and transmission coefficients of a cylindrical wave im-

pinging on a cylindrical dielectric interface. Once this problem is solved, we apply its solutions to the evanescent field tail of the guided wave in the cladding. In this way, we obtain approximate field expressions for the field reaching into the lossy jacket. It is now a simple matter to calculate the amount of power flowing from the guided mode into the lossy jacket and to use it to determine the waveguide losses.

In the following sections, we outline the mathematical details of our approximate procedure. The only difficulty encountered consists in producing the cylinder functions of large order and, at least inside the lossy jacket, of complex argument.

III. BENT SLAB WAVEGUIDE WITH INFINITELY WIDE CLADDING

We are interested in the TE modes of the curved slab. Using the coordinates indicated in Fig. 1, we can express the z component of the electric field in the three regions as

$$E_z = \begin{cases} AJ_\nu(n_2kr)e^{-i\nu\phi} & 0 < r < a_1 \\ [BJ_\nu(n_1kr) + CN_\nu(n_1kr)]e^{-i\nu\phi} & a_1 < r < a_2 \\ FH_\nu^{(2)}(n_2kr)e^{-i\nu\phi} & a_2 < r < \infty. \end{cases} \tag{3}$$

The z coordinate is directed perpendicular to the plane of the figure. The Bessel and Neumann functions of order ν are J_ν and N_ν. The free space propagation constant is defined as

$$k = \frac{2\pi}{\lambda} = \omega\sqrt{\epsilon_0\mu_0}. \tag{4}$$

The r and ϕ components of the magnetic fields are obtained from the E_z component by differentiation.[4]

$$H_r = \frac{i}{\omega\mu_0}\frac{1}{r}\frac{\partial E_z}{\partial\phi} = \frac{1}{\omega\mu_0}\frac{\nu}{r}E_z \tag{5}$$

$$H_\phi = \frac{-i}{\omega\mu_0}\frac{\partial E_z}{\partial r}. \tag{6}$$

The remaining field components E_r, E_ϕ, and H_z vanish. Since the waves travel along the curved slab in ϕ direction, we can define the propagation constant of the guided mode

$$\beta = \frac{\nu}{R}. \tag{7}$$

The requirement of continuity of the E_z and H_ϕ components at the core boundaries $r = a_1$ and $r = a_2$ lead to the determination of the

amplitude coefficients

$$A = \frac{n_1 J_\nu(x_{11}) N'_\nu(x_{11}) - n_1 J'_\nu(x_{11}) N_\nu(x_{11})}{n_1 J_\nu(x_{21}) N'_\nu(x_{11}) - n_2 J'_\nu(x_{21}) N_\nu(x_{11})} B, \tag{8}$$

$$C = -\frac{n_1 J_\nu(x_{21}) J'_\nu(x_{11}) - n_2 J'_\nu(x_{21}) J_\nu(x_{11})}{n_1 J_\nu(x_{21}) N'_\nu(x_{11}) - n_2 J'_\nu(x_{21}) N_\nu(x_{11})} B, \tag{9}$$

and

$$F = \frac{1}{H_\nu^{(2)}(x_{22})} [B J_\nu(x_{12}) + C N_\nu(x_{12})]. \tag{10}$$

The definition

$$x_{ij} = n_i k a_j \tag{11}$$

was used. The prime indicates the derivative of the functions with respect to the argument.

The Bessel and Neumann functions are real. The Hankel function of the second kind appearing in (3) and (10) is complex,

$$H_\nu^{(2)} = J_\nu - iN_\nu. \tag{12}$$

Because of the complex value of $H_\nu^{(2)}$, the eigenvalue equation (that results from the requirement that the determinant of the equation system for the determination of A, B, C, and F vanish) is itself complex, leading to complex solutions for ν. However, for well-guided modes we have

$$\nu > n_2 k a_2 \gg 1. \tag{13}$$

The inequality (13) results, in turn, in

$$|J_\nu(x_{22})| \ll |N_\nu(x_{22})|. \tag{14}$$

The Hankel function is thus predominantly imaginary with a very small real part. By replacing the Hankel function with the approximation

$$H_\nu^{(2)} = -iN_\nu \tag{15}$$

in (3) and (10), we obtain the real eigenvalue equation

$$\begin{aligned}[n_2 J_\nu(x_{11}) J'_\nu(x_{21}) - n_1 J_\nu(x_{21}) J'_\nu(x_{11})] \\ \times [n_1 N_\nu(x_{22}) N'_\nu(x_{12}) - n_2 N'_\nu(x_{22}) N_\nu(x_{12})] \\ + [n_1 J_\nu(x_{21}) N'_\nu(x_{11}) - n_2 J'_\nu(x_{21}) N_\nu(x_{11})] \\ \times [n_1 J'_\nu(x_{12}) N_\nu(x_{22}) - n_2 J_\nu(x_{12}) N'_\nu(x_{22})] = 0.\end{aligned} \tag{16}$$

This eigenvalue equation has real solutions of ν ignoring radiation losses caused by waveguide curvature. However, the mode problem that we have formulated describes the distorted fields in the curved

waveguide accurately. The curvature losses are obtained later by accounting for the amount of power that is lost in the form of radiation.

The power carried by the modes can be expressed as

$$P = \frac{1}{4\omega\mu_0}\left\{A^2\left[x\left(J_{\nu+1}\frac{\partial J_\nu}{\partial \nu} - J_\nu\frac{\partial J_{\nu+1}}{\partial \nu}\right) + J_\nu^2\right]_0^{x_{21}}\right.$$
$$+ B^2\left[x\left(J_{\nu+1}\frac{\partial J_\nu}{\partial \nu} - J_\nu\frac{\partial J_{\nu+1}}{\partial \nu}\right) + J_\nu^2\right]_{x_{11}}^{x_{21}}$$
$$+ 2BC\left[x\left(J_{\nu+1}\frac{\partial N_\nu}{\partial \nu} - J_\nu\frac{\partial N_{\nu+1}}{\partial \nu}\right) + J_\nu N_\nu\right]_{x_{11}}^{x_{12}}$$
$$+ C^2\left[x\left(N_{\nu+1}\frac{\partial N_\nu}{\partial \nu} - N_\nu\frac{\partial N_{\nu+1}}{\partial \nu}\right) + N_\nu^2\right]_{x_{11}}^{x_{12}}$$
$$\left. + |F|^2\left[x\left(N_{\nu+1}\frac{\partial N_\nu}{\partial \nu} - N_\nu\frac{\partial N_{\nu+1}}{\partial \nu}\right) + N_\nu^2\right]_{x_{22}}^{\infty}\right\}. \quad (17)$$

The notation $[\ ]_{x_1}^{x_2}$ indicates that the value of the bracket evaluated at x_1 must be subtracted from the expression evaluated at x_2. Since the ratios of the amplitude coefficients are real quantities, A, B, and C are assumed real. However, with approximation (15), F becomes imaginary. The contributions of the lower limit 0 of the first bracket and of the upper limit ∞ of the last bracket may be neglected since the fields decrease rapidly with increasing distance from the waveguide core.

IV. REFLECTION AND TRANSMISSION OF A WAVE AT A CYLINDRICAL INTERFACE

Our solution of the mode problem of the bent slab waveguide ignored radiation losses caused by the curvature and losses resulting from the presence of the lossy jacket. We calculate these losses by accounting for the outflow of power from the curved waveguide. To obtain expressions for the power outflow, we study the problem of a cylindrical wave that is impinging on a cylindrical interface between two dielectric media.

Ignoring, for the moment, the presence of the waveguide core and the jacket region that contains the center of curvature, we consider a cylindrical wave in the region to the left of the interface between the media with refractive indices n_2 and n_3,

$$E_z = [GH_\nu^{(2)}(n_2kr) + IH_\nu^{(1)}(n_2kr)]e^{-i\nu\phi} \qquad R + d < r < R + D. \quad (18)$$

According to time dependence (2), the Hankel function of the second kind describes the incident cylindrical wave, while the Hankel function

of the first kind belongs to the reflected wave. Inside the jacket we have a transmitted wave

$$E_z = KH_\nu^{(2)}(n_3kr)e^{-i\nu\phi} \qquad r > R + D. \tag{19}$$

The corresponding magnetic field components follow again from (5) and (6). Continuity of the E_z and H_ϕ components is achieved if the following relations hold between the three amplitude coefficients:

$$I = \frac{n_3H_\nu^{(2)}(y_2)H_\nu^{(2)'}(y_3) - n_2H_\nu^{(2)'}(y_2)H_\nu^{(2)}(y_3)}{n_2H_\nu^{(1)'}(y_2)H_\nu^{(2)}(y_3) - n_3H_\nu^{(2)'}(y_3)H_\nu^{(1)}(y_2)} G \tag{20}$$

and

$$K = \frac{n_2H_\nu^{(1)'}(y_2)H_\nu^{(2)}(y_2) - n_2H_\nu^{(1)}(y_2)H_\nu^{(2)'}(y_2)}{n_2H_\nu^{(1)'}(y_2)H_\nu^{(2)}(y_3) - n_3H_\nu^{(1)}(y_2)H_\nu^{(2)'}(y_3)} G, \tag{21}$$

with

$$y_i = n_ik(R + D). \tag{22}$$

It remains to relate the amplitude G to the amplitude F of the evanescent field tail of the guided mode in the curved slab. Our treatment is, of course, not exact, since multiple reflections of the wave between core and jacket are ignored. However, if the refractive index differences remain small, multiple reflections are unimportant. Furthermore, the field intensity decays exponentially with increasing distance from the waveguide core. The incident wave $GH_\nu^{(2)}(n_2kr)$ is, thus, an evanescent wave in most cases so that the effect of the core cladding boundary on this wave is only very slight. Whether the incident wave is an evanescent or a propagating wave depends on the distance between core and jacket. If this distance is small, the guided mode field behaves predominantly as an evanescent wave. If the distance between core and jacket is large, the evanescent wave has converted itself to a traveling wave before the jacket is reached. Our approximate procedure works in either case for most cases of practical interest.

To obtain the relation between the amplitude G and the amplitude F of the guided wave, we consider the field in the immediate vicinity of the core boundary and equate the fields (3) and (18)

$$FH_\nu^{(2)}(n_2kr) = GH_\nu^{(2)}(n_2kr) + IH_\nu^{(1)}(n_2kr). \tag{23}$$

It was explained earlier that we may approximate the Hankel function of the second kind by (15). Likewise, we use the approximation

$$H_\nu^{(1)} = iN_\nu. \tag{24}$$

Using (15) and (24), we obtain from (23)

$$G = \frac{F}{1 - \frac{I}{G}}. \tag{25}$$

The ratio I/G is given by (20). We thus have determined the amplitude of the wave that is incident on the jacket (at least approximately) and can now compute the amount of power that is carried into the jacket.

V. CALCULATION OF THE LOSSES

The amount of power outflow in r direction per unit length along the waveguide axis (and also per unit length in z direction) is given by the r component of the Poynting vector

$$S_r = -\tfrac{1}{2} \operatorname{Re} \{E_z H_\phi^*\}. \tag{26}$$

If we denote by α the amplitude attenuation coefficient of the guided wave, we obtain the power attenuation coefficient 2α from the relation

$$2\alpha = \frac{S_r}{P}. \tag{27}$$

This relation holds since P is by definition the amount of power carried by the guided mode per unit length (in z direction). Using (6), (19), and (26) we obtain

$$2\alpha = \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{|K|^2}{2P} \operatorname{Im} \{n_3^* H_\nu^{(2)}(y_3) H_\nu^{(2)'*}(y_3)\}. \tag{28}$$

The asterisk indicates complex conjugation and Im ( ) designates that the imaginary part of the complex expression in brackets is to be taken. The argument y_3 is defined by (22).

A small amount of power also flows into the jacket on the other side of the waveguide, the side facing the center of curvature. However, for reasonably strongly curved guides, this power outflow is orders of magnitude smaller than the power outflow included in (28) so that we may safely neglect it.

The solution of the loss problem is now reduced to a determination of the cylinder functions appearing in our equations. We evaluate (28) by using (8) through (10), (17), (21), and (25). The order of the cylinder functions is determined as a solution of the eigenvalue equa-

tion (16). As stated earlier, our method has the advantage that no complex eigenvalue equation need be solved. Owing to the difficulty of computing accurate values for the cylinder functions of large complex order and large complex argument, a direct determination of the losses with the help of the complex eigenvalue equation is hard to achieve. Our method is straightforward in principle. We only face the computational difficulty of determining the cylinder functions of large real order and, at least for some functions, of large complex argument. However, our present method does not require knowledge of these functions to extreme accuracy.

In two limiting cases, the attenuation formulas for the curved slab waveguide are known. For a straight slab with lossy jacket, we use eq. (10.3–14), p. 420, of Ref. 4.

$$2\alpha = \frac{8\kappa^2\gamma^3 \operatorname{Im}(\rho) e^{-2\gamma(D-d)}}{\beta(1+\gamma d)(\kappa^2+\gamma^2)|\gamma+\rho|^2} \tag{29}$$

with

$$\kappa^2 = n_1^2 k^2 - \beta^2, \tag{30}$$

$$\gamma^2 = \beta^2 - n_2^2 k^2, \tag{31}$$

and

$$\rho^2 = \beta^2 - n_3^2 k^2. \tag{32}$$

The propagation constant β is obtained as a solution of the eigenvalue equation

$$\tan \kappa d = \frac{\gamma}{\kappa} \tag{33}$$

for even modes and from

$$\tan \kappa d = -\frac{\kappa}{\gamma} \tag{34}$$

for odd modes.

For a curved slab without lossy jacket but infinitely wide cladding, eq. (9.6–27), p. 404, of Ref. 4 is available,

$$2\alpha = \frac{\kappa^2\gamma^2}{\beta(1+\gamma d)(\kappa^2+\gamma^2)} e^{2\gamma d} \exp\left\{-\frac{2}{3}\frac{\gamma^3}{\beta^2} R\right\}. \tag{35}$$

If we use the eigenvalue β obtained from (33) or (34), we obtain good results only for single mode guides or for very large radii of curvature. Better agreement with numerical evaluations of (28) is obtained if we use solutions of the eigenvalue equation (16) and calculate β with the help of (7) and the other parameters from (30) and (31).

VI. NUMERICAL EXAMPLES

The principal problem of evaluating the formulas of our theory consists in generating the Bessel function of large real order and large (sometimes) complex argument. The Hankel functions can be expressed in terms of Bessel and Neumann functions. These latter functions are approximated by using the asymptotic formulas (9.3.7) through (9.3.17) on p. 366 of Ref. 5 and eqs. (9.3.23) and (9.3.24) on p. 367 of the same reference. It is not stated clearly in any reference book on Bessel functions that these asymptotic formulas are valid for complex arguments. [This statement refers to the functions given by (9.3.7) through (9.3.17).] However, the first terms of these expressions can easily be derived either by using the integral representation of Bessel and Neumann functions and the method of steepest descent or [for $J_\nu(x)$ with $\nu > x$] by using approximate solutions obtained directly from the differential equation. Either method clearly holds also for complex arguments. It may be that the convergence behavior and the error estimates available for real arguments may not apply to

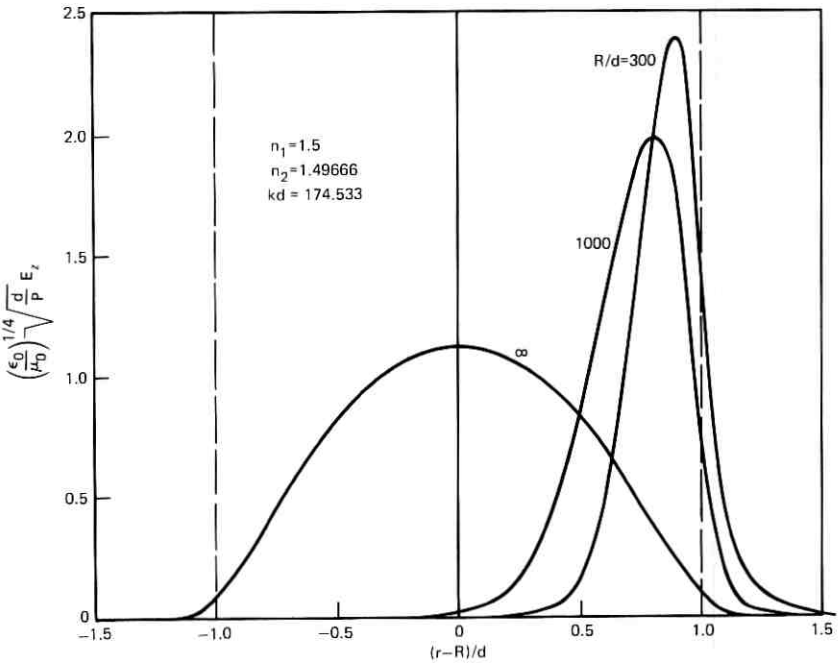

Fig. 2a—Normalized E_z component of the first guided mode for different radii of curvature. The refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$.

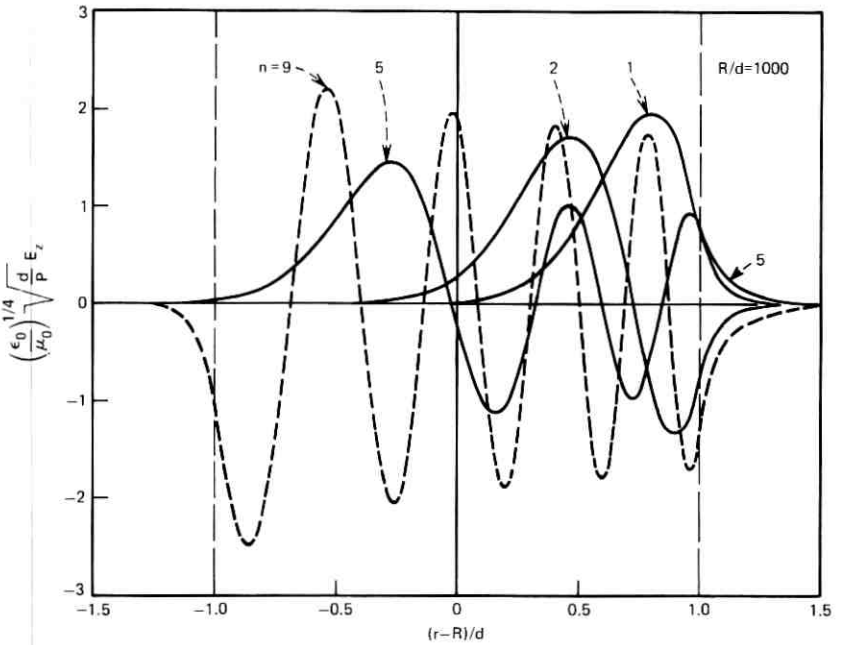

Fig. 2b—Distribution of the electric field for four guided modes, $n = 1, 2, 5$, and 9 for $R/d = 1000$.

complex arguments, but at least the first terms of the asymptotic formulas can be justified for functions with complex argument. For this reason, these formulas were used even if the argument of the cylinder functions is complex. This procedure appears even more valid when we consider that in all cases of practical interest the phase angle of the complex argument remains very small.

The derivatives with respect to the order number were generated by taking the derivatives of the asymptotic formulas. Our method of generating the necessary cylinder functions seems justified by the excellent agreement that was obtained with formulas (29) and (35) in all instances where such agreement could be tested.

Since the arguments of the cylinder functions are of the form nkr, there are practical limits to the size of the radius of curvature of the waveguide axis. For ratios of R/d in excess of 1000, exponent overflow was encountered in the numerical calculations so that the limiting case of a straight slab could not be approached very closely.

The distortion of the field distribution caused by waveguide curvature is dramatically evident from the curves of Figs. 2(a) and 2(b).

Figure 2(a) shows the shape of the normalized E_z component of the lowest order mode, labeled $n = 1$, for several values of R/d. The curve for $R/d = \infty$ was obtained from eqs. (8.3–9), (8.3–12), and (8.3–18) of Ref. 4. It is apparent that the core cladding boundary on the side facing the center of curvature does not contribute to guiding the lowest-order mode in case of sharp bends. It is also evident that substantial mode conversion must result if a curved waveguide section is joined to a straight waveguide without tapering the curvature. Finally, we see from the figure that the field is forced far deeper into the cladding region by the waveguide curvature so that it tends to interact more strongly with the lossy jacket.

Figure 2(b) shows the distribution of the E_z fields for several modes. Both figures were drawn for the following parameters: $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$. The important V parameter defined by

$$V = \sqrt{n_1^2 - n_2^2}kd \tag{36}$$

assumes the value $V = 17.46$. The straight slab is thus able to support 11 TE modes. Figure 2(b) shows plots for the modes $n = 1$, 2, 5, and 9. We see that the higher-order modes occupy more of the available space inside the waveguide core. The period of oscillation becomes shorter toward the side of the core opposite the center of curvature. However, the field amplitudes are largest on the side nearest the center of curvature.

With regard to the normalization used for the electric field component,

$$\left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{4}} \sqrt{\frac{d}{P}}\, E_z, \tag{37}$$

we must remember that the parameter P stands for the power carried by the slab waveguide per unit length (in z direction).

All numerical examples discussed (with the exception of Fig. 12) are based on the waveguide parameters given above. The propagation constants β obtained from (16) and (7) are listed in Table I for all TE modes that can be supported by the guide for $R/d = 300$, 1000, and ∞. The values for $R/d = \infty$ were obtained from the eigenvalue equations (33) and (34) for even and odd TE modes of the straight slab waveguide. The table shows that the number of guided modes decreases as the curvature of the guide increases.

Figure 3 shows the normalized loss coefficient $2\alpha d$ as a function of d/R for several modes of a slab without jacket. The horizontal dotted lines appearing in this and all subsequent figures indicate the level of 1 dB/km and 10 dB/km loss for a guide with the slab half width

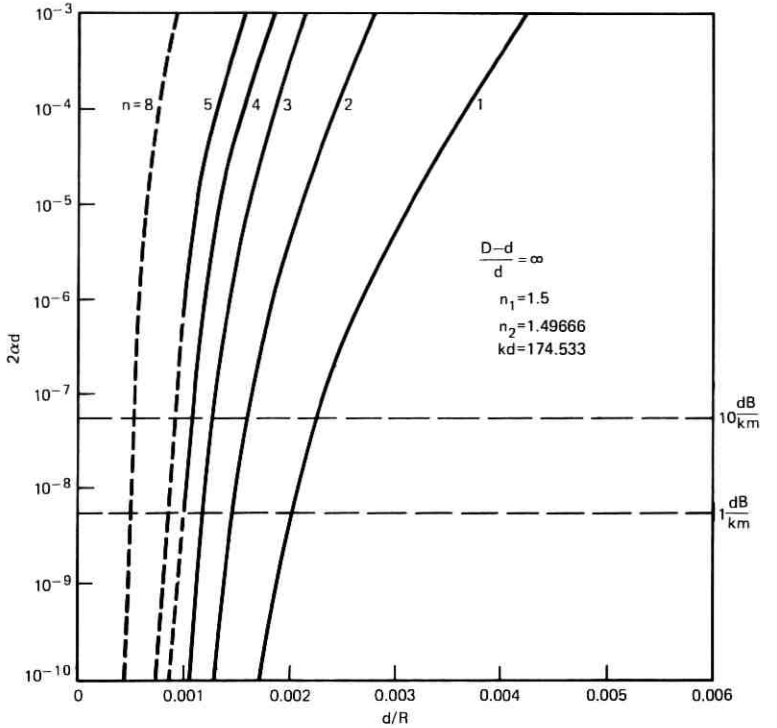

Fig. 3—Curvature losses of a slab with infinitely thick jacket as a function of the inverse radius of curvature for several TE modes. The refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$.

Table I — Values of the normalized propagation constant $\beta_n d$ for all the TE modes of a curved slab ($n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$)

n	$\beta_n d$		
	$R/d = 300$	$R/d = 1000$	$R/d = \infty$
1	262.461	261.958	261.795
2	262.270	261.870	261.783
3	262.111	261.797	261.762
4	261.840	261.734	261.732
5	261.498	261.676	261.694
6		261.621	261.648
7		261.569	261.594
8		261.513	261.532
9		261.436	261.463
10		261.381	261.386
11		261.224	261.303

$d = 25$ μm. It is apparent how very strongly the curvature losses depend on the radius of curvature of the waveguide axis. The losses in decibels are obtained by dividing the numerical values, that are read off the vertical axis of the figure, by the slab half width d and multiplying by 4.34 (to convert the result to decibels).

For a comparison with formula (35), we state that the loss value of the lowest-order mode for $d/R = 0.001$ is $2\alpha d = 3.41 \times 10^{-19}$ as computed with the help of the theory presented in this paper. From Table I we find for $n = 1$, $\beta_1 d = 261.958$, so that from (31) we obtain $\gamma d = 19.695$. If we try to compute κ^2 from (30) we find a negative value. Therefore, we use the far-from-cutoff approximation $\kappa d = \pi/2$. Using these values in (35), we find $2\alpha d = 3.37 \times 10^{-19}$ in excellent agreement with the value obtained from our theory. For the high loss values appearing in Fig. 3, the agreement is not as good.

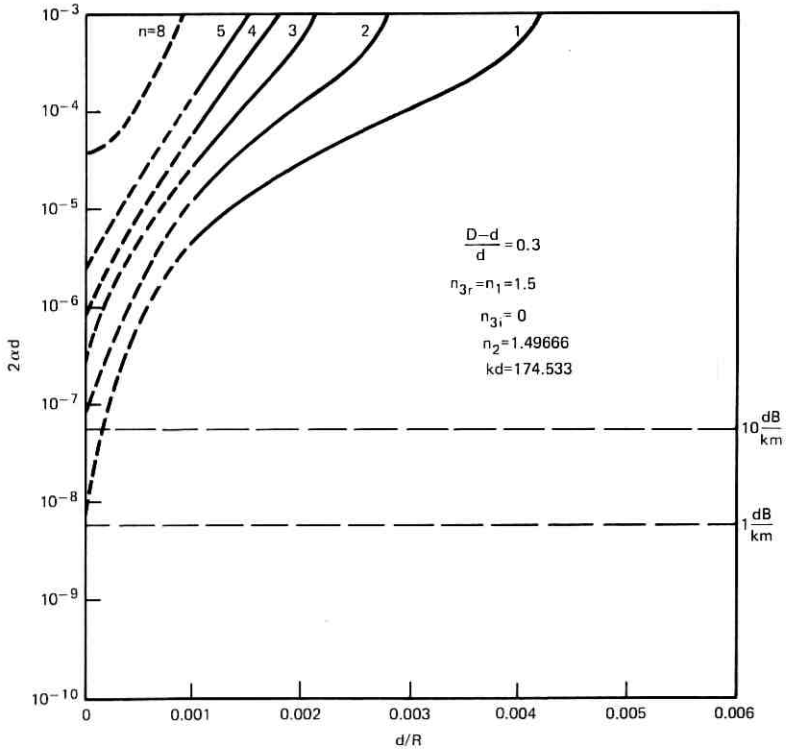

Fig. 4—Curvature losses in the presence of a lossless jacket. The normalized cladding thickness is $(D - d)/d = 0.3$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_1$, $n_{3i} = 0$.

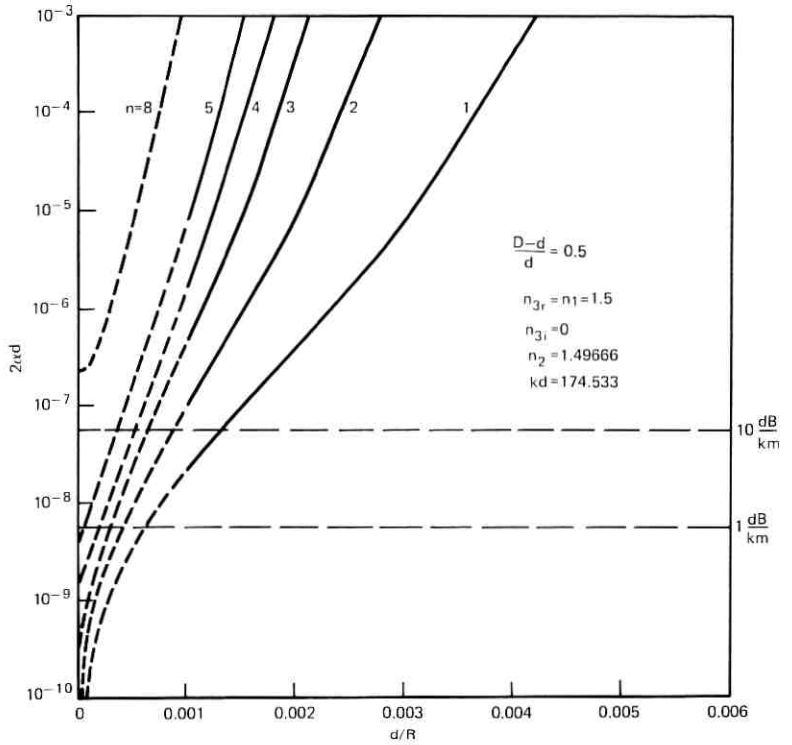

Fig. 5—Curvature losses in the presence of a lossless jacket. The normalized cladding thickness is $(D - d)/d = 0.5$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_1$, $n_{3i} = 0$.

To gain insight into the effect that the jacket has on the curvature losses, we have plotted the loss values that result if we use a lossless jacket whose refractive index equals that of the waveguide core. Even though the lossless jacket does not dissipate power, it causes the portion of the evanescent field tail reaching the jacket to turn into a propagating wave and thus to radiate away. Figures 4 through 6 show the curvature losses in the presence of the "high-index" jacket as a function of d/R for several modes and for different values of the relative cladding thickness $(D - d)/d$. Comparison of Figs. 3 and 4 shows clearly the dramatic increase in the curvature losses for a thin cladding with $(D - d)/d = 0.3$. As the cladding becomes thicker, the influence of the jacket decreases, as seen in Fig. 5. The upper parts of the curves in Fig. 6 already coincide with the curves of Fig. 3 for an infinitely thick cladding. In this case, the field detaches itself from the guide

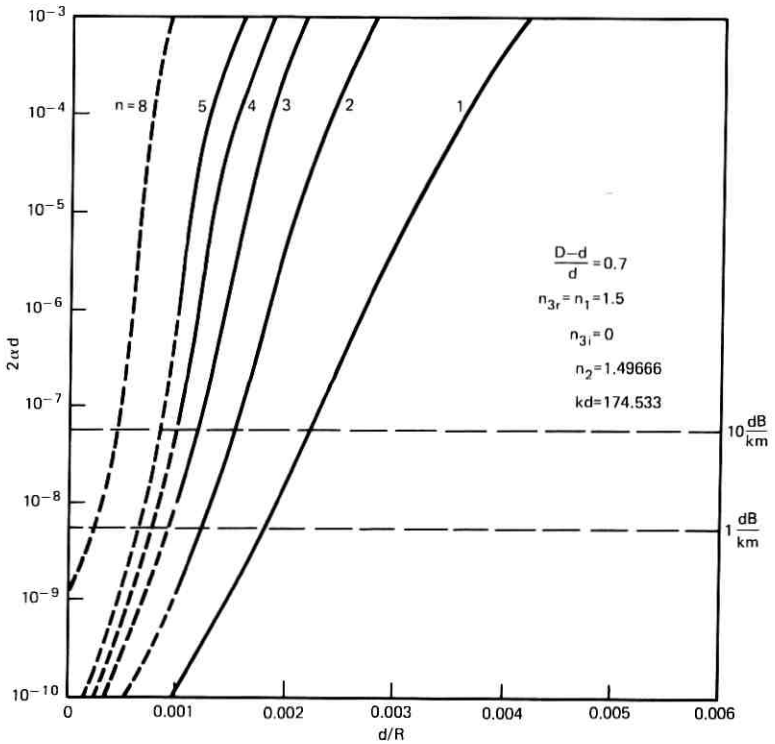

Fig. 6—Curvature losses in the presence of a lossless jacket. The normalized cladding thickness is $(D - d)/d = 0.7$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_1$, $n_{3i} = 0$.

inside the cladding so that the jacket no longer converts an evanescent field tail into a radiation field, but simply modifies the radiation field in an almost imperceptible way. These curves show that it is very necessary to maintain the "high-index jacket" at a sufficient distance from the waveguide core.

The dotted lines in these and all following figures are estimated curves. We pointed out that the computer program fails to function for very large values of R/d. The solid lines are the results of the numerical evaluation of our theory. The end points of the dotted curves at $d/R = 0$ were computed from (29). The region between $d/R = 0$ and $d/R = 0.001$ was bridged by the estimated dotted lines.

The curves for mode 8 shown in these and subsequent figures have a special meaning. We want to use our slab model to gain information about round fibers. If we consider a fiber with core radius $a = d$ and

the same refractive indices used for the slab, the total number M of fiber modes is proportional to the square of the total number N of slab modes, $M = KN^2$. It is instructive to consider fibers capable of transmitting at least half their total number of guided modes. The corresponding number of slab modes is $N' = N/\sqrt{2}$. With our numerical values we have a slab supporting $N = 11$ modes. $N' = 11/\sqrt{2} = 8$ is thus the mode number that corresponds to half the total number of fiber modes. If the losses of mode $n = 8$ are just tolerable, but all higher-order modes suffer too much loss, we know that we have found operating conditions that would cause half the total number of fiber modes to be lost. For this reason, we have included mode $n = 8$ in our figures to be able to estimate the conditions that would allow half the fiber modes to be transmitted. Figure 4 shows that only a very small number of modes can propagate with low losses in a fiber whose

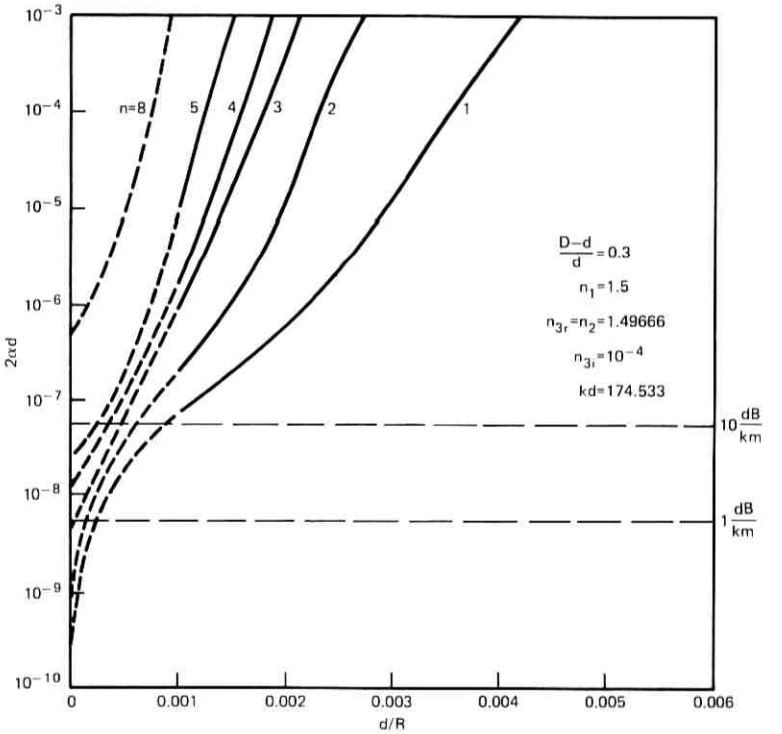

Fig. 7—Curvature losses in the presence of a lossy jacket. The normalized cladding thickness is $(D - d)/d = 0.3$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_2$, $n_{3i} = 10^{-4}$.

core thickness is $D - d = 0.3d$. For $D - d = 0.5d$, we see from Fig. 5 that more than half the fiber modes would suffer losses in excess of 10 dB/km even in the straight guide. This estimate is based on a jacket with large refractive index, $n_{3r} = n_1$. For jackets with lower index, the losses would be reduced. But to be on the safe side, it seems advisable to design a jacket so that it does not cause excessive loss even in the worst possible case. The conditions corresponding to Fig. 6 show that well over half the fiber modes are transmitted with low loss as long as $d/R < 0.0004$.

Figures 7 through 9 apply to the case of a jacket with a refractive index whose real part is matched to the cladding, $n_{3r} = n_2 = 1.49666$. The imaginary part of the jacket index is $n_{ri} = 0.0001$. This modest value of the imaginary part of the refractive index results in a plane

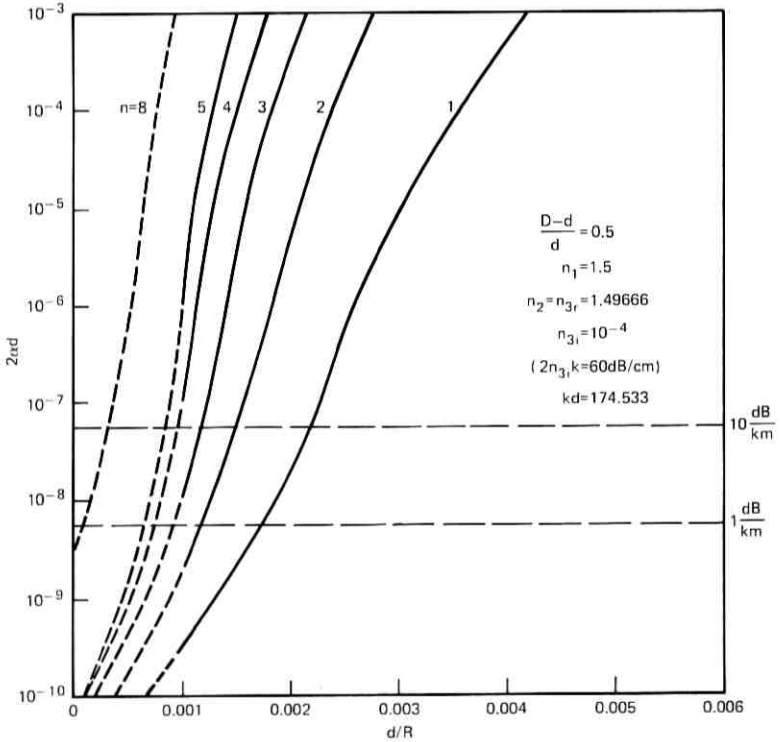

Fig. 8—Curvature losses in the presence of a lossy jacket. The normalized cladding thickness is $(D - d)/d = 0.5$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_2$, $n_{3i} = 10^{-4}$.

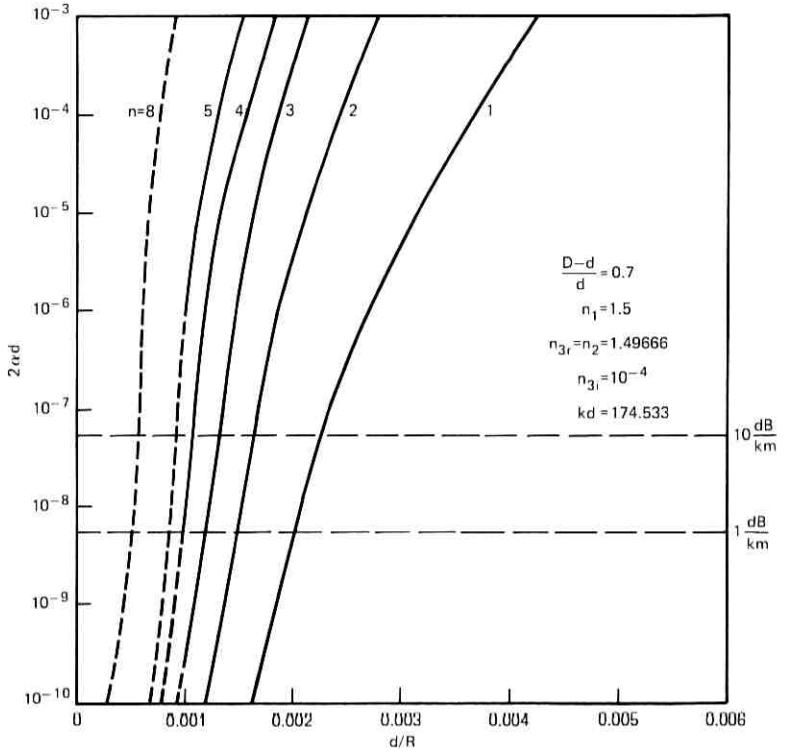

Fig. 9—Curvature losses in the presence of a lossy jacket. The normalized cladding thickness is $(D - d)/d = 0.7$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_2$, $n_{3i} = 10^{-4}$.

wave loss in the jacket material that is given by

$$2\alpha_{\text{jacket}}d = 2n_{3i}kd. \tag{38}$$

For our particular example, we have $2\alpha_{\text{jacket}}d = 0.035$. For $d = 25$ μm, this cladding loss amounts to 61 dB/cm.

Comparison of Figs. 7 through 9 with Fig. 3 for the case of the infinitely thick cladding shows that the lossy jacket has a considerable influence if it is located too close to the waveguide core. However, even for $(D - d)/d = 0.5$, its influence on the curvature losses is only slight and all but vanishes for $(D - d)/d = 0.7$.

A lossy jacket with $n_{3r} < n_2$ has only a very slight influence on the waveguide losses, since the evanescent field tail decays even more

rapidly in a medium of low refractive index. Therefore, no curves are provided for this case.

Figures 10 and 11 show the influence of the imaginary part of the refractive index of the jacket on the curvature losses. Figure 10 applies to the lowest-order mode, $n = 1$, and shows the dependence of the curvature loss on the logarithm of n_{3i} for $R/d = 500$ and $R/d = 1000$. It is apparent that the dependence of the loss on n_{3i} is linear in regions of the curve that are dominated by the losses in the jacket. For very small values of n_{3i}, the losses of the jacket become immaterial and the curves approach asymptotically the curvature loss of a waveguide with lossless, infinitely thick cladding.

The two curves in Fig. 11 dramatize this behavior. For $R/d = 1000$, the losses of the third mode are still dominated by the loss of the

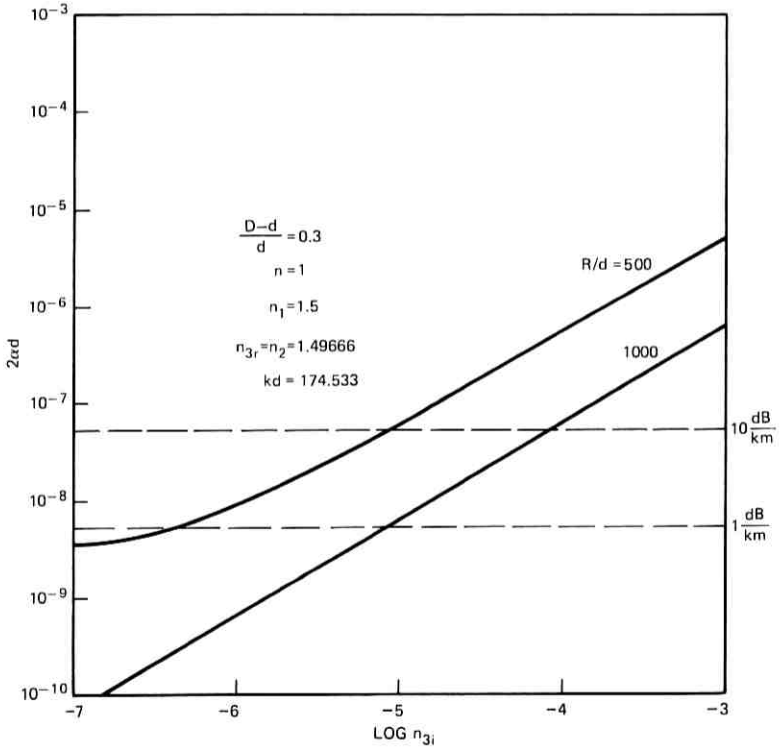

Fig. 10—Dependence of the curvature losses of the first mode on the imaginary part of the refractive index of the jacket material. The normalized cladding thickness is $(D - d)/d = 0.3$ and the refractive indices are $n_1 = 1.5$, $n_2 = 1.49666$, $kd = 174.533$, $n_{3r} = n_2$.

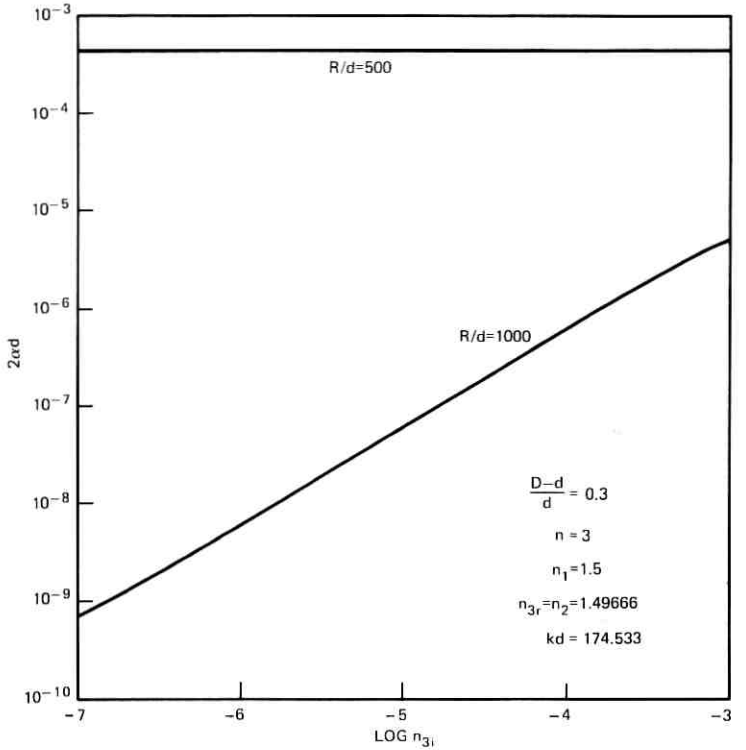

Fig. 11—Dependence of the curvature losses of the third mode on the imaginary part of the refractive index of the jacket material. The waveguide parameters are the same as in Fig. 10.

jacket. For $R/d = 500$, the field already radiates away in the space between core and jacket so that the losses are independent of the power dissipation in the jacket. Both figures are drawn for $(D - d)/d = 0.3$.

Finally, we discuss briefly a slab waveguide supporting $N = 24$ TE modes. We use once more $kd = 174.533$ and $n_1 = 1.5$, but choose the cladding index $n_2 = 1.485$. Figure 12 shows loss curves as functions of d/R for the mode $n = 17$ for several values of the cladding thickness $D - d$. The jacket is of the "high-index" type, with $n_{3r} = n_1$ since this condition results in high losses. Mode $n = 17$ separates the corresponding modes of the fiber in equal halves. We see from Fig. 12 that even for a straight guide half the fiber modes have losses in excess of 5 dB/km if the cladding thickness is $D - d = 0.3d$. For $D - d = 0.4d$, the losses of the straight guide are reasonably low; they become im-

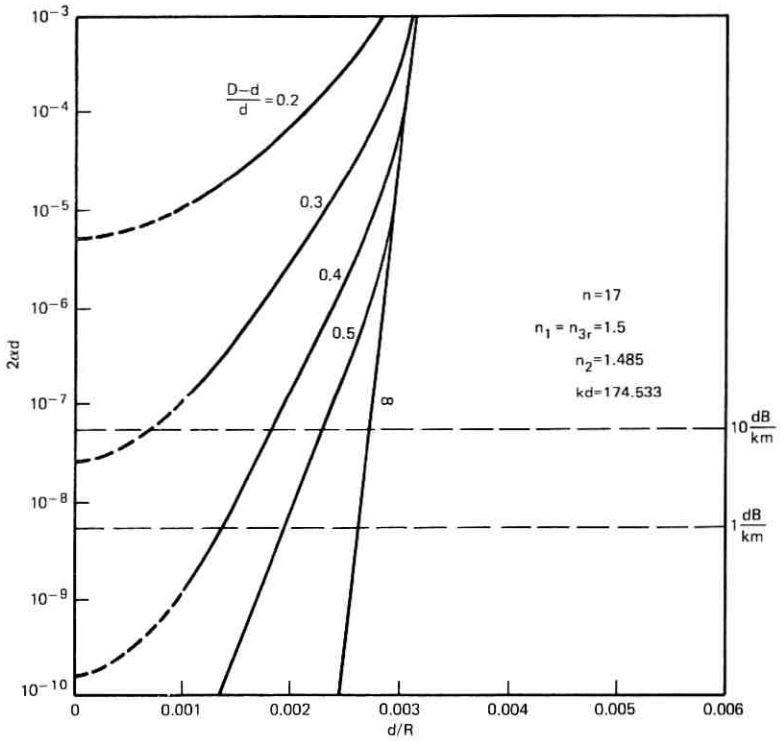

Fig. 12—Curvature losses in the presence of a jacket with $n_{3r} = n_1$ for mode $n = 17$ of a slab supporting 24 TE modes. The other parameters are $n_1 = 1.5$, $n_2 = 1.485$, $kd = 174.533$.

portant for $d/R > 0.0015$. The larger index difference of this example has the effect of allowing us to use a slightly thinner cladding and lower radii of curvature compared to the previous case with $n_2 = 1.49666$.

VII. CONCLUSIONS

The presence of a jacket can increase the curvature losses of dielectric optical waveguide. It is thus important to keep the jacket sufficiently far from the waveguide core. The worst possible case is that of a jacket whose refractive index is slightly higher than the index of the cladding. However, we see an increase of the curvature losses caused by the power dissipation in the lossy jacket even if the real part of the index of the jacket is equal to the cladding index.

A first indication of trouble can be obtained by computing the losses caused by the presence of the jacket from formula (29) for a straight slab waveguide. (A corresponding formula for the HE_{1n} modes of the round optical fiber can be found in eq. (10.4–22), p. 426, of Ref. 4.) Even if the presence of a lossy jacket does not seem to increase the losses of the straight guide above a certain tolerable level, it is important to keep in mind that waveguide curvature will increase the values of the loss coefficient by orders of magnitude for sufficiently tight bends.

The discussion of curvature losses in the presence of a lossy jacket was based on considering TE modes of a slab waveguide. The general behavior of the losses is expected to be the same for round optical fibers. Since experience has shown that even the numbers obtained from a slab model give the correct order of magnitude for round fibers, the numerical example discussed in this paper may be used to estimate the curvature losses of a round fiber with lossy jacket.

REFERENCES

1. D. Marcuse, "Attenuation of Unwanted Cladding Modes," B.S.T.J., *50*, No. 8 (October 1971), pp. 2565–2583.
2. E. A. J. Marcatili, "Bends in Optical Dielectric Waveguides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2103–2132.
3. M. A. Miller and V. I. Talanov, "Electromagnetic Surface Waves Guided by a Boundary with Small Curvature," Zh. Tekh. Fiz., *26*, No. 12, 1956, p. 2755.
4. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold, 1972, p. 290.
5. M. Abromovitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, Vol. 55, Washington D.C.: National Bureau of Standards, 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

Spectrum of a Binary Signal Block Coded for DC Suppression

By L. J. GREENSTEIN

(Manuscript received January 8, 1974)

This paper analyzes a block-coding scheme designed to suppress spectral energy near $f = 0$ for any binary message sequence. In this scheme, the polarity of each block is either maintained or reversed, depending on which decision drives the accumulated digit sum toward zero. The polarity of the block's last digit informs the decoder as to which decision was made.

Our objective is to derive the average power spectrum of the coded signal when the message is a random sequence of $+1$'s and -1's and the block length (M) is odd. The derivation uses a mixture of theoretical analysis and computer simulation. The theoretical analysis leads to a spectrum description in terms of a set of correlation coefficients, $\{\rho_q\}$, $q = 1, 2$, etc., with the ρ_q's functions of M. The computer simulation uses FFT algorithms to estimate the power spectrum and autocorrelation function of the block-coded signal. From these results, $\{\rho_q\}$ is estimated for various M. A mathematical approximation to ρ_q in terms of q and M is then found which permits a closed-form evaluation of the power spectrum. Comparisons between the final formula and simulation results indicate an accuracy of ± 5 percent (± 0.2 dB) or better.

The block-coding scheme treated here is of particular interest because of its practical simplicity and relative efficiency. The methods used to analyze it can be applied to other block-coding schemes as well, some of which are discussed here for purposes of comparison.

I. INTRODUCTION

1.1 Block coding

In its most general meaning, block coding consists of dividing a digital sequence into time-contiguous blocks and performing a separate coding operation on each block. In actual usage, the term is

most often applied to cases in which both the original and block-coded sequences are binary and the coding serves either (*i*) to enable error detection and/or correction or (*ii*) to shape the digital sequence spectrum.

The most widely used spectrum-shaping codes are those that suppress the energy near zero frequency. This suppression enables the use of transformers and ac-coupled amplifiers in processing the digital signal and, in modulation applications, provides for a null region near the carrier frequency to facilitate carrier extraction. Our concern here is with this kind of block coding.

In particular, we examine a block-coding approach invented (in analog form) by F. K. Bowers.[1] The digital version of this scheme has been treated separately by Carter[2] and Pierce,[3] and implemented recently by Ruthroff and Bodtmann.[4,5] The scheme can be used with blocks of either odd or even length (M), but our attention here is confined to odd M. Our objective is to derive the average power spectrum of a sequence so coded when the original message sequence is totally random. This problem has been partially studied by Rice[6] for the same block coding with M even (to which case our method of analysis is equally applicable), and by Slepian[7] and Franklin and Pierce[8] for other dc-suppressing block codes. Results for some of these cases are given later.

1.2 Description

Unless otherwise specified, the term block coding means the process we describe here, with the aid of Figs. 1 and 2.

The original sequence of binary digits is divided into blocks of length $(M - 1)$, and a $+1$ digit (the so-called code digit) is added to the end of each block (Fig. 1a). A resettable counter measures the digit sum A_k in each block k, omitting the code digit if M is even and including that digit if M is odd. In either case, this count can take on only odd values. It is compared with the sum over all previous output digits, B_k, and a decision is made as follows: If A_k and B_k have the same polarity, all pulses in block k are inverted; if A_k and B_k have opposite polarity, the pulses in block k are unaltered; and if $B_k = 0$, its polarity is taken to be that of its most recent nonzero value, i.e., B_{k-1}, or B_{k-2}, etc.* In the decoder, each received block is inverted if the polarity of

* There are other ways to resolve the case $B_k = 0$ (e.g., by random decisions, as suggested by Rice), and the ultimate choice should be dictated by practical considerations.

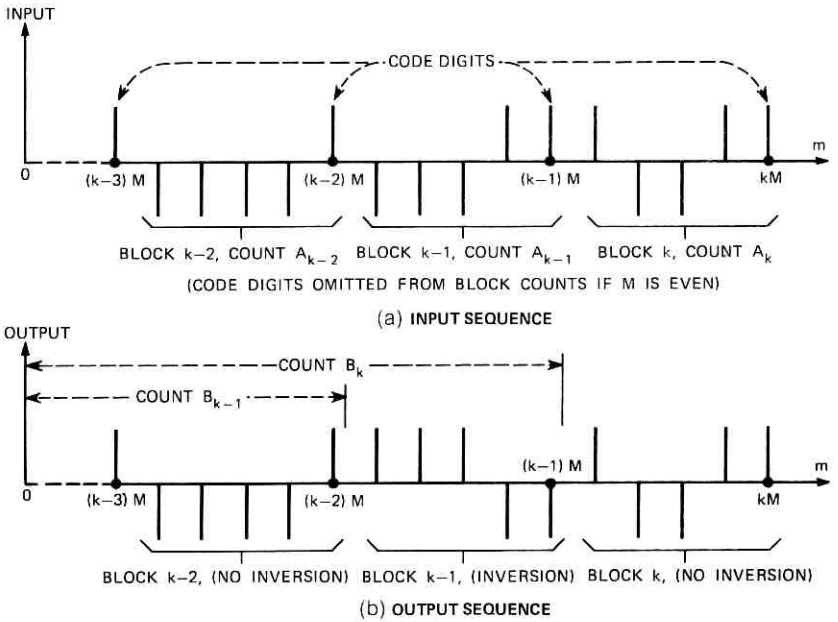

Fig. 1.—Sample input and output digital sequences.

the recovered code digit is negative and is not inverted if that digit is positive.

Figure 2 depicts the logical process just described and is a simplified diagram of how block coding is actually implemented. The identification of the code digit in the decoder is accomplished with the help of framing, which is not depicted (or treated) here. The penalties in this form of block coding are a $100/M$-percent reduction in information rate and twofold increase in the random error rate.

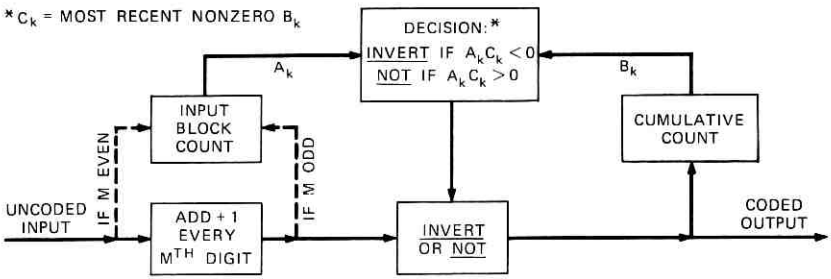

Fig. 2—Coding process.

1.3 Scope of the paper

It is easy to see that the accumulated sum of the output digits at the end of any block is limited in magnitude to M (i.e., $|B_k| \leqq M$), and that the magnitude of this sum at *any* instant in time is limited by $(3M - 1)/2$ if M is odd and $3M/2$ if M is even. For this reason, the output sequence has no spectral energy—either discrete or continuous—at $f = 0$. At the same time, the total sequence "power" is unchanged by the coding, since every digit has the same "energy," regardless of polarity.* Obviously, then, the suppressed energy near $f = 0$ is redistributed over the rest of the frequency range, and it is of more than passing interest to know how.

The answer, of course, depends on the nature of the message sequence being encoded. In this study, we assume a totally random sequence (all digits independent, with equally likely polarities) and derive the block-coded signal spectrum for odd values of M. The derivation uses a mixture of theoretical analysis and computer simulation and leads to a closed-form expression for the spectrum which compares quite favorably with simulation data. Section II gives the purely theoretical part of the derivation, Section III describes the simulation study, and Section IV gives the final result and some examples.

II. ANALYSIS

2.1 General form of the spectrum

We represent the uncoded message sequence as a binary stream of pulses at a rate $1/T$,

$$s_i(t) = \sum_{n=-\infty}^{\infty} a_n p(t - nT), \tag{1}$$

where

$$a_n = +1 \text{ or } -1 \text{ with equal probabilities,}$$

$$\overline{a_n a_m} = \begin{cases} 1 \text{ if } n = m \\ 0 \text{ if } n \neq m, \end{cases} \tag{2}$$

and $p(t)$ is a pulse centered on $[0, T]$ of arbitrary shape, area T, and Fourier transform $P(f)$. The first step in the coding consists of opening up a one-pulse slot after every $M - 1$ message pulses and injecting a positive pulse, $+p(t)$. The new sequence, with positive code pulses

* In the ensuing analysis, the digital sequence of Fig. 1 is replaced by a pulse stream at a rate $1/T$, with each pulse having an area of magnitude T.

every M time slots, is then block coded to produce an output signal

$$s_o(t) = \sum_{n=-\infty}^{\infty} b_n p(t - nT). \tag{3}$$

It is the average power spectrum of $s_o(t)$ that we wish to evaluate.

We define the autocorrelation function of the coded signal to be

$$R(\tau) = \lim_{T_o \to \infty} \left\{ \frac{1}{2T_o} \int_{-T_o}^{T_o} s_o(t) s_o(t + \tau) dt \right\} \tag{4}$$

and the average power spectrum to be the Fourier transform of $R(\tau)$,

$$S(f) = \mathfrak{F}\{R(\tau)\} = \int_{-\infty}^{\infty} R(\tau) \exp(-j\omega\tau) d\tau. \tag{5}$$

To simplify the derivation of $S(f)$, it is convenient to express $s_o(t)$, eq. (3), as the convolution

$$s_o(t) = \underbrace{\left\{ \sum_{n=-\infty}^{\infty} b_n \delta(t - nT) \right\}}_{s_u(t;\, T)} * p(t), \tag{6}$$

where $\delta(t)$ is the unit impulse function. It is now obvious that $S(f)$ is the product

$$S(f) = S_u(f; T) |P(f)|^2, \tag{7}$$

where $S_u(f; T)$ is the average power spectrum of $s_u(t; T)$ in (6), and $P(f)$ is the Fourier transform of $p(t)$.

We can obtain $S_u(f; T)$ by applying (4) and (5) to $s_u(t; T)$. In so doing, we make use of the fact that the convolution between two unit impulse functions separated by mT seconds is a unit impulse function $\delta(t - mT)$. It is then easy to show that

$$S_u(f; T) = \mathfrak{F}\underbrace{\left\{ \frac{1}{T} \sum_{m=-\infty}^{\infty} \left[\lim_{N \to \infty} \frac{1}{2N} \sum_{n=-N}^{N} \overline{b_n b_{n+m}} \right] \delta(\tau - mT) \right\}}_{R_u(\tau;\, T)}, \tag{8}$$

where $\overline{b_n b_{n+m}}$ is an average over the sequence ensemble and the bracketed term is the further averaging over the time position of b_n. Because $R_u(\tau; T)$ is a sequence of uniformly spaced impulses, $S_u(f; T)$ is periodic in frequency with a repetition interval $1/T$. The shaping of this spectrum by the nonperiodic pulse spectrum function $|P(f)|^2$ leads to the overall spectral characteristic of the block-coded signal.

2.2 Analysis of $R_u(\tau; T)$

This analysis is aimed at simplifying $R_u(\tau; T)$, eq. (8), by finding a description for $\overline{b_n b_{n+m}}$ and its average over n. To do this, we make two important observations about the block-coded binary sequence $\{b_n\}$:

(*i*) The 2^M possible digital sequences within each code block of $\{b_n\}$ are equiprobable, even though the last digit is a code digit.

(*ii*) The correlation between b_n and b_{n+m} depends only on the number of blocks separating these two digits, i.e., on the number (q) of code digits in the interval $[n, n + m)$.

The first observation is easily proved: The first $M - 1$ digits of each block at the coder input, which are assumed to be totally random, form one of 2^{M-1} equiprobable sequences. With the addition of the $+1$ code digit, there are still only 2^{M-1} realizable sequences per block. The possible inversion of the block by the coder, however, produces another 2^{M-1} realizable sequences (the original 2^{M-1} sequences with -1 instead of $+1$ for the last digit), leading to a total of 2^M. Further, since the probability of a block inversion is $\frac{1}{2}$ for a random input sequence, the 2^M realizable output sequences are equiprobable. The significance of this is that $\{b_n\}$ is statistically the same as if all M digits in each input block were derived by random selection.

The second observation depends on the first. For in the absence of block inversions and with all input digits randomly derived, there would be no correlation between any two digits of the digital stream. Any correlations in the block-coded sequence, therefore, are due solely to the inversions. It follows that $\overline{b_n b_{n+m}}$ depends, at most, on the number of possible block inversions (or code digits, q) between b_n and b_{n+m}.

We conclude that $R_u(\tau; T)$ can be expressed in terms of a set of numbers $\{\rho_q\}$, ρ_q being the correlation between any two digits having q code digits between them.* To reduce (8) to such a representation, we first observe that, if $|m| = lM + p$, where $1 \leqq p \leqq M$, then q is either l or $l + 1$, depending on the position of b_n within the block containing it. By letting n vary from the first to the last block position, we can see that $q = l$ for a fraction $[(M - p)/M]$ of all possible positions, and $q = l + 1$ for a fraction (p/M) of all possible positions. We can therefore express the bracketed quantity in (8) as

$$\left[\lim_{N\to\infty} \frac{1}{2N} \sum_{n=-N}^{N} \overline{b_n b_{n+m}}\right] = \left(\frac{M-p}{M}\right)\rho_l + \left(\frac{p}{M}\right)\rho_{l+1}, \qquad (9)$$

* It is obvious that $\rho_0 = 0$, because any two digits in the same input block are taken to be uncorrelated, and this fact is not altered by the coding.

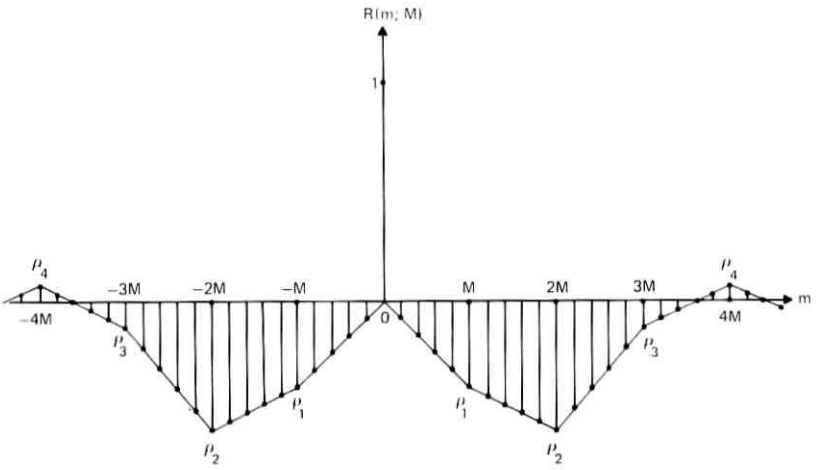

Fig. 3—Representation of $R(m; M)$.

where

$$l \triangleq \text{greatest integer in } \frac{|m|}{M}. \tag{10}$$

Since $p = |m| - lM$, the right side of (9) can be re-expressed as

$$R(m; M) \equiv \frac{(l+1)M - |m|}{M} \rho_l + \frac{|m| - lM}{M} \rho_{l+1}, \tag{11}$$

with l related to m and M by (10).

It is clear from the statistical symmetry of the b's in (9) that $R(-m; M) = R(m; M)$, and (11) reflects this fact. It is also clear from (9) that $R(0; M) = 1$. We can thus express $R_u(\tau; T)$ in (8) as

$$R_u(\tau; T) = \frac{1}{T}\left[\delta(\tau) + \sum_{m=1}^{\infty} R(m; M)\{\delta(\tau + mT) + \delta(\tau - mT)\}\right]. \tag{12}$$

Although the mathematical description for $R(m; M)$, eq. (11), seems complicated, it has the very simple graphical interpretation shown in Fig. 3. The value of $R(m; M)$ for q complete block separations (i.e., $m = \pm qM$) is just ρ_q,* and the variation between $m = qM$ and $m = (q+1)M$ is a linear progression from ρ_q to ρ_{q+1}. This result can now be used to derive $S_u(f; T)$.

---

* The one exception to this is the singular case $q = 0$, where $R(m = 0; M) = \rho_0 + 1$, with $\rho_0 = 0$.

2.3 Expression for $S_u(f; T)$

If the lines in Fig. 3 are envisioned as impulses of area $1/T$ separated by T seconds, then $S_u(f; T)$ can be found as the Fourier transform of this sequence. The algebra is straightforward but somewhat tedious, and so we give just the final result:

$$S_u(f; T) = \frac{1}{T}\left[1 + 2M\left(\frac{\sin M\omega T/2}{M \sin \omega T/2}\right)^2 \sum_{q=1}^{\infty} \rho_q \cos \omega MT\right]. \quad (13)$$

The periodicity of this function, with repetition interval $1/T$, is easy to see. The ρ_q's are functions of M so that a complete description of $S_u(f; T)$, and thus of $S(f)$ as given by (7), reduces to knowing the array of functions $\{\rho_q(M)\}$. Unfortunately, there is no apparent way to determine these functions from purely theoretical considerations. One useful bit of information, however, is that $S_u(0; T) = 0$ by virtue of the block coding.* This being the case, we see from (13) that

$$\sum_{q=1}^{\infty} \rho_q(M) = -\frac{1}{2M}. \quad (14)$$

Beyond (7), (13), and (14), we have little information about the block-coded signal spectrum on theoretical grounds. Using the methods of computer simulation, however, it is possible to estimate the ρ_q's for various M, and to seek functional descriptions for them that permit a closed-form evaluation of (13). This task constitutes the remainder of the development.

III. SIMULATION STUDY

3.1 Computer programs

The computer programs used to derive $\{\rho_q(M)\}$ empirically are depicted in Fig. 4. The routine called BLOCK generates random sequences $\{a_n\}$ having the properties described by (2) and, for specified M, converts them into block-coded sequences $\{b_n\}$ by emulating the logic in Fig. 2. These coded sequences are supplied on demand to the main program, labelled SIMULATION.

The SIMULATION program operates in the following manner: In each of N_T trials, it accepts an N-term sequence from BLOCK and performs an N-point discrete Fourier transform (DFT[9]), producing complex spectral samples at $f = k/NT$, $k = 0, N - 1$. The squared magnitude of the kth sample (normalized by N) represents a one-trial

* This is so because the long-time integration of the coded sequence is bounded in magnitude (specifically, by $3M/2$).

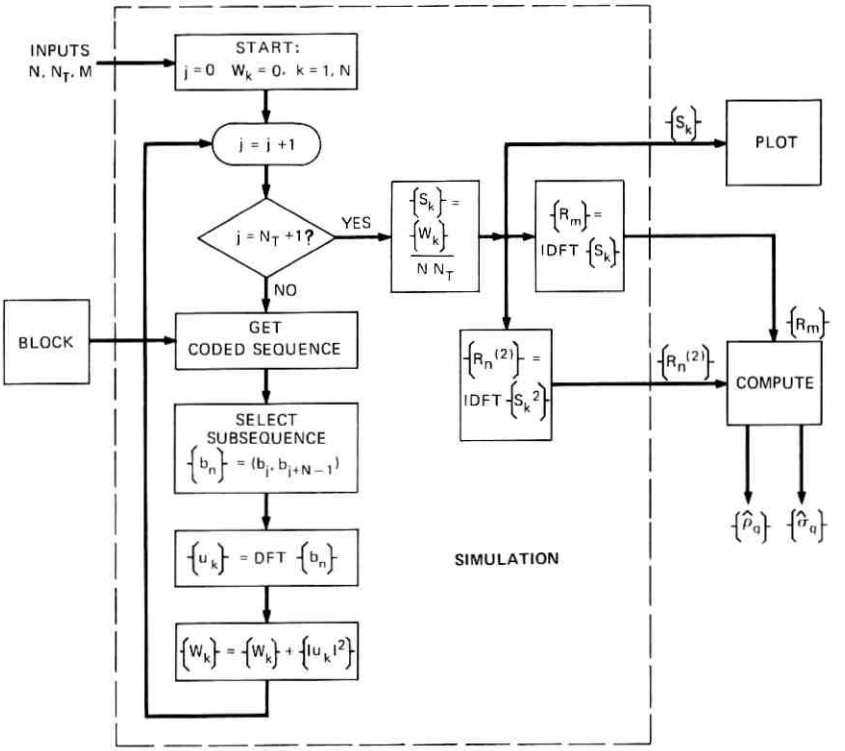

Fig. 4—Computer simulation program.

estimate of $TS_u(f = k/NT; T)$. These estimates for the N_T trials are averaged to produce the array $\{S_k\}$, $k = 0, N - 1$. This procedure is made efficient by implementing the DFT's with fast Fourier transform (FFT) algorithms.[10] To maximize efficiency, N is constrained in all simulations to be an integral power of 2.

The full benefit of the multitrial averaging is obtained by enforcing statistical independence among the N_T sequences supplied by BLOCK, and also by effectively randomizing the time phase of the sequences analyzed. The latter is accomplished by means of the function labelled "SELECT SUBSEQUENCE . . ." (Fig. 4), which causes the starting time of the analyzed sequences to vary uniformly among the M possible positions within a block. To accommodate this feature, the independent sequences supplied by BLOCK have a total length $N + N_T - 1$ or greater.

There are three output arrays produced in the SIMULATION program. One is $\{S_k\}$, which approximates $TS_u(f; T)$ at the N frequencies

k/NT, $k = 0, N - 1$. This array is applied to the PLOT routine, Fig. 4, where it is plotted on the same graph as the mathematical function given by (28). This function is derived using numerical parameters extracted from $\{R_m\}$, the inverse DFT of $\{S_k\}$, which is the second output array, Fig. 4. The elements of this array represent estimates of $R(m; M)$, i.e., $R_m \doteq R(m; M)$, and they are applied to the COMPUTE routine to produce estimates of ρ_q, i.e., the array $\{\hat{\rho}_q\}$. These are the quantities used to derive (28) from (13). To evaluate the accuracy of these estimates, the COMPUTE routine also produces the array $\{\hat{\sigma}_q\}$, where $\hat{\sigma}_q$ is the approximate rms error in $\hat{\rho}_q$. The computation of $\{\hat{\sigma}_q\}$ involves the array $\{R_n^{(2)}\}$, the inverse DFT of $\{S_k^2\}$, which is the third output of the SIMULATION program. The formulas relating $\{R_m\}$, $\{\hat{\rho}_q\}$, $\{R_n^{(2)}\}$, and $\{\hat{\sigma}_q\}$ are presented in Section 3.3.

3.2 Choices of N and N_T

The difference between the computer-derived spectral sample S_k and the quantity it approximates, $TS_u(k/NT; T)$, contains two distinct components, (*i*) a deterministic error due to the finiteness (N) of the sample length and (*ii*) a random error due to the finiteness (N_T) of the number of independent simulations. Similar remarks apply to the difference between R_m and $R(m; M)$. We now apply these considerations to the choices of N and N_T.

Because N is finite, the normalized spectrum estimated by the computer program is not $TS_u(f; T)$, but the convolution between $TS_u(f; T)$ and the function

$$F(f) = NT\left(\frac{\sin \pi NfT}{\pi NfT}\right)^2. \tag{15}$$

Thus, S_k is an estimate of the quantity

$$TS_u'\left(\frac{k}{NT}; T\right) = \int_{-\infty}^{\infty} TS_u(f; T)F\left(f - \frac{k}{NT}\right)df. \tag{16}$$

Since the area of $F(f)$ is unity, the approximation of $TS_u(k/NT; T)$ by $TS_u'(k/NT; T)$ is very good if $TS_u(f; T)$ changes negligibly over the main lobe of $F(f - k/NT)$. The difference is the deterministic error in S_k; the inverse DFT of the k-sequence of these errors gives the deterministic errors in the estimates of $R(m; M)$.

By considering the interference between the peak of $TS_u(f; T)$ (which occurs near $f = 1/4MT$) and the sidelobes of $F(f)$ (which decrease as $1/f^2$), we have determined a rule of thumb for which worst-case deterministic errors are negligibly small. The rule constrains N

to the region

$$N \geqq 100\, M, \tag{17}$$

a constraint we have used throughout this study.

The random errors in estimating $R(m; M)$ also decrease with increasing N, as we shall see in Section 3.3. This consideration, added to (17) and the requirement that N be an integral power of 2, has helped to shape the final choices of N for different values of M. Typically, we have used $N = 512$ for $M = 3$, $N = 2048$ for $M = 9$, and $N = 4096$ for $M = 17$.

We shall also see in Section 3.3 that the random errors in estimating $S_u(f; T)$ and $R(m; M)$ decrease with increasing N_T. For example, the fractional rms error in S_k is accurately given by $1/\sqrt{N_T}$. In deriving estimates for $R(m; M)$, we have used 400 trials to achieve the accuracies desired, while, to obtain precise spectral estimates for comparison with the final formula, we have used 1600 trials (corresponding to ± 2.5 percent accuracy).

3.3 Analysis of computer results

The simulation estimates of $R(m; M)$ for $m > 0$ and $M = 3$, 9, and 17 are given by the points in Figs. 5, 6, and 7. The existence of straight-line variations between $m = qM$ and $m = (q + 1)M$ for $q = 0$, 1, 2, etc., as predicted by the analysis of Section II, is evidenced here. The deviations of the points from straight lines are due to statistical fluctuations in the finite simulation, and the straight lines shown are derived from the points by least-squares techniques. The pertinent error analyses and estimation procedures used to obtain these straight lines and further data reductions are now summarized. We assume from here on that deterministic errors are made negligible by the choice of N, i.e., that all errors in S_k and R_m are random errors due to finite N_T. Table I lists the symbols to be used.

3.3.1 Error correlations

We now establish the error correlations $\overline{\delta_k \delta_l}$ and $\overline{\epsilon_m \epsilon_p}$ with the aid of the definitions in Table I. As N becomes very large, the real and imaginary parts of u_k become more and more like independent gaussian variates (central limit theorem[11]), and we assume this to be the case here. The importance of this assumption is that the definitions of δ_k, S_k, and $\overline{S_k}$ in Table I can then be used to obtain

$$\overline{\delta_k \delta_l} = \frac{\overline{|u_k u_l^*|^2}}{N^2 N_T}. \tag{18}$$

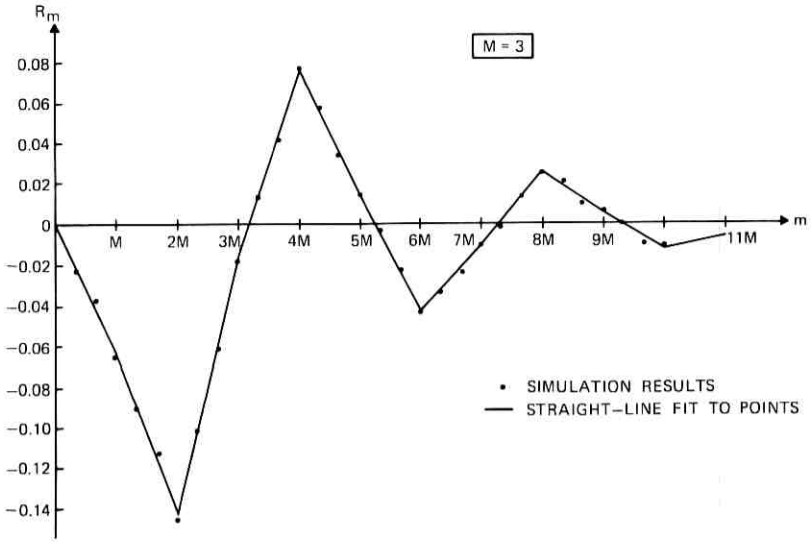

Fig. 5—Simulation results for $M = 3$ ($N = 512$, $N_T = 400$ trials).

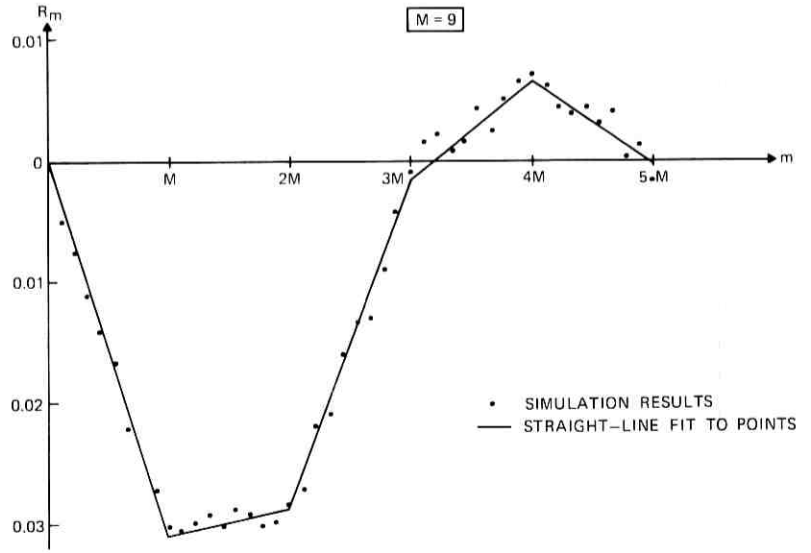

Fig. 6—Simulation results for $M = 9$ ($N = 2048$, $N_T = 400$ trials).

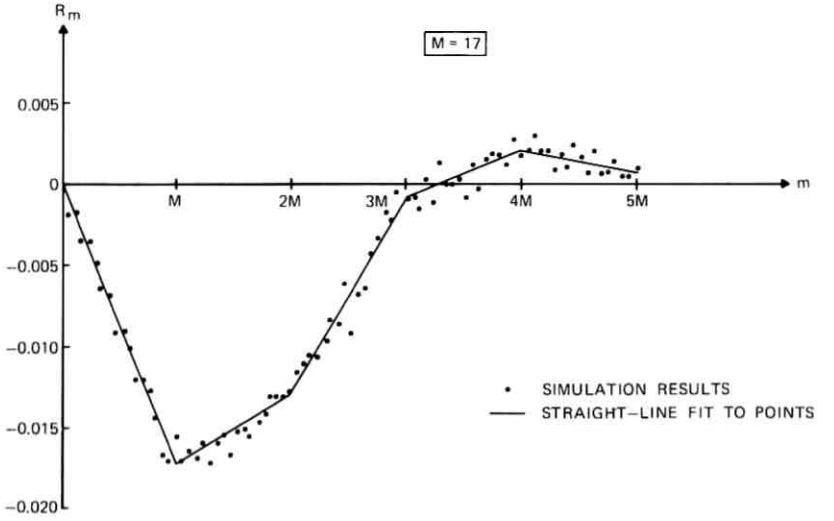

Fig. 7—Simulation results for $M = 17$ ($N = 4096$, $N_T = 400$ trials).

Clearly, $\overline{\delta_k^2} = (\overline{S_k})^2/N_T$, i.e., the rms error in estimating $\overline{S_k}$ is $\overline{S_k}/\sqrt{N_T}$. Unfortunately, $\overline{\delta_k \delta_l}$ for $l \neq k$ is not zero for the block-coded signal under study. Nevertheless, these correlations are found to be sufficiently weak (particularly as $|k - l|$ increases) that we can ignore them without any first-order effects on the results. The benefit of this is a considerable simplification in the mathematics. Accordingly, we shall assume that

$$\overline{\delta_k \delta_l} \doteq \begin{cases} (\overline{S_k})^2/N_T; & l = k \\ 0; & l \neq k. \end{cases} \tag{19}$$

Table I — Symbols used in error analysis

Symbol	Meaning
$\{u_k\}$, $k = 0, N-1$	DFT of $\{b_n\}$, $n = 0, N-1$
$S_u(f; T)$	Average power spectrum of infinitely long block-coded signal
S_k	Simulation estimate of $S_u(f = kT/N; T)$: $S_k = \underset{N_T}{\text{Ave}}(\lvert u_k \rvert^2/N)$
$\overline{S_k}$	Limiting value of S_k as $N_T \to \infty$: $\overline{S_k} = \overline{\lvert u_k \rvert^2}/N$
δ_k	Random error in S_k: $\delta_k = S_k - \overline{S_k}$
$R(m; M)$	Coefficient, at separation m, of autocorrelation function of block coded signal
R_m	Simulation estimate of $R(m; M)$: $\{R_m\} = \text{IDFT}\{S_k\}$
ϵ_m	Random error in R_m: $\epsilon_m = R_m - R(m; M) = \text{IDFT}\{\delta_k\}$
ρ_q, $\hat{\rho}_q$	Exact, estimated values of $R(m = qM; M)$

To establish $\overline{\epsilon_m \epsilon_p}$, we begin with the fact that $\{\epsilon_m\}$ is the inverse DFT of $\{\delta_k\}$. Using (19) and a modest amount of manipulation, we can then show that

$$\overline{\epsilon_m \epsilon_p} = \frac{2}{NN_T} \underbrace{[\text{IDFT}\{(\overline{S_k})^2\}|_{m+p}}_{R^{(2)}_{m+p}} + \underbrace{\text{IDFT}\{(\overline{S_k})^2\}|_{|m-p|}]}_{R^{(2)}_{|m-p|}}. \quad (20)$$

In the SIMULATION program (see Fig. 4), the array $\{R_n^{(2)}\}$ is estimated by computing the inverse DFT of $\{S_k^2\}$, and is then used to determine the rms errors in the estimates for $\rho_q(M)$.

3.3.2 Estimations of $\rho_q(M)$

Given the simulation estimates $\{R_m\}$, an obvious way to estimate $\rho_q (\equiv R(m = qM; M))$ is by $\hat{\rho}_q = R_{m=qM}$. However, the values of R_m for $m = qM - 1,\ qM - 2,\ \cdots qM - (M - 1)$ can be included to yield more accurate estimates of ρ_q, as the following analysis shows.

Using Table I and Fig. 3, we can express R_m in the general form

$$R_m = \rho_{q-1} + \frac{\rho_q - \rho_{q-1}}{M}[m - (q-1)M] + \epsilon_m; \quad (q-1)M < m \leqq qM$$
$$q = 1, 2, \text{ etc.} \quad (21)$$

Now suppose that we estimate ρ_q as a linearly weighted sum of R_m over $[(q-1)M + 1, qM]$. Using the substitution $m' = m + (q-1)M$, this sum can be written as

$$\hat{\rho}_q = \sum_{m=1}^{M} w_m R_{m'} = \rho_{q-1} \sum_{m=1}^{M} w_m + \frac{\rho_q - \rho_{q-1}}{M} \sum_{m=1}^{M} m w_m + \sum_{m=1}^{M} w_m \epsilon_{m'}. \quad (22)$$

This estimate is made unbiased by choosing $\{w_m\}$ so that

$$\sum_{m=1}^{M} w_m = \begin{cases} \text{Arbitrary}, & q = 1 \\ 1, & q > 1 \end{cases} \quad \text{and} \quad \sum_{m=1}^{M} m w_m = M. \quad (23)$$

(For the singular case $q = 1$, there is no constraint on $\sum w_m$ because ρ_0 is known to be zero.) To see how to choose $\{w_m\}$ $(m = 1, M)$ within these constraints, we combine (20), (22), and (23) and obtain the following mean-square error for $\hat{\rho}_q$:

$$\sigma_q^2 = \sum_{m=1}^{M}\sum_{p=1}^{M} w_m w_p \overline{\epsilon_{m'} \epsilon_{p'}} = \frac{2}{NN_T} \sum_{m=1}^{M}\sum_{p=1}^{M} w_m w_p [R^{(2)}_{m'+p'} + R^{(2)}_{|m'-p'|}]. \quad (24)$$

To a first approximation, the dominant component of σ_q^2 is $2R_0^{(2)}(\sum w_m^2)/$

Table II — Estimates of $\rho_q(M)$

q \ M	3 ($\sigma \doteq 3.1\times10^{-3}$)	5 ($\sigma \doteq 1.8\times10^{-3}$)	9 ($\sigma \doteq 1.0\times10^{-3}$)	13 ($\sigma \doteq 6.1\times10^{-4}$)	17 ($\sigma \doteq 5.3\times10^{-4}$)
1	−0.06321	−0.04787	−0.03098	−0.02271	−0.00172
2	−0.14363	−0.06625	−0.02870	−0.01751	−0.00128
3	−0.01754	−0.00632	−0.00168	−0.00140	−0.00081
4	0.07654	0.02712	0.00659	0.00398	0.00212
5	0.01431	0.00515	−0.00018	0.00024	0.00066
6	−0.04212	−0.01120	−0.00198	−0.00077	−0.00091
7	−0.01074	−0.00200	−0.00075	0.00075	−0.00041
8	0.02653	0.00701	0.0	0.00002	0.00103
9	0.00546	−0.00018	0.00069	0.00039	0.00088
10	−0.01244	−0.00483	−0.00026	0.00069	0.00020
11	−0.00600	−0.00040	−0.00016	−0.00009	−0.00050
12	0.01124	0.00185	−0.00103	0.00017	0.00061
13	0.00275	0.00143	0.00183	0.00024	−0.00031
14	−0.00471	−0.00096	0.00009	0.00032	−0.00017
15	−0.00062	0.00022	−0.00008	0.00003	0.00018
16	0.00038	0.00200	−0.00062	0.00061	0.00027
17	0.00171	0.00016	0.00013	−0.00023	0.00017
18	0.00027	−0.00407	−0.00087	−0.00050	0.00004
19	0.00254	−0.00134	0.00039	0.00047	−0.00019
20	0.00024	0.00201	0.00057	0.00051	0.00024

NN_T, because $|R_n^{(2)}|$ tends to be small for $n \neq 0$. Using this fact, an approximate least-squares approach is to derive the sequence $\{w_m\}$ for which $\sum w_m^2$ is a minimum within the constraints of (23). Using Lagrangian multipliers, it is straightforward to show that the solution is

$$w_m \underset{(m=1,M)}{=} \begin{cases} -\dfrac{2}{M} + \dfrac{6m}{M(M+1)}\,; & q > 1 \\ \dfrac{6m}{(M+1)(2M+1)}\,; & q = 1. \end{cases} \tag{25}$$

We assume that, for practical purposes, (25) represents the least-squares coefficient array for the error given by (24). It was used in the COMPUTE routine of Fig. 4 to obtain $\{\hat{\rho}_q\}$ [based on (22) and the estimates $\{R_m\}$] and to estimate $\{\sigma_q\}$ [based on (24) and the estimates of $\{R_n^{(2)}\}$].

The results are shown in Table II for several values of M and for $q = 1, 20$. It is found that σ_q is fairly constant with q, except for σ_1, which tends to be lower by 10 to 30 percent. The quantity σ in each column heading of Table II is the average of the computed σ_q's from $q = 2$ to $q = 20$.* These rms errors are lower than those obtained

* Note that, for all M, $\hat{\rho}_q$ is in the simulation "noise" (i.e., $|\hat{\rho}_q| \leqq \sigma$) for $q \geqq 20$.

by estimating ρ_q as $R_{m=qM}$ [which is equivalent to using $\{w_m\} = (0, 0, \cdots, 1)$], the improvement factor increasing with M and having a value near 2.5 for $M = 17$.

3.4 Reduction to mathematical descriptions

From the data of Table II, a useful and valid description for $\hat{\rho}_q(M)$ can be shown to be

$$\hat{\rho}_q(M) \doteq \frac{A_q}{M} + \frac{B_q}{M^2}, \tag{26}$$

where A_q and B_q are functions solely of q. For each q, raw estimates of A_q and B_q are derivable from the $\hat{\rho}_q$ values at any two values of M. To satisfy (14) for all M, however, it is necessary that these estimates be

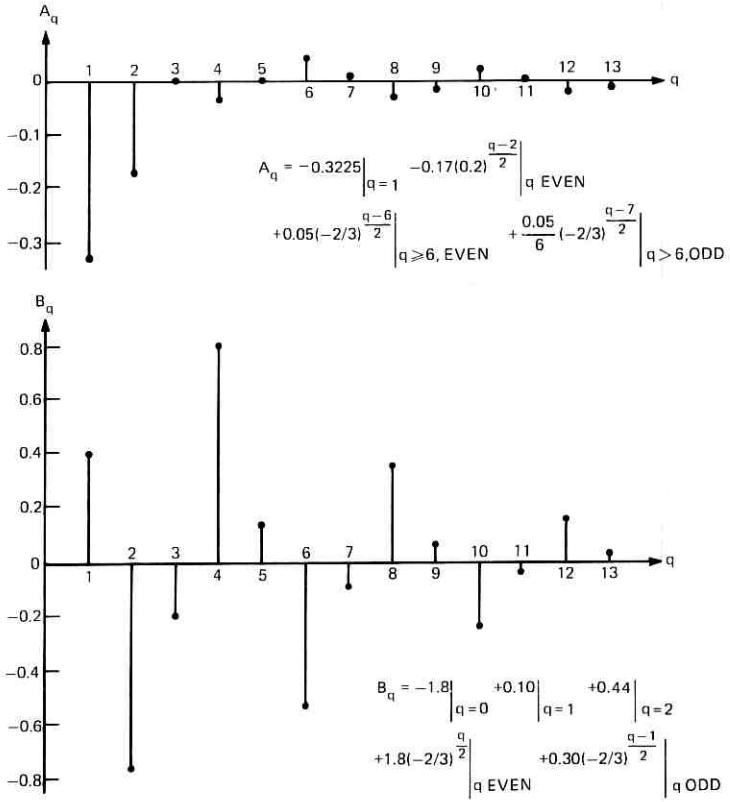

Fig. 8—Possible solutions for $\{A_q\}$ and $\{B_q\}$.

refined to satisfy

$$\sum_{q=1}^{\infty} A_q = -\frac{1}{2M}, \qquad \sum_{q=1}^{\infty} B_q = 0. \tag{27}$$

Moreover, these refined estimates should express A_q and B_q in mathematical forms that will permit a closed-form evaluation of (13). One possible solution that satisfies all these requirements is Fig. 8. Using these results in (26) and comparing with the values in Table II, agreement is found to be quite good: Nearly all the new estimates of $\hat{\rho}_q$ obtained in this way lie within a standard deviation ($\pm\sigma$) of the tabulated values; also, the values of $\hat{\rho}_q$ for $q > 20$ lie within $\pm\sigma$ about zero, decaying in magnitude with q as expected from physical reasoning.

IV. FINAL RESULT AND EXAMPLES

4.1 Expression for $S_u(f; T)$

Combining Fig. 8 with (26) and (13), it is possible to obtain a closed-form expression for $S_u(f; T)$. Once again, the algebra is tedious but straightforward, so we merely state the result:

$$TS_u(f; T) = 1 - \left(\frac{\sin(M\omega T/2)}{M \sin(\omega T/2)}\right)^2 F(\omega T; M), \tag{28a}$$

where

$$\begin{aligned} F(\omega T; M) &= 0.645 \cos(M\omega T) - \frac{0.88}{M}\cos(2M\omega T) \\ &\quad - 0.85\,\frac{\cos(2M\omega T) - 0.2}{\cos(2M\omega T) - 2.6} - \frac{1}{13 + 12\cos(2M\omega T)} \\ &\quad \times \left[\frac{2}{M}[-7.2 + 6.4\cos(M\omega T) - 10.8\cos(2M\omega T) \right. \\ &\quad + 0.6\cos(3M\omega T)] + 0.05[12\cos(4M\omega T) + 2\cos(5M\omega T) \\ &\quad \left. + 18\cos(6M\omega T) + 3\cos(7M\omega T)]\right]; \quad M \text{ odd.} \end{aligned} \tag{28b}$$

Rather than do an error analysis of this result (e.g., based on the σ's in Table II), we have compared this formula with fresh simulation results for $\{S_k\}$ based on 1600 trials (± 2.5 percent rms error). The PLOT routine shown in Fig. 4 plots the simulation data as points and plots the formula as a solid line. Figures 9 through 13 give the results for $M = 3$, 5, 9, 13, and 17. Given the rms errors of the simulations and the scatter about the solid curves, we estimate from these comparisons that the formula is accurate to within ± 5 percent (± 0.2 dB) or better for all M and ω. The accuracy is especially good in the all-

important rising portions near $f = 0$, where the simulation points are seen to lie very close to—or within the line thickness of—the solid curves.

4.2 Comparison with even-M block code

It is tempting to extrapolate the new formula to the case of even M, although Fig. 2 warns us that the coding schemes for odd and even M are qualitatively different. Figure 14 shows simulation points, along with a solid curve derived from the new formula, for $M = 4$. Figures 15 and 16 do the same for $M = 8$ and 16. Although N and N_T are lower in these simulations, the consequent increases in the deterministic and random errors do not account for the observed discrepancies. It is concluded that the new formula is not accurate for low even values of M, but that its accuracy improves as M increases to large even values.

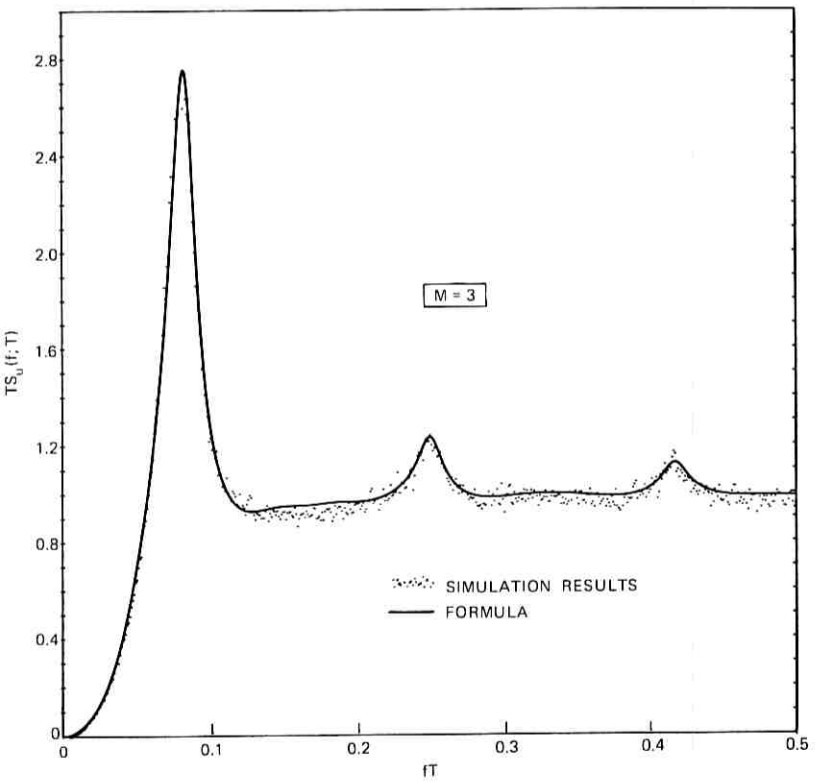

Fig. 9—Comparison of formula with simulation results for $M = 3$ ($N_T = 1600$).

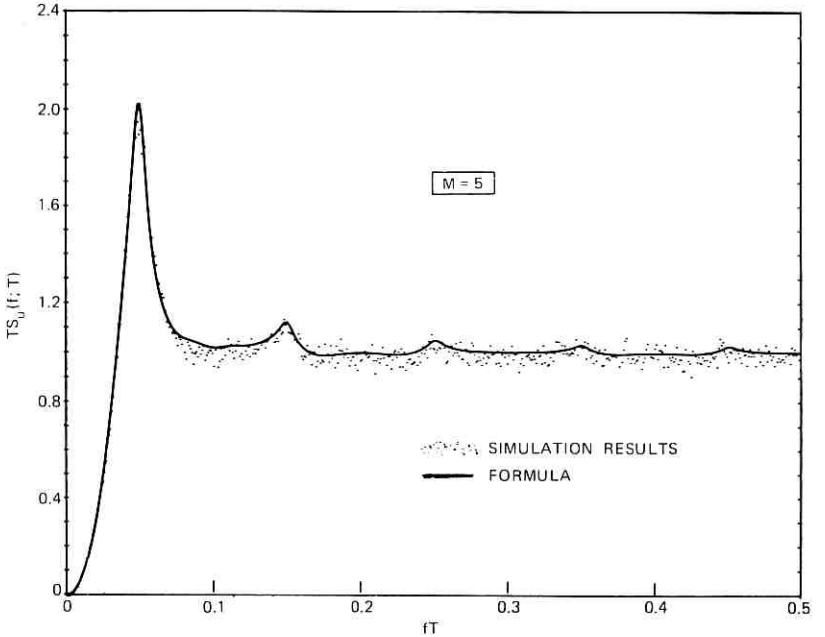

Fig. 10—Comparison of formula with simulation results for $M = 5$ ($N_T = 1600$).

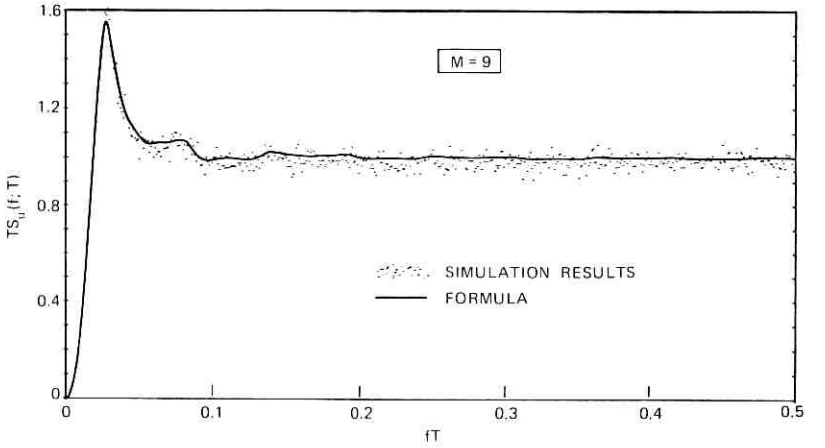

Fig. 11—Comparison of formula with simulation results for $M = 9$ ($N_T = 1600$).

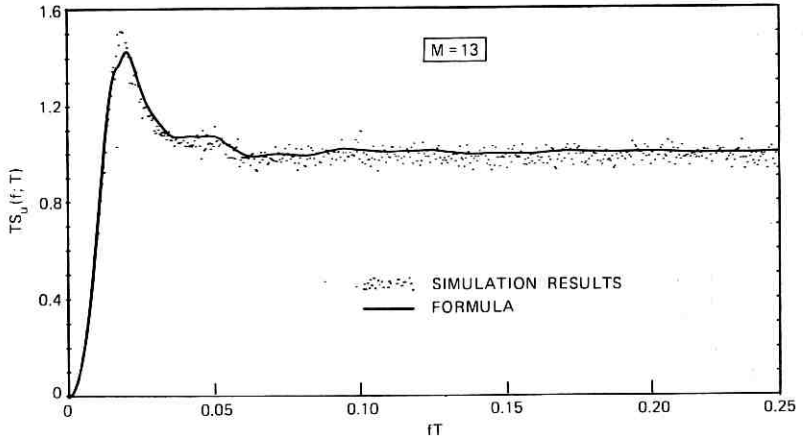

Fig. 12—Comparison of formula with simulation results for $M = 13$ ($N_T = 1600$).

4.3 Comparison with zero-disparity code

A different kind of block-coding scheme is one in which M is even and each block is constrained to have an equal number of positive and negative digits.[2] For a given M, this so-called zero-disparity code is less efficient in information rate than the one studied here (see Franklin and Pierce[8]), but has superior spectral properties, as we now show.

For the zero-disparity code, Franklin and Pierce show that $TS_u(f;T)$ is $M/(M-1)$ times the function (28a), with $F(\omega T;M)$ replaced by 1.

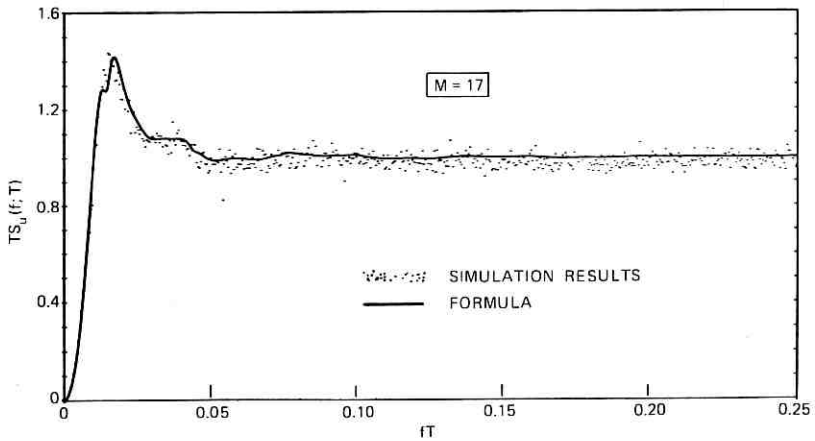

Fig. 13—Comparison of formula with simulation results for $M = 17$ ($N_T = 1600$).

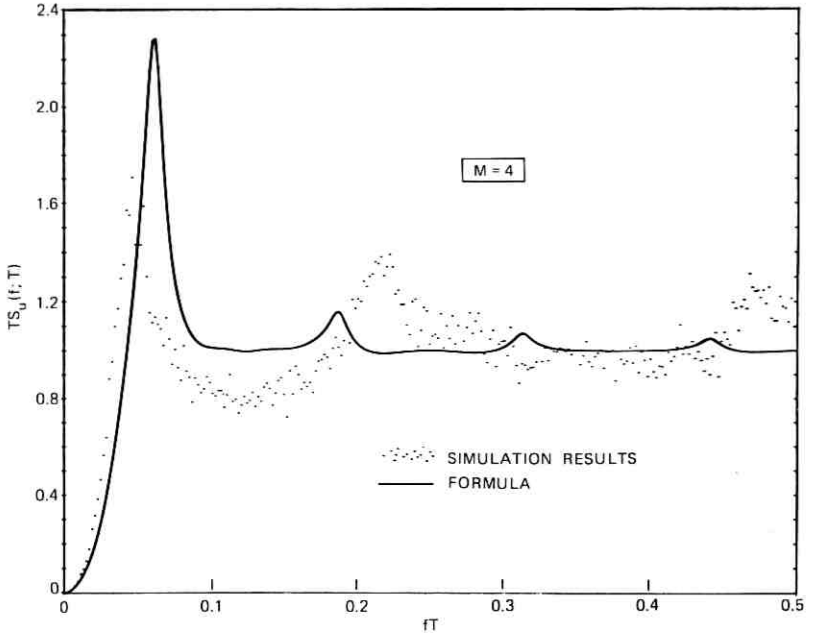

Fig. 14—Comparison of formula with simulation results for $M = 4$ ($N_T = 400$).

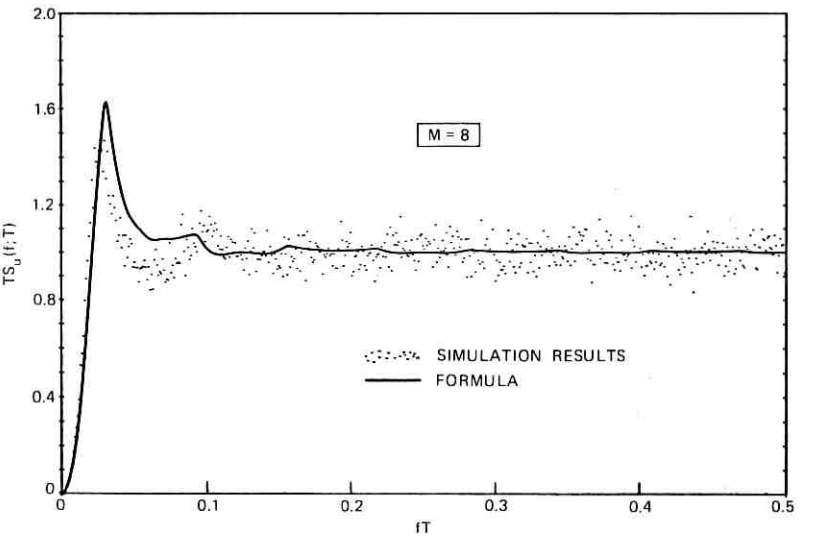

Fig. 15—Comparison of formula with simulation results for $M = 8$ ($N_T = 400$).

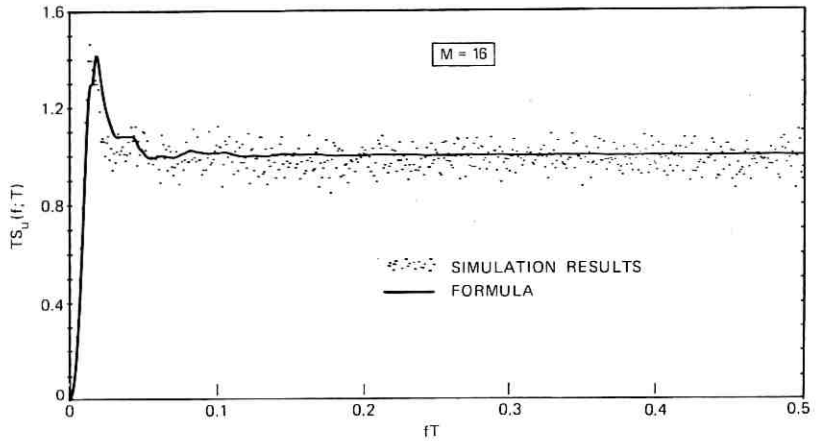

Fig. 16—Comparison of formula with simulation results for $M = 16$ ($N_T = 400$).

For $f \leqq 1/2T$ and $M \geqq 4$, this result can be represented to within 0.3 dB as follows:

$$\frac{M-1}{M}\, TS_u(f;T) \doteq 1 - \left(\frac{\sin\,(M\omega T/2)}{M\omega T/2}\right)^2. \tag{29}$$

This function is shown (dashed curve) in Fig. 17 and compared with

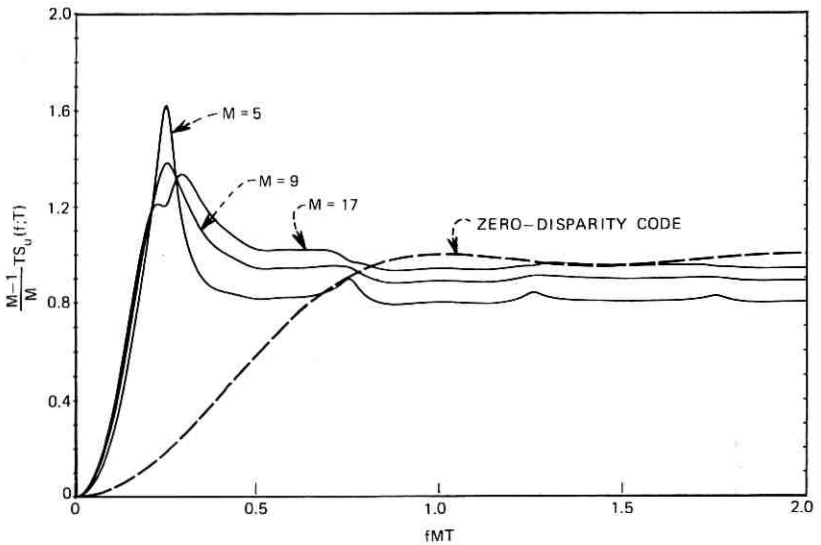

Fig. 17—Comparisons with zero-disparity block code.

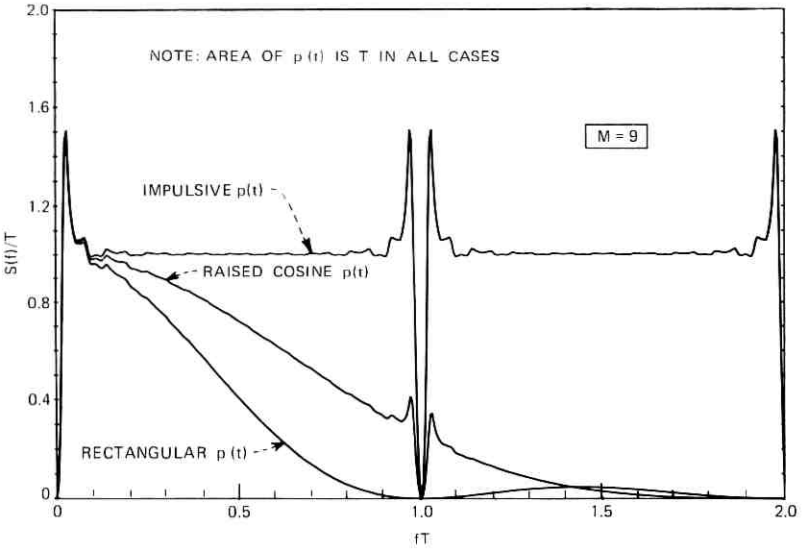

Fig. 18—Block-coded signal spectrum for various pulse shapes ($M = 9$).

some corresponding results for the block code studied here. The latter results, based on the new formula, are for $M = 5$, 9, and 17. The suppressed energy near $f = 0$ is seen to be redistributed to higher frequencies (and in a more uniform way) by the zero-disparity code, permitting more relaxed requirements on ac-coupled processing stages for a given M. If the quantity held fixed is information rate, however, the zero-disparity code must use a larger value of M for which its spectral superiority all but vanishes.

4.4 Effects of pulse shape

The overall spectrum of the block-coded signal must take into account the spectrum of the pulse shape $p(t)$. Figure 18 gives some results for $S(f)$, (7), for the case $M = 9$. The relative differences due to pulse shape are identical to those that occur without block coding. The effect of the block coding is to force spectral nulls near $f = n/T$, ($n = 0, 1, 2,$ etc.) and "bumps" within $\pm 1/4MT$ of each null.

V. CONCLUSION

The general analysis of Section II leading to (7) and (13) applies to a wide class of block-coding schemes aside from the one treated here. The simulation/analysis procedures described in Section III can like-

wise be applied to these schemes to find the ρ_q's. Unfortunately, the computer costs involved in accurately estimating the ρ_q's and comparing the resulting formula with simulation data can be quite high.

Aside from computer cost considerations, a strictly theoretical solution to this kind of problem would be more accurate and provide more insight into the correlation factors influencing this kind of random process. Although qualitative explanations can be given for the oscillating behavior of ρ_q with q (Table II), the approach described here requires and offers little insight into such phenomena.

In strictly practical terms, however, the result of this study provides a spectrum description which is quite accurate and fairly simple to use. For studies involving the passage of block-coded signals through ac-coupled amplifiers, or carrier extraction from signals modulated with block-coded sequences, such descriptions are highly useful.

VI. ACKNOWLEDGMENT

Numerous people have contributed to this study with suggestions and assistance. I am particularly indebted to S. O. Rice for his suggestions on combining theoretical analysis and computer simulation, to Steve Michael for his collaboration in designing the simulation program, and to Diane Vitello for her patience and skill in executing the various computer programs.

REFERENCES

1. F. K. Bowers, "Pulse Code Transmission System," Patent No. 2,957,947, issued Oct. 25, 1960.
2. R. O. Carter, "Low-Disparity Binary Coding System," Elec. Letters, *1*, No. 3 (May 1965), pp. 67–68.
3. J. R. Pierce, unpublished work.
4. C. L. Ruthroff and W. F. Bodtmann, "Adaptive Coding for Coherent Detection of Digital Phase Modulation," B.S.T.J., *53*, No. 3 (March 1974), pp. 449–466.
5. W. F. Bodtmann, unpublished work.
6. S. O. Rice, unpublished work.
7. D. Slepian, unpublished work.
8. J. N. Franklin and J. R. Pierce, "Spectra and Efficiency of Binary Codes without DC," IEEE Trans. on Comm., *COM-20*, No. 6 (Dec. 1972), pp. 1182–1184.
9. B. Gold and C. M. Rader, *Digital Processing of Signals*, New York: McGraw-Hill, 1969, Sec. 6.2.
10. Ref. 9, Secs. 6.4–6.6.
11. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965, Sec. 8.6.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

An Algorithm for Locating the Beginning and End of an Utterance Using ADPCM Coded Speech

By L. H. ROSENTHAL, R. W. SCHAFER, and L. R. RABINER

(Manuscript received December 10, 1973)

We describe a simple algorithm for locating the beginning and end of a speech utterance. The algorithm is based on the fact that the code words for an adaptive differential (ADPCM) representation of speech exhibit considerable variation among all quantization levels during both voiced and unvoiced speech intervals while, because of a constraint on the minimum step size, during silent intervals the code words vary only slightly within the smallest quantization steps.

The use of the algorithm is illustrated for automatically locating the beginning and end of vocabulary entries for a computer voice response system.

I. INTRODUCTION

The need to automatically locate the beginning and end of a speech utterance frequently arises in speech processing for automatic speech recognition and speaker verification. We have also encountered this problem in implementing an automatic vocabulary preparation scheme for a multiline computer voice response system.[1] Since our solution to the problem of automatically locating the endpoints of an utterance is based on some unique properties of the adaptive differential PCM (ADPCM) representation of speech waveforms,[2] we must first discuss the fundamentals of ADPCM waveform coding.

II. ADPCM SPEECH CODING

Two characteristics of speech signals that are of concern in digital coding are the wide range of amplitudes of speech sounds and the redundancy of the speech signal. The ADPCM coder depicted at the top of Fig. 1 is based on a conventional differential PCM structure

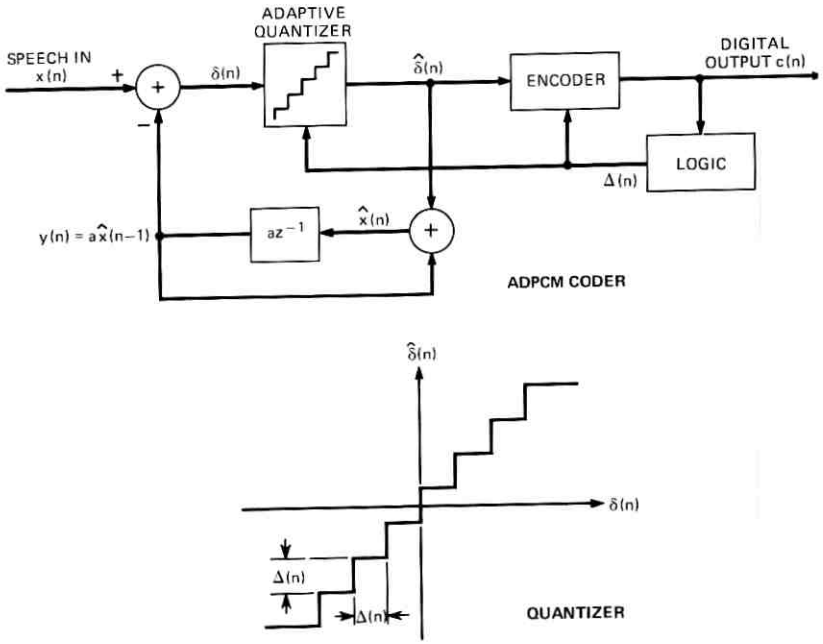

Fig. 1—Adaptive differential PCM (ADPCM) coder showing adaptive quantizer.

with a first-order fixed predictor in the feedback loop. To remove some of the redundancy, we form the difference between the input speech sample, $x(n)$, and an estimate of the input, $y(n)$. If the estimate of the input is good, then the difference should be small and, thus, more accurately represented by a fixed number of bits than the input samples. The difference signal is quantized and encoded for transmission or storage in a computer memory. An approximation to the input speech, $\hat{x}(n)$, is reconstructed by adding the quantized difference to the estimate $y(n)$. The next estimate of the input speech is obtained by linear prediction based on the previous value of the reconstructed signal, $\hat{x}(n-1)$.

The difference signal, although somewhat less redundant, still has a wide range of amplitudes. To make most efficient use of the quantization levels, the peak excursion of the signal should be matched to the range of the quantizer. Thus, for low-level signals such as fricatives, the absolute amplitude value of the step size should be small compared to that required for high-level voiced sounds. In our hardware implementation, we use a four-bit or 16-level quantizer; however, for simplicity we show a three-bit quantizer at the bottom of Fig. 1.

The block labelled LOGIC monitors the coded output and provides for adaptation of the step size on the basis of the most recent quantizer output. For example, if the previous code word corresponds to one of the extreme levels, the quantizer is overloaded and the step size should be increased. On the other hand, if the previous code word corresponds to one of the lowest levels, the step size should be decreased. The step size $\Delta(n)$ satisfies the following equation:

$$\Delta(n) = M \cdot \Delta(n-1),$$

where $M > 1$ if it is determined that the step size should increase and $M < 1$ if the step size should decrease. The details (see Ref. 2) of implementing such an adaptation strategy need not concern us here, except to note the rather important fact that there are strong practical reasons for imposing limits on how large or how small the step size may be; i.e., the step size satisfies the equation

$$\Delta_{\min} \leqq \Delta(n) \leqq \Delta_{\max}.$$

The step-size adaptation acts effectively to compress the amplitude variations so that the quantizer treats low-level unvoiced speech signals much the same as high-level voiced speech signals. The objective is that each quantizer level be used a significant portion of time regardless of the absolute amplitude level of the speech. However, when the input amplitude is on the order of the minimum step size, the adaptation logic insures that the step size will seek its minimum value and the difference signal will then fall within the very lowest quantization levels. Thus, when no speech is present at the input, it is expected that the code words will vary only slightly. It is this feature of ADPCM speech coding that is the basis of our endpoint location scheme.

III. THE ENDPOINT LOCATION ALGORITHM

Figure 2 shows a typical code-word sequence at the beginning of a word. Since the sampling rate is 6 kHz and there are 256 samples per line, each line corresponds to roughly 40 milliseconds of the signal. We note that, for the top line and most of the second line, the code words show little activity, remaining for the most part within the middle four quantization levels. This first part of the sequence corresponds to silence. However, at the end of the second line and then for the remaining two lines, the code-word sequence fluctuates much more rapidly and with greater amplitude. This segment corresponds to

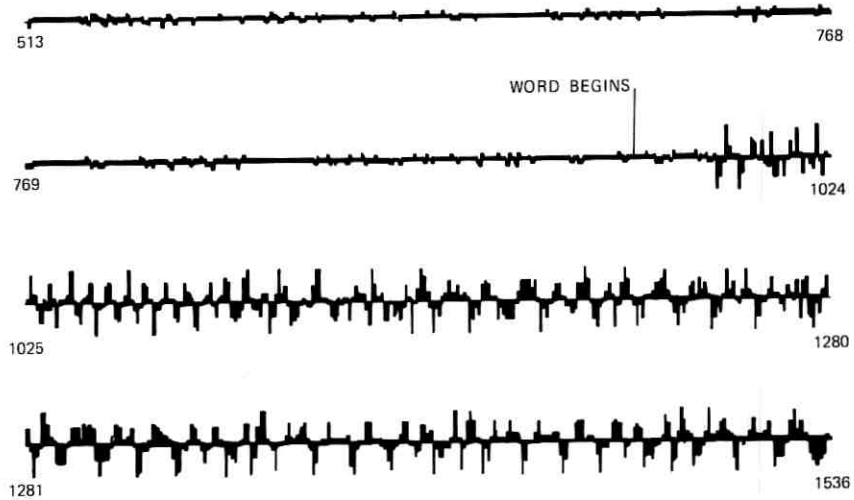

Fig. 2—Code-word sequence for the utterance /o/ showing coded silence followed by voicing.

voiced speech, as is evident from Fig. 3, which shows the decoded speech waveform corresponding to the previous code-word sequence.

These properties of the ADPCM representation of speech are reflected in what we call the *code-word energy*, defined as the sum of squares of the code words over a 101 sample, or 16-millisecond window

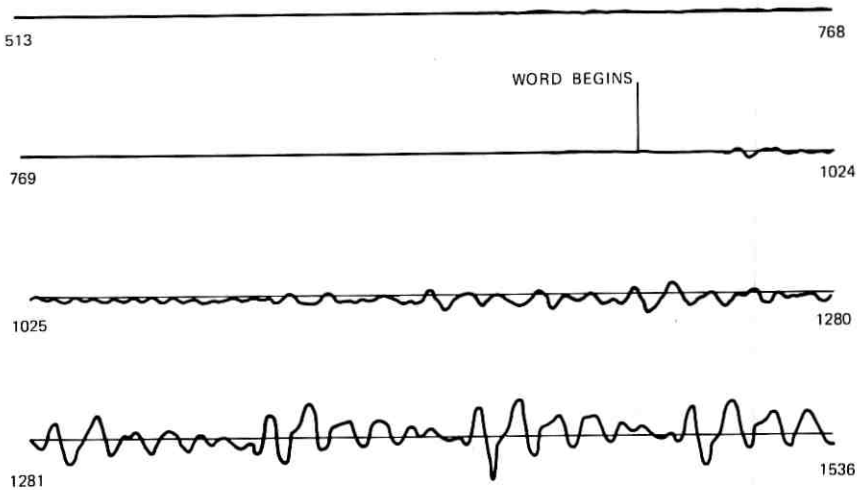

Fig. 3—Decoded speech waveform corresponding to code-word sequence of Fig. 2.

centered around the current sample. That is, the code-word energy $E(n)$ is

$$E(n) = \sum_{i=n-50}^{n+50} [c(i) - 7.5]^2.$$

In our hardware implementation (see Ref. 2), the largest negative quantization level is represented by the binary code word 0000, while the largest positive quantization level is represented by the binary code word 1111, or the decimal number 15. Thus, 7.5 is subtracted from the code words to make the dc level of the code words equal to zero. We have found that performance is only slightly degraded if $|c(i) - 7.5|$ is used instead of $[c(i) - 7.5]^2$. The use of only the magnitude leads to simplifications in hardware implementations of the algorithm.

Using either of these definitions of code-word energy, the endpoint location algorithm is as follows. The code-word energy is computed at each sample and compared with a threshold which is set midway between the measured energy of silence and the average for speech. When the code-word energy exceeds this threshold for 300 consecutive samples or 50 milliseconds, the point at which the energy first exceeds the threshold is recorded as the beginning of an utterance. The energy comparison is continued, and when the code-word energy falls below the threshold for 1000 consecutive samples, or 160 milliseconds, the point at which the energy first falls below threshold is recorded as the end of the utterance. The 160-millisecond criterion ensures that a stop consonant within a word or phrase will not be mistaken for the end of the utterance.

An example of the operation of the above algorithm is illustrated by the sequence of waveforms in Figs. 4 to 9. Figure 4 shows the sequence of code words for the beginning of the word /three/. The left half of the first line shows very little code-word variation and corresponds to low-level tape noise. The right half of the first line and the next two lines, corresponding to the initial fricative /th/, show markedly greater variation, as does the last line which corresponds to the beginning of voicing. The marker in the middle of the first line denotes the beginning point as located by the algorithm just described. Figure 5 shows the code-word energy plotted as a function of time. The marker again denotes the point at which the energy exceeded threshold and remained above for at least 50 milliseconds. Note that the code-word energy is roughly the same for both the voiced and unvoiced segments, while it is significantly lower when no speech is present. These assertions are confirmed by Fig. 6, which shows the actual

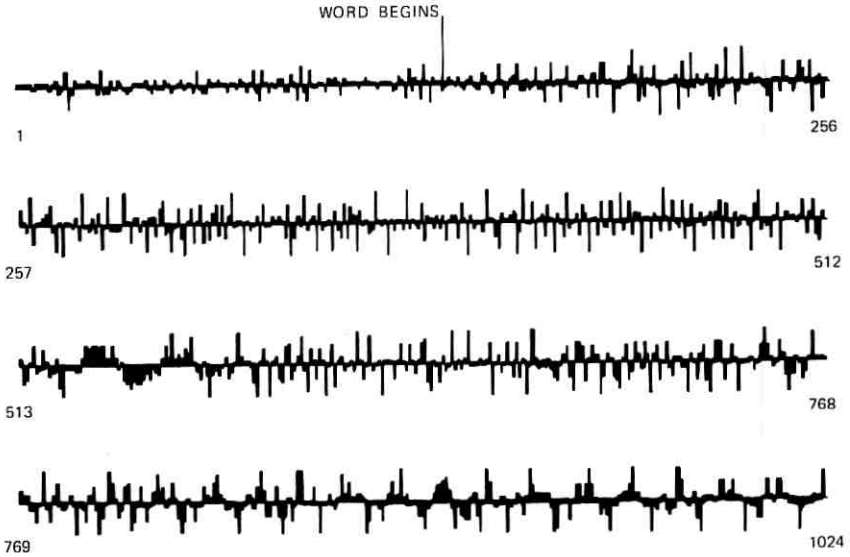

Fig. 4—Code-word sequence for the beginning of the utterance /three/.

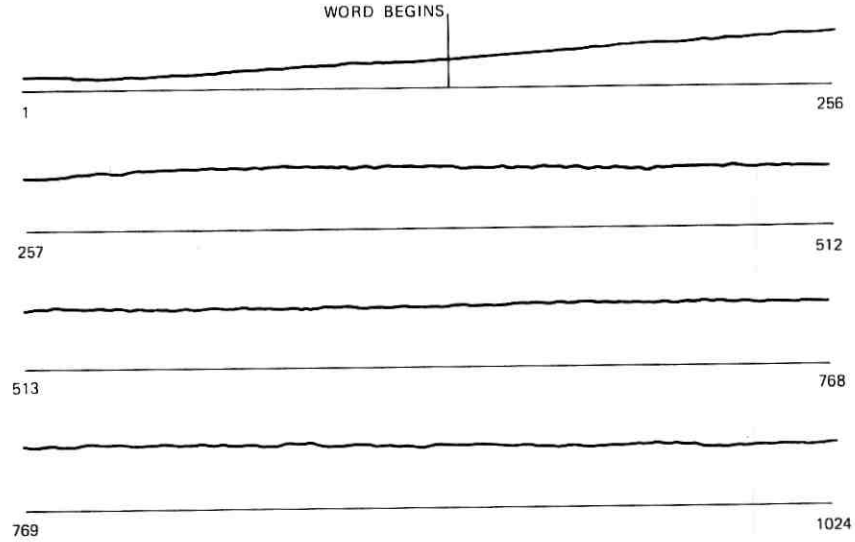

Fig. 5—Code-word energy for the code-word sequence of Fig. 4.

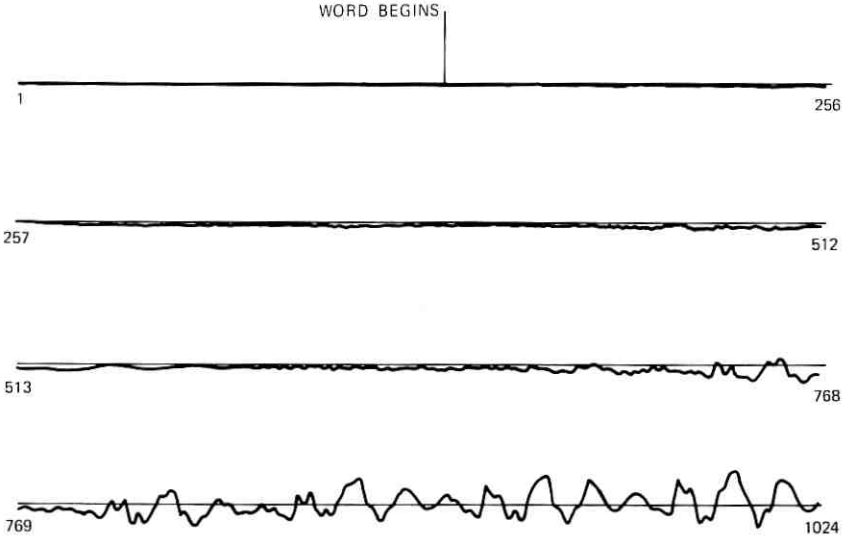

Fig. 6—Decoded speech waveform corresponding to the code-word sequence of Fig. 4.

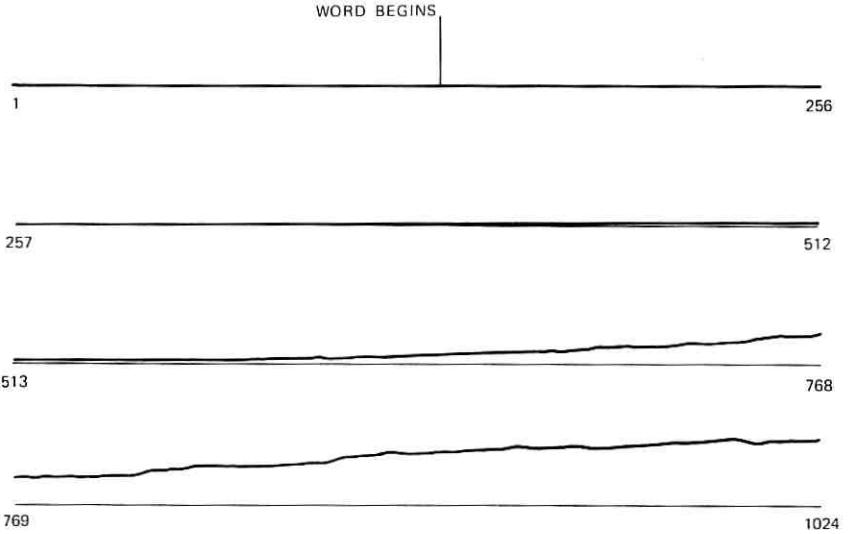

Fig. 7—Energy of speech waveform of Fig. 6.

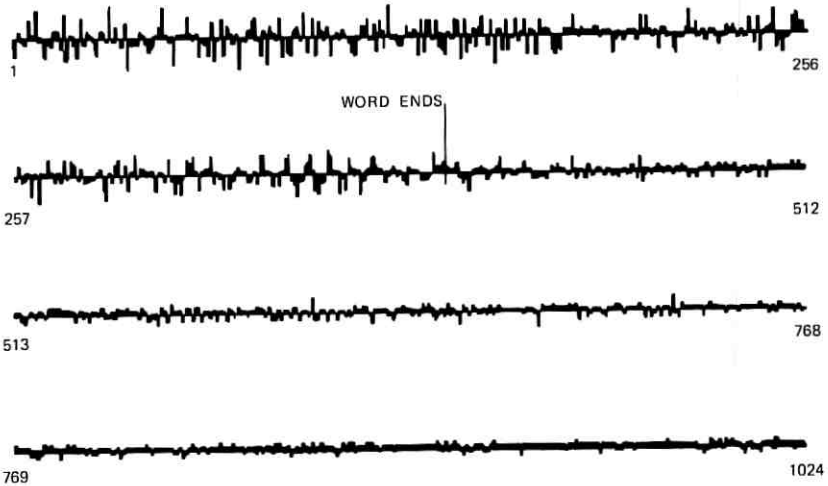

Fig. 8—Code-word sequence at end of the utterance /three/.

speech waveform represented by the previous code-word sequence. We see the beginning unvoiced segment and the following voiced segment. The actual beginning of the word is not nearly as evident as in the code-word sequence. Figure 7, which shows the energy of the speech waveform, emphasizes the fact that the simple algorithm that we have proposed would not be effective when operating upon uncoded samples

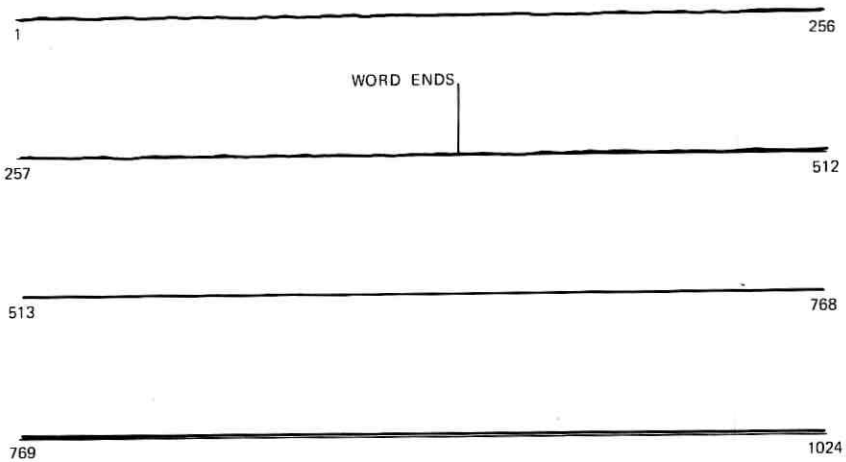

Fig. 9—Decoded speech waveform corresponding to the code-word sequence of Fig. 8.

of the speech waveform. Figure 8 shows the code-word sequence at the end of the word /three/. Also shown is the automatically determined endpoint; i.e., the point at which the code-word energy first fell below threshold and remained below for a period of 160 milliseconds. Figure 9 shows the corresponding decoded speech waveform. The endpoint, which was clearly in evidence in the code-word sequence, is much less prominent in the speech waveform itself.

IV. CONCLUSION

This scheme was tested on a large number of typical entries for a voice response vocabulary with no errors. Auditory and visual inspection indicated no evidence of shortening or inclusion of extra silence for any of the words. The performance of this simple scheme has allowed us to implement a completely automatic system for cataloging words for a computer voice response system.[1]

REFERENCES

1. L. H. Rosenthal, "An Automatic Voice Response System Utilizing Adaptive, Differential Pulse-Code Modulation," M. S. Thesis, Department of E.E., Massachusetts Institute of Technology, June 1974.
2. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

Combining Intraframe and Frame-to-Frame Coding for Television

By J. O. LIMB, R. F. W. PEASE, and K. A. WALSH

(Manuscript received October 9, 1973)

A method of frame-to-frame coding is proposed in which the changes from one frame to the next are first detected and then transmitted as an intraframe coded signal rather than as frame-to-frame differences. A coder was constructed to test the proposal using DPCM for the intraframe encoding.

Three aspects of the coder design presented particular problems. They were:

(*i*) *Movement detection (as a result of the increase in frame-to-frame noise caused by the intraframe coding).*
(*ii*) *Smooth reduction of bit-rate and picture quality so as to take advantage of the reduction in spatial quality that a viewer tolerates when areas are moving fast.*
(*iii*) *Control strategy for linking the operation of the buffer, the movement detector, and the operating state of the coder.*

The coder gave good picture quality at a transmission rate of 1.5 megabits per second (0.75 bit per picture element), except in extreme situations where the moving area covered almost the entire screen. The performance is described in detail at bit rates of 2.0, 1.5, and 0.5 megabits per second.

The experimental coder has a number of desirable properties from an overall systems point of view when compared with transmission of frame differences. These include high tolerance to transmission errors and small frame storage requirements.

I. INTRODUCTION

More than forty years ago it was first realized that channel capacity requirements could be significantly reduced by transmitting only those parts of a television signal that represent the changes from one frame of an image to the next.[1] However, only recently technology has been

available to store a complete frame of video information to enable such a system to become practicable.[2,3]

In addition to the high correlation from frame to frame (temporal correlation), quite high correlation also exists from line to line and between adjacent elements along a line. It is these spatial forms of correlation which have been most widely exploited in coding television signals. For example, within a single frame we can switch between previous element prediction and previous line prediction, depending on whether there is more horizontal or more vertical similarity between adjacent picture elements.[4] Similarly, in frame-to-frame coding the element in the previous frame corresponding to the element being encoded is a good prediction when an object is moving slowly, whereas a spatially adjacent element in the same frame is a better prediction of the current element when the object is moving fast.

In an ideal situation, it is easy to determine the changeover point at which the element difference is smaller than the frame difference. Consider an image moving horizontally at a constant speed of one picture element per frame period (pef). This speed is quite slow; it would take about 8 seconds for an object to cross from one side of the screen to the other. During one frame an element moves so as to occupy the position occupied by the element adjacent to it in the previous frame. Consequently, at this speed the element-difference signal equals the frame-difference signal: at greater speeds the frame-difference signal is larger.[5]

One early scheme for frame-to-frame coding, called Conditional Picture-Element Replenishment, updated the changed picture elements with a new PCM value.[3] We refer to this as CR/PCM coding. The efficiency of this scheme can be improved significantly by transmitting the difference between a stored reference frame and the new frame (CR/FF). The changes can be transmitted with little more than four bits per element, on the average, rather than between six and eight bits for PCM transmission.[6]

In conditional replenishment (CR) schemes, data are generated at a very uneven rate, and therefore it becomes necessary to use a buffer to smooth the peaks if a constant transmission bit rate is required. In general, while the buffer can smooth data within the field, it is not practicable to smooth from one activity peak to the next because the size of the buffer would need to be very large.* Further, in the video-

* For example, if a movement lasted for a duration of 1 second, between 3 and 6 megabits of data could easily be generated, most of which would need to be stored (Ref. 7).

telephone situation, the signal delay inherent in a large buffer becomes intolerable to a user. Consequently, the efficiency of a coding scheme is highly dependent on the peak data generation rate. However, the coding of moving areas by intraframe techniques becomes more efficient with faster movement. This is in contrast to most other frame-to-frame coding schemes in which the efficiency decreases with the speed of movement. There are other advantages to coding the moving parts as an intraframe signal:

(*i*) In many video-telephone situations, only the intraframe coded signal is available and, in general, transmitting the intraframe signal minimizes requantization effects.
(*ii*) Such a scheme lends itself very well to economizing on frame storage requirements by storing only intraframe differences.

A conditional replenishment system using intraframe coding of the changed parts of the signal (CR/IR) was first demonstrated in 1970.[8] This paper describes that system and subsequent improvements associated with movement detection and the control strategy. Related work is described by Wendt[9] and Kanaya is currently investigating a CR/IR type system.[10]

The concept of CR/IR coding is illustrated in Fig. 1. The output of an intraframe coder is stored locally in a frame-memory loop. If a significant difference is detected between the input signal and the decoded version of the stored signal, the two switches move to position 1 and new data are entered into the frame memory and at the same time transmitted to a frame memory at the receiver. It is also necessary to

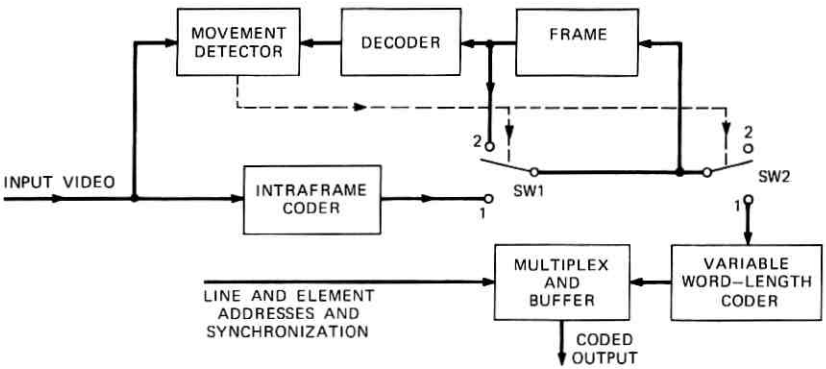

Fig. 1—Basic concept of the conditional replenishment intraframe (CR/IR) coder.

transmit addresses so that the receiver can insert the coded signal in the correct place.

Figure 1 is deceptively simple, and a large combination of techniques is needed to implement such a coder successfully. However, the configuration we describe should not be regarded as a complete system, but rather as the result of an experiment, first, to evaluate the feasibility of transmitting an intraframe-coded signal in moving areas and, second, to explore methods of varying and controlling the horizontal accuracy with which the intraframe signal is coded.

A brief description of the CR/IR coder is given in the next section, while more details are given in the appendix, Section A.1. Section III describes the performance of the coder and Section IV discusses, first, some additional techniques which could be used for further improvement and, second, some implications of CR/IR coding for overall system design.

II. DESCRIPTION OF CONDITIONAL REPLENISHMENT INTRAFRAME CODING TECHNIQUES

2.1 Switching between "stationary" and "moving" signals

Let us be specific and assume that the intraframe coder is a differential quantizer[11] (differential pulse-code-modulation coder). The scheme of Fig. 1 works satisfactorily if the switch is operated (closed or opened) only when the digital value of the decoded form of the coded signal is the same at both the output of the frame memory and the output of the intraframe coder. If this condition is not met, an error term is added to the coded signal which is equal to the difference between the decoded value of the two signals incident at switch 1 at the instant of switching. This would result in a streaky picture with streaks similar to those produced by transmission errors. Figure 2 illustrates this lack of tracking between the intraframe coder and the CR/IR decoder when the switches of Fig. 1 change position to accept new data.

To permit the switches to change position only when there is no difference (or a very small difference) between the decoded values of the two signals arriving at switch 1 (Fig. 1) would be very restrictive and would probably result in a significant increase in the area to be transmitted, particularly if the input signal is at all noisy.

This difficulty is overcome with the configuration of Fig. 3.* The switch now handles normal (accumulated PCM) values rather than

* Notice that the input to the coder is in intraframe coded form. We imagine the CR/IR coder as being one stage of a hierarchy of coders in which each stage would probably be at different physical locations (Section 4.2).

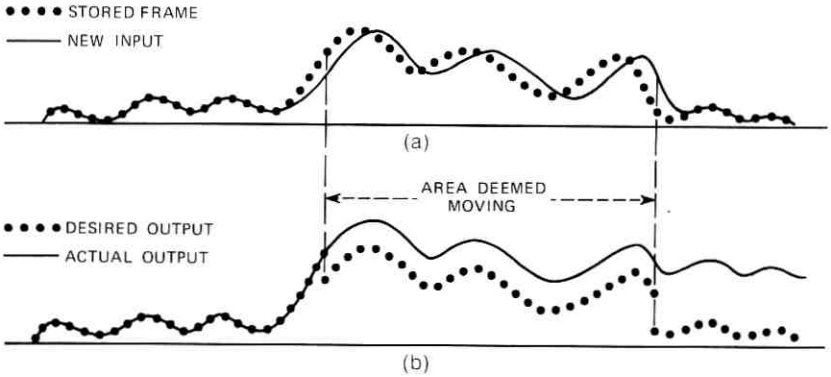

Fig. 2—Waveforms showing operation of conditional replenishment coder of Fig. 1. (a) Dotted line: Decoded value of stored signal (in frame memory); solid line: New incoming signal which is shifted to the right in the moving area because of a change in position of subject. (b) Solid line: Output of conditional coder of Fig. 1. Notice the offset at the instant of switching caused by addition of a new element-difference signal to the old (stored) decoded signal; dotted line: Desired representation of the combined input and stored signals.

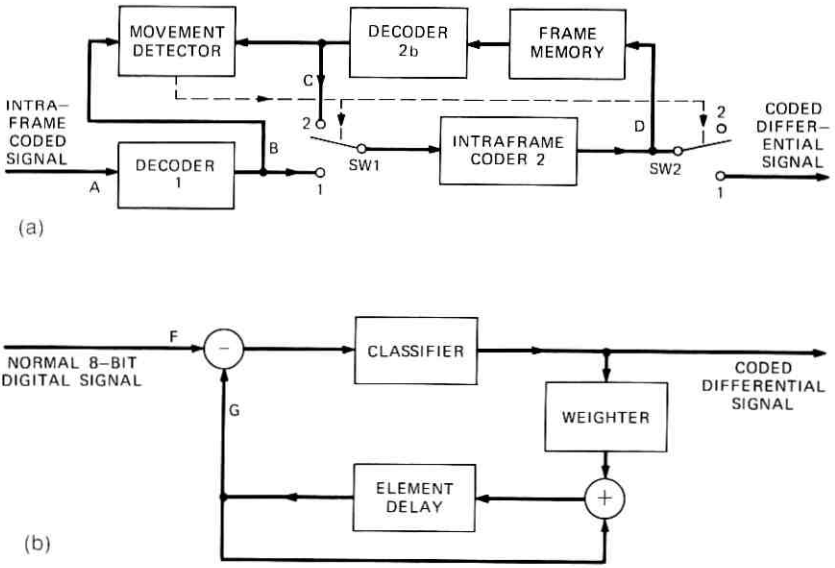

Fig. 3—(a) Diagram of the CR/IR coder. Notice the change from Fig. 1: Coder 2 and decoder 2b are added to the loop so that the offset problem shown in Fig. 2 is eliminated. (b) Diagram of intraframe coder (DPCM).

differential values, and the signal is recoded before it is stored in the frame memory. While there is no detected movement, the signal circulates through the frame memory, decoder, and coder without change. If the switch changes position while the PCM values entering switch 1 are identical, then the coded signal will not be changed after passing through decoder 1 and being recoded (i.e., the signals at A and D will be the same). On the other hand, if the switch changes when the two PCM values are different, a small amount of recoding noise will occur while the new signal is corrected.

The operation of Fig. 3 is probably best appreciated by a numerical example shown in Table I. Let us assume a five-level differential quantizer with decision levels ± 1, ± 4, and representative levels 0, ± 2, ± 6 (see, for example, Ref. 12).

If row A represents the input to decoder 1, then row B represents the output given that the value of the accumulator is 32 before decoding.* Let C represent decoder 2b output. Row D represents the output of intraframe coder 2 and is the same as the output from the frame memory before decoding up to the point that the switches change from position 2 to position 1. Row E is the accumulated value of D and represents the signal at the receiver. Just before switching, the difference between the two values B and C at switch 1 is 6. After switching, the input to intraframe coder 2 is 44 (signal F, Fig. 3(b)), while the value in the accumulator is 36 (signal G). The difference is $+8$ which is coded as a 6 (therefore, E is 42). Coding continues with $B = F$ as the input and row D as the coder output. On the fourth sample after switching, the two signals B and E are the same, and signals A and D will remain locked together until the switch returns to position 2. Thus, the coding noise at the switching point is confined to three samples and has the values -2, $+2$, $+2$ (obtained by subtracting row B from row E). The time to lock in depends on the quantizing characteristic, the input waveform, and the amount of difference at the instant of switching; in many instances, lock-in is immediate. The average lock-in time for the quantizer used in this study was measured

* We are dealing with many different types of signals in connection with the differential quantizer, and it is important to have a clear description of the terms used. A signal can be either analog or digital (i.e., PCM). The signal will be called "normal" (e.g., normal digital) if it is directly related to the amplitude of the video signal (signals B and C of Fig. 3(a)). Similarly, a signal will be called differential if it is directly related to some form of difference-signal (signals A and D of Fig. 3(a)). A standard 8-bit PCM signal will be referred to as a normal digital signal. A signal that has passed through both a differential coder and decoder will be called normal-differentially-quantized (signal B of Fig. 3(a)), while if it has only passed through the coder it will be referred to as a coded-differential signal (signals A and D of Fig. 3(a)).

Table I — Numerical example of the operation of the coder shown in Fig. 3

	Successive Picture Elements in Horizontal Direction										
A		+2	+2	0	+6	+2	+2	−6	0	0	+2
B	32	34	36	36	42	44	46	40	40	40	42
C	32	34	36	36	36	X	X	X	X	X	X
D		+2	+2	0	0	+6	+6	−6	−2	0	+2
E	32	34	36	36	36	42	48	42	40	40	42
						↑					
						Switching Point					

at 0.11 element per transition of switch 1 (Fig. 3(a)) for the case when there was virtually no movement and the small amount of updating that was occurring was triggered primarily by noise. Where there was a significant amount of movement, the average lock-in was 0.30 element.* A similar lock-in time is required when the switch 1 moves from 2 to 1 (return to stored signal).

2.2 *Moving area detection*

Accurate detection of changed areas within the picture is important for efficient coding. This is straightforward when working with a high-quality digital signal.[13,6] However, as can be seen from Fig. 3, we are detecting the changed areas from a signal that has been intraframe coded and is therefore relatively noisy, particularly at edges where the coarse outer levels of the differential quantizer are used. This means that more sophisticated movement-detection techniques are required to obtain adequate detection. Reference 14 derives some correlation properties of the types of frame-difference signals generated in conditional replenishment encoding and Ref. 15 describes the implementation of a previous design. The movement detection used in this study is similar in principle to that described in Ref. 15. The difference between the stored frame and the current frame is: (*i*) spatially and temporally filtered; (*ii*) applied to a varying threshold which is under control of a modified element-difference signal (this compensates for the larger errors introduced in high-detail areas by the differential quantizer) and (*iii*) "blocked-in," an operation which both produces a more contiguous moving area and rejects small isolated changes.

* These figures were obtained with the coder control circuit locked in mode 1 for the first figure ("no movement") and mode 2 for the second figure ("movement"). See Section 2.4 for a description of the various operating modes. We suspect that the short lock-in times result partly from the fact that the second representative level is twice the value of the first (see Table IV).

Specific details of the movement-detector used for this study are given in appendix Section A.2.

2.3 Reduction of resolution

As the speed of a moving object increases, the resolution of the resulting image in the direction of movement decreases because of the light-integrating action of the camera target. For horizontal movement, this in turn reduces the amplitude of the element-to-element differences, and the entropy of the associated intraframe coded signal decreases. Figure 4 is a picture of the unquantized element-difference signal of a moving object against a stationary background at two different speeds. The reduction in contrast is quite obvious in the moving area as the speed goes from one-half element per frame to four elements per frame.

Although it appears that the eye can detect smearing of the picture because of camera target integration, an observer is reasonably tolerant of this type of degradation and, in fact, we would like to take the process a little further. As we can see from Fig. 4 (see also Ref. 14), the effect of target integration is to reduce the bandwidth of the spatial signal in moving areas; this, in turn, reduces the first-order entropy of the coded signal. But relying solely on the first-order entropy reduction of the intraframe coded signal at full sampling rate does not take full advantage of the redundancy in the signal at high speeds, when the signal is essentially oversampled.*

Smoothly reducing the sampling rate as the speed increases would be very effective but is impracticable. Switching to a submultiple of the sampling rate is quite practicable, but the difference in picture quality in going from the full sampling rate to half sampling rate is quite large, especially for differential quantization. Thus, the change in quality at the instant of switching is noticeable.

A coding technique called receiver-model coding was developed partially for this application.[17] It enables properties of the observer to be incorporated into the coding process. A particularly simple form of receiver-model coding (referred to as "level variable sampling" in Ref. 18) is 2:1 horizontal conditional subsampling, in which every second point in the picture is differentially quantized in the normal manner. The alternate points (conditional points) are extrapolated from the previous point (zero-order hold) unless the error incurred by so doing exceeds a predetermined threshold (in which case, they are

* The results of Bobilin show how rise-time and edge-busyness change as the ratio between sample rate and bits per sample (for DPCM) is altered (Ref. 16).

(a)

(b)

Fig. 4—Reduction in amplitude of element differences with increase in speed. (a) Head moving at a speed of 0.5 pefs. (b) Head moving at a speed of 4.0 pefs. Reduction is caused by integration of light falling on camera target for duration of one frame.

also differentially quantized in the normal manner). When the threshold is low, nearly all points are coded normally. As the threshold is increased, more and more conditional points are extrapolated until, if the threshold is high enough, the signal is effectively subsampled. To have a bit-rate advantage with horizontal conditional subsampling, we need to use a variable-length code since information is transmitted about all points, including the conditional points unless the signal is fully subsampled (see appendix Section A.3.1).

The coder used in this study did not have a continuous threshold control, but could be switched to give one of five "operating states" starting with normal differential quantization and going to 4:1 horizontal subsampling, which gave a picture quality that was scarcely adequate even in very fast moving areas.

2.4 Control strategy

There are two different ways in which the data-generation rate may be reduced. One is by reducing the accuracy and resolution with which the moving area is coded as described above. The other method is to reduce the size of the moving area by demanding that the difference (measured in some way) between the stored signal and the incoming signal in a given area be larger before that area is regarded as moving. Raising the criteria for movement detection is most effective for areas that are moving slowly.

Two possible control strategies are;

(*i*) Use a measure of the speed of the moving object in the picture to reduce the resolution and, therefore, the data generation rate in the moving areas, but not so much that picture quality will be significantly affected. Data may still be generated at a rate that exceeds the channel rate, especially when large areas are moving slowly.

(*ii*) Use a measure of the buffer fullness to reduce the resolution and size of the moving area.*

At the time of this study, a speed-measurement circuit was not available and so the buffer alone was used to control both the spatial resolution within the moving area and the size of the moving area.†

---

* These types of control are quite different in effect (Ref. 7).

† Some relatively simple techniques for determining the approximate speed of the moving area are currently being evaluated by the first author and J. A. Murphy of Bell Laboratories.

Table II — Bit-rate control modes—summary of the bit-rate reduction techniques for each mode

Mode (Section A.3.4)	Level Variable Sampling (Section A.3.1)				Moving Area Detector Threshold Select (Section A.3.2)					Single Point Threshold (Section A.3.2)	
	Levels deleted 2:1			4:1							
	±1	±1 and ±2	All	All	T_1	T_2	T_3	T_4	T_5	Low	High
1					X					X	
2	X					X				X	
3	X						X			X	
4		X					X			X	
5			X				X				X
6			X					X			X
7				X					X		X
8	Frame Repeat										

Feedback from the buffer progressively reduces spatial resolution and increases thresholds for moving area detection in a sequence of eight steps with the last step being the prevention of all updating.

We have built a system based on the scheme of Fig. 3 using a simulated buffer with the buffer-control strategy described above. The equipment is described in detail in the appendix, and the feedback modes are summarized in Table II. The experiments carried out and the results obtained are described below.

III. EXPERIMENTS AND RESULTS

The functional blocks of the coder interact in a complex manner, making it difficult to evaluate the separate contribution of each block. Furthermore, transitions between modes can occur very rapidly so that in certain instances the coder may oscillate between adjacent modes at line rate. We first report the performance (picture quality and bit rate) of the operating states applied to the whole picture (with no movement detection or feedback control). Next, we describe the additional effect of movement detection still without feedback control. Finally, we describe the performance of the overall coder at different transmission rates.

A head-and-shoulders view was used with the subject covering slightly less than half of the viewing area. Thus, with the size of the subject constant, varying the speed at which he or she moved across the screen varied the data rate. The subject was wearing relatively

low-detail clothing; when high-detail clothing is worn, the data rates are a little higher.

3.1 Resolution reduction: effect of changing operating states

Table IIIa shows the performance of the coder with the various operating states applied to the whole of a stationary picture. The bit rate represents the amplitude bits per picture element and, of course, does not include addressing, etc. There is a bit-rate reduction of 45 percent in going from full sampling to 2:1 sampling accompanied by a gradual decrease in picture quality.

3.2 Effect of moving area detector

To show the effect on bit rate of each mode (described in Table II), the speed of a subject was chosen so that when only mode 1 is used (feedback-control inhibited and manually selecting mode 1), the bit rate needed for transmission was approximately 2.0 megabits per second. While the subject conditions are kept constant, each remaining mode was manually activated and the resulting bit rate recorded (Table IIIb). Here the bit rate is a total system bit rate (appendix Section A.3.3). There is about a 10:1 drop in average bit rate in going from mode 1 to mode 7. The reduction in bit rate in going from mode 2 to mode 3 and from mode 5 to mode 6 is a result only of a reduction in the moving area (see Table II). These measurements are not an exact indication of the bit-rate reduction of each mode, since in actual

Table IIIa — Bit rate and picture quality for each operating state with movement detection disconnected (coding applied to whole picture)

Operating State	Level Deletion	Bit Rate (bits/picture element)	Picture Quality
1	None	3.07	Very good. Limited only by the quantization process.
2	Level ± 1	2.58	Very good. There is a just-noticeable increase in noise in areas having fine detail and low contrast.
3	Levels ± 1 and ± 2	2.36	Good. The increase in random noise is more noticeable than state 2, and some fine detail with low contrast is lost.
4	2:1 subsampling	1.69	Fair. Sharp edges become serrated and fine detail is blurred.
5	4:1 subsampling	0.89	Poor.

Table IIIb — Bit rate for each control mode

Mode	1	2	3	4	5	6	7
Bit rate (Mbits/s)	2.01	1.60	0.88	0.80	0.64	0.46	0.19

operation the speed and size of the moving area would be different for each mode.

3.3 Performance at different transmission rates

3.3.1 Performance at 1.5 megabits per second

Table IIIc gives the performance of the system operating at a transmission rate of 1.5 megabits per second. To enable detailed observation and measurement of the effect of each mode, the coder was locked to each mode. Then the picture quality and amplitude bits per transmitted element were recorded for the type of movement appropriate to that mode. The picture quality depends strongly on the size of the moving area; as noted, the moving subject filled approximately half the picture. With smaller moving areas, the higher modes are used less frequently and the picture quality is better; the situation reverses in larger moving areas. In the table, conversational movements are considered movements of the face and gentle head movements. The X denotes that these modes cannot be activated only by side-to-side body motion.

3.3.2 Performance at 2.0 megabits per second and 500 kilobits per second

With the coder operating normally, the picture quality was observed at transmission rates of 2.0 megabits per second and 500 kilobits per second.

At 2.0 megabits per second, very slow (1 pef) to moderate (3 pef) side-to-side movements cause mode 1 to be used continuously. This provides good picture quality and also good moving area detection. Only during very fast motion does mode 3 come into use, which reduces the accuracy of the moving area detection and subsampling on the inner pair of levels. Mode 5 is used only for violent changes such as panning the camera or walking in front of the camera. The noticeable defect is a coarse structured effect in the moving areas produced by the 2:1 subsampling and the reduced accuracy of the moving area detector.

Table IIIc — Picture quality and bit rate for each control mode at 1.5 megabits per second

Mode	Amplitude Bits per Transmitted pel	Mode Activated by		Buffer Level (fraction of full capacity)	Picture Quality
		Speed of Movement (pef)	Type of Movement		
1	3.3	$<\frac{1}{2}$	Normal conversational	$<\frac{1}{8}$	Very good. Slight increase in edge noise at very slow speeds because of movement detector. Picture quality is better than with normal DPCM because the quantizing noise is less visible when it is "frozen."
2	2.9	$1\frac{1}{2}$	Active conversational	$\frac{1}{8}$	Good. A higher moving area detector threshold is chosen, and thus the area detected is reduced. The result is a slight increase in low detail noise.
3	—	3	Active conversational with hand and arm movements	$\frac{1}{4}$	Fair. A higher threshold and the elimination of temporal feedback cause some low detail areas in motion to appear somewhat contoured or "patchy." These effects are only marginally noticeable.
4	2.2	5	Very active body, hand, and arm movements	$\frac{3}{8}$	Fair. The threshold is the same as above; however, the added noise caused by the level variable sampling causes movement detection to become slightly worse.
5	1.7	X	Violent motion or standing up in front of camera	$\frac{1}{2}$	Acceptable. The threshold remains the same, but the added noise of the sampling and the FOS inhibit (see section A.3.1) reduce movement-detection accuracy. Also, at some speeds (multiples of 2 pef) the sampling produces a striped structure.
6	—	X	Motion such as walking in front of camera	$\frac{5}{8}$	Marginally acceptable. A higher threshold is chosen which produces a noticeable dirty window effect.
7	—	X	Motion such as panning the camera	$\frac{7}{8}$	Poor.

With the transmission rate limited to 500 kilobits per second and the subject in very slow side-to-side motion (0.5 pef) or in normal conversational movements (i.e., gentle lip and head movements), the system uses only the first four modes and the quality picture is still good. At a speed of 1 pef, modes 5 and 6 are used in which 2:1 subsampling is employed and the movement-detector uses the higher thresholds. The result is a slightly more noisy picture with the movement detector producing either a "dirty window" or a patchy effect.

At a speed of 2 pefs, mode 6 is mostly used. At this point the picture quality is probably unacceptable with the major degradations being: (*i*) the coarse structured effect caused by poor movement detection, (*ii*) the noisy edges caused by the 2:1 subsampling, and (*iii*) the general increase in noise.

At a speed of 3 pef, mode 7 is used more frequently and the picture becomes unacceptable, with the major degradations being poor moving area detection and a "column" effect produced at some speeds by the 4:1 subsampling.

IV. DISCUSSION

The above experiments are only a start in investigating the techniques of CR/IR coding. However, even at this stage we can see the encouraging performance for fast moving scenes. For example, at a transmission rate of 2 megabits per second, motion such as panning the camera only invokes mode 5; i.e., neither 4:1 subsampling nor the highest levels of the movement detector are used. In a previously described CR/FF coder, motion such as panning the camera invoked frame repeating.[6] Further work is needed to examine related techniques that could significantly improve coder performance. One example is an evaluation of intraframe coding techniques that are more efficient and better suited to CR/IR operation. In addition, we should investigate the application of known frame-to-frame coding techniques; we discuss some of these below.

4.1 Add-on techniques

The vertical resolution can be reduced by transmitting only alternate lines in each field and filling in the missing lines by vertically averaging. In this study, the horizontal resolution was reduced by up to a factor of 4. This is inferior to spreading the resolution reduction more equally between the vertical and horizontal dimensions. A horizontal resolution reduction of 4:1 is acceptable in very fast moving areas, but if the mode is invoked at lower speeds, for example, where the camera is

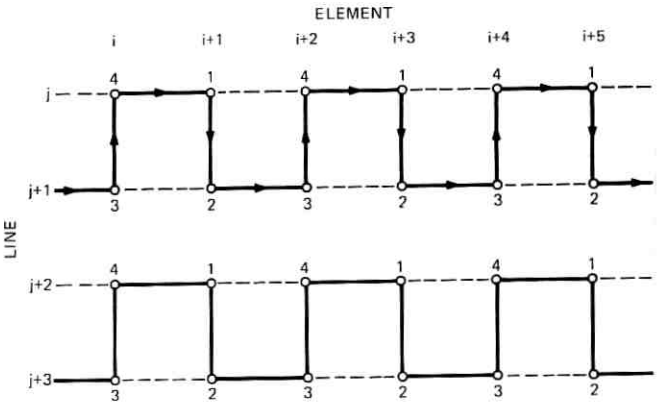

Fig. 5—Merli scan path in which elements are taken alternately from adjacent scan lines. Picture elements are processed in quads.

being panned slowly, then serious degradation results. One method for smoothly reducing the resolution in both dimensions by a combined factor of 4 would be to apply receiver-model coding to the Merli scanning algorithm.[17,19] In this coding scheme two adjacent scan lines are coded simultaneously by following the notched path of Fig. 5. The elements are processed in quads with the number 1 elements always coded with full precision. An attempt is made to represent the number 2, 3, and 4 elements as linear interpolations based only on the number 1 elements. The interpolation error is calculated and filtered to approximate the liminal vision of the human observer. If the filtered error signal at a particular point exceeds the allowed threshold for a given quality, then the point is updated.

As the threshold is raised, fewer conditional elements (numbers 2, 3, and 4) are transmitted. If the threshold is raised far enough, a 4:1 subsampled picture is obtained with a reduction of 2:1 in both the vertical and horizontal directions. Horizontal subsampling reduces the number of amplitude bits that have to be transmitted without affecting the number of address bits or line synchronizing bits.* By using the Merli algorithm, on the other hand, the line address and synchronizing bits would be almost halved since a line now contains twice as many elements as it previously did.

Conditional-vertical subsampling is a technique that is applied from field to field. Alternate fields are obtained by a four-way average of

* Actually, one bit could be dropped from the address word when 2:1 subsampling is used.

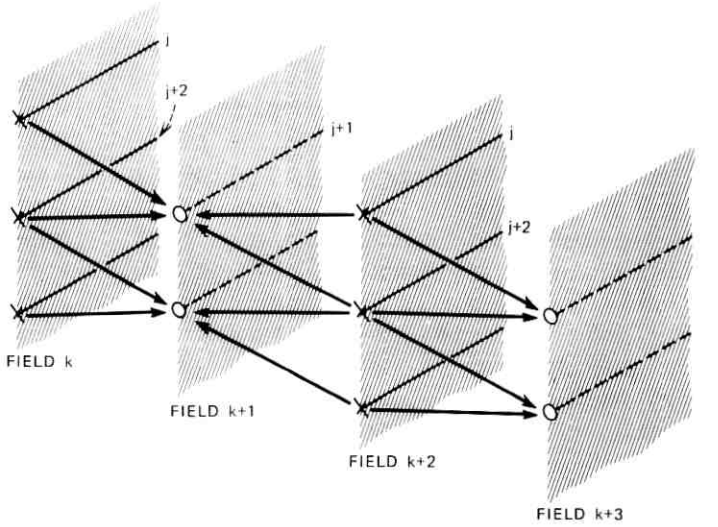

Fig. 6—Four-way averaging in which alternate fields are not transmitted. At the receiver, the missing fields (even-numbered fields) are replaced by a four-way average of elements in the adjacent fields.

samples in the immediately preceding and succeeding fields as shown in Fig. 6. Should the average fail badly for a particular element, then an additional correction signal may be transmitted, depending on the quality that is required. This four-way field averaging reduces both spatial and temporal resolution by a small amount.[20]

More severe temporal averaging can be employed by using what may be called conditional frame-to-frame subsampling. Such techniques are most useful where large areas are moving slowly, the particular condition which is handled poorly by the CR/IR coder and quite easily by the CR/FF encoder. However, if there is a significant reduction in temporal resolution, it is important that it be under the control of a speed-indicator circuit so that it can be switched out when the speed starts to increase.

4.2 System implications

In a practical visual communication system, transmission links will vary greatly in length. As a consequence, on short links a simple inexpensive coder would be appropriate, whereas on longer links more expensive frame-to-frame encoding might be suitable. Now in a complex switched system, we may well want to pass through a number of digital links in tandem, some being short and others long. Thus, it is

important to have a family of coders that are compatible in the sense that they can operate in tandem without unduly degrading the system. We could envisage at least four stages of coding: (*i*) a simple differential quantizer stage; (*ii*) a more efficient intraframe encoder using a variable-length code on the output of 1; (*iii*) an interframe coding stage and (*iv*) a channel-sharing stage where a number of users share a high capacity channel, trading on the fact that there is a low probability of all users being active simultaneously (as in TASI).[21,22] The conditional intraframe coder is well suited for this type of multistage tandem operation. As we have seen, the frame-to-frame coding stage does not add quantizing noise to the signal except in elements adjacent to the points of switching between stationary and moving areas or when feedback from the buffer decreases the accuracy of the intraframe coder in the storage loop. If the signal is converted back to the intraframe form and frame-to-frame encoded for a second time, then the second frame-to-frame encoding will give a signal that is identical to the first frame-to-frame encoding if one prerequisite is met: the position of the switching points between moving and stationary areas are indicated in the intraframe signal. This would increase the intraframe data rate by approximately 2 percent.

If an improvement is made in the performance of the intraframe encoding stage, this improvement will carry right through to the frame-to-frame channel-sharing stages.*

The fact that an intraframe coder is connected to a CR/IR coder will tend to affect the type of algorithms that we employ in the intraframe stage. For example, techniques that complicate the encoder design but require a simple decoder will be preferred because there are more decoders in the system than there are encoders (see Fig. 3). Notice that the conditional horizontal subsampling studied here requires no modification of the decoder design.

4.2.1 Feedback control

The different coding stages of the overall coding hierarchy would normally be at different switching offices. This almost certainly rules out any feedback from one stage to a previous stage of coding, since to incorporate feedback would considerably increase the overall complexity. For this reason, the feedback control to achieve level-deletion was kept within the frame-to-frame coder (Coder 2 of Fig. 3) rather

* Of course, changes in the intraframe encoder may well necessitate changes in the encode and decode blocks of the frame-to-frame coder.

than operating on the primary encoder. There are two consequences of this restriction for the simple type of receiver-model coding employed here. First, the effective threshold used to delete components must jump from decision-level to decision-level rather than increase smoothly because by precoding the signal in the primary encoder the element-to-element changes are restricted to the small set of values allowed by the differential quantizer. Second, there is a small increase in coding noise since the two tandem intraframe encodings are different when level-deletion is used in coder 2. In practice, however, the smoothness of control is quite adequate.* The increase in coding noise when compared with feedback to the primary encoding stage is just noticeable in a stationary picture but is virtually impossible to detect in the operation of the overall system.

Recoding noise resulting from feedback control could become a problem with, for example, higher quality systems. However, there are intraframe coders that would virtually eliminate the problem. These coders transmit two or more separate signals which represent different components of the signal so that when one component is deleted the coding of the other component is unaffected. In one system of this type,[23,24] every second sample is transmitted as PCM or DPCM and the alternate samples are transmitted as a correction signal between an estimate based on the first set of signals and the actual input. Thus, the correction signal may be deleted without interfering with the coding of the main signal. Another example of such an encoding is the Hadamard transformation applied to a small block of picture elements;[25] higher-order components can be deleted without interfering with the decoding of the lower-order components.

4.2.2 Error performance

In achieving the improved performance of CR/FF coding over CR/PCM coding, certain system advantages were lost. These advantages are partially regained with CR/IR coding. Consider, first, the effect of transmission errors on picture quality.

Since a separate interframe decoder has not been constructed, experiments on the behavior of the CR/IR coder-decoder in the presence of channel errors have not been possible. However, some intuitive predictions can be made by considering the effect of different types of errors.

---

* We only used two intermediate steps (level ± 1 delete, level ± 1 and ± 2 delete) out of a possible six.

If an amplitude word (as distinct from an address word) is in error, a noise streak will be introduced into the picture which will probably extend to the end of a line unless predictor leak is used. When there is a lot of movement there is a high probability that the error will be eliminated in the next frame since, by the usual nature of movement, the segment in error will likely be updated in the next frame and the updated segment builds only on information in the corresponding line of the stored frame to the left of the segment. With no movement or slow movement, there is much less chance that a segment in error will be "written over" in the next frame and the line in error would persist in the picture.

The signal can be made significantly more robust by transmitting a six- or seven-bit normal digital signal value at the start of a segment along with the addressing. In this way, updated segments would not build on the past values in any way. Based on an average of three segments per line, the additional amplitudes would require 0.145 megabit per second. The transmission of the additional values would terminate the effect of transmission errors already introduced and, by comparing the amplitude with the decoded value, errors could be detected. Once detected, substitution techniques could replace the line in error with a best estimate. This estimate would then last until the area was again updated. If, instead, the moving area addressing information is in error, then a large unpredictable section of a line will be in error. The effect of an error in the element address will be similar to an amplitude error, but on the average should affect a larger section of line.

In a practical system, we would want to send the line address word very securely and the start-of-frame word even more securely. The latter poses no problem since, as it occurs so rarely, it requires a negligible increase in bit rate to assign a large number of bits to the word.

It is interesting to consider what would happen if both frame and line synchronization were completely lost. Assume the receiver was aware of the loss and that it reset the frame memory to zero. Then, as soon as the person moved at the transmitting end the area in movement would be relayed faithfully to the receiver and the background would be inserted in the newly revealed area.

Although no experiments have yet been performed to determine channel error response, it appears that by transmitting an amplitude word before the start of each moving-area segment and using error detection and substitution techniques, the conditional intraframe en-

coder could be made to give acceptable performance at error rates as high as 10 or 20 per frame (an error rate of 2 to 4 $\times$ 10^{-4}). Forced updating would probably not be necessary.

4.2.3 Data Interleaving

Data interleaving is a scheme for using the frame memory to achieve a degree of smoothing of the coded data, thus considerably reducing the size of the buffer store required or eliminating it altogether.[26] It has been shown that, unless a very large buffer is used, the main smoothing effect is already achieved with a buffer large enough to smooth the irregular data over a field.[7] A 4:1 interleaving of data is achieved, for example, by transmitting lines in the order 1, 65, 33, 97; 2, 66, 34, 98; 3, 67 $\cdots$. Now if the signal stored in the frame memory can easily be converted to the coded transmission signal, then taps can be placed on the frame memory and the data can be transmitted in an interleaved manner. Two examples of coders in which the signal is stored in the frame memory in a form similar to the transmitted signal is the CR/PCM coder and the CR/IR coder. Note, however, that the frame memory stores the whole picture and we need to know which components are to be transmitted. This information would have to be included in the stored signal and would probably result in a 5-percent increase in the size of the frame memory. If a four-bit word were used to represent the differential signal, one combination could be reserved to denote a change, either from a nonupdated to an updated segment or in the reverse direction.

Data interleaving is shown applied to the CR/IR coder and decoder in Fig. 7. Code words are inserted at the coder to denote changes between updated and nonupdated segments before the signal is stored in the frame memory; these words are disregarded by the local decoder. Switch A selects lines according to the required sequence and the moving area selector interprets the marker words and selects those segments for transmission that have been newly updated. The main decoder loop [Fig. 7(b)] operates on the data just as it is received; that is, the signal in the frame memory is stored in interleaved form. The signal is de-interleaved and decoded in order to obtain an output.

Notice that such a scheme would not work if the intraframe algorithm operated on more than one line at a time since the decoder is not processing consecutive lines. Such a restriction would not apply to the vertical processing of the Merli algorithm where a line, in essence, is twice as long as a normal line.

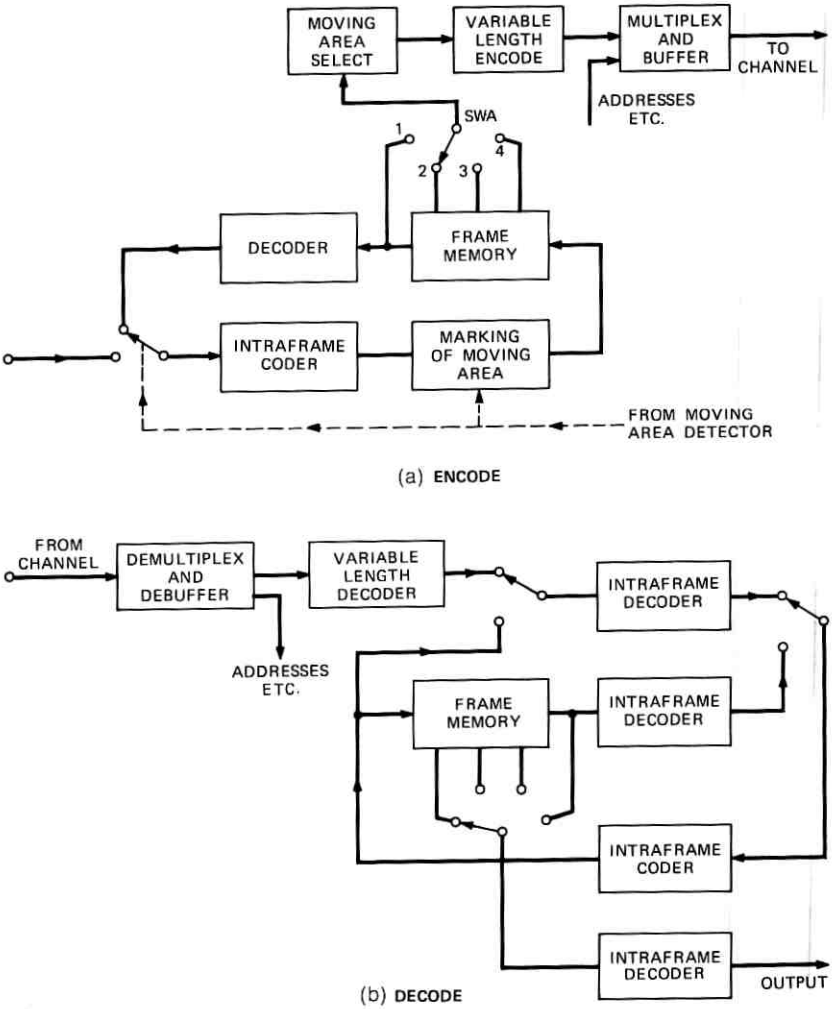

Fig. 7—Data interleaving applied to the CR/IR coder. Data interleaving reduces buffer size by shifting much of the smoothing operation from buffer to frame memory. Note that data in the decoder frame memory are in interleaved form.

4.2.4 Channel sharing

Haskell has simulated a channel-sharing and buffering scheme in which a number of encoder outputs are combined and transmitted over one high data-rate channel with one large buffer.[22] He shows that in this way the channel requirements are more than halved when 20 encoder outputs are combined. In an actual system, in the unlikely

event that a large number of users were simultaneously active, there would be feedback from the channel-sharing circuit to the encoders to reduce the data-generation rate by reducing picture quality in some manner. Although this would occur very rarely, the situation must be accommodated since we cannot arbitrarily discard data without seriously affecting picture quality: to ensure that the situation never occurs could require a significant increase in channel rate.*

Ideally, we would like to insert channel-sharing at multiple points in the transmission path, and these points may be quite remote from the encoder.[22] In this situation, feedback from the channel-sharing stage to the frame-to-frame coder would considerably complicate the overall system design. However, the CR/IR coder would enable data to be discarded with little effect on picture quality, since each new segment does not build on the past coded signal (assuming that a starting amplitude is transmitted with each segment as discussed in Section 4.2.2).

Thus, if overload of the channel-sharing stage were imminent, the whole line could be deleted except for the line addressing word (required for receiver synchronization) and a further special code word that would be inserted to inform the receiver that the line had been deleted. The receiver would then make a best estimate of the missing line based on the signal that it already has and the current control mode of the receiver (see, for example, Ref. 27). The line would be corrected by normal updating of the moving area. One would like to use a channel-sharing strategy that fairly evenly distributes deleted lines among the updated lines of all users and thus minimizes the possibility of deleting consecutive lines from one source.

4.3 Comments on conditional element-difference vs. conditional frame-difference coding

As mentioned in the introduction, transmission of element differences and transmission of frame differences are complementary in many ways. When transmitting frame differences, it is easier to control smoothly the temporal resolution since we are working directly with frame differences. We can still achieve a similar result when trans-

---

* The results of Haskell indicate that the variation in channel-rate requirements is only about 10 percent for 20 sources. However, there are a number of reasons why the variation could increase significantly in an actual system: (*i*) Channel-sharing schemes which minimize the buffering requirements would increase the variation. (*ii*) The interframe coders feeding the channel-sharing unit may be of different types. (*iii*) The channel-sharing unit may be designed for fewer sources or may have priority channels with different types of signal (e.g., data).

mitting element-difference signals: essentially the same signals as for frame-difference transmission are available to the transmitter on which to base a decision as to how elements should be coded. Similarly, it is easier to smoothly control spatial resolution when transmitting element differences, although we can achieve similar ends with frame-difference transmission (e.g., the horizontal subsampling used by Candy et al.).[6] Probably of most importance is the effect on the overall system of using one type of signal or another.

One tempting technique would be transmission of a frame difference in stationary or slowly moving areas and an element difference in fast moving areas or transmission of an element-difference-of-a-frame-difference.[9] Either method would tend to increase complexity associated with system considerations such as recoding, error mitigation, and channel sharing. It is also interesting that Wendt's results suggest to him that transmission of an intraframe coded signal is preferable to either transmission of a frame-to-frame coded signal or transmission of both signal types.

V. SUMMARY

We have described techniques for frame-to-frame coding in which the moving areas are transmitted as an *intraframe* coded signal (rather than as a PCM or frame-to-frame difference signal). This approach permits the intraframe encoding to efficiently adapt to the spatial resolution requirements of the moving area as the speed of an object changes. A coder has been constructed which uses a differential quantizer (DPCM coder) as the intraframe coder, and a strategy was developed for merging the new differentially quantized signal from the moving area with the old differentially quantized and stored signal from the stationary area with only transient error.

Because of inherent noise in the input signal and the error introduced in the initial coding, adequate detection of moving areas requires relatively complex processing involving a nonlinear, time-varying filter with an impulse response that extends temporally and spatially. The bit rate is kept within the capacity of the channel by feedback from the buffer to both the intraframe coder and the movement detection logic. As the buffer fills, the feedback reduces the accuracy of the intraframe encoding (and hence the bit rate) in four steps by a method referred to in an earlier paper as "level-variable sampling."[18] The feedback to the movement detector involves changing not only the level of significant frame-to-frame difference but also the parameters of the spatio-temporal filter contained in the movement detector.

The experimental study used a head-and-shoulders view occupying slightly less than half the field of view and a visible raster size of 255 lines by 220 elements. For a bit rate of 1.5 megabits per second, the picture quality sank below "fair" only for motion covering the entire field such as occurs when the subject stands up in front of the camera.

Transmission of an intraframe coded signal in the moving area leads to a number of advantages from the overall systems point of view when compared with the transmission of frame differences. By starting each transmitted segment within a line with a PCM value, updating becomes independent of previously transmitted data. Thus, errors will not propagate from frame to frame within the moving area. This also has implications for sharing a high-rate channel with a number of users where it would occasionally be necessary to delete segments of data. The signal is stored in the frame memory at the coder and decoder in intraframe coded form. This means that the frame memory need be only approximately half of that required to store the PCM signal. Further, since the stored signal can be simply converted to the form of the transmitted signal, we can use the data-interleaving technique to significantly reduce buffer requirements.

VI. ACKNOWLEDGMENTS

This study has in many ways grown out of the experience of, and discussions with, our colleagues over quite a long period of time. We express our thanks in particular to J. C. Candy, D. J. Connor, B. G. Haskell, F. W. Mounts, and W. G. Scholes.

APPENDIX

Description of Conditional Replenishment Intraframe (CR/IR) Coder

The picture format used in this study is similar to that used in the *Picturephone*® visual telephone system. There are 271 lines per frame, of which 255 are visible; 248 elements per line, of which 223 are visible; and 30 frames per second with 2:1 interlace.

The coding system that has been simulated consists of two parts, the primary intraframe encoding stage which is an element-differential quantizer and the secondary encoding stage which uses interframe techniques (Fig. 8). The output signal from the primary encoding stage is in normal differentially quantized form rather than coded differential form, thus avoiding the need for an additional decoder before the secondary encoder.

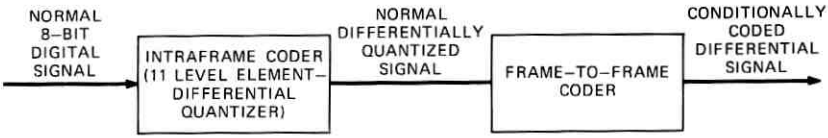

Fig. 8—Configuration of experimental CR/IR coder system.

A.1 Primary intraframe coder

The encoder used in this loop is an 11-level element-differential quantizer whose input, for our purposes, is a normal eight-bit digital (PCM) signal from an A/D converter, but in other respects is similar to that described in Ref. 12. In an actual system, the input would be analog rather than digital, but for experimental purposes it is more convenient to work with the digital signal. As shown in Fig. 9, it contains a decoder section whose output is the normal differentially quantized signal.

In the experiments to be described here, the accumulator loop has no "leak." However, the integrator is reset to a fixed value at the beginning of each line. The quantizer decision and representative levels are given in Table IV.

A.2 Movement detection

Since the outputs of decoder 1 and decoder 2b (Fig. 3) are separated in time by exactly one frame, they are used to form a frame-difference signal.* Frame differences caused by noise (negatively correlated in the moving area)[14] can be separated from those caused by motion

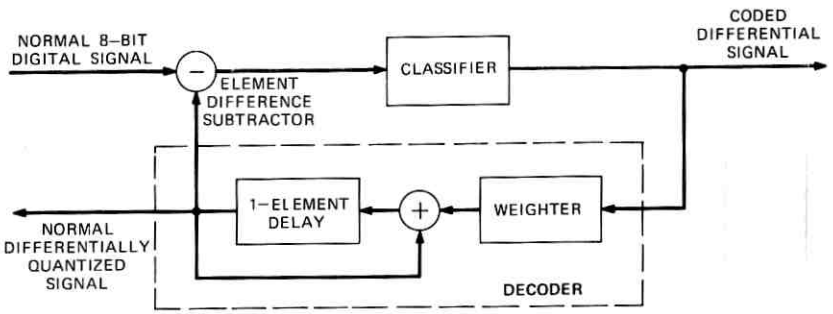

Fig. 9—Differential quantizer (DPCM coder) used as primary intraframe coder.

* For convenience, the term "frame difference signal" will be used, although it is actually the difference between a stored frame and a new frame, both of which have been differentially quantized.

Table IV — Quantizer level settings used by differential quantizer

Level No.	Decision Level (out of 256)	Representative Weight (out of 256)
0	2	0
± 1 inner levels	6	4
± 2	12	8
± 3	22	16
± 4	36	28
± 5 outer levels		44

(mostly positively correlated) by employing various spatial and temporal filtering operations. In addition, certain compensations are made for the nonlinear nature of the coding noise.

A block diagram of the moving area detector used in this experiment is shown in Fig. 10. The frame-difference signal is first fed to a spatial filter which provides a simple average over four adjacent elements along a line (4 × 1 filter). Temporal, single-pole filtering is then provided by placing the spatial filter in a feedback loop with a field delay. Since noise in the frame-difference signal is negatively correlated only in the updated area,[14] we would like to use temporal low-pass filtering only when updating occurs. This is achieved by closing the feedback loop (via switch 1) only when movement is detected. Since it would be expensive to delay a six-bit signal for the duration of one field, a different method was used. A three-bit dither signal was added (adder 1) to the output of the 4 × 1 spatial filter and the resulting sign-bit was used as a one-bit representation of the signal.* The field-delayed signals from the line above and below the current line are added and then assigned a "value" or "weight" before being added (adder 3) back into the frame-difference signal. The loop-gain, or the amount of temporal filtering, is controlled by means of the weighter. The spatio-temporal impulse response of this filter is rather unusual, spreading vertically as well as temporally and horizontally because a field delay, rather than a frame delay, is used (see Fig. 11).

The output of the spatio-temporal filter is then converted from 2's complement to sign-magnitude form (Fig. 10). A modified version of

* Other more complex one-bit representations could have been used; one-bit companded delta modulation, four-bit PCM samples at one-fourth of the sample rate.

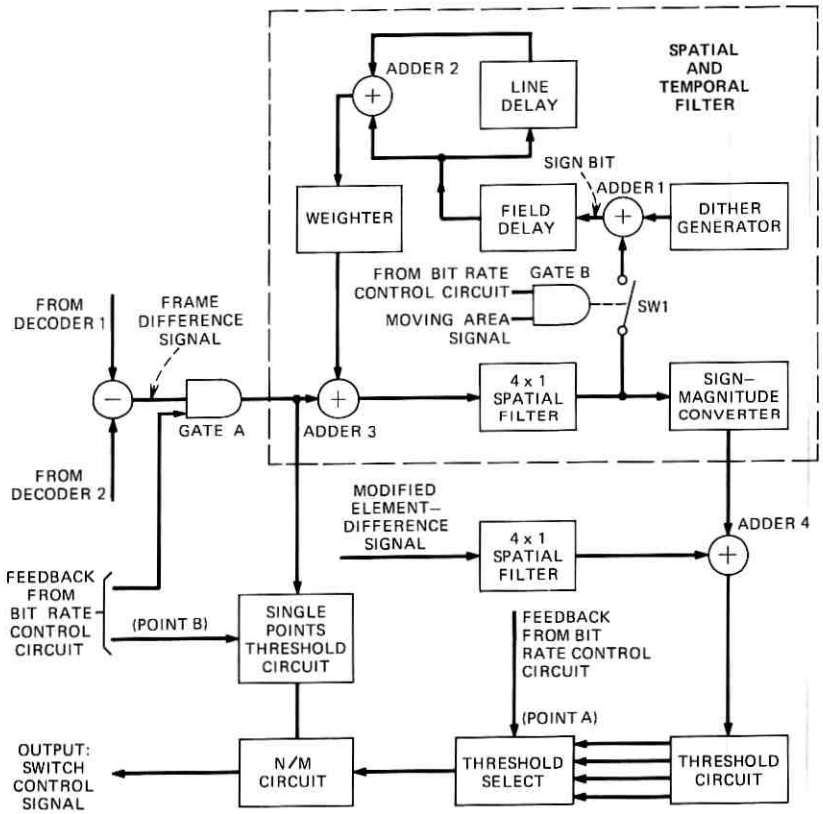

Fig. 10—Diagram of the moving area detector. The detector first filters the frame-difference signal spatially and temporally and then applies compensation for intra-frame coding noise. The filtered signal is tested against one of several thresholds, and the resulting binary signal is blocked in using the N/M circuit. The output is used to select moving areas of the picture for transmission.

the coded-differential signal is added (adder 4) to the output of the sign-magnitude converter.[15] The purpose of this signal is to compensate for areas of the picture where more coding noise is likely to appear, namely at sharp edges where the outer decision levels of the quantizer are used.

Next, the output of adder 4 is fed to a circuit consisting of several thresholds. One of these thresholds is then chosen (depending on the bit-rate control strategy being used) as the input to an N/M circuit.[13] The function of the N/M circuit is to block in the moving area; that is, adjacent but noncontiguous points along a line are joined together to form one longer segment. In this way, the overall data rate is

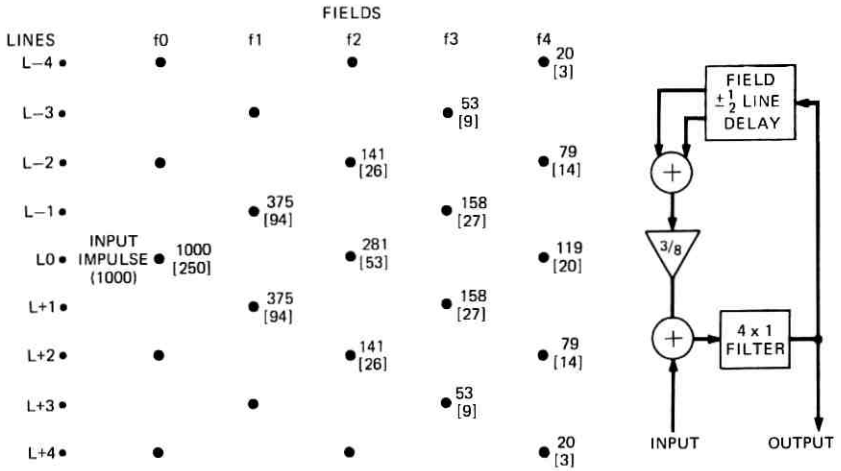

Fig. 11a—Impulse response of the spatio-temporal filter used in the moving area detector. The impulse response is a function of three dimensions, and the figure shows the response only in the vertical and temporal directions. The upper figures represent the area under the horizontal impulse response for each affected line in five fields. The lower (bracketed) figures represent the maximum value of each horizontal impulse response.

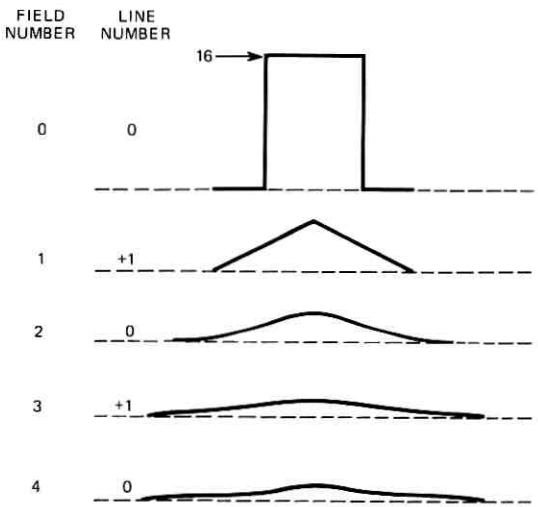

Fig. 11b—Horizontal impulse response of the spatio-temporal filter shown in Fig. 11a. The waveshape is given for line 0 and line 1 for each field of Fig. 11a.

reduced since each segment, however short, is allocated a 12-bit start-stop code, whereas the number of bits required to specify the amplitude of an element may be only 1 or 2.

By using the above-mentioned filtering techniques, large moving areas are easily detected; however, small isolated moving objects cause frame differences that, because of their short duration, are filtered out. Small moving objects of high contrast are detected by thresholding the unfiltered frame-difference signal with a large threshold value. The threshold signal is combined logically with the main signal path in the N/M circuit. The output of the N/M circuit is the final output of the moving area detector and controls the selection and transmission of new data (switches 1 and 2, Fig. 3).

A.3 Bit-rate control

The data-generation rate is matched to the transmission-bit rate by monitoring the level of fill of the transmission buffer and then applying controls to reduce the data-generation rate accordingly. These controls are applied to two parts of the system: the secondary element-differential encoder and the movement detector shown in Fig. 12.

A.3.1 Coder control

To reduce data in the encoder, a technique referred to as level-variable sampling is used.[17,18] The filtered energy in the error signal is important to the visibility of the quantizing error. Thus, close spacing of the inner levels insures that, where the input signal is fairly constant

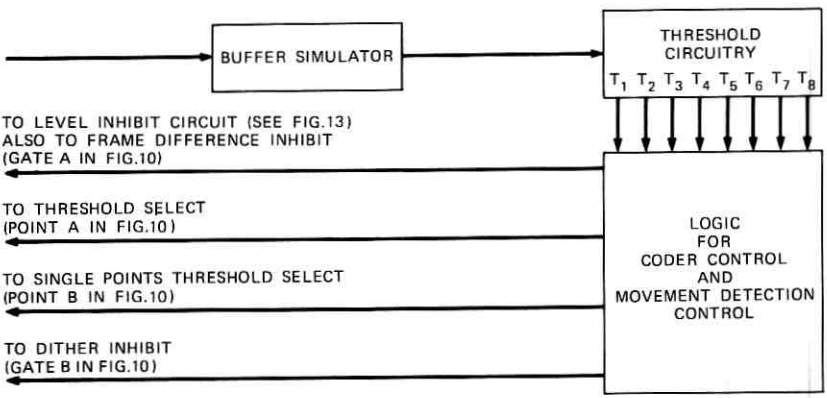

Fig. 12—Bit-rate control system. The system selects one or a combination of several bit-rate reduction techniques, depending on the level of the buffer simulator.

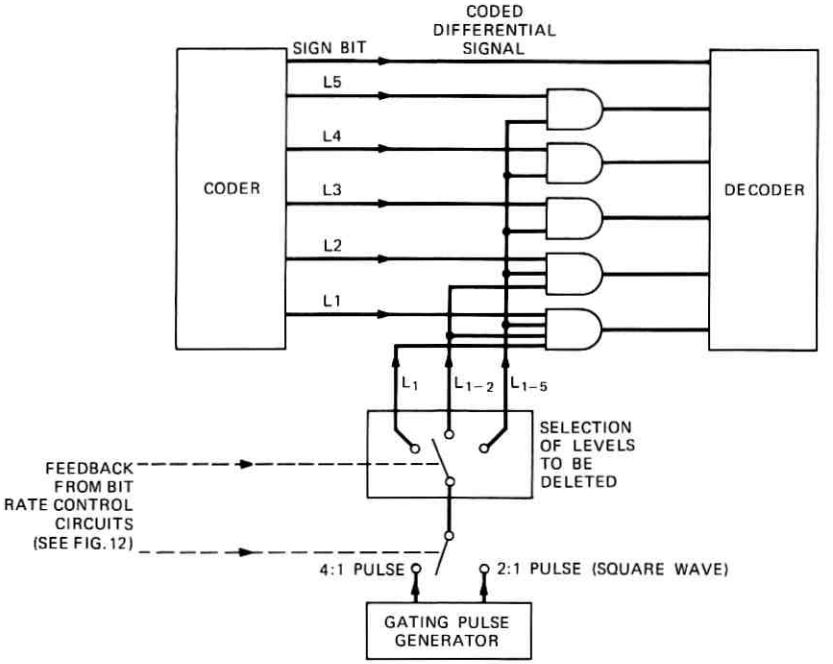

Fig. 13—Level-variable sampling. One or more of the coded differential level pairs is inhibited on every alternate element or, as an extreme measure, the signals are inhibited on three out of every four elements.

(low-detail area), the output signal will approximate the input very closely. However, such precision is not needed on every picture element. Consequently, the inner levels can be used less frequently than the outer levels.[17]

Figure 13 shows how level-variable sampling is performed. If, for example, we subsample just the inner pair of levels, L_1 is inhibited on every alternate element along the line. The effect of inhibiting a level is to change the quantizer scale, for that element, from an 11-level to a 9-level quantizer, as shown in Fig. 14. Two steps are taken to minimize the visibility of the resulting distortion: the subsampling pattern is synchronized to the horizontal rate; the pattern is staggered (by one element for 2:1 and two elements for 4:1 subsampling) so that subsampled elements are offset relative to the subsampled elements of the lines above and below (which are in the other field).

Control of the amount of data-rate reduction is achieved by switching between five different coder states. They are: (*i*) full sampling; (*ii*)

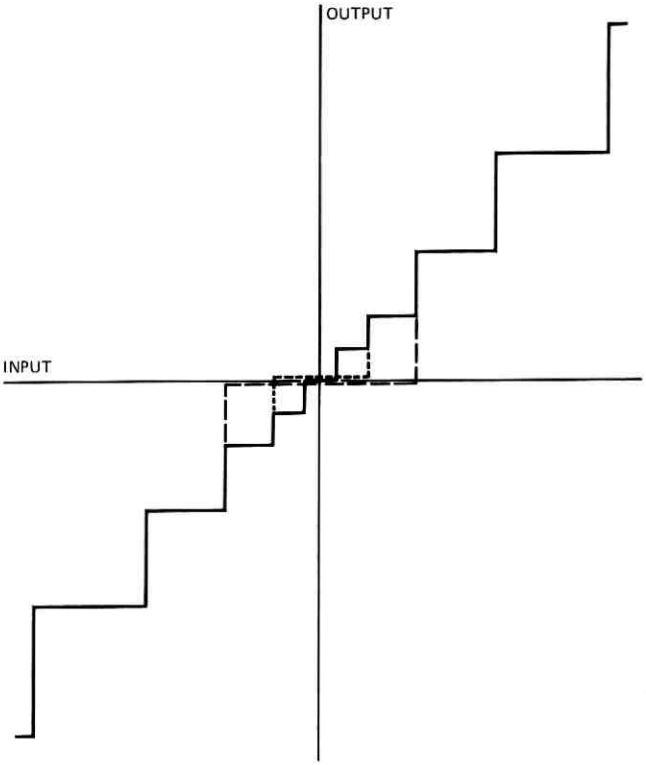

Fig. 14—Change in quantizer scale because of level variable sampling. Inhibiting the inner pair of coded differential levels (± 1) changes the quantizer characteristics to that shown by the short dashed line. Inhibiting the two inner pairs of levels changes the characteristics to that shown by the long dashed line. On the elements in which no level inhibition takes place, the quantizer scale returns to normal (solid line).

subsampling on only the inner pair of levels (levels ± 1); (*iii*) subsampling on the inner two pairs of levels (± 1 and ± 2); (*iv*) subsampling on all levels at a 2:1 rate; (*v*) subsampling on all levels at a 4:1 rate (1 element in 4 is sampled). A description of the variable word-length coding and the efficiencies achieved is given in Section A.3.3.

Subsampling introduces additional noise into the coding operation, particularly on the elements that are not sampled. This makes detection of the moving area more difficult especially since the signal from the primary coder is not subsampled. This problem is partially alleviated by setting the frame difference signal to zero on the unsampled elements by means of gate A in Fig. 10.

A.3.2 Movement detector control

As shown in Fig. 10, the movement-detector circuit generates a filtered frame-difference signal. Moving areas are detected by testing to see if the amplitude of this signal is above a certain threshold. By raising that threshold, less area will be detected as moving and less data will have to be transmitted. This is done, however, at the expense of reducing the quality of the picture in moving areas. Five different thresholds are used, as shown in Table II, so that the data rate can be reduced gradually.

Two other methods are used to reduce the amount of area detected by the movement detector: first, the feedback from the temporal filter is inhibited by gate B and switch 1 in Fig. 10, and second, the number of single points in the moving area is reduced.

The combined effect of the movement-detector controls is to gradually reduce the data rate and also to adapt the movement detector as the speed of movement increases so as to maximize its efficiency. Normally, when the subject is moving slowly, the amount of data being generated is small and the first mode of the movement-detector is used [i.e. (*i*) temporal filtering and (*ii*) low single-point threshold]. As the speed increases, the higher modes are used; the feedback from the temporal filter is inhibited, the filtered threshold is raised, and the single-point threshold is raised. Notice that the frame-difference signal resulting from faster movement is also larger.

A.3.3 Buffer simulator

A buffer simulator circuit was built to simulate operation at many different data rates. It is assumed that a variable-length code is used to transmit the coded differential signal. The lengths assigned to each classifier output are given in Table Va. Notice that the code changes depending on the particular coding mode that is being used. For example, for the mode where levels ± 1 are deleted on alternate samples, the fully coded samples use code D with a maximum code-word length of 4 bits, whereas the alternate samples use code B with a maximum code-word length of 5.

Because the quantizer level usage changes from picture to picture, the codes of Tables Va and Vb will not always be optimum. In order to determine what could be gained by paying more attention to the code assignment (e.g., an adaptive strategy), we have calculated the efficiency of these codes for two different head-and-shoulders scenes. The entropy, bit rate, and efficiency for codes A, B, and C in one case and code D in another are given in Table VI. The results are given for four

Table Va — Four variable-word-length codes used in the CR/IR code

Level	Code Word Length			
	Code *A*	Code *B*	Code *C*	Code *D*
0	1	1	1	3
±1	3	—	—	3
±2	4	3	—	3
±3	5	4	3	4
±4	6	5	4	4
±5	6	5	4	4

Table Vb — The particular code used by each bit-rate control mode

Mode 1 Full Sampling	Modes 2, 3 Levels ± 1 Delete		Mode 4 Levels ± 1, ± 2 Delete		Modes 5, 6, 7 2:1 and 4:1
	Unconditional Samples	Conditional Samples	Unconditional Samples	Conditional Samples	
Code *A*	Code *D*	Code *B*	Code *D*	Code *C*	Code *D*

modes corresponding to (see Table II): (*i*) full sampling; (*ii*) deletion of levels ±1; (*iii*) deletion of levels ±1 and ±2; and (*iv*) 2:1 subsampling. The asterisk denotes the codes that are actually used in the implementation. Picture *X* has somewhat more detail than picture *Y*. It would have been slightly more efficient to use code *D* for mode 1 rather than code *A* for these particular pictures.

The entropies are rather high for an 11-level differential quantizer. The reason for this in mode 1 is that the moving area detector will update moving edges and highly detailed areas more frequently than low-detail areas, resulting in higher usage of the outer levels which, in turn, increases the first-order entropy. For the other modes, the tendency for the entropy of the unconditional picture elements to increase because of the deletion of levels on the alternate elements is almost balanced by the reduction in the amplitude of the element-to-element difference caused by camera integration.

In all cases except one, the efficiency of the variable-length code is greater than 90 percent. For the conditional elements in mode 4, the distribution is very peaked and the entropy is less than 1 bit per

Table VI — Entropy, bit rate, and efficiency for pictures X and Y

Mode	Scene	Entropy	Code A,B,C (as applicable)	Efficiency (%)	Code D	Efficiency (%)
1	X	2.946	3.246*	90.8	3.057	96.4
	Y	3.146	3.433*	91.6	3.297	95.4
2	X Unconditional	3.072	3.517	87.3	3.277*	93.7
	Y Unconditional	3.344	3.847	86.9	3.396*	98.5
	X Conditional	2.080	2.147*	96.9	3.067	67.8
	Y Conditional	2.551	2.727*	93.5	3.210	79.5
4	X Unconditional	2.987	3.105	96.2	3.004*	96.2
	Y Unconditional	2.972	3.304	90.0	3.253*	91.4
	X Conditional	0.822	1.301*	63.2	2.854	28.8
	Y Conditional	0.866	1.286*	67.3	3.000	28.9
5	X	3.146	3.752	83.8	3.383*	93.0
	Y	3.157	3.449	91.5	3.309*	95.4

* Code used in implementation.

element. To improve efficiency, it would be necessary to code elements in groups rather than singly. However, in this case the overall gain would be small.

The heart of the buffer simulator is an accumulator loop. For each transmitted sample the accumulator is incremented by an amount equal to the length of the corresponding code word. In addition, a count of 12 is added every time a new segment is transmitted to the receiver; this could be eight bits for a start-of-run address and four bits for an end-of-run code word. A count of 12 is also added to the accumulator at the start of each line to permit the decoder to synchronize at the start of line. No allocation is made for a start-of-frame code word (if a 50-bit code word were used, we would have to say that we are operating at 1.503 megabits per second rather than 1.500 megabits per second). The accumulator is decremented at a constant rate depending on the particular transmission rate that is being simulated. Thus, the output of the buffer simulator shows how full a buffer would be if it were actually used to transmit data to the receiver.

A circuit similar to this is used to monitor the data-generation rate. The accumulator is incremented with the same signal as the buffer simulator, but at the end of each line the contents are strobed into a

commercial counter which enables us to integrate the data-rate count over any desired period.

A.3.4 Bit-rate control system

The bit-rate control system (Fig. 12) monitors the buffer simulator. The output range of the buffer is divided into eight regions, and a control mode is selected depending upon which region the buffer is in. Each mode uses a combination of the two previously described bit-rate reduction techniques (i.e., coder control and movement-detector control). The function of each mode is given in Table II.

In the lower modes, little or no level-dependent sampling occurs and the movement detector uses a low threshold. The movement detector completely covers the moving areas, but may also respond to a small amount of residual noise so that some stationary areas of the picture may also be detected. The result is that for limited subject activity a relatively large amount of data is generated and, correspondingly, the quality is little different from the primary encoder output.

As the buffer level increases, the intermediate modes (modes 2 to 4) are used. In these modes, level-variable sampling is used on the inner one or two pairs of levels; the moving-area detector operates on a higher threshold. As the buffer level increases further, the high modes (modes 5 to 7) are used. Subsampling is used on all classifier levels: at first in a 2:1 ratio and then finally (in mode 7) in a 4:1 ratio. The moving area detector coverage is reduced in two ways: first, the single point threshold is raised so that fewer single points are detected; second, in each consecutive mode the moving area detector threshold is raised. Normally, when the high modes are used it is because the subject is moving fast. Under these conditions the effects of these bit-rate reduction modes is somewhat masked because of the nature of human vision. Furthermore, since large frame-difference signals are generated, the moving area can still be accurately defined even though high moving-area-detector thresholds are used.

If the buffer level continues to rise, transmission of data is stopped. In this case, the receiver repeats the information from the previous frame (stored in its frame memory) until such time as the transmitter buffer level reduces sufficiently to allow new data to be transmitted.

REFERENCES

1. R. D. Kell, "Improvements Relating to Electronic Picture Transmission Systems," British Patent No. 341811, April 1929.
2. A. J. Seyler, "An Experimental Frame Difference Signal Generator for the Analysis of Television Signals," Proc. I.R.E. (Aust.), *24*, 1963, pp. 797–807.

3. F. W. Mounts, "A Video Encoding System Employing Conditional Picture-Element Replenishment," B.S.T.J., *48*, No. 7 (September 1969), pp. 2545–2554.
4. R. E. Graham, "Predictive Quantizing of Television Signals," IRE WESCON Convention Record, Pt. 4, 1958, pp. 142–157.
5. K. Teer, "Investigations into Redundancy and Possible Bandwidth Compression in Television Transmission," Philips. Res. Rep., *14* (1959), pp. 501–556 and *15* (1960), pp. 30–96.
6. J. C. Candy, M. A. Franke, B. G. Haskell, and F. W. Mounts, "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1889–1917.
7. J. O. Limb, "Buffering of Data Generated by the Coding of Moving Images," B.S.T.J., *51*, No. 1 (January 1972), pp. 239–259.
8. B. G. Haskell, J. O. Limb, and R. F. W. Pease, "Frame-to-Frame Redundancy Reduction System Which Transmits an Intraframe Coded Signal," U. S. Patent No. 3,767,847, October 23, 1973.
9. H. Wendt, Interframe-Codierung fur Videosignale Internat. Elektron. Rundschau, *27*, No. 1 (January 1973), pp. 2–6.
10. F. Kanaya, Nippon Telegraph and Telephone, private discussions.
11. C. C. Cutler, "Differential Quantization of Communication Signals," U. S. Patent No. 2,605,361, July 29, 1952.
12. J. O. Limb and F. W. Mounts, "Digital Differential Quantizer for Television," B.S.T.J., *48*, No. 7 (September 1969), pp. 2583–2599.
13. R. F. W. Pease and J. O. Limb, "Exchange of Spatial and Temporal Resolution in Television Coding," B.S.T.J., *50*, No. 1 (January 1971), pp. 191–200.
14. D. J. Connor and J. O. Limb, "Properties of Frame Difference Signals Generated by Moving Images," to be published.
15. B. G. Haskell, D. J. Connor, and F. W. Mounts, "A Frame-to-Frame *Picturephone* Coder for Signals Containing Differential Quantizing Noise," B.S.T.J., *52*, No. 1 (January 1973), pp. 35–51.
16. R. T. Bobilin, "A Study of Prefilters and Sampling Rate for Intraframe *Picturephone* Coders," unpublished memorandum.
17. J. O. Limb, "Picture Coding: The Use of a Viewer Model in Source Encoding," B.S.T.J., *52*, No. 8 (October 1973), pp. 1271–1302.
18. J. O. Limb, "Adaptive Encoding of Picture Signals," M.I.T. Conference on Bandwidth Compression, April, 1969; Published in T. S. Huang and O. J. Tretiak, *Picture Bandwidth Compression*, New York: Gordon and Breach, 1972, pp. 341–382.
19. J. O. Limb, "A Picture Coding Algorithm for the Merli Scan," IEEE Transactions on Communications, *COM-21*, No. 4 (April 1973), pp. 300–305.
20. R. F. W. Pease, "Conditional Vertical Subsampling—A Technique to Assist in the Coding of Television Signals," B.S.T.J., *51*, No. 4 (April 1972), pp. 787–802.
21. E. F. O'Neill, "TASI: Time Assignment Speech Interpolation," Bell Laboratories Record, *37*, No. 3 (March 1959), pp. 83–87.
22. B. G. Haskell, "Buffer and Channel Sharing by Several Interframe *Picturephone* Coders," B.S.T.J., *51*, No. 1 (January 1972), pp. 261–289.
23. E. R. Kretzmer, "Reduced Bandwidth Transmission System," U. S. Patent No. 2,949,505, August 16, 1960.
24. K. Fukushima and H. Ando, "Television Band Compression by Multimode Interpolation," J. Inst. Elec. Commun. Eng. Jap., *47*, 1964, pp. 55–64.
25. P. A. Wintz, "Transform Picture Coding," Proc. IEEE, *60*, No. 7 (July 1972), pp. 809–820.
26. J. C. Candy and F. W. Mounts, "Redundancy Reduction System with Data Editing," U. S. Patent No. 3,571,807, March 23, 1971.
27. D. J. Connor, "Techniques for Reducing the Visibility of Transmission Errors in Digitally Encoded Video Signals," IEEE Trans. Commun., *COM-21*, No. 3 (June 1973).

Contributors to This Issue

Harry W. Astle, A.A.S., 1966, Hartford State Technical College, Bell Laboratories, 1966—. Mr. Astle has worked with optical gas lenses and is presently engaged in the fabrication and evaluation of optical fibers.

Mary-Jane Cross, A.B. (Mathematics), 1966, University of Massachusetts; M.S. (Mathematics), 1972, New York University; Bell Laboratories, 1966—. Ms. Cross is presently working on network modeling. Member, Phi Beta Kappa, Phi Kappa Phi, SIAM, MAA, MAA Committee on High School Lecturers.

Larry J. Greenstein, B.S.E.E., 1958, M.S.E.E., 1961, and Ph.D. (E.E.), 1967, Illinois Institute of Technology; Bell Laboratories, 1970—. Since joining Bell Laboratories, Mr. Greenstein has engaged in studies of digital encoding, processing, and transmission. He is supervisor of a group responsible for research in digital radio repeaters. Member, AAAS, IEEE.

Peter Kaiser, Diplom Ingenieur, 1963, Technical University, Munich, West Germany; M.S., 1965, and Ph.D., 1966, University of California, Berkeley; Bell Laboratories, 1966—. At Berkeley Mr. Kaiser worked on frequency-independent antennas. At Bell Laboratories, Mr. Kaiser has been engaged in optical transmission research, including the design and testing of gas-lens beam waveguides and, more recently, in the development and characterization of low-loss optical fibers. Member, IEEE, American Optical Society.

John O. Limb, B.E.E., 1963, and Ph.D., 1967, University of Western Australia; Research Laboratories, Australian Post Office, 1966–1967; Bell Laboratories, 1967—. Mr. Limb has worked on the coding of picture signals to reduce channel capacity requirements involving intraframe coding, frame-to-frame coding, and the coding of color signals. He currently heads the Visual Communication Research Department. Member, IEEE, Association for Research in Vision and Opthalmology, Optical Society of America.

Dietrich Marcuse, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, and studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of three books. Fellow, IEEE; member, Optical Society of America.

J. A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has been doing research in a variety of problems in mathematical physics and applied mathematics. His recent interests have included the theory of stochastic differential equations and propagation in random media. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester of 1968. Member, American Mathematical Society, SIAM, Sigma Xi.

R. F. W. Pease, B.A., 1960, M.A. and Ph.D., 1964, University of Cambridge; Bell Laboratories, 1967—. Before joining Bell Laboratories, Mr. Pease held a faculty appointment at the University of California at Berkeley, where he worked on electron microscopy. At Bell Laboratories he has worked on the digital encoding of television signals. Presently, he is engaged in using electron beams to make integrated circuits.

Lawrence R. Rabiner, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; Member of the IEEE G-ASSP Technical Committee on Digital Signal Processing; President of the G-ASSP AdCom, associate editor of the G-ASSP Transactions; member of the technical committees on speech communication of both the IEEE and Acoustical Society.

Lewis H. Rosenthal, S.B. (Electrical Engineering) and S.M. (Electrical Engineering), 1974, Massachusetts Institute of Technology; Bell Laboratories, 1971—. Mr. Rosenthal has worked in the areas of loop transmission, digital terminal design, and automatic voice response. Member, Eta Kappa Nu, Tau Beta Pi.

Ronald W. Schafer, B.S. (E.E.), 1961, and M.S. (E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Laboratories, 1968—. Mr. Schafer has been engaged in research on digital waveform processing techniques and speech communication. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America, and the IEEE G-ASSP Technical Committees on Digital Signal Processing and Speech Communication; associate editor of the IEEE Transactions on Acoustics, Speech, and Signal Processing.

K. K. Thornber, B.S., 1963, M.S. (E.E.), 1964, Ph.D. (E.E.), 1966, California Institute of Technology; Research Associate, Stanford Electronics Laboratories, 1966–68; Research Assistant, Physics Department, University of Bristol, 1968–69; Bell Laboratories, 1969–. Mr. Thornber is a member of the Unipolar Integrated Circuit Laboratory. Member, Sigma Xi, Tau Beta Pi.

Kenneth A. Walsh, Associate Degree (electrical engineering technology), Kent State University, Salem, Ohio, 1969; Bell Laboratories, 1969—. Mr. Walsh's work at Bell Laboratories has been mainly concerned with efficient coding of digital *Picturephone*® transmissions, including the use of a buffered, bit-rate controlled coder, the use of channel sharing, and transmission of color *Picturephone* signals.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 6, July–August 1974
Printed in U.S.A.

B.S.T.J. BRIEF

Stripline Downconverter With Subharmonic Pump

By M. V. SCHNEIDER and W. W. SNELL, JR.

(Manuscript received April 19, 1974)

I. INTRODUCTION

The process of frequency conversion and its applications are well known and have been extensively treated in the literature.[1–3] The conversion is usually performed by pumping a nonlinear resistive or reactive element embedded in a linear network and by extracting the sum or difference frequencies that are generated by the signal and the pump frequency. The purpose of this Brief is to describe a novel thin-film converter* which has the following properties:

(*i*) The pump frequency required for efficient upconversion or downconversion is a submultiple of that needed in conventional frequency converters.
(*ii*) The circuit does not require a dc return path.
(*iii*) The separation of the signal and the pump frequency is readily obtained and the loss in the signal path is small.

The new converter consists of two stripline filters and two Schottky barrier diodes, which are shunt mounted with opposite polarities in a strip transmission line. The conversion loss measured at a signal frequency of 3.5 GHz is 3.2 dB for a pump frequency of 1.7 GHz and 4.9 dB for a pump frequency of 0.85 GHz. The circuit looks attractive for use at millimeter-wave frequencies where stable pump sources with low FM noise are not readily available.

* After the manuscript for this Brief was completed, it was learned that M. Cohn, J. E. Degenford, and B. A. Newman at Westinghouse Electric Corp., Baltimore, Md., have begun independent work along similar lines.

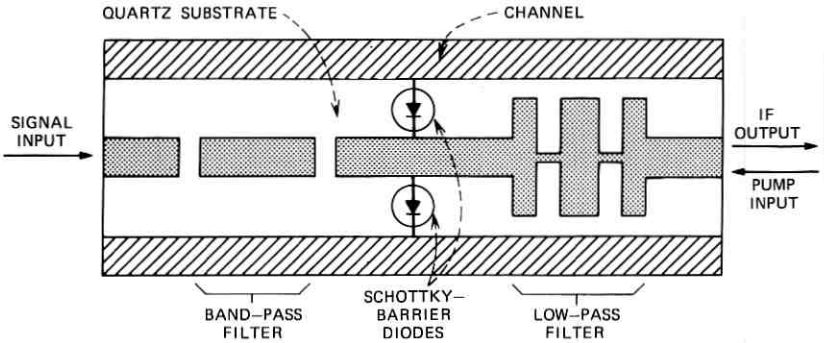

Fig. 1—Microstrip conductor pattern on quartz substrate in a metal channel. The diode pair is shunt mounted with opposite polarities to the ground on opposite sides of the strip transmission line.

II. DESCRIPTION OF STRIPLINE CIRCUIT

A top view of the stripline conductor pattern used in the down-converter is shown in Fig. 1 and a cross-sectional view is shown in Fig. 2. A strip transmission line is used because the conversion from the hybrid TEM mode to the first-order waveguide mode (longitudinal section magnetic mode) is substantially reduced compared to the conversion obtained with other transmission line circuits such as microstrip lines.[4] This approach eliminates noise contributions from undesired bands near the harmonics of the pump frequency.

The conductor pattern consists of a 50-ohm line section at the signal input, a half-wavelength resonator for the bandpass filter, a five-element low-pass filter, and a 50-ohm line section for the pump input and the IF output. Two Schottky-barrier diodes with opposite polarities are connected to the section between the filters at opposite

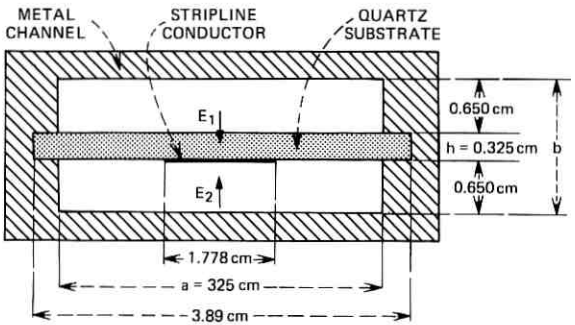

Fig. 2—Cross-sectional view of shielded stripline with symmetrically suspended quartz substrate and stripline conductor.

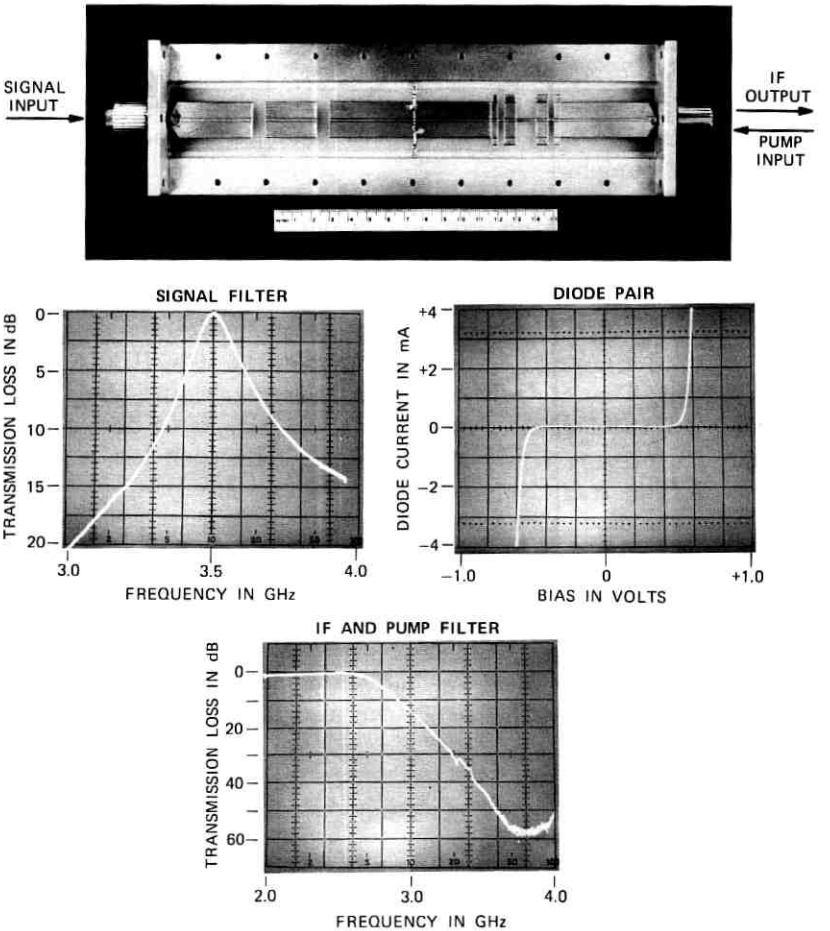

Fig. 3—Photograph of downconverter showing top view of stripline conductor pattern in rectangular channel. The characteristics of the band-pass filter, the low-pass filter, and the diode pair are displayed at the bottom of the photograph.

sides of the stripline conductor. Coupling to undesired waveguide modes above the cutoff frequency of the metal channel is suppressed because the electric field vectors in the top and the bottom section of the stripline are of opposite polarity as indicated in Fig. 2. A photograph of the downconverter is shown in Fig. 3. The figure also shows the measured transmission characteristics of the stripline band-pass filter and low-pass filter, and also the current-voltage characteristics of the diode pair. The current-voltage characteristics of the diode

pair are symmetrical with respect to the origin. This results in a current waveform that has only odd-order harmonics and a conductance waveform with even-order harmonics. The second feature combined with the low conversion to waveguide modes results in a converter that has a good conversion loss and a low noise figure for subharmonic pumping

III. PERFORMANCE OF STRIPLINE CONVERTER

The measured single-sideband noise figure for the stripline converter of Fig. 3 is plotted in Fig. 4 as a function of the signal frequency ω_s for $m = 2$ and $m = 4$, where m is the harmonic integer defined by $m = (\omega_{\text{signal}} \pm \omega_{\text{IF}})/\omega_{\text{pump}}$. The noise figure of the 100-MHz IF amplifier is 1.7 dB. The total single-sideband noise figure, including the IF amplifier noise at a signal frequency of 3.455 GHz, is 4.9 dB for $m = 2$ and 6.6 dB for $m = 4$. The corresponding conversion loss is 3.2 dB for $m = 2$. This result approaches the theoretically predicted loss of 2.1 dB for the diode pair with a series resistance $R_s = 2$ ohms and a zero bias capacitance of $C_0 = 0.45$ pF for each diode.[5]

The new harmonically pumped stripline circuit can be readily scaled to higher microwave frequencies and particularly to millimeter-wave frequencies where solid-state oscillators are only available at subharmonics of the local oscillator frequency. The basic design principles discussed in this paper can also be applied to other con-

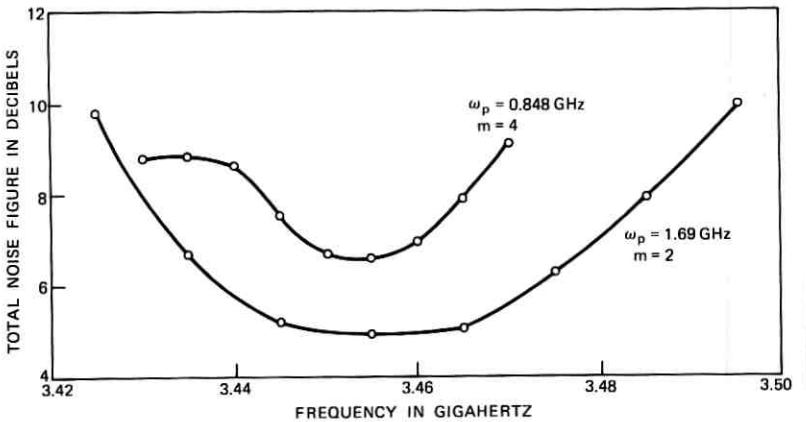

Fig. 4—Single-sideband noise figure including IF amplifier noise for downconverter pumped at the second subharmonic ($m = 2$) and the fourth subharmonic ($m = 4$). The noise figure of the 100-MHz IF amplifier is 1.7 dB.

verters in the electromagnetic spectrum, such as upconverters, harmonic generators, and parametric amplifiers.

REFERENCES

1. C. T. Torrey and C. A. Whitmer, *Crystal Rectifiers*, MIT Radiation Laboratory Series, *15*, New York: McGraw Hill, 1948.
2. C. Dragone, "Amplitude and Phase Modulations in Resistive Diode Mixers," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1967–1998.
3. A. A. M. Saleh, *Theory of Resistive Mixers*, Research Monograph No. 64, Cambridge, Mass.: Massachusetts Institute of Technology Press, 1971.
4. M. V. Schneider and B. S. Glance, "Suppression of Waveguide Modes in Strip Transmission Lines," Proc. IEEE, to be published.
5. C. Dragone, private communication.

Fiber Ribbon Optical Transmission Lines

By R. D. STANDLEY

(Manuscript received April 2, 1974)

This brief proposes the use of fiber ribbons consisting of a linear array of fibers embedded in a thin, flexible supporting medium as components of a cable for fiber transmission systems. With the progress that has been made in drawing low-loss fibers, the physical form used to cable the fibers has become a truly relevant problem and is presently being pursued at several laboratories.

Figure 1 shows some of the structures of interest. The value of ribbons in a transmission cable was initially conceived as relating well to planar technology for connector and repeater circuitry fabrication. A natural layout for repeater electronics is an input consisting of a linear array of detectors with a similar emitter array for the output.

Fiber ribbons should also be easier to handle than conventional bundles. In the event of cable breakage, the ribbon resolves the problem of fiber identification; coding is simple. Ribbons may be easily stacked to form higher-capacity cables. The geometry lends itself well to connector design. For example, suppose the supporting medium to be some sort of plastic. To make fiber separation easy, we cut the ribbon, then we dissolve a portion of the supporting medium to free the fiber ends. The ends are then placed in the connector, which is finally recoated with the plastic.

verters in the electromagnetic spectrum, such as upconverters, harmonic generators, and parametric amplifiers.

REFERENCES

1. C. T. Torrey and C. A. Whitmer, *Crystal Rectifiers*, MIT Radiation Laboratory Series, *15*, New York: McGraw Hill, 1948.
2. C. Dragone, "Amplitude and Phase Modulations in Resistive Diode Mixers," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1967–1998.
3. A. A. M. Saleh, *Theory of Resistive Mixers*, Research Monograph No. 64, Cambridge, Mass.: Massachusetts Institute of Technology Press, 1971.
4. M. V. Schneider and B. S. Glance, "Suppression of Waveguide Modes in Strip Transmission Lines," Proc. IEEE, to be published.
5. C. Dragone, private communication.

Fiber Ribbon Optical Transmission Lines

By R. D. STANDLEY

(Manuscript received April 2, 1974)

This brief proposes the use of fiber ribbons consisting of a linear array of fibers embedded in a thin, flexible supporting medium as components of a cable for fiber transmission systems. With the progress that has been made in drawing low-loss fibers, the physical form used to cable the fibers has become a truly relevant problem and is presently being pursued at several laboratories.

Figure 1 shows some of the structures of interest. The value of ribbons in a transmission cable was initially conceived as relating well to planar technology for connector and repeater circuitry fabrication. A natural layout for repeater electronics is an input consisting of a linear array of detectors with a similar emitter array for the output.

Fiber ribbons should also be easier to handle than conventional bundles. In the event of cable breakage, the ribbon resolves the problem of fiber identification; coding is simple. Ribbons may be easily stacked to form higher-capacity cables. The geometry lends itself well to connector design. For example, suppose the supporting medium to be some sort of plastic. To make fiber separation easy, we cut the ribbon, then we dissolve a portion of the supporting medium to free the fiber ends. The ends are then placed in the connector, which is finally recoated with the plastic.

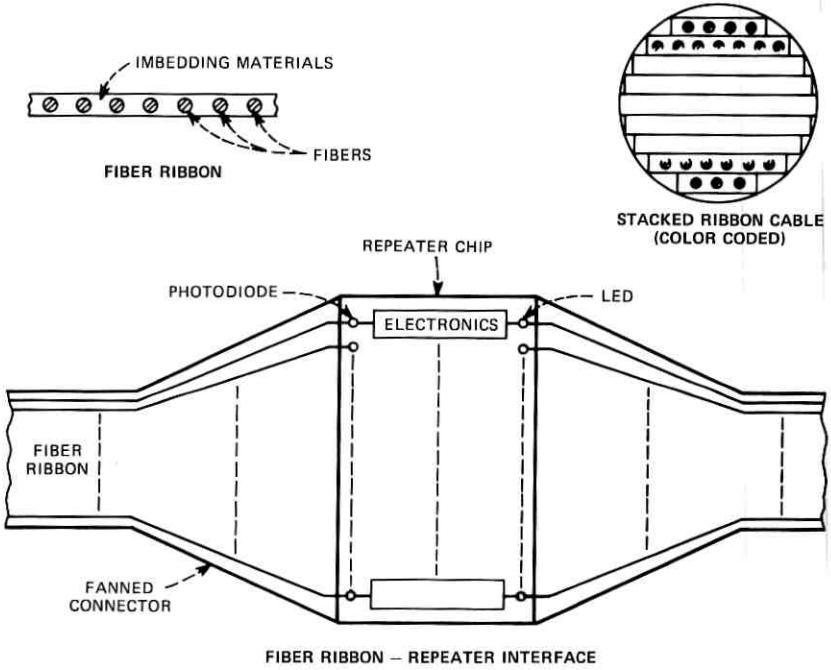

Fig. 1—Some fiber ribbon structures.

Many forms can be envisioned for the connector. For example, consider a glass plate whose refractive index is less than that of the fiber. Using conventional photolithographic techniques, one can etch channels in the glass. The fibers may then be placed in the channels and covered with a second glass plate or a plastic similar to the ribbon support. The output end of the connector can be polished to clean up the fiber ends if necessary.

Finally, the manufacture of ribbons should be straightforward. Two methods are described in the literature.[1,2]

As stated previously, the purpose of this brief has been to describe concepts of fiber ribbon transmission line accessories. It is recognized that practical difficulties will ensue when attempting to reduce any of the concepts to the hardware stage. For example, mechanical tolerances, which will generally be dependent upon the fiber core diameter, are of prime importance in any hardware for any fiber optic transmission line. However, we believe that the naturally planar form of the fiber ribbon, associated connectors, and circuitry described above

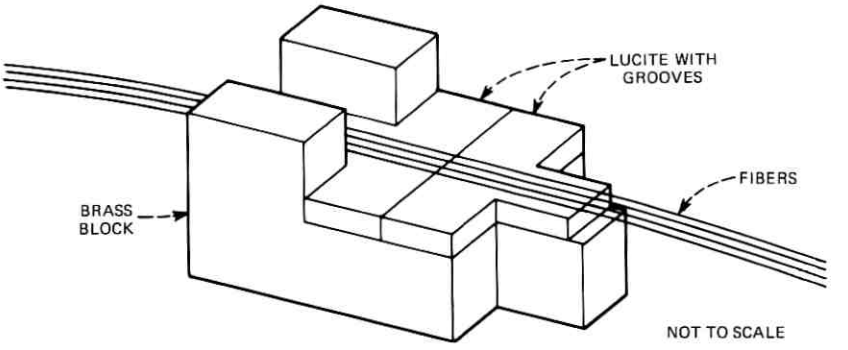

Fig. 2—Grooved lucite planar connector for fiber ribbon transmission line.

would permit excellent dimensional control. Experimental work would be necessary to define quantitative limits.

Some years ago we did experiments on fiber connectors having the form shown in Fig. 2. Here grooves were hot-pressed into lucite blocks using fibers of the same size as those to be mounted as templates. Fibers were then inserted into the grooves and held in place by cement. Typical loss achieved upon disassembly and reassembly was 1 dB $\pm$ 0.5 dB, which was considered acceptable for such a crude structure. In another experiment, one lucite block was made mechanically movable to form a single-pole, double-throw switch; loss variation upon operating the switch was again about 1 dB.

The prospects for near-term use of optical fibers in communications systems are indeed good; what is hoped is that the above concepts will stimulate others in the pursuit of a useful and economic cabling method and, thus, lead to a more rapid application of fibers in practical systems. Recently, a method was proposed for splicing fiber ribbons of the type described above.[3]

REFERENCES

1. U. S. Patents No. 3,247,755 and No. 3,272,063.
2. M. Borner, "Fiber Component Design," Seminar on Optical Fibers in Telecommunications, April 9–13, 1973, Bologna, Italy.
3. E. A. J. Marcatili, U. S. Patent No. 3,789,099, "Methods for Splicing Optical Fibers."

verters in the electromagnetic spectrum, such as upconverters, harmonic generators, and parametric amplifiers.

REFERENCES

1. C. T. Torrey and C. A. Whitmer, *Crystal Rectifiers*, MIT Radiation Laboratory Series, *15*, New York: McGraw Hill, 1948.
2. C. Dragone, "Amplitude and Phase Modulations in Resistive Diode Mixers," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1967–1998.
3. A. A. M. Saleh, *Theory of Resistive Mixers*, Research Monograph No. 64, Cambridge, Mass.: Massachusetts Institute of Technology Press, 1971.
4. M. V. Schneider and B. S. Glance, "Suppression of Waveguide Modes in Strip Transmission Lines," Proc. IEEE, to be published.
5. C. Dragone, private communication.

Fiber Ribbon Optical Transmission Lines

By R. D. STANDLEY

(Manuscript received April 2, 1974)

This brief proposes the use of fiber ribbons consisting of a linear array of fibers embedded in a thin, flexible supporting medium as components of a cable for fiber transmission systems. With the progress that has been made in drawing low-loss fibers, the physical form used to cable the fibers has become a truly relevant problem and is presently being pursued at several laboratories.

Figure 1 shows some of the structures of interest. The value of ribbons in a transmission cable was initially conceived as relating well to planar technology for connector and repeater circuitry fabrication. A natural layout for repeater electronics is an input consisting of a linear array of detectors with a similar emitter array for the output.

Fiber ribbons should also be easier to handle than conventional bundles. In the event of cable breakage, the ribbon resolves the problem of fiber identification; coding is simple. Ribbons may be easily stacked to form higher-capacity cables. The geometry lends itself well to connector design. For example, suppose the supporting medium to be some sort of plastic. To make fiber separation easy, we cut the ribbon, then we dissolve a portion of the supporting medium to free the fiber ends. The ends are then placed in the connector, which is finally recoated with the plastic.

verters in the electromagnetic spectrum, such as upconverters, harmonic generators, and parametric amplifiers.

REFERENCES

1. C. T. Torrey and C. A. Whitmer, *Crystal Rectifiers*, MIT Radiation Laboratory Series, *15*, New York: McGraw Hill, 1948.
2. C. Dragone, "Amplitude and Phase Modulations in Resistive Diode Mixers," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1967–1998.
3. A. A. M. Saleh, *Theory of Resistive Mixers*, Research Monograph No. 64, Cambridge, Mass.: Massachusetts Institute of Technology Press, 1971.
4. M. V. Schneider and B. S. Glance, "Suppression of Waveguide Modes in Strip Transmission Lines," Proc. IEEE, to be published.
5. C. Dragone, private communication.

Fiber Ribbon Optical Transmission Lines

By R. D. STANDLEY

(Manuscript received April 2, 1974)

This brief proposes the use of fiber ribbons consisting of a linear array of fibers embedded in a thin, flexible supporting medium as components of a cable for fiber transmission systems. With the progress that has been made in drawing low-loss fibers, the physical form used to cable the fibers has become a truly relevant problem and is presently being pursued at several laboratories.

Figure 1 shows some of the structures of interest. The value of ribbons in a transmission cable was initially conceived as relating well to planar technology for connector and repeater circuitry fabrication. A natural layout for repeater electronics is an input consisting of a linear array of detectors with a similar emitter array for the output.

Fiber ribbons should also be easier to handle than conventional bundles. In the event of cable breakage, the ribbon resolves the problem of fiber identification; coding is simple. Ribbons may be easily stacked to form higher-capacity cables. The geometry lends itself well to connector design. For example, suppose the supporting medium to be some sort of plastic. To make fiber separation easy, we cut the ribbon, then we dissolve a portion of the supporting medium to free the fiber ends. The ends are then placed in the connector, which is finally recoated with the plastic.

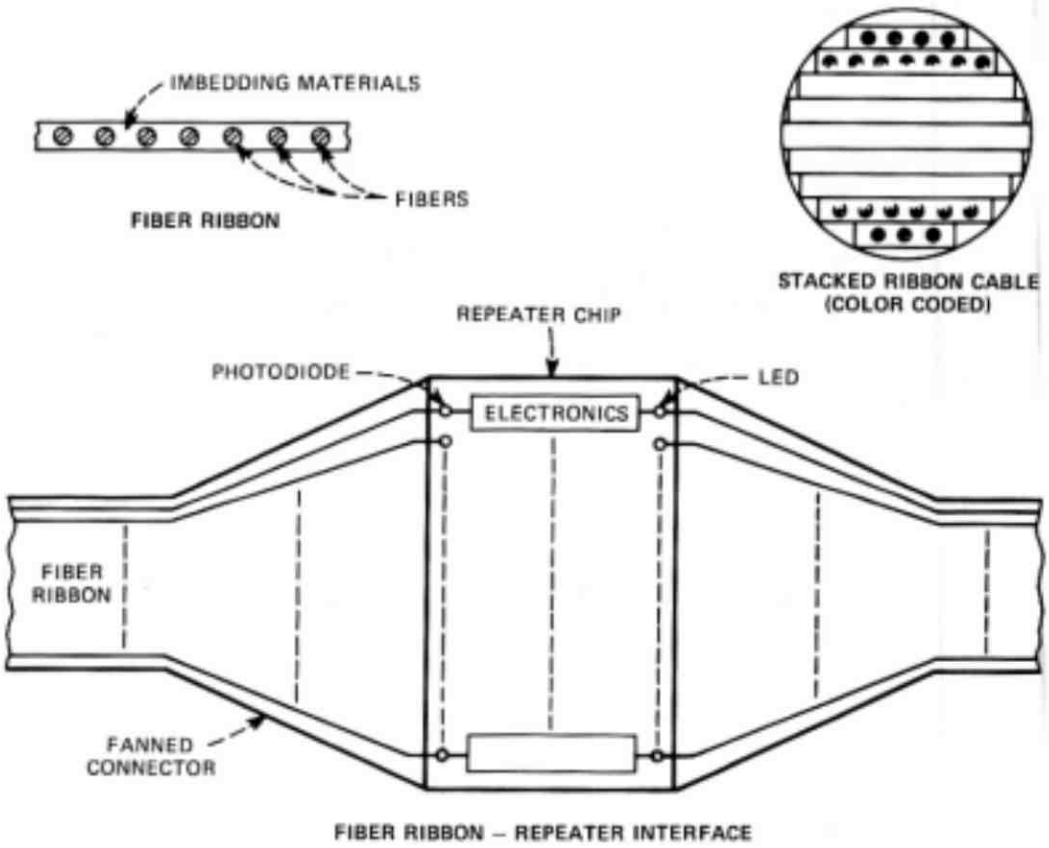

Fig. 1—Some fiber ribbon structures.

Many forms can be envisioned for the connector. For example, consider a glass plate whose refractive index is less than that of the fiber. Using conventional photolithographic techniques, one can etch channels in the glass. The fibers may then be placed in the channels and covered with a second glass plate or a plastic similar to the ribbon support. The output end of the connector can be polished to clean up the fiber ends if necessary.

Finally, the manufacture of ribbons should be straightforward. Two methods are described in the literature.[1,2]

As stated previously, the purpose of this brief has been to describe concepts of fiber ribbon transmission line accessories. It is recognized that practical difficulties will ensue when attempting to reduce any of the concepts to the hardware stage. For example, mechanical tolerances, which will generally be dependent upon the fiber core diameter, are of prime importance in any hardware for any fiber optic transmission line. However, we believe that the naturally planar form of the fiber ribbon, associated connectors, and circuitry described above

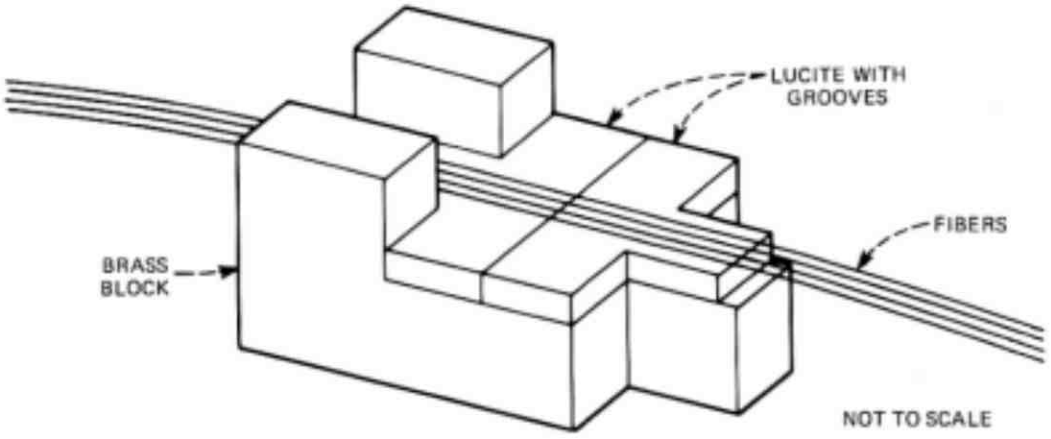

Fig. 2—Grooved lucite planar connector for fiber ribbon transmission line.

would permit excellent dimensional control. Experimental work would be necessary to define quantitative limits.

Some years ago we did experiments on fiber connectors having the form shown in Fig. 2. Here grooves were hot-pressed into lucite blocks using fibers of the same size as those to be mounted as templates. Fibers were then inserted into the grooves and held in place by cement. Typical loss achieved upon disassembly and reassembly was 1 dB $\pm$ 0.5 dB, which was considered acceptable for such a crude structure. In another experiment, one lucite block was made mechanically movable to form a single-pole, double-throw switch; loss variation upon operating the switch was again about 1 dB.

The prospects for near-term use of optical fibers in communications systems are indeed good; what is hoped is that the above concepts will stimulate others in the pursuit of a useful and economic cabling method and, thus, lead to a more rapid application of fibers in practical systems. Recently, a method was proposed for splicing fiber ribbons of the type described above.[3]

REFERENCES

1. U. S. Patents No. 3,247,755 and No. 3,272,063.
2. M. Borner, "Fiber Component Design," Seminar on Optical Fibers in Telecommunications, April 9–13, 1973, Bologna, Italy.
3. E. A. J. Marcatili, U. S. Patent No. 3,789,099, "Methods for Splicing Optical Fibers."