THE BELL SYSTEM

Technical Journal

THE BELL SYSTEM TECHNICAL JOURNAL

Integrated Optics: An Introduction

By STEWART E. MILLER

This paper outlines a proposal for a miniature form of laser beam circuitry. Index of refraction changes of the order of 10^{-2} or 10^{-3} in a substrate such as glass allow guided laser beams of width near 10 microns. Photolithographic techniques may permit simultaneous construction of complex circuit patterns. This paper also indicates possible miniature forms for a laser, modulator, and hybrids. If realized, this new art would facilitate isolating the laser circuit assembly from thermal, mechanical, and acoustic ambient changes through small overall size; economy should ultimately result.

I. INTRODUCTION

Laboratory work and experimental repeater work at laser wavelengths (0.4 to 10 + μm) has been carried out by interconnecting the oscillators, modulators, detectors, and so on, using a form of extremely short-range radio. A freely propagating beam has been reflected around corners, occasionally refocused with lenses to avoid energy loss resulting from beam spreading, and often sheltered by tubular enclosures from refractive distortions resulting from thermal gradients in the ambient air. Typical separations between components range from a few centimeters to a foot; aggregations of apparatus in a single-channel experimental laser repeater are measured

in square feet. The resulting apparatus is sensitive to ambient temperature gradients, to absolute temperature changes, to airborne acoustical effects, and to mechanical vibrations of the separately mounted parts. All of these effects are understood and are susceptible to appropriate engineering design; but one naturally looks for alternatives.

Looking ahead, one sees the possibility of guiding laser beams on miniature transmission lines, analogous to the hollow rectangular waveguide or coaxial cable used extensively in lower frequency repeaters. Accompanying papers report contributions leading toward the new form of laser circuitry.[1-3] This paper gives a general view of the proposal and indicates specific component possibilities.

II. LASER BEAM GUIDANCE

We visualize a dielectric waveguide wherein a region having an index of refraction n_2 is surrounded by a region of index n_1, as in Fig. 1a. Then a two-dimensional analysis shows that the energy in the lowest-order guided wave is confined almost entirely to the n_2 region if

$$n_1 = n_2(1 - \Delta),\tag{1}$$

where

$$\Delta \cong \frac{3}{4} \left(\frac{\lambda}{a}\right)^2\tag{2}$$

λ = free space wavelength
a = half-width of n_2 region, $(\lambda/an_2) \ll 1$.

Table I, calculated from equations (1) and (2) for $\lambda = 0.6328$ μm, shows that only a very small change in index Δn_2 is needed to provide the desired guidance. Some higher order modes are above cutoff using these parameters; more exact theory can be used to calculate the smaller guide width which restricts the guidance to a single mode at

(a)   (b)
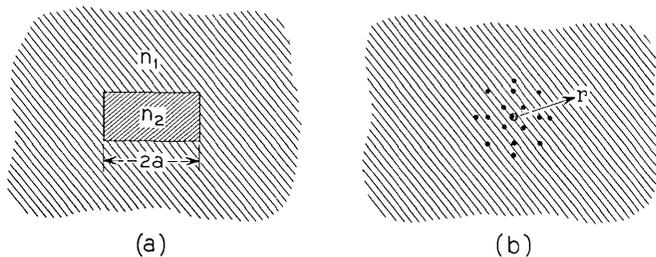
Fig. 1 — Waveguide cross sections: (a) rectangular shape, index $n_2 > n_1$. (b) round shape.

TABLE I—VALUES OF Δ FOR VARIOUS
OPTICAL BEAM WIDTHS

Optical Beam Width 2a	Δ
1 mm	10^{-6}
0.1 mm	10^{-4}
0.01 mm	10^{-2}

the expense of having a larger field component at the n_2 to n_1 interface where dimensional irregularities may occur.[1-4] Values of Δ larger than tabulated for a particular guide width $2a$ do not appreciably change the field distribution for the lowest order mode in the n_2 region but would allow more propagating modes.

It is not important that there be a sharp step in index as in the n_2 to n_1 transition of Fig. 1a. Alternatively, the index can taper smoothly from a maximum at the waveguide's center to a lower value at radius r according to[*]

$$n = n_2[1 - c(r/a)^p]$$ (3)

with

$$c = 0.16\left(\frac{\lambda}{a}\right)^2$$

$2a$ = laser beam width, provided $a \gg \lambda$. (4)

The exponent p can have any even positive value; the lowest order mode field always has an approximately cosinusoidal shape in the region $0 < r < a$ with about 1/10 peak value at $r \cong a$ and with approximately exponentially decaying magnitude for $r > a$.

The square law index variation, given by $p = 2$ in equation (3), has the well-known property that phase constant differences for the various propagating modes are independent of frequency.[6,7] The square law medium is free of delay distortion resulting from mode conversion and is unique in that property.[5,8]

We can anticipate guiding beams around relatively sharp bends as summarized in Table II. The Δ's associated with these beam widths may be obtained from equation (2) or Table I. By using a guide which confines the beam to a 5 to 10μm width (implies a Δ of 0.04 to 0.01) the bend radius can be in the 1.8 to 14.5 mm (70 to 570 mils) region, which could facilitate very small circuitry.

---

[*] A somewhat more accurate expression is given as equation (59) in Ref. 5. This permits a series of terms in $(r/a)^p$ to represent the index variation.

TABLE II—ESTIMATED BENDING RADIUS

Laser Beam Width $2a$ in mm	Estimated Acceptable Bending Radius in m* ($\lambda = 0.633$ μm)
1	14,500
0.1	14.5
0.01	0.0145
0.005	0.0018

* This estimate is obtained using equation (33) of Ref. 9, and includes an allowance of 0.25 dB maximum loss resulting from a bend of any angle.

III. FABRICATION OF SMALL WAVEGUIDES

Tiny laser guides can be fabricated in the form of glass fibers. Previous work on fiber-optics for image transmission or incoherent light sensing has provided a considerable body of experience on which to build, not all of which is applicable. So-called "clad" fibers have two discrete regions of index as in Fig. 1a. The n_1 region (which carries little light) must be as thin as possible in image-transmitting fibers to minimize the "dead" region in the output image. For modulated laser beam transmission the cladding must be much thicker and the "core" (n_2 of Fig. 1a) much smaller to yield well-isolated single mode transmission.

Whereas glass fibers may be used to connect repeater components and certainly are convenient as flexible connections, we can use another form of dielectric waveguide for miniature laser circuitry. Fig. 2
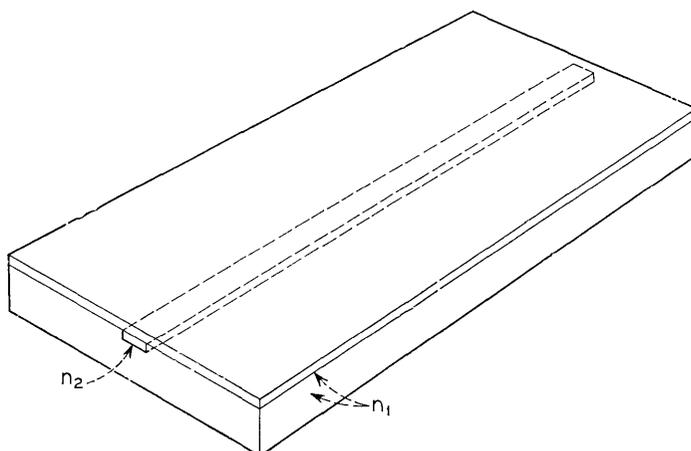

Fig. 2 — Planar waveguide formed using photolithographic techniques.

shows a channel of index n_2 surrounded by a region of index n_1, which would serve as a dielectric waveguide of the type discussed in connection with Fig. 1. This might be created in glass using a series of steps as follows. A mask could be used to expose selectively a light-sensitive photo-resist previously placed on a sheet of glass, followed by washing and selective deposition (if needed) of a more durable material for masking purposes. Then a diffusion, bombardment, or ionic replacement process could be used to change the index of refraction of the glass, thereby creating the n_2 channel imbedded in the n_1 substrate. Finally the top layer of n_1 material could be sputtered on the entire top surface.

Using photolithographic techniques which are currently evolving for low frequency integrated circuit applications, channel widths in the 2 to 5 μm range may be achievable and dimensions on the order of 10 μm are readily held. Complicated masking patterns may in time be made, leading to the possibility of simultaneously making complicated laser circuits using combinations of elements such as those described in the following paragraphs.*

This description is intended to be a broad indication of possible feasibility rather than a blueprint. However, relevant contributions are appearing. G. M. C. Fisher and A. D. Pearson have reported processes which reduce or increase the index of refraction of glass by as much as 0.7 per cent.[10] F. K. Reinhart, D. F. Nelson, and J. McKenna have reported the existence of an index increase in gallium phosphide junctions which is effective as a light guide at zero bias.[11-13] Optical waveguides formed by proton irradiation have been reported.[14] Further contributions may be anticipated.[15]

Some relevant work on two-dimensional light guides has been reported.[16-20] In this work one transverse dimension of the guided wave was in the 10 to 100 μm region; but the other transverse dimension was orders of magnitude larger. We seek waveguides tightly guided in both transverse dimensions in order to make possible the components proposed in Section IV.

IV. INTEGRATED-CIRCUIT LASER

The transmission line of Fig. 2 becomes a resonator when mirrors are placed at the ends, or when a series of partially reflecting trans-

---

*Complicated masking patterns are feasible now where the area involved is small; depth-of-focus problems may require advances in masking to produce the large area patterns we need.

verse lines are spaced at an odd quarter-wave multiple apart to rein-
force reflections at the resonator's peak frequency (Fig. 3). The partial
reflectors are analogous to layered dielectric mirrors and are large
enough in the transverse plane to intercept most of the guided-wave
energy; they may be increased index regions placed in the sheet as
noted in Section III, empty grooves, or minute grooves coated with
metal.

By adding a small concentration of neodymium ions and by pro-
viding a pump, the resonant cavity becomes a laser. Fig 4 shows, in
cross section, two possible ways the pump might be applied. In Fig 4a
the active material (such as neodymium) can be applied only in the
vicinity of the n_2 waveguide channel (by sputtering on the surface,
beneath the SnO_2 film, for example) or might be distributed through-
out the substrate. The spherical reflector confines the pump energy
near the waveguide where the laser field is a maximum. The electro-
luminescent material (for example, doped zinc sulphide) is selected to
provide radiation at a pumping line for the active lasing materials.

In Fig. 4b, ac (kilohertz rate) excitation of the electroluminescent
pumping material is implied; the electroluminescent material is dis-
tributed throughout the glass substrate. Relatively low power laser
sources might be produced in similar structures, the order of 0.1 watt
being adequate for many communication applications.
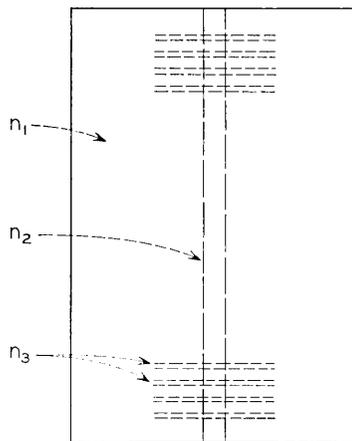
Fig. 3 — Resonator using planar waveguide.
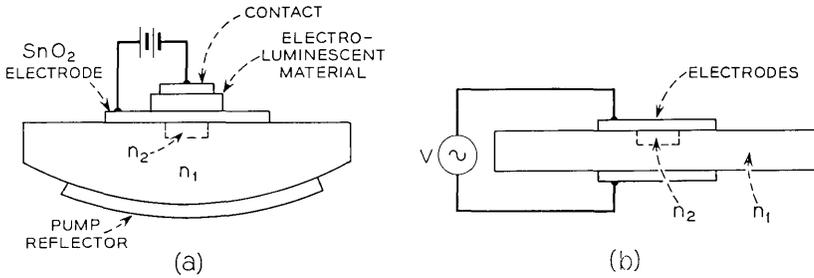
Fig. 4 — Cross sections of possible lasers in planar waveguide: (a) external pump (b) pump ions imbedded in laser circuit.

V. MODULATOR

Figure 5 shows a possible phase modulator for a guided laser beam. The electrooptic material might be the substrate or might be applied as a thin surface layer adjacent to the guiding index region n_2. Using photolithographic techniques, it should be possible to use spacing between the metallic electrodes of about 25 μm which would yield large modulating fields with only a few volts of modulator drive.
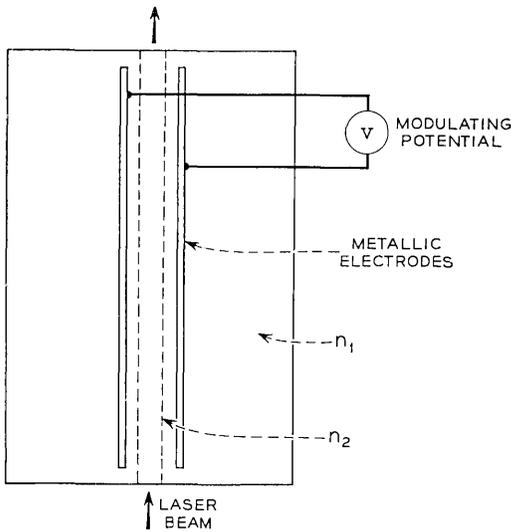
Fig. 5 — Phase modulator.

VI. HYBRID

Figure 6 shows the directional coupler form of hybrid. The exponentially decaying fields, propagating in the n_1 region of Fig. 2, overlap for the two parallel guides of Fig. 6, providing continuous distributed coupling. Reference 1 gives approximate expressions for calculating the guide spacing and needed coupling length.

Figure 7 shows the partially reflecting mirror form of hybrid; the reflecting line may be a narrow groove coated with a metal film, an empty groove, or a high index dielectric region created by a masking and diffusion or ionic replacement process. A single empty groove, an odd quarter of a wavelength thick, in the direction of propagation would give a coupling loss of about 9 dB.

VII. FREQUENCY-SELECTIVE FILTERS

Using techniques familiar at lower frequencies, hybrids and resonant circuits can be combined to form filters, a needed component in frequency-division multiplex systems. Figure 8 shows such an arrangement, where band pass cavities C_1 and C_2 are used to separate f_a from f_b and f_c; hybrids divide and recombine the energy to form a constant
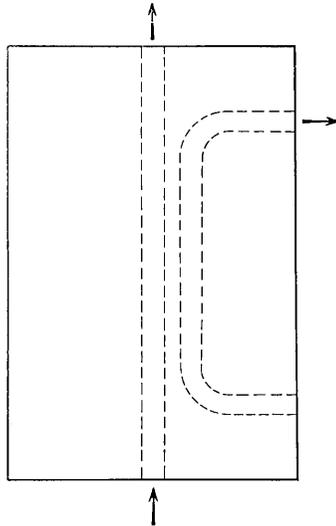
Fig. 6 — Directional coupler type hybrid.
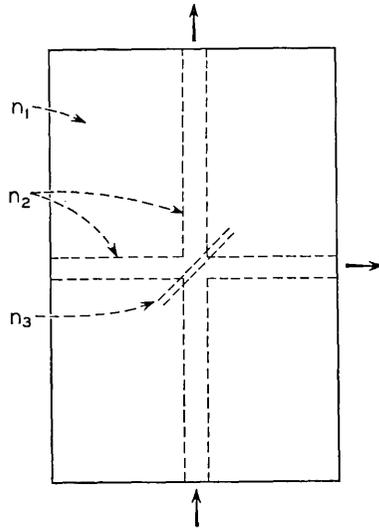
Fig. 7 — Junction type hybrid.
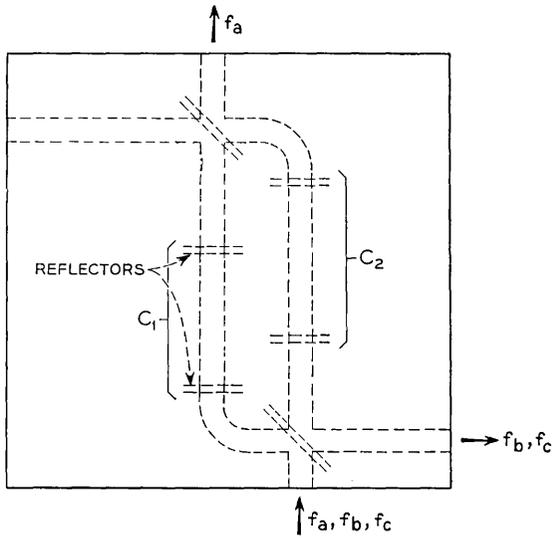
Fig. 8 — Channel dropping filter.

resistance filter. Alternatively, a multiple-line grating could be used in place of the resonant cavities as the reflecting element to reflect f_a only, and the output positions of f_a and f_b, f_c would interchange.

In filters of this kind the intrinsic loss of the substrate is of course important. Good quality glasses have bulk losses as low as 1 dB per m, which corresponds to an intrinsic Q of about 30 million; this would allow filters with band widths of a few hundred megacycles in the visible region; therefore, intrinsic substrate loss should not be too limiting.

VIII. CONCLUSIONS

This paper outlines a prospect for laser circuitry and devices which, if realized, would have many attractive features. Photolithographic processes would simplify reproducing complicated circuits, once the original was developed. Small size would facilitate isolating the completed circuit assembly from thermal, mechanical, and acoustic ambient changes. For communication purposes, low laser power levels are adequate so that the heat to be dissipated hopefully will not be large. In the very small laser beam cross sections, nonlinear effects needed for modulation and frequency changing should be achievable with only a few volts of drive.

Finally, a word of caution is needed. Work is just beginning in the directions indicated, and we have identified goals rather than accomplishments. We recognize these are difficult goals; but we believe they are worth the serious effort required to achieve them.

IX. ACKNOWLEDGEMENT

REFERENCES

1. Marcatili, E. A. J., "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," B.S.T.J., this issue, pp. 2071–2102.
2. Goell, J. E., "A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides," B.S.T.J., this issue, pp. 2133–2160.
3. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., this issue, pp. 2103–2132.
4. Schlosser, W., and Unger, H. G., "Partially Filled Waveguides and Surface Waveguides of Rectangular Cross Section," Advances in Microwaves, New York: Academic Press, 1966, pp. 319–387.
5. Miller, S. E., "Light Propagation in Generalized Lens-like Media," B.S.T.J., 44, No. 9 (November 1965), pp. 2017–2064.
6. Pierce, J. R., "Modes in Sequences of Lenses," Proc. Nat. Acad. Sci., 47, No. 11 (November 1961), pp. 1808–13.

7. Marcatili, E. A. J., "Modes in a Sequence of Thick Astigmatic Lens-like Focusers," B.S.T.J., *43*, (November 1964), pp. 2887–2904.
8. Gordon, J. P., "Optics of General Guiding Media," B.S.T.J., *45*, No. 2 (February 1966), p. 321.
9. Marcatili, E. A. J., and Miller, S. E., "Improved Relations Describing Directional Control in Electromagnetic Wave Guidance," B.S.T.J., this issue, pp. 2161–2188.
10. Fisher, G. M. C., and Fritz, T. C., unpublished work; see also: Pearson, A. D., French, W. G., and Rawson. E. G., "Preparation of a Light-Focusing Rod by Ion Exchange Techniques," Appl. Phys. Letters, *15*, No. 2 (July 15, 1969), pp. 76–77.
11. Reinhart, F. K., Nelson, D. F., and McKenna, J., "Electrooptic and Waveguide Properties of Reverse-Biased Gallium Phosphide p-n Junctions," Phys. Rev. *177*, No. 3 (January 15, 1969), pp. 1208–1221.
12. Ashkin, A., and Gershenzon, M., "Reflection and Guiding of Light at p-n Junctions," J. Appl. Phys., *34*, No. 7 (July 1963), p. 2116.
13. Nelson, D. F., and Reinhart. F. K., "Light Modulation by the Electro-Optic Effect in Reverse-Biased GaP p-n Junction," Appl. Phys. Letters, *5*, No. 7 (October 1, 1964), p. 148.
14. Schineller, E. Ronald, Flam, R. P., and Wilmot, D. W., "Optical Waveguides Formed by Proton Irradiation of Fused Silica," J. Opt. Soc. of Amer., *58*, No. 9 (September 1968), pp. 1171–1176.
15. Tien, P. K., Ulrich, R., and Martin, R. J., "Modes of Propagating Light Waves in Thin Deposited Semiconductor Films," Appl. Phys. Letters, *14*, No. 9 (May 1, 1969), pp. 291–294.
16. Kaplan, R. A., "Optical Waveguide of Macroscopic Dimension in Single-mode Operation," Proc. IEEE, *51*, No. 8 (August 1963), p. 1144.
17. Schineller, E. Ronald, "Summary of the Development of Optical Waveguides and Components," Wheeler Laboratories Report #1471, (April 1967).
18. Anderson, D. B., and Shaw, C. B. Jr., "Dielectric Waveguide for Infrared Wavelengths," Digest of 1968 G-MTT Int. Microwave Symp. Detroit. Michigan, May 20-22, 1968.
19. Shubert, R., and Harris, Jay H., "Plane Waveguide Mode Effects in the Visible Spectrum," Digest of 1968 G-MTT Int. Microwave Symp., Detroit, Michigan, May 20–22, 1968.
20. Shaw, C. B. Jr., French, B. T., and Warner, Charles III, "Further Research on Optical Transmission Lines," Report #2 AD 652501, Contract AF49(638)-1504, May 1967.

Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics

By E. A. J. MARCATILI

We study the transmission properties of a guide consisting of a dielectric rod with rectangular cross section, surrounded by several dielectrics of smaller refractive indices. This guide is suitable for integrated optical circuitry because of its size, single-mode operation, mechanical stability, simplicity, and precise construction.

After making some simplifying assumptions, we solve Maxwell's equations in closed form and find, that, because of total internal reflection, the guide supports two types of hybrid modes which are essentially of the TEM *kind polarized at right angles. Their attenuations are comparable to that of a plane wave traveling in the material of which the rod is made.*

If the refractive indexes are chosen properly, the guide can support only the fundamental modes of each family with any aspect ratio of the guide cross section. By adding thin lossy layers, the guide presents higher loss to one of those modes. As an alternative, the guide can be made to support only one of the modes if part of the surrounding dielectrics is made a low impedance medium.

Finally, we determine the coupling between parallel guiding rods of slightly different sizes and dielectrics; at wavelengths around one micron, 3-dB directional couplers, a few hundred microns long, can be achieved with separations of the guides about the same as their widths (a few microns).

I. INTRODUCTION

Proposals have been made for dielectric waveguides capable of guiding beams in integrated optical circuits very much as waveguides and coaxials are used for microwave circuitry.[1-3] Figure 1 shows the basic geometries for these waveguides. The guide is a dielectric rod of refractive index n immersed in another dielectric of slightly smaller refractive index $n(1 - \triangle)$; both are in contact with a third dielectric which may be air (Fig. 1a) or a dielectric of refractive index $n(1 - \triangle)$,
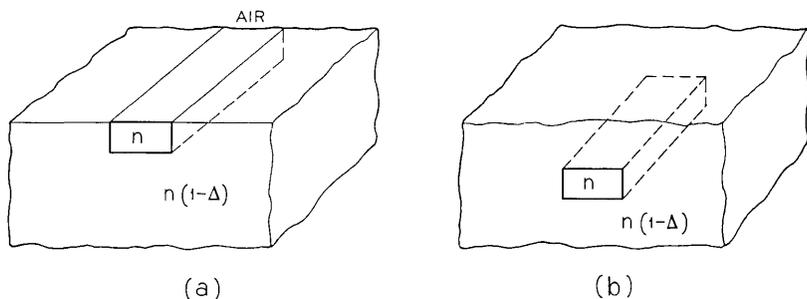
Fig. 1 — Dielectric waveguides for integrated optical circuitry.

(Fig. 1b). These geometries are attractive not only because of simplicity, precision of construction, and mechanical stability, but also because by choosing $\triangle$ small enough, single-mode operation can be achieved with transverse dimensions of the guide large compared with the free space wavelengths, thus relaxing the tolerance requirements.

Even though in a real guide the cross section of the guiding rod is not exactly rectangular and the boundaries between dielectrics are not sharply defined, as in Fig. 1, it is worth finding the characteristics of the modes in the idealized structure and the requirements to make it a single-mode waveguide.

Furthermore, directional couplers made by bringing two of those guides close together, Fig. 2, may become important circuit components.[1,2] In this paper we study the transmission through such a coupler; the modes in a single guide result as a particular case, when the separation between the two guides is so large that the coupling is negligible. Through use of a perturbation technique, we also find the coupler properties when the two guides are slightly different.
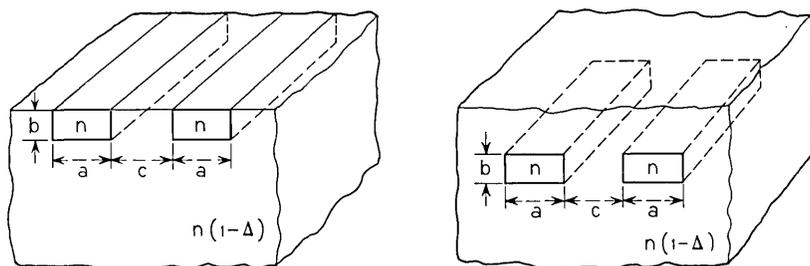
Fig. 2 — Directional couplers.

The guiding properties of the rectangular cross section guide immersed in a single dielectric are compared with those derived through computer calculations by Goell.[4] Similarly, the coupling properties of two guides of square cross section immersed in a single dielectric are compared with those of two guides of circular cross section derived by Jones and by Bracey and others.[5,6] In both comparisons agreement is quite good.

II. FORMULATION OF THE BOUNDARY VALUE PROBLEM

For analysis, we redraw in Fig. 3 the cross section of the coupler subdivided in many areas. Nine of the areas have refractive indexes n_1 to n_5; we do not specify the refractive indexes in the six shaded areas. The reasons for these choices will become obvious.

A rigorous solution to this boundary value problem requires a computer;[4,7] nevertheless, it is possible to introduce a drastic simplification which enables one to get a closed form solution. This simplification arises from observing that, for well-guided modes, the field decays exponentially in regions 2, 3, 4, and 5; therefore, most of the power travels in regions 1, a small part travels in regions 2, 3, 4, and 5, and even less travels in the six shaded areas. Consequently, only a small error should be introduced into the calculation of fields in regions 1 if one does not properly match the fields along the edges of the shaded areas.

The matching made only along the four sides of regions 1 can be achieved assuming simple field distribution. Thus the field components in regions 1 vary sinusoidally in the x and y direction; those in 2 and 4 vary sinusoidally along x and exponentially along y; and those in regions 3 and 5 vary sinusoidally along y and exponentially along x. The propagation constants k_{x1}, k_{x2}, and k_{x4} along x in media 1, 2, and
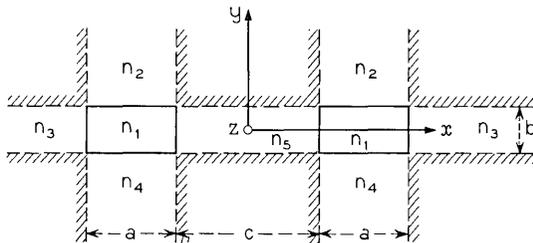
Fig. 3 — Coupler cross section subdivided for analysis.

4 are identical and independent of y. Similarly, the propagation constants k_{y1}, k_{y3}, and k_{y5} along y in the regions 1, 3, and 5 are also identical and independent of x.

In the appendix we calculate these propagation constants and find, as expected, that all the modes are hybrid and that guidance occurs because of total internal reflection. Nevertheless, because of another approximation which consists of choosing the refractive indexes n_2, n_3, n_4, and n_5 slightly smaller than n_1, total internal reflection occurs only when the plane wavelets that make a mode impinge on the interfaces at grazing angles.* Consequently, the largest field components are perpendicular to the axis of propagation; the modes are essentially of the TEM kind and can be grouped in two families, E_{pq}^x and E_{pq}^y. The main field components of the members of the first family are E_x and H_y, while those of the second are E_y and H_x. The subindex p and q indicate the number of extrema of the electric or magnetic field in the x and y directions, respectively. Naturally, E_{11}^x and E_{11}^y are the fundamental modes; we concentrate on them as we discuss the transmission properties of different structures.

III. GUIDE IMMERSED IN SEVERAL DIELECTRICS

The guide immersed in several dielectrics (Fig. 4a) is derived from Fig. 3 by choosing

$$c = \infty. \tag{1}$$

It supports a discrete number of guided modes which we group in two families E_{pq}^x and E_{pq}^y plus a continuum of unguided modes.[8,9]

3.1 The E_{pq}^y Modes

The main transverse field components of the E_{pq}^y modes are E_y and H_x. They are depicted in solid and broken lines, respectively, in Fig. 4a for the fundamental mode E_{11}^y. Within the guiding rod each component varies sinusoidally both along x and along y. Outside the guide each component decays exponentially. Such functional dependence is given in equation (38) and depicted in Fig. 4b. We assume $n_2 \neq n_3 \neq n_4 \neq n_5$; consequently the field distributions are not symmetric with respect to the planes $x = 0$ and $y = 0$. In Fig. 5a we assume $n_2 = n_4$ and $n_3 = n_5$; the E_{pq}^y modes depicted are either symmetric or antisymmetric with respect to the same planes. These modes look similar to those in laser

---

* This approximation is not very demanding. Even when n_1 is 50 percent larger than n_2, n_3, n_4, and n_5, the results are valid.
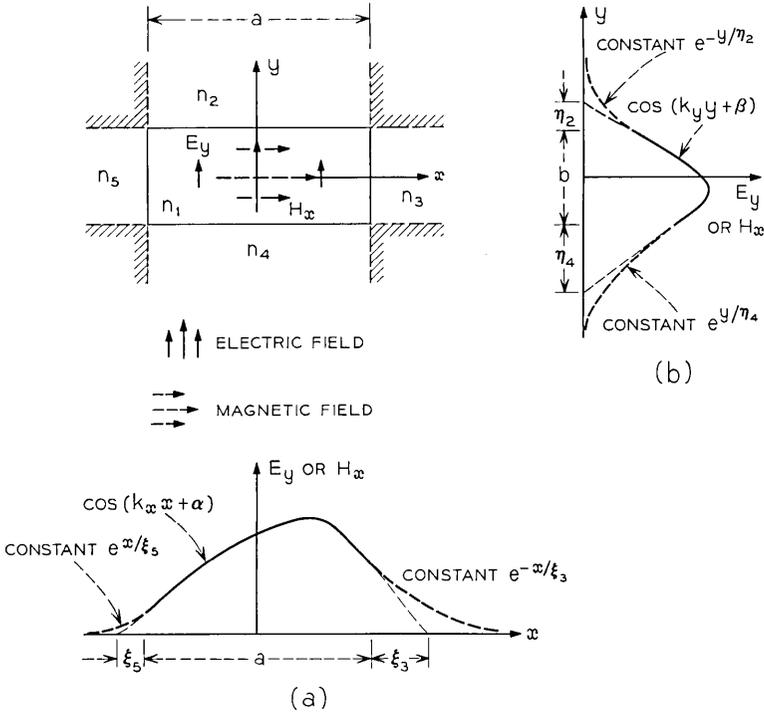
Fig. 4 — Guide immersed in different dielectrics: (a) cross section and (b) field distribution of the fundamental mode $E_{11}{}^y$.

cavities with rectangular flat mirrors, but our nomenclature is different.[10] The subindexes p and q indicate the number of extrema each component has within the guide.

Now we describe these modes quantitatively by reproducing the propagation constants found for each medium in Section A.1 of the appendix. Let us call k_z the axial propagation constant and $k_{x\nu}$ and $k_{y\nu}$ the transverse propagation constants along the x and the y directions, respectively, in the νth medium ($\nu = 1, 2, \cdots 5$). Furthermore, let us call

$$k_\nu = k n_\nu = \frac{2\pi}{\lambda} n_\nu \tag{2}$$

the propagation constant of a plane wave in a medium of refractive index n_ν and free-space wavelength λ.

According to equations (39) through (52)

$$k_z = (k_1^2 - k_x^2 - k_y^2)^{\frac{1}{2}} \tag{3}$$
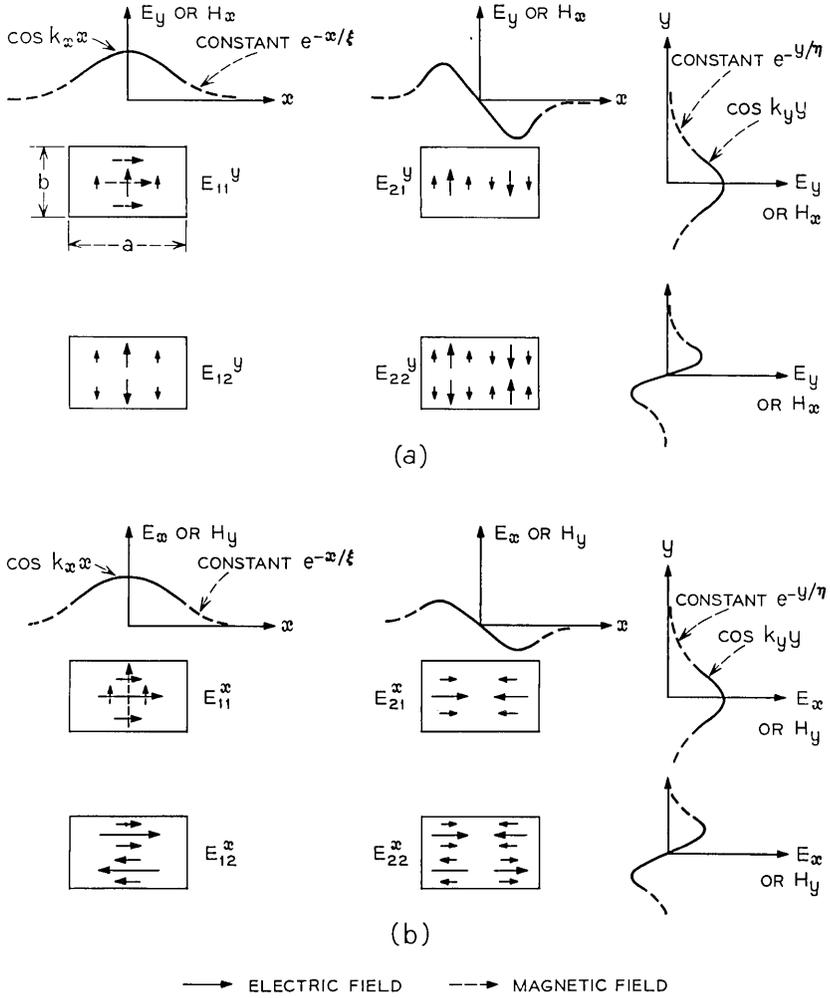
Fig. 5 — (a) Field configuration of $E_{pq}{}^y$ modes. (b) Field configuration of $E_{pq}{}^x$ modes.

in which

$$k_x = k_{x1} = k_{x2} = k_{x4} \tag{4}$$

and

$$k_y = k_{y1} = k_{y3} = k_{y5} . \tag{5}$$

This means that the fields in media 1, 2, and 4 have the same x

dependence and similarly those in media 1, 3, and 5 have identical y dependence. These transverse propagation constants are solutions of the transcendental equations:

$$k_x a = p\pi - \tan^{-1} k_x \xi_3 - \tan^{-1} k_x \xi_5 \tag{6}$$

$$k_y b = q\pi - \tan^{-1} \frac{n_2^2}{n_1^2} k_y \eta_2 - \tan^{-1} \frac{n_4^2}{n_1^2} k_y \eta_4 \tag{7}$$

in which

$$\xi_{3 \atop 5} = \left| \frac{1}{k_{x3 \atop 5}} \right| = \frac{1}{\left[\left(\frac{\pi}{A_{3 \atop 5}} \right)^2 - k_x^2 \right]^{\frac{1}{2}}} \tag{8}$$

$$\eta_{2 \atop 4} = \left| \frac{1}{k_{y2 \atop 4}} \right| \quad \frac{1}{\left[\left(\frac{\pi}{A_{2 \atop 4}} \right)^2 - k_y^2 \right]^{\frac{1}{2}}} \tag{9}$$

and

$$A_{2,3,4,5} = \frac{\pi}{(k_1^2 - k_{2,3,4,5}^2)^{\frac{1}{2}}} = \frac{\lambda}{2(n_1^2 - n_{2,3,4,5}^2)^{\frac{1}{2}}}. \tag{10}$$

In the transcendental equations (6) and (7), a and b are the transverse dimensions of the guiding rod, and the $\tan^{-1}$ functions are to be taken in the first quadrant.

What are the physical meanings of $\xi_{3 \atop 5}$, $\eta_{2 \atop 4}$, and $A_{2,3,4,5}$? The amplitude of each field component in medium 3 (Fig. 4) decreases exponentially along x. It decays by $1/e$ in a distance $\xi_3 = 1/ \mid k_{x3} \mid$. Similarly ξ_5, η_2, and η_4 measure the "penetration depths" of the field components in media 5, 2, and 4, respectively.

The meaning of A_2 is the following. Consider a symmetric slab derived from Fig. 4 by choosing $a = \infty$ and $n_2 = n_4$. The maximum thickness for which the slab supports only the fundamental mode is A_2.

Expressions (3), (8), and (9) contain k_x and k_y, which are solutions of the transcendental equations (6) and (7). These cannot be solved exactly in closed form. Nevertheless, for well-guided modes, most of the power travels within medium 1, implying

$$\left(\frac{k_x A_3}{\pi} \right)^2 \ll 1 \quad \text{and} \quad \left(\frac{k_y A_2}{\pi} \right)^2 \ll 1. \tag{11}$$

It is possible then to solve those transcendental equations in closed,

though approximate, form. Their solutions are

$$k_x = \frac{p\pi}{a}\left(1 + \frac{A_3 + A_5}{\pi a}\right)^{-1} \tag{12}$$

$$k_y = \frac{q\pi}{b}\left(1 + \frac{n_2^2 A_2 + n_4^2 A_4}{\pi n_1^2 b}\right)^{-1}. \tag{13}$$

For large a and b, the electrical width, $k_x a$, and the electrical height, $k_y b$, of the guide are close to $p\pi$ and $q\pi$, respectively.

Substituting equations (12) and (13) in equations (3), (8), and (9), we obtain explicit expressions for k_z, ξ_3, ξ_5, η_2, and η_4:

$$k_z = \left[k_1^2 - \left(\frac{\pi p}{a}\right)^2\left(1 + \frac{A_3 + A_5}{\pi a}\right)^{-2} - \left(\frac{\pi q}{b}\right)^2\left(1 + \frac{n_2^2 A_2 + n_4^2 A_4}{\pi n_1^2 b}\right)^{-2}\right]^{\frac{1}{2}} \tag{14}$$

$$\xi_3 = \frac{A_3}{\pi}\left[1 - \left[\frac{p A_3}{a} \frac{1}{1 + \frac{A_3 + A_5}{\pi a}}\right]^2\right]^{-\frac{1}{2}} \tag{15}$$

$$\eta_2 = \frac{A_2}{\pi}\left[1 - \left[\frac{q A_2}{b} \frac{1}{1 + \frac{n_2^2 A_2 + n_4^2 A_4}{\pi n_1^2 b}}\right]^2\right]^{-\frac{1}{2}}. \tag{16}$$

3.2 The E_{pq}^x Modes

Except for the fact that the main transverse components are E_x and H_y, the E_{pq}^x modes are qualitatively similar to the E_{pq}^y modes (Fig. 5b); they differ quantitatively. Distinguishing with bold-face type the symbols corresponding to E_{pq}^x modes, the axial propagation constant and the "penetration depth" in media 2, 3, 4, and 5 are, according to equations (60), (63), and (64),

$$\mathbf{k}_z = (k_1^2 - \mathbf{k}_x^2 - \mathbf{k}_y^2)^{\frac{1}{2}} \tag{17}$$

$$\boldsymbol{\xi}_3 = \frac{1}{|\mathbf{k}_{x3}|} = \frac{1}{\left[\left[\frac{\pi}{A_3}\right]^2 - \mathbf{k}_x^2\right]^{\frac{1}{2}}} \tag{18}$$

$$\mathbf{n}_2 = \frac{1}{|\mathbf{k}_{y2}|} = \frac{1}{\left[\left[\frac{\pi}{A_2}\right]^2 - \mathbf{k}_y^2\right]^{\frac{1}{2}}} \tag{19}$$

in which $\mathbf{k}_x$ and $\mathbf{k}_y$ are solutions of the transcendental equations

$$\mathbf{k}_x a = p\pi - \tan^{-1} \frac{n_3^2}{n_1^2} \mathbf{k}_x \xi_3 - \tan^{-1} \frac{n_5^2}{n_1^2} \mathbf{k}_x \xi_5 \tag{20}$$

$$\mathbf{k}_y b = q\pi - \tan^{-1} \mathbf{k}_y \mathbf{n}_2 - \tan^{-1} \mathbf{k}_y \mathbf{n}_4 . \tag{21}$$

The approximate closed form solutions of these equations are

$$\mathbf{k}_x = \frac{p\pi}{a} \left(1 + \frac{n_3^2 A_3 + n_5^2 A_5}{\pi n_1^2 a}\right)^{-1} \tag{22}$$

and

$$\mathbf{k}_y = \frac{q\pi}{b} \left(1 + \frac{A_2 + A_4}{\pi b}\right)^{-1}. \tag{23}$$

Substituting these expressions in equations (17), (18), and (19), we derive the explicit results:

$$\mathbf{k}_z = \left[k_1^2 - \left(\frac{\pi p}{a}\right)^2\left(1 + \frac{n_3^2 A_3 + n_5^2 A_5}{\pi n_1^2 a}\right)^{-2} - \left(\frac{\pi q}{b}\right)^2\left(1 + \frac{A_2 + A_4}{\pi b}\right)^{-2}\right]^{\frac{1}{2}} \tag{24}$$

$$\xi_3 = \frac{A_3}{\pi} \left[1 - \left[\frac{p A_3}{a} \frac{1}{1 + \frac{n_3^2 A_3 + n_5^2 A_5}{\pi n_1^2 a}}\right]^2\right]^{-\frac{1}{2}} \tag{25}$$

$$\mathbf{n}_2 = \frac{A_2}{\pi} \left[1 - \left[\frac{q A_2}{b} \frac{1}{1 + \frac{A_2 + A_4}{\pi b}}\right]^2\right]^{-\frac{1}{2}}. \tag{26}$$

If

$$\frac{1}{n_1} \left|n_1 - n_2 \atop {3 \atop 4 \atop 5}\right| \ll 1,$$

these results coincide with those in equations (14), (15), and (16), indicating that the E_{pq}^x and E_{pq}^y modes become degenerate.

3.3 *Examples*

The axial propagation constants k_z and $\mathbf{k}_z$, given in equations (3) and (17) and properly normalized, have been plotted in Figs. 6a through k as a function of the normalized height of the guide

$$\frac{b}{A_4} = \frac{2b}{\lambda} (n_1^2 - n_4^2)^{\frac{1}{2}}$$
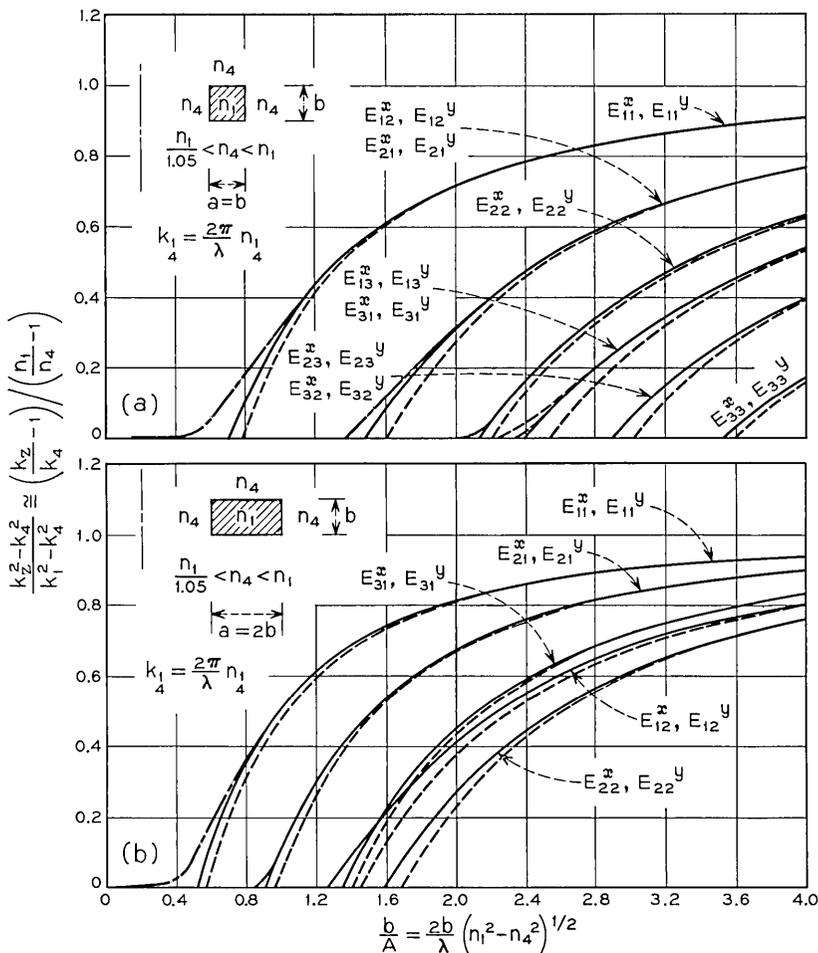
Fig. 6 — Propagation constant for different modes and guides. ——————— transcendental equation solutions; ———— closed form solutions; —·—·— Goell's computer solutions of the boundary value problem.

for several geometries and surrounding media.* The ordinate in each of these figures is

$$\frac{k_z^2 - k_4^2}{k_1^2 - k_4^2} \; ;$$

it varies between 0 and 1. It is 0 when $k_z = k_4$, that is, when the guide

---

* In these figures we use the same symbol k_z for both the $E_{pq}{}^y$ and the $E_{pq}{}^x$ modes.

is so small that the mode under consideration becomes unguided or, in other words, the "penetration depth" in medium 4 is ∞. It is 1 when the guide is so large that $k_z = k_1$, which means that all the field travels within the guiding rod and the "penetration depths" in media 2, 3, 4, and 5 are zero.

The solid curves have been obtained using the exact numerical solutions of the transcendental equations (6), (7), (20), and (21); for the transverse propagation constants k_x and k_y; the dashed lines have been derived using the closed form approximations (12), (13), (22), and (23). In Figs. 6a, 6b, 6e, and 6f, for comparison, we have also included the dotted-dashed lines which are the results obtained by Goell as computer solutions of the boundary value problem.[4]

The three solutions coincide even for moderately large values of b.

(e)

(f)

(g)

$$\frac{k_z^2 - k_4^2}{k_1^2 - k_4^2}$$

$$\frac{b}{A} = \frac{2b}{\lambda}\left(n_1^2 - n_4^2\right)^{1/2}$$
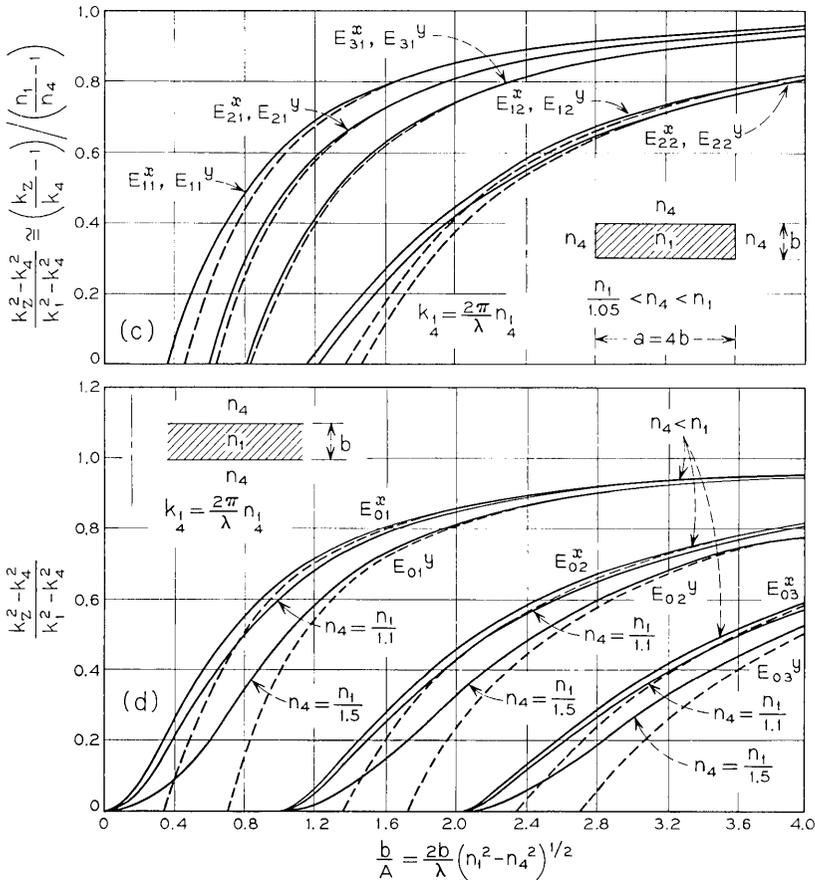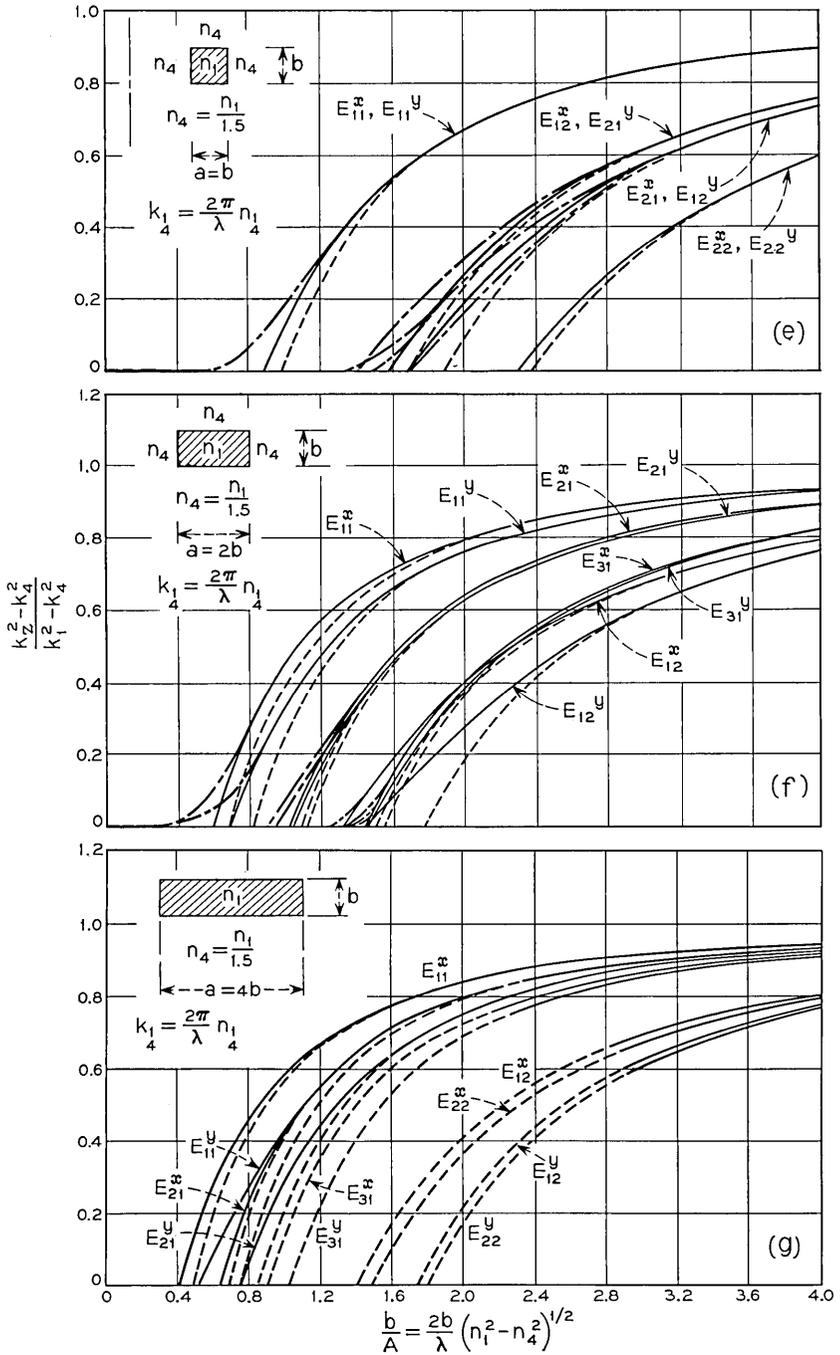
Thus, for a guide and mode for which

$$\frac{k_z^2 - k_4^2}{k_1^2 - k_4^2} \gtreqqless 0.5,$$

the closed form approximation is within a few percent of the exact value. This gives us confidence to use our results in guides with an aspect ratio $a/b > 2$, in guides surrounded by several dielectrics and in directional couplers for which there are no computer calculations available.

The largest discrepancy between our results and Goell's occurs for

$$\frac{k_z^2 - k_4^2}{k_1^2 - k_4^2} \cong 0$$

$$\frac{b}{A} = \frac{2b}{\lambda} \left(n_1^2 - n_4^2 \right)^{1/2}$$
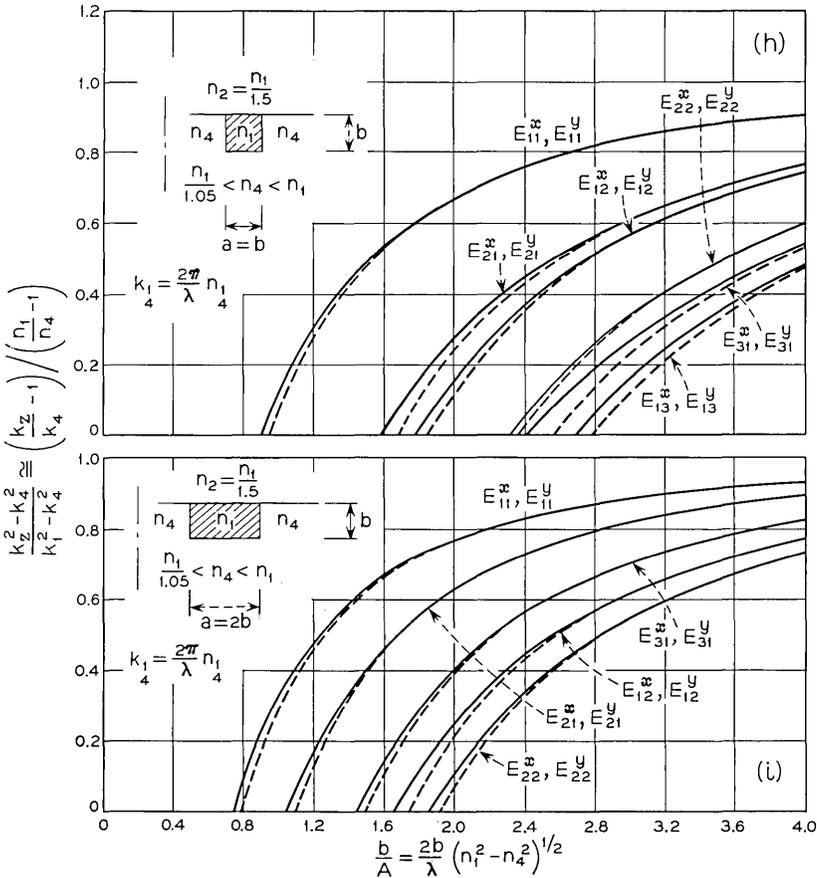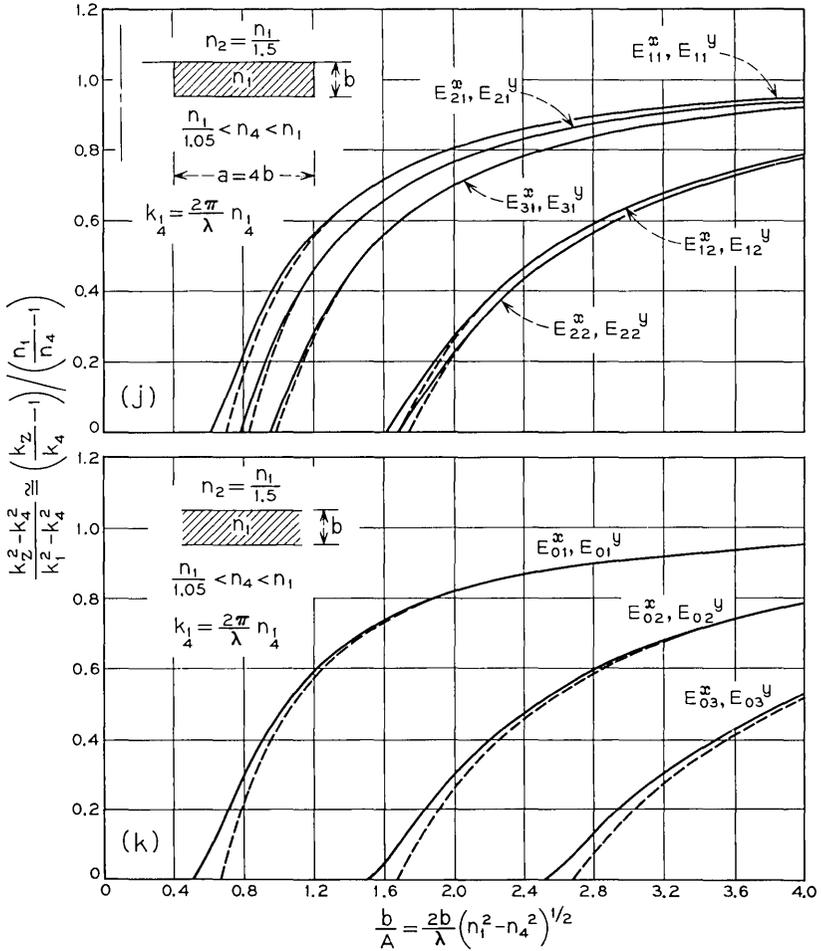
and especially for the fundamental modes E_{11}^x and E_{11}^y. Our approximate theory is incapable of predicting the fact that these modes remain guided no matter how small the guide's cross section.

Figures 6a through d cover the cases of rectangular guides totally embedded in a single dielectric of slightly lower refractive index. For all practical purposes, given p and q, the E_{pq}^x and E_{pq}^y modes are degenerate, and the square cross section provides the widest separation between modes.

Figures 6e through g also consider rectangular guides embedded in a single dielectric, but the external refractive index is 1.5 times smaller

than the internal one. A glass rod immersed in air is an example. The substantial difference of refractive indexes breaks the degeneracy for any rectangular cross section. Rectangular waveguides as in Fig. 1a, with three sides in contact with slightly lower refractive indexes and the fourth side in contact with air, are covered in Fig. 6h through k.

The approximate dispersion relation (14) for E_{pq}^y modes, in a rectangular guide surrounded by four different dielectrics, has been put in graphical form in Fig. 7 by plotting the equivalent equation

$$p^2 X + q^2 Y = 1 \tag{27}$$

in which

$$X = \left(\frac{\pi}{a}\right)^2 \left(1 + \frac{A_3 + A_5}{\pi a}\right)^{-2} (k_1^2 - k_z^2)^{-1} \tag{28}$$

and

$$Y = \left(\frac{\pi}{b}\right)^2 \left(1 + \frac{n_2^2 A_2 + n_1^2 A_4}{\pi n_1^2 b}\right)^{-2} (k_1^2 - k_z^2)^{-1}. \tag{29}$$
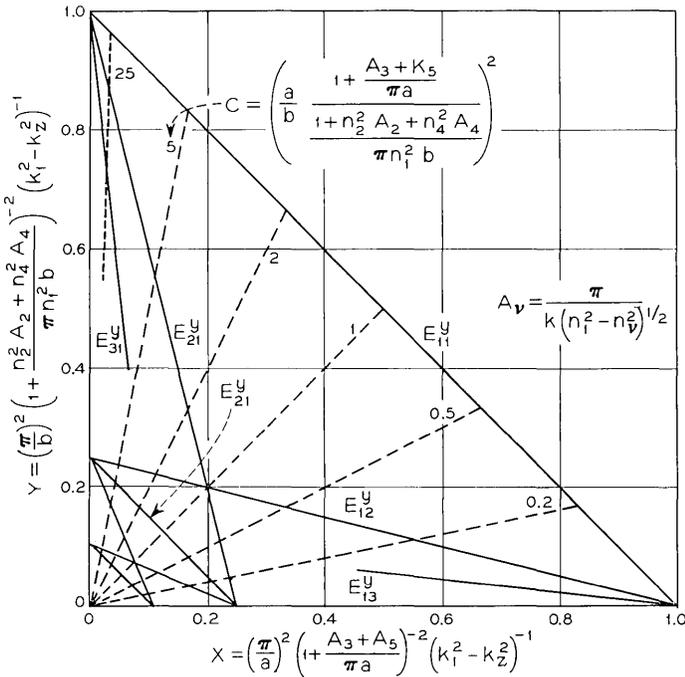
Fig. 7 — Nomograph to dimension a guide immersed in several dielectrics in such a way that it supports any prescribed number of modes.

The curves plotted for different values of p and q are straight lines (solid lines); since the values of X and Y are physically meaningful when they are positive, the plots are kept within the first quadrant.

In Fig. 7 the dotted lines depict the equation

$$\frac{Y}{X} = \left[\frac{a}{b} \frac{1 + \dfrac{A_3 + A_5}{\pi a}}{1 + \dfrac{n_2^2 A_2 + n_4^2 A_4}{\pi n_1^2 b}} \right]^2 = C. \tag{30}$$

Given any guide, we can calculate C which is a function of the dimensions, refractive indexes, and wavelength. The corresponding dotted line intersects all the solid lines representing the different modes. The abscissa or ordinate of each intersection yields, after some algebra, the propagation constant k_z of each particular mode. If the resulting k_z is smaller than the smallest k_ν, that mode is not guided.

Another way of using the graph is this: Suppose one wants a guide with such dimensions that at a given wavelength only the E_{11}^y mode is supported. Picking $k_z = k_{\nu\min}$, any combination of n_1, n_2, n_3, n_4, n_5, a, and b represented by a point within the triangle limited by the solid lines E_{11}^y, E_{12}^y, and E_{21}^y will satisfy the proposed single-mode requirement.

In the graph it is enough to substitute a by b and everything we said about E_{pq}^y modes is applicable to E_{pq}^x modes.

Figures 6a through k have been used to determine dimensions for several guides. All of them have the maximum dimensions compatible with exclusive guidance of the E_{11}^x and E_{11}^y modes. The results are collected in Table I.

In general, the geometry with $n_2 < n_4$ requires a larger waveguide cross section than with $n_2 = n_4$. This means reducing the refractive index on one side of the guide reduces its ability to guide. The explanation of this paradox is found in the known fact that a symmetric slab indeed guides "better" than an asymmetric one. Comparing, for example, Figs. 6d and 6k, in which the solid curves have been drawn solving Maxwell's equations exactly, the E_{p1}^x and E_{p1}^y modes can be guided by the symmetric slab (Fig. 6d) no matter how small the thickness b; there is a minimum thickness required for the asymmetric slab (Fig. 6k) to guide the same modes.[9]

Consider the guide immersed in a single dielectric. In general, the guide's height b is inversely proportional to

$$\frac{1}{(n_1^2 - n_4^2)^{\frac{1}{2}}}.$$

TABLE I—TYPICAL DIMENSIONS FOR SEVERAL GUIDES*
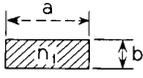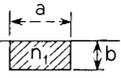
	$\frac{n_1}{n_4}=1.001$	$\frac{n_1}{n_4}=1.01$	$\frac{n_1}{n_4}=1.05$	$\frac{n_1}{n_4}=1.5$	$\frac{n_1}{n_2}=1.5;$ $\frac{n_1}{n_4}=1.001$	$\frac{n_1}{n_2}=1.5;$ $\frac{n_1}{n_4}=1.01$	$\frac{n_1}{n_2}=1.5;$ $\frac{n_1}{n_4}=1.05$
$a = b$	15.3†	4.9	2.25	0.92	17.7	5.6	2.6
$a = 2b$	19	6.1	2.8	1.21	23.2	7.4	3.4
$a = 4b$	26.8	8.5	3.8	1.37	34.9	11	4.9

* Dimensions are for guides capable of supporting only the fundamental modes $E_{11}{}^x$ and $E_{11}{}^y$.
† All numbers in the table must be multiplied by λ/n_1.

For $n_1 = 1.5$, $n_4 = 1$, and $\lambda = 1\mu$, the largest guide height corresponds to the square cross section, and $b = a = 0.61\mu$. This dimension may be too small and difficult to control. The tolerance requirements may be relaxed by choosing $n_1 - n_4 \ll 1$. Nevertheless, this difference cannot be made arbitrarily small because the guide loses its ability to negotiate sharp bends.[11]

In all these examples the fundamental modes E_{11}^x and E_{11}^y are almost degenerate, so symmetry imperfections of the guide tend to couple these modes. A lossy layer, added to one of the interfaces between guiding rod and surrounding dielectrics, should attenuate the mode with polarization parallel to that interface. As an alternative, the guide can be made to support only the fundamental mode E_{11}^y by substituting medium 2 with a low impedance medium such as a dielectric with large refractive index or a metal.

An example of such a guide and the propagation constant of its modes are shown in Fig. 8. By choosing

$$a < \frac{0.7\lambda}{(n_1^2 - n_4^2)^{\frac{1}{2}}}$$

only the E_{11}^y mode is guided. If the metal is not perfect, there is power leakage into the low impedance medium. The smaller that impedance, the smaller the leakage.

Guides for integrated optics may be easier to build with $a/b \gg 1$. We can use Fig. 7 to design a guide of arbitrary dimensions a and b which is
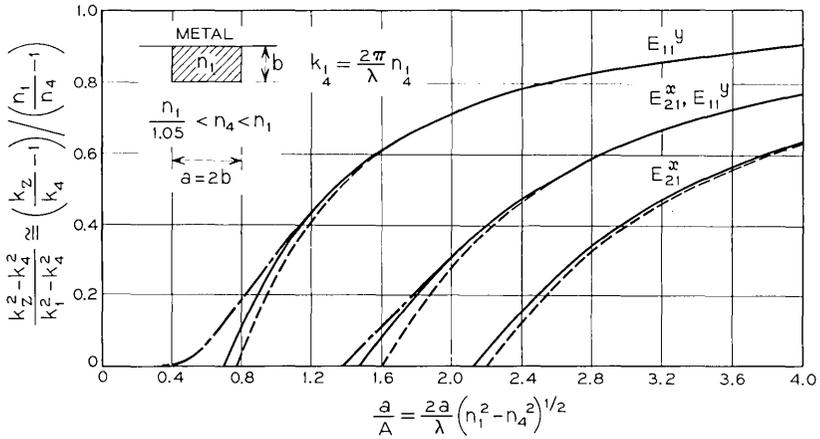
Fig. 8 — Propagation constant for modes in a guide surrounded by metal and dielectrics. ——————— transcendental equation solutions; — — — — closed form solutions; —·—·— Goell's computer solutions of the boundary value problem.

still capable of supporting only the E_{11}^x and E_{11}^y modes. An as example, let us calculate what the values

$$n_3 = n_5 = n_1(1 + \Delta) \quad \text{and} \quad n_2 = n_4 = n_1(1 + \Delta')$$

should be, assuming

$$\Delta, \Delta' \ll 1, \quad \text{and} \quad \frac{a}{b} = 5.$$

Choosing

$$\left(\frac{\Delta'}{\Delta}\right)^{\frac{1}{2}} = \frac{a}{b} = 5, \tag{31}$$

one derives from Fig. 7

$$C = \left(\frac{a}{b}\right)^2 = 25.$$

The curve corresponding to $C = 25$ has been plotted as a dotted line in Fig. 7. It intercepts the E_{21}^y line at

$$Y = \left[\frac{b}{\pi} + \frac{1}{\pi k n_1}\left(\frac{2}{\Delta'}\right)^{\frac{1}{2}}\right]^{-2}(k_1^2 - k_z^2)^{-1} = 0.88.$$

In this expression, by making

$$k_z = kn_1(1 - \Delta),$$

the guide supports only the E_{11}^y and E_{11}^x modes; its height is then

$$b = 1.66 \frac{\lambda}{n_1(\Delta')^{\frac{1}{2}}}. \qquad (32)$$

We can choose b arbitrarily by the proper selection of Δ'.
For

$$\lambda = 1\mu \ n_1 = 1.5, \text{ and } b = 5\mu,$$

from equations (31) and (32) we obtain

$$a = 25\mu, \ \Delta = 0.002, \text{ and } \Delta' = 0.05.$$

IV. DIRECTIONAL COUPLER

In general, the directional coupler can transmit E_{pq}^x and E_{pq}^y modes; but if the sides a and b of the guides are selected small enough, only the fundamental modes E_{11}^x and E_{11}^y are guided. Let us concentrate on the E_{11}^y mode. The coupler guides two kinds of E_{11}^y modes: one is symmetric (Fig. 9c) while the other is antisymmetric (Fig. 9d). Both are essentially TEM modes with main field components E_y and H_x. The electric and magnetic field intensity profiles for both modes are depicted qualitatively in Figs. 9b, c, and d.

Ignoring the small effects introduced by the loose coupling, the electrical width $k_x a$ and height $k_y b$ of each guide, as well as the field penetrations ξ_3 and η_2, coincide with those of the guide described in
$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad 5 \quad\quad\quad 4$
Section III. Similar reasoning applies to the E_{11}^x mode.

The coupling coefficient K between the two guides and the length L necessary for complete transfer of power from one to the other are, according to equations (56) and (59),[12]

$$-iK = \frac{\pi}{2L} = 2\frac{k_x^2 \, \xi_5}{k_z \, a} \frac{\exp\,(-c/\xi_5)}{1 + k_x^2\xi_5^2}. \qquad (33)$$

For E_{pq}^y modes, k_z and ξ_5 are given in equations (3) and (8), and k_x is the solution of equation (6). Similarly, for E_{pq}^x modes, k_z, ξ_5, and k_x are obtained from equations (17), (18), and (20). As expected, the coupling decreases exponentially with the ratio c/ξ_5 between the guide's separation and the field penetration in medium 5.

The normalized coupling coefficient

$$\frac{|K| \, a}{\left[1 - \left(\dfrac{n_5}{n_1}\right)^2\right]^{\frac{1}{2}}} \frac{k_z}{k_1} = \frac{\pi}{2}\frac{a}{L} \frac{1}{\left[1 - \left(\dfrac{n_5}{n_1}\right)^2\right]^{\frac{1}{2}}} \frac{k_z}{k_1}$$

$$= 2\left(\frac{k_x A_5}{\pi}\right)^2\left[1 - \left(\frac{k_x A_5}{\pi}\right)^2\right]^{\frac{1}{2}} \exp\left\{-\pi \frac{c}{A_5}\left[1 - \left(\frac{k_x A_5}{\pi}\right)^2\right]^{\frac{1}{2}}\right\} \quad (34)$$
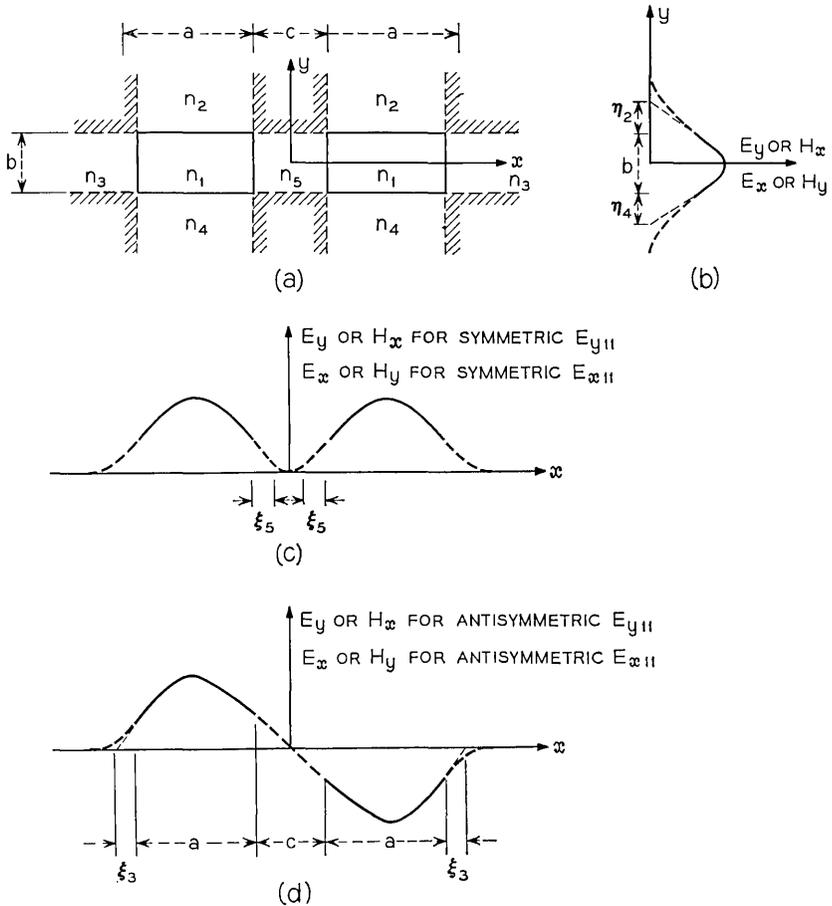
Fig. 9 — Directional coupler immersed in several dielectrics: (a) cross section, (b), (c), and (d) field distributions.

derived from equation (33) by substituting ξ_5 for its value given in equation (8) has been plotted in Fig. 10 for the E_{1a}^y mode, assuming $n_3 = n_5$ and n_1/n_5 is arbitrary. The solid and dotted lines were obtained using the exact solution of (6) and the approximate expression (12), respectively, for k_x. Both sets of curves are close to each other, especially for $2a/\lambda(n_1^2 - n_5^2)^{\frac{1}{2}} \geqq 1$.

The dashed-dotted lines are the couplings obtained by A. L. Jones[5] for two parallel cylinders of refractive index $n_1 = 1.8$ embedded in a medium $n_5 = 1.5$.[5] As expected, if the diameters of the round guides are

equal to the widths of the rectangular guides, and if the separations are the same, the coupling between the round guides should be slightly smaller than that between the rectangular ones.

The normalized coupling equation (34) for the E_{1q}^x mode has been plotted in Fig. 11, using for k_x the exact solution of equation (20). For n_1/n_5 close to unity, the lines get close to the solid curves in Fig. 10 as the E_{1q}^y and E_{1q}^x modes approach degeneracy. The influence of the height b of the guides, the refractive indices n_2 and n_4, and the value of q in the coupling of either mode is not important since they only affect k_z.

To work some examples, assume

$$n_1 = 1.5, \qquad n_2 = n_3 = n_4 = n_5 = \frac{1.5}{1.01}, \quad \text{and} \quad a = 2b.$$
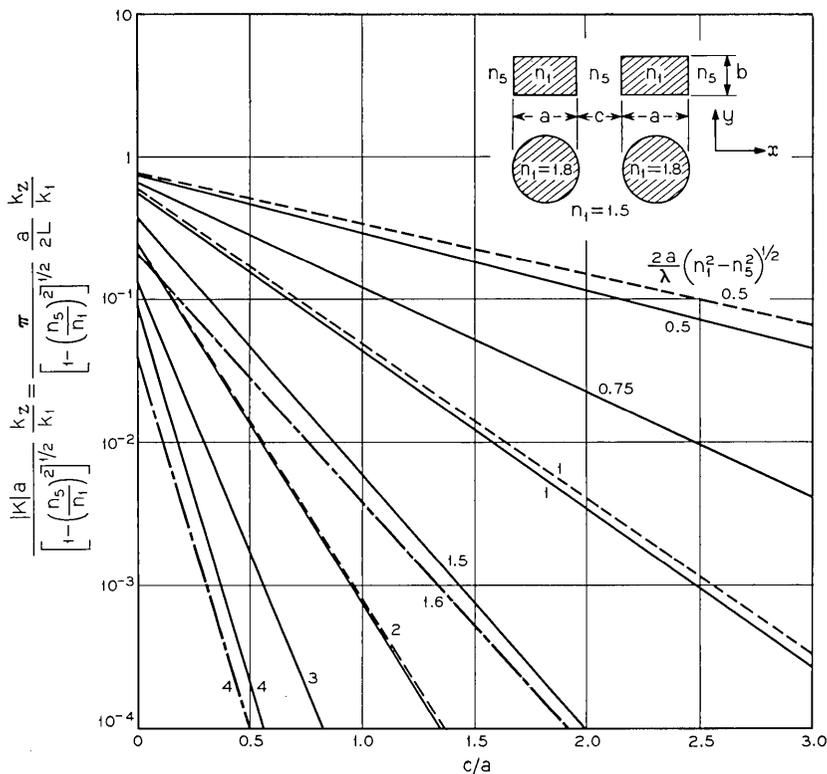
Fig. 10 — Coupling coefficient for $E_{1q}{}^y$ modes. ———— coupling calculated from trancendental equations; — — — — — closed form approximations; —·—·— coupling between two cylindrical rods (A. L. Jones[5]).
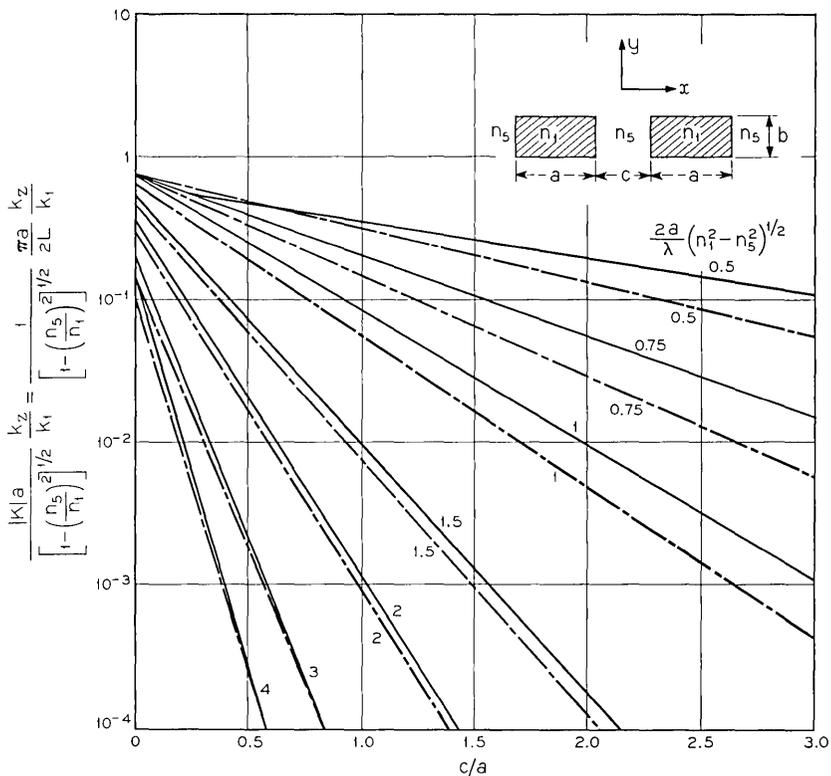
Fig. 11 — Coupling coefficient for $E_{1q}{}^x$ modes. ———— $E_{1q}{}^x$ coupling for $n_1/n_5 = 1.5$; —·—·— $E_{1q}{}^x$ coupling for $n_1/n_5 = 1.1$.

To insure that each guide only supports the E_{11}^x and E_{11}^y modes, the normalized dimension b according to Fig. 6b must be chosen to be

$$\frac{2b}{\lambda} \left(n_1^2 - n_4^2\right)^{\frac{1}{2}} = 0.75.$$

Consequently

$$b = 1.77\lambda, \qquad a = 3.54\lambda, \quad \text{and} \quad \frac{k_z}{k_1} \cong 1.$$

From Fig. 10 we obtain the coupler length L for complete power transfer:

$$L = 6540\lambda \quad \text{for} \quad c = a \quad \text{and} \quad L = 262\lambda \quad \text{for} \quad c = \frac{a}{4}.$$

How far apart should two guides of length l be spaced to have small coupling? If the transfer coefficient $|T| = l\,|K| \ll 1$, from equation (33) we derive

$$c = \xi_5 \log\left[2\,\frac{l}{|T|}\,\frac{k_x^2}{k_z}\,\frac{\xi_5}{a}\,\frac{1}{1 + k_x^2\xi_5^2}\right].$$ (35)

For the same guide dimensions of the previous example and for

$$l = 1 \text{ cm}, \quad \lambda = 1\mu, \quad \text{and} \quad T = 0.01,$$

we derive, from either equation (35) or Fig. 10, that $c/a = 2.5$. Consequently, both guides 3.54μ wide and 1 cm long would couple -40dB if their separtion is 8.9μ.

Now we evaluate how a small change of the refractive index between the guides modifies their coupling. Such would be the case if the medium between the guides is, for example, an electrooptic material and we change the applied field to modulate or switch the output.

For E_{11}^x and E_{11}^y modes, assuming well-guided modes $(k_xA_5/\pi \ll 1)$ and $n_1 - n_5/n_1 \ll 1$, the ratio between couplings for two values of refractive index in medium 5 (for example, n_5 and $n_5(1 + \delta)$), result from equations (34) and (12):

$$\frac{K_1}{K_2} = \frac{L_2}{L_1} = \exp\left\{-\pi\left(\frac{n_1^2}{n_5^2} - 1\right)^{-1}\frac{c\delta}{A_5}\left[1 - \left(\frac{2}{\pi} + \frac{a}{A_5}\right)^{-2}\right]^{\frac{1}{2}}\right\}.$$ (36)

That ratio is 1/2 if

$$\delta = 0.22\left(\frac{n_1^2}{n_5^2} - 1\right)\frac{A_5}{c}\left[1 - \left(\frac{2}{\pi} + \frac{a}{A_5}\right)^{-2}\right]^{-\frac{1}{2}}.$$ (37)

A directional coupler with coupling coefficient K_1 and length $L = \pi/|2K_1|$ would transfer all the power from one guide to the other. If the refractive index of the medium between the guides was changed from n_5 to $n_5(1+\delta)$ such that equation (37) is satisfied, the power would emerge at the end of the input guide. The larger the separation c of the guides, and the smaller the difference of refractive indexes $n_1 - n_5$, the smaller the change of refractive index required.

Following the example above, for

$$n_1 = 1.5, \qquad n_2 = n_3 = n_4 = n_5 = \frac{1.5}{1.01},$$

$$a = 1.5A_5 = 3.54\lambda, \quad \text{and} \quad c = a,$$

the percentage change of index required is only $\delta = 0.0033$.

V. DIRECTIONAL COUPLER MADE WITH SLIGHTLY DIFFERENT GUIDES

Consider the directional coupler of Fig. 12 in which the two guides have slightly different heights: one measures $b + h$ and the other $b - h$.

Let us qualitatively plot the coupling coefficient as a function of h, Fig. 13. Because of simple arguments of symmetry, the absolute value of coupling coefficient is stationary (first derivative zero) around $h = 0$. Therefore, the coupling coefficient between two guides of height b_1 and b_2 is the same as that of the coupling between two identical guides of height $1/2(b_1 + b_2)$, provided that $| b_1 - b_2 |$ is small enough.

This reasoning applies to guides with different widths, heights, and refractive indices, provided that the differences are small enough. Unfortunately, as in most perturbation analysis, we don't know what "small enough" is unless we calculate the next higher order term.

VI. SUMMARY AND CONCLUSIONS

A dielectric rod (Fig. 4a) of rectangular cross section a by b surrounded by different dielectrics supports, through total internal reflection, two families of hybrid modes. They are essentially TEM modes polarized either in the x or the y direction; we call them E^x_{pq} and E^y_{pq}. The subindices state the number of extrema (p in the x direction and q in the y direction) of the magnetic or electric transverse field components.

Dispersion curves for guides of different proportions and different surrounding dielectric are plotted in Figs. 6a through k. Typical dimensions for several guides capable of supporting only the fundamental modes E^x_{11} and E^y_{11} are contained in Table I.

By picking dielectrics with similar indexes, the guide dimensions can be made large compared with λ, thus reducing the tolerance requirements. The dimensions a and b can be picked arbitrarily and still achieve
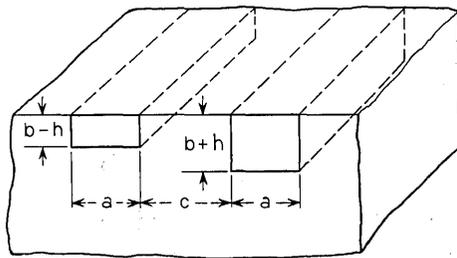
Fig. 12 — Directional coupler with guides of different heights.
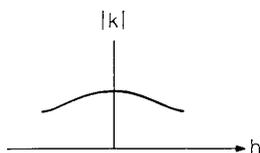
Fig. 13 — Qualitative behavior of the coupling coefficient as a function of h.

a guide which supports only the fundamental modes if one can choose the refractive indexes. The design is achieved with the help of either equation (14) or Fig. 7.

The penalty one pays with most of these guides is that the fundamental modes are almost degenerate; consequently, symmetry imperfections tend to couple them. A lossy layer added to the interface $y = b/2$ (Fig. 4a) should attenuate the E_{11}^x mode more than the E_{11}^y. As an alternative, the guide can be made to support only the E_{11}^y mode by metalizing the same interface. Dispersion curves are shown in Fig. 8.

Since the field is not confined, there is coupling between two of these guides (Fig. 3). Design curves for directional couplers are given in Figs. 10 and 11.

Typically, for $n_1 = 1.5$, $n_2 = n_3 = n_4 = n_5 = 1.5/1.01$, $a = 3.54\lambda$, $b = a/2 = 1.77\lambda$, and $c = a/4 = 0.88\lambda$, according to equation (33) the length necessary for 3dB coupling is $L/2 = 131\lambda$. This length increases exponentially with the separation between the guides.

Increasing the refractive index between the guides by a 3 per thousand doubles the coupling.

What is a reasonable separation to prevent coupling? Using the numbers of the previous example, two parallel guides 1 cm long separated by 2.5 times the width of each guide have a coupling of −40 dB.

The dielectric waveguides and the directional couplers described show great promise as basic elements for integrated optical circuitry because they:

(*i*) Can be made single mode even though their transverse dimensions can be large compared with the free space wavelength of operation. Consequently, the tolerance requirements can be relaxed.

(*ii*) Permit the building of compact optical components.

(*iii*) Are mechanically stable and alignment problems are minimized.

(*iv*) Are relatively simple structures and lend themselves to being fabricated with high precision integrated circuit techniques.

(*v*) Can include active devices of comparable small dimensions.

APPENDIX A

Field Analysis of the Directional Coupler

We solve Maxwell's equations for the directional coupler whose cross section is depicted in Fig. 3. The structure is symmetric with respect to the $x = 0$ plane; therefore, the modes have electric fields which are either symmetric or antisymmetric with respect to that plane. Consequently, the guide we have to study is simpler (Fig 14): if the plane $x = 0$ is an electric short circuit, the modes of the coupler propagating along z are antisymmetric; if the plane $x = 0$ is a magnetic short circuit, the modes are symmetric. As is known, it is the interaction of these symmetric and antisymmetric modes traveling with different phase velocities along z that represents the effect of coupling.

As discussed in Section II, by neglecting the power propagating through the shaded areas, the fields must be matched only along the sides of region 1. We find that two families of modes can satisfy the boundary conditions; we call them E_{pq}^x and E_{pq}^y. Each mode in the first family has most of its electric field polarized in the x direction, while each mode of the second family has the electric field almost completely polarized in the y direction. The subindexes p and q characterize the member of the family by the number of extrema that these transverse field components have along the x and y directions, respectively. For example, the E_{11}^x mode has its electric field virtually along x, its magnetic field along y; the amplitudes of the field have one maximum in each direction.

Each family of modes will be studied separately.

A.1 E_{pq}^y *Modes: Polarization Along y*

The field components in the νth of the five areas in Fig. 14 are:[13]

$$
H_{x\nu} = \exp\left(-ik_zz + i\omega t\right)\begin{cases} M_1 \cos\left(k_xx + \alpha\right)\cos\left(k_yy + \beta\right) & \text{for } \nu = 1 \\ M_2 \cos\left(k_xx + \alpha\right)\exp\left(-ik_{y2}y\right) & \text{for } \nu = 2 \\ M_3 \cos\left(k_yy + \beta\right)\exp\left(-ik_{x3}x\right) & \text{for } \nu = 3 \\ M_4 \cos\left(k_xx + \alpha\right)\exp\left(ik_{y4}y\right) & \text{for } \nu = 4 \\ M_5 \cos\left(k_yy + \beta\right)\sin\left(k_{x5}x + \gamma\right) & \text{for } \nu = 5 \end{cases}
$$

$$H_{y\nu} = 0$$

$$H_{z\nu} = -\frac{i}{k_z}\frac{\partial^2 H_{x\nu}}{\partial x\,\partial y} \tag{38}$$
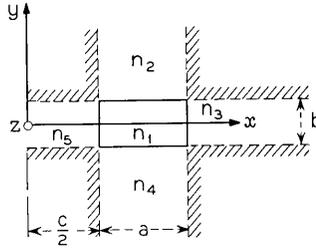
Fig. 14 — Coupler cross section with plane $x = 0$ either an electric or magnetic short circuit.

$$E_{x\nu} = -\frac{1}{\omega\epsilon n_\nu^2 k_z} \frac{\partial^2 H_{x\nu}}{\partial x \, \partial y}$$

$$E_{y\nu} = \frac{k^2 n_\nu^2 - k_{y\nu}^2}{\omega\epsilon n_\nu^2 k_z} H_{x\nu}$$

$$E_{z\nu} = \frac{i}{\omega\epsilon n_\nu^2} \frac{\partial H_{x\nu}}{\partial y}$$

in which M_ν determines the amplitude of the field in the νth medium; α and β locate the field maxima and minima in region 1; γ equal to $0°$ or $90°$ implies that the plane $x = 0$ is an electric (antisymmetric mode) or magnetic (symmetric mode) short circuit, respectively; ω is the angular frequency; ϵ and μ (appearing in $k^2 = \omega^2\epsilon\mu$) are the permittivity and permeability of free space.

In the νth medium the refractive index is n_ν, and the propagation constants $k_{x\nu}$, $k_{y\nu}$, and k_z are related by

$$k_{x\nu}^2 + k_{y\nu}^2 + k_z^2 = \omega^2\epsilon\mu n_\nu^2 = k_\nu^2 . \tag{39}$$

To match the fields at the boundaries between the region 1 and the regions 2 and 4, we have assumed in equation (38)

$$k_{x1} = k_{x2} = k_{x4} = k_x \tag{40}$$

and similarly to match the fields between media 1, 3, and 5,

$$k_{y1} = k_{y3} = k_{y5} = k_y . \tag{41}$$

Before finding the characteristic equations, let us assume the refractive index n_1 of the guide to be slightly larger than the others. That is

$$\frac{n_1}{n_2} - 1 \ll 1. \tag{42}$$

$$\begin{matrix} 3 \\ 4 \\ 5 \end{matrix}$$

As a consequence only modes made of plane wavelets impinging at grazing angles on the surface of medium 1 are guided. Since this implies that

$$\begin{matrix} k_x \\ k_y \end{matrix} \ll k_z \; , \tag{43}$$

the field components E_x in equation (38) can be neglected.

Now we match the remaining tangential components along the edges of region 1 and from equation (38) we obtain

$$\tan\left(k_y \frac{b}{2} \pm \beta\right) = i \frac{n_1^2}{n_2^2} \frac{k_{y2}}{k_y}. \tag{44}$$

$$\tan\left[k_x \begin{bmatrix} \frac{c}{2} \\ a + \frac{c}{2} \end{bmatrix} + \alpha\right] = i \frac{k_{x5}}{\frac{k_x}{3}} \begin{bmatrix} ictn\left(k_{x5}\frac{c}{2} + \gamma\right) \\ 1 \end{bmatrix}. \tag{45}$$

Where there are two choices, the upper ones go together and the lower ones go together.

T. Li pointed out that each of these equations considered separately is the characteristic equation of a boundary value problem simpler than that of Fig. 14.[8, 9] Thus for a dielectric slab infinite in the x and z directions and refractive indexes as depicted in Fig. 15a, the characteristic equation for modes with no H_y component coincides with
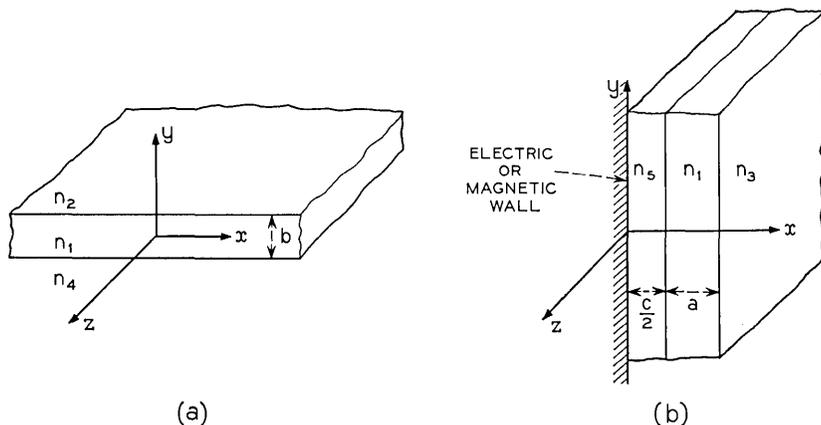
Fig. 15 — Dielectric slabs.

equation (44). Similarly, for two slabs infinite in the y and z directions and limited at $x = 0$ by an electric or magnetic short as in Fig. 15b, the characteristic equation for modes with $E_x = 0$ is equation (45).

A similar technique has been used by Schlosser and Unger to find the transmission properties of a rectangular dielectric guide immersed in another dielectric.[7] If the two guiding rods are so far apart that the coupling between them is a perturbation, then

$$| k_{x5} c | \gg 1 \tag{46}$$

and we can rewrite the characteristic equations (44) and (45) with the help of equations (39) and (46), making a and b explicit, as

$$k_y b = q\pi - \tan^{-1} \frac{n_2^2}{n_1^2} k_y \eta_2 - \tan^{-1} \frac{n_4^2}{n_1^2} k_y \eta_4 \tag{47}$$

$$k_x a = k_{x0} a \left[1 + \frac{2\xi_5}{a} \frac{\exp\left(-\dfrac{c}{\xi_5} - i2\gamma\right)}{1 + k_{x0}^2 \xi_5^2} \right] \tag{48}$$

where k_{x0} is the solution of

$$k_{x0} a = p\pi - \tan^{-1} k_{x0} \xi_3 - \tan^{-1} k_{x0} \xi_5 , \tag{49}$$

$$\eta_{2\atop 4} = \frac{1}{\left| k_{y2\atop 4} \right|} = \frac{1}{\left[\left[\dfrac{\pi}{A_{2\atop 4}} \right]^2 - k_y^2 \right]^{\frac{1}{2}}} \tag{50}$$

$$\xi_{3\atop 5} = \frac{1}{\left| k_{x3\atop 5} \right|} = \frac{1}{\left[\left[\dfrac{\pi}{A_{3\atop 5}} \right]^2 - k_{x0}^2 \right]^{\frac{1}{2}}} \tag{51}$$

and

$$A_{2,3,4,5} = \frac{\pi}{(k_1^2 - k_{2,3,4,5}^2)^{\frac{1}{2}}} = \frac{\lambda}{(n_1^2 - n_{2,3,4,5}^2)^{\frac{1}{2}}}. \tag{52}$$

In the transcendental equations (47) to (49), p and q are the arbitrary integers characterizing the order of the propagating mode, and the $\tan^{-1}$ functions are to be taken in the first quadrant. The angles $k_x a$ and $k_y b$ measure the phase shift of any field component across the guiding rod in the x and y directions respectively, or in other words, the electrical width and height of each guide of the coupler. On the other hand, $k_{x0} a$ is the electrical width of each guide assuming no interaction between the guides, that is assuming $c \to \infty$.

Let us find the physical significance of $\eta_{2,4}$ and $\xi_{3,5}$. The amplitude of each field component in medium 2 (Fig 14) decreases exponentially along y. It decays by $1/e$ in a distance η_2 given by equation (50). Similarly η_4, ξ_3, and ξ_5 measure the "penetration depth" of the field components in media 4, 3, and 5, respectively.

The propagation constant along z for each mode of the coupler is, according to equations (39), (40), and (41),

$$k_z = (k_1^2 - k_x^2 - k_y^2)^{\frac{1}{2}}. \tag{53}$$

With the help of equation (48), the slightly different propagation constants of the symmetric ($\gamma = 90°$) and antisymmetric modes ($\gamma = 0$) are

$$\left.\begin{matrix} k_{zs} \\ k_{za} \end{matrix}\right\} = k_{z0}\left[1 \pm 2 \frac{k_{x0}^2}{k_{z0}^2} \frac{\xi_5}{a} \frac{\exp{(-c/\xi_5)}}{1 + k_{x0}^2\xi_5^2} \right]. \tag{54}$$

In this expression

$$k_{z0} = (k_1^2 - k_{x0}^2 - k_y^2)^{\frac{1}{2}} \tag{55}$$

is the propagation constant of the E_{pq}^y mode of a single guide ($c \to \infty$).

The coupling coefficient K between the two guides and the length L necessary for complete transfer of power from one to the other are related to the propagation constants k_{zs} and k_{za} by[12]

$$-iK = \frac{\pi}{2L} = \frac{k_{zs} - k_{za}}{2} = 2 \frac{k_{x0}^2}{k_{z0}} \frac{\xi_5}{a} \frac{\exp{(-c/\xi_5)}}{1 + k_{x0}^2\xi_5^2}$$

$$= \frac{2}{\pi} \frac{A_5 k_{x0}^2}{a k_{z0}} \left[1 - \left(\frac{k_{x0}A_5}{\pi}\right)^2 \right]^{\frac{1}{2}} \exp\left\{ -\frac{\pi c}{A_5}\left[1 - \left(\frac{k_{x0}A_5}{\pi}\right)^2 \right]^{\frac{1}{2}} \right\}. \tag{56}$$

As expected, the coupling increases exponentially both by decreasing c and by increasing the penetration depth ξ_5 in medium 5.

All these formulas contain either k_{x0} or k_y, which are solutions of the transcendental equations (47) and (49). For well-guided modes, most of the power travels within medium 1 and consequently

$$\left(\frac{k_{x0}A_3}{\pi}\right)^2 \ll 1 \tag{57}$$

and

$$\left(\frac{k_y A_2}{\pi}\right)^2 \ll 1. \tag{58}$$

It is possible then to solve those transcendental equations in a closed though approximate form by expanding the $\tan^{-1}$ functions in power of those small quantities and keeping the first two terms of the expansions. The explicit solutions of equations (47), (49), (50), (51), (55), and (56) are given in Section III.

A.2 E^x_{pq} *Modes: Polarization in the x Direction*

The field components and propagation constants can be derived from those in Section A.1 by changing E to H and μ to $-\epsilon$, and vice versa. Except for their polarizations, the E^x_{pq} and E^y_{pq} modes are very similar and have comparable propagation constants. Using boldface type to distinguish the symbols corresponding to E^x_{pq} modes, from equations (56), (55), (47), (49), (50), and (51), we obtain

$$-i\mathbf{K} = \frac{\pi}{2\mathbf{L}} = 2\frac{\mathbf{k}^2_{x0}}{\mathbf{k}_{x0}}\frac{\boldsymbol{\xi}_5}{a}\frac{\exp\left(-c/\boldsymbol{\xi}_5\right)}{1 + (\mathbf{k}_{x0}\boldsymbol{\xi}_5)^2} \tag{59}$$

where

$$\mathbf{k}_{x0} = (k^2_1 - \mathbf{k}^2_{x0} - \mathbf{k}^2_y)^{\frac{1}{2}} \tag{60}$$

and $\mathbf{k}_{x0}$ and $\mathbf{k}_y$ are solutions of the transcendental equations

$$\mathbf{k}_y b = q\pi - \tan^{-1}\mathbf{k}_y\mathbf{n}_2 - \tan^{-1}\mathbf{k}_y\mathbf{n}_4 \tag{61}$$

and

$$\mathbf{k}_{x0}a = p\pi - \tan^{-1}\frac{n^2_3}{n^2_1}\mathbf{k}_{x0}\boldsymbol{\xi}_3 - \tan^{-1}\frac{n^2_5}{n^2_1}\mathbf{k}_{x0}\boldsymbol{\xi}_5 \tag{62}$$

in which

$$\mathbf{n}_{2\atop 4} = \frac{1}{\left[\left[\frac{\pi}{A_{2\atop 4}}\right]^2 - \mathbf{k}^2_y\right]^{\frac{1}{2}}} \tag{63}$$

and

$$\boldsymbol{\xi}_{3\atop 5} = \frac{1}{\left[\left(\frac{\pi}{A_{3\atop 5}}\right)^2 - \mathbf{k}^2_{x0}\right]^{\frac{1}{2}}} \tag{64}$$

As in Section A.1, the transcendental equations (61) and (62) can be solved in closed, though approximate, form provided that

$$\left[\frac{\mathbf{k}_{x0}A_{3\atop 5}}{\pi}\right]^2 \ll 1 \tag{65}$$

and

$$\left[\frac{\mathbf{k}_y A_2}{\pi}\right]^2 \ll 1. \tag{66}$$

The explicit results are given in Section III.

REFERENCES

1. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., this issue, pp. 2059–2069.
2. Schineller, E. R., "Summary of the Development of Optical Waveguides and Components," Report 1471, Wheeler Laboratories, Inc., April 1967.
3. Kaplan, R. A., "Optical Waveguide of Macroscopic Dimension in Single-Mode Operation," Proc. IEEE, 51, No. 8 (August 1963), pp. 1144–1145.
4. Goell, J. E., "A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides," B.S.T.J., this issue, pp. 2133–2160.
5. Jones, A. L., "Coupling of Optical Fibers and Scattering in Fibers," J. Opt. Soc. Amer., 55, No. 3 (March 1965), pp. 261–271.
6. Bracey, M. F., Cullen, A. L., Gillespie, E. F. F., and Staniforth, J. A., "Surface-Wave Research in Sheffield," I.R.E. Trans. Antennas and Propagation, AP-7, Special Supplement (December 1959), pp. S219–S225.
7. Schlosser, W., and Unger, H. G., "Partially Filled Waveguides and Surface Waveguides of Rectangular Cross Section," Advances in Microwaves, New York: Academic Press, 1966, pp. 319–387.
8. Collin, R. E., Field Theory of Guided Waves, New York: McGraw-Hill, 1966, pp. 470–477.
9. Nelson, D. F., and McKenna, J., "Electromagnetic Modes of Anisotropic Dielectric Waveguides at p-n Junctions," J. Appl. Phys., 38, No. 10 (September 1967), pp. 4057–4074.
10. Fox, A. G., and Li, T., "Resonant Modes in a Maser Interferometer," B.S.T.J., 40, No. 2 (March 1961), pp. 453–488.
11. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., this issue, pp. 2103–2132.
12. Miller, S. E., "Coupled Wave Theory and Waveguide Applications," B.S.T.J., 33, No. 3 (May 1954), pp. 661–719.
13. Schelkunoff, S. A., Electromagnetic Waves, New York: D. van Nostrand, 1943, p. 94.

Bends in Optical Dielectric Guides

By E. A. J. MARCATILI

Light transmission through a curved dielectric rod of rectangular cross section embedded in different dielectrics is analyzed in closed, though approximate form. We distinguish three ranges:

(i) Small cross section guides such as a thin glass ribbon surrounded by air—Making its width 1 percent of the wavelength, most of the power travels outside of the glass; the attenuation coefficient of the guide is two orders of magnitude smaller than that of glass, and the radius of curvature that doubles the straight guide loss is around 10,000λ.

(ii) Medium cross section guide for integration optics—It is only a few microns on the side and capable of guiding a single mode either in low loss bends with short radii of curvature or in a high Q closed loop useful for filters. Q's of the order of 10^8 are theoretically achievable in loops with radii ranging from 0.04 to 1 mm, if the percentage refractive index difference between guide and surrounding dielectric lies between 0.1 and 0.01.

(iii) Large cross section guides—They are multimode and are used in fiber optics. Conversion to higher order modes are found more significant than radiation loss resulting from curvature.

I. INTRODUCTION

A dielectric rod, embedded in one or more dielectrics of lower refractive index, is the basic ingredient of three types of optical waveguide which differ only in their relative dimensions and consequently in their guiding properties.

The first is a small cross section guide which supports only the fundamental mode; most of the power travels in a lower loss external medium. Thus, the attenuation of the mode is smaller than if all the power flowed through the higher loss internal medium. Tiny rods, thin ribbons, or films made of glass or other substances embedded in either air or low loss liquids are typical examples.[1-3]

The second is a medium size guide capable of supporting only a few

modes; most of the power travels in the internal medium. Such a guide, (Fig. 1 of Ref. 10) has been proposed as the building block of passive and active components for integrated optical circuitry.[4-6] Lasers, modulators, directional couplers, and filters are some of the many devices which could be built in a single substrate utilizing the high precision techniques available from integrated circuitry; consequently they would be compact, mechanically stable, and reproducible.

The third, a large size guide (clad fiber) which can support many modes, is used typically in fiber optics.[7]

These basic guides, having round or rectangular cross section and straight axis, have been studied both analytically and through computer calculations.[8-13] Also the directional coupler (Fig. 2 of Ref. 10) obtained by running two guides of rectangular or circular cross sections parallel to each other, has been analyzed.[10,12,14]

To my knowledge, though, little is known quantitatively about the ability of any of the three types of guides to negotiate bends, or about the radiation losses in loops, such as the one depicted in Fig. 1 as part of a channel dropping filter. This paper should supply such information.

In Section II the boundary value problem is discussed, and the fundamental modes of each polarization are described. Section III contains a discussion of the results and numerical examples. Conclusions are drawn in Section IV and all the mathematical derivations are exiled to the appendix.

II. FORMULATION OF THE BOUNDARY VALUE PROBLEM

Figure 2 depicts, in perspective, the basic geometry of the curved guide with radius of curvature R. The cross section is a rectangle whose sides are a and b. The refractive index of the guide is n_1, and the refractive indices around the guide are n_2, n_3, n_4, and n_5, all of which are smaller than n_1. Furthermore, for reasons which become apparent later, we do not specify the refractive indices in the four shaded areas.

This boundary value problem is solved in closed, though approximate form in the appendix, by introducing the same simplification used in solving the problem of transmission in the straight guide.[10] That simplification arises from solving Maxwell's equations only for guide dimensions such that a small percentage of the total power flows through the shaded areas and consequently a negligible error is expected if one does not match properly the fields along their edges.

Two types of hybrid modes propagate through this curved guide;

LOOP RESONANT AT
FREQUENCY f$_2$

$f_1 f_3 \cdots$

f_2

$f_1 f_2 f_3 \cdots$
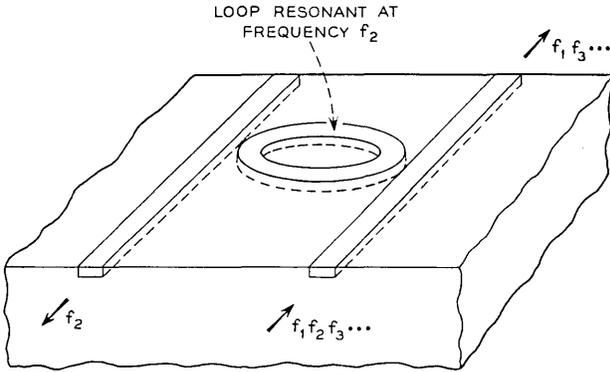
Fig. 1 — Channel dropping filter (ring type).

each one has six field components. But since some of the refractive indices n_2, n_3, n_4, and n_5 are chosen close to n_1, guidance occurs through total internal reflection only when the plane wavelets that make a mode impinge on the interfaces at grazing angles. Consequently, the only large field components are perpendicular to the curved z axis (Fig. 2). The modes are then of the TEM kind and we group them in two families, E_{pq}^x and E_{pq}^y. The main field components of the members of the first family are E_x and H_y, while those of the second are E_y and H_x.
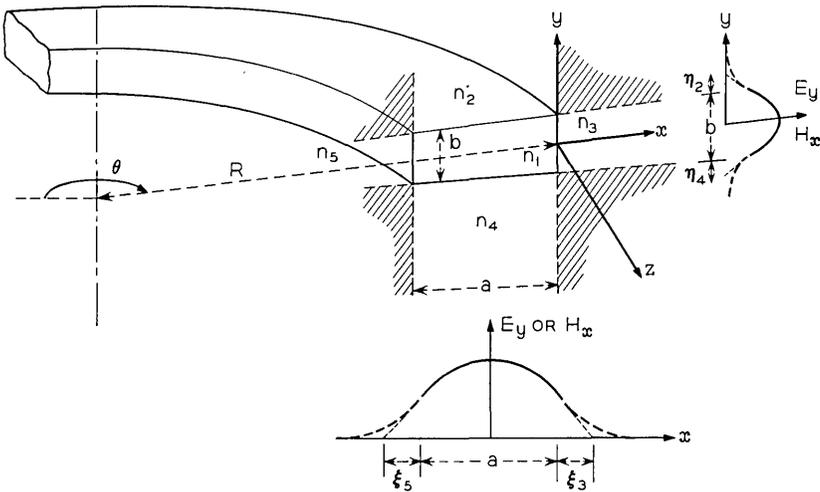
Fig. 2 — Curved dielectric guide.

Virtually every one of these components varies sinusoidally along x and y within the guiding medium 1 and decays exponentially in the surrounding media 2, 3, 4, and 5 (Fig. 2). The subindices p and q represent the number of extrema of each field component in the x and y directions, respectively. The field configurations of some members of the two families in straight guides are depicted in Fig. 5 of Ref. 10; section 2.1 describes the influence of a finite radius of curvature on those field configurations.

General expressions for the different phase and propagation constants in each medium of the curved guide are calculated in the appendix, for arbitrary modes and for $n_2 \neq n_3 \neq n_4 \neq n_5$. In the text, we consider only the fundamental modes of each family E_{11}^x and E_{11}^y; furthermore, we choose

$$n_3 = n_5 \tag{1}$$

and leave n_2 and n_4 arbitrary. This choice of refractive indexes encompasses the most interesting cases.

2.1 E_{11}^x Mode

We first study the E_{11}^x mode. As we said before, the main components are E_x along the x direction and H_y along y. Both components have a single maximum located within medium 1 and drop sinusoidally toward the edge of it. Outside of the medium, the decay is exponential.

The axial propagation constant is according to equation (47)

$$k_z = (k_1^2 - k_x^2 - k_y^2)^{\frac{1}{2}}, \tag{2}$$

where $k_1 = kn_1 = (2\pi/\lambda)n_1$ and λ is the free space wavelength, k_x is the propagation constant along x in media 1, 2, and 4, and k_y is the propagation constant along y in media 1, 3, and 5. This means that the electrical width of media 1, 2, and 4 is the same and equal to $k_x a$, and the electrical height of 1, 3, and 5 is also the same and equal $k_y b$.

The transverse propagation constant k_y is independent of the radius of curvature R and can be found from the transcendental equation (37)

$$k_y b = \pi - \tan^{-1}\left[\left(\frac{\pi}{k_y A_2}\right)^2 - 1\right]^{-\frac{1}{2}} - \tan^{-1}\left[\left(\frac{\pi}{k_y A_4}\right)^2 - 1\right]^{-\frac{1}{2}} \tag{3}$$

in which

$$A_2 = \frac{\lambda}{2\left(n_1^2 - n_2^2\right)^{\frac{1}{2}}}. \tag{4}$$

If the height of the guide b is selected so large that

$$\frac{A_2 + A_4}{\pi b} \ll 1, \tag{5}$$

only a small percentage of the power carried by the mode travels in media 2 and 4; and equation (3) can be solved approximately, yielding

$$k_y = \frac{\pi}{b}\left(1 + \frac{A_2 + A_4}{\pi b}\right)^{-1}.$$

According to equation (49), the other transverse propagation constant

$$k_x = k_{x0}\left[1 + \frac{2c}{ak_{x0}} - i\frac{k_{z0}\alpha_c}{k_{x0}^2}\right] \tag{6}$$

is valid if

$$\frac{c}{ak_{x0}} \ll 1 \tag{7}$$

$$\alpha_c R \ll 1.$$

The first term in equation (6), k_{x0}, is the propagation constant in the x direction of the guide without curvature; the second and third terms, which according to equation (7) must be small, are perturbations related to the change of field profile and to radiation loss, both of which are introduced by the curvature. More precisely, α_c is the attenuation coefficient of the curved guide, $\alpha_c R$ is the attenuation per radian, that is the attenuation in a length of guide equal to R, and c is a conversion loss coefficient such that, at a junction between a straight and a curved section of the same guide, c^2 measures the power that the fundamental mode in the straight section would couple to modes higher than the fundamental in the curved section. The fact that equation (6) is valid if $c \ll 1$ requires the radius of curvature R to be so large that the field profiles of the fundamental modes in the straight and curved guides are quite similar. Later in this section we consider formulas applicable when $c \cong 1$.

The axial propagation constant, k_{z0}, of the straight guide is related to k_{x0} and k_y by the expression

$$k_{z0} = (k_1^2 - k_{x0}^2 - k_y^2)^{\frac{1}{2}}; \tag{8}$$

and k_{x0} is the solution of the transcendental equation (55)

$$k_{x0}a = \pi - 2\tan^{-1}\frac{n_3^2}{n_1^2}\left[\left(\frac{\pi}{k_{x0}A}\right)^2 - 1\right]^{-\frac{1}{2}}. \tag{9}$$

The length

$$A = \frac{\lambda}{2(n_1^2 - n_3^2)^{\frac{1}{2}}} \tag{10}$$

is used as a normalizing dimension. What does it measure? If one assumes $b = \infty$, the guide becomes a slab of width a. If a $\leq A$, only the fundamental mode is guided; if a $> A$, the slab is multimode.

Figure 3 is a graph of the electrical width, $k_{x0}a$, of the straight guide as a function of a/A. The solid curve is the solution of equation (9) assuming $n_1/n_3 = 1.5$, while the dotted one is the solution for $n_1/n_3 = 1$. For thin guides, $a/A \ll 1$, the electrical width is proportional to a; for thick guides, $a/A \gg 1$, the electrical width goes asymptotically to π.

The attenuation per radian $\alpha_c R$ and the conversion coefficient c, obtained from equations (50) and (51) with $n_3 = n_5$ are

$$\alpha_c R = \frac{1}{2}\left(1 - \frac{n_3^2}{n_1^2}\right)^{-\frac{1}{2}}\left(\frac{n_3 k_{x0}a}{n_1}\right)^2\left(\frac{A}{\pi a}\right)^3\left[1 - \left(\frac{k_{x0}A}{\pi}\right)^2\right]^{\frac{1}{2}}$$

$$\cdot\frac{\Re \exp\left\{-\frac{\Re}{3}\left[1 - \left(\frac{k_{x0}A}{\pi}\right)^2\left(1 + \frac{2c}{ak_{x0}}\right)^2\right]^{\frac{3}{2}}\right\}}{1 - \left(1 - \frac{n_3^4}{n_1^4}\right)\left(\frac{k_{x0}A}{\pi}\right)^2 + 2\frac{n_3^2 A}{n_1^2 a}\left[1 - \left(\frac{k_{x0}A}{\pi}\right)^2\right]^{-\frac{1}{2}}} \tag{11}$$

and

$$c = \frac{1}{2k_{x0}a}\left(\frac{\pi a}{A}\right)^3\frac{1}{\Re}, \tag{12}$$

where

$$\Re = \frac{2\pi^3 R}{k_{x0}^2 A^3} = 2\frac{k_1^3}{k_{x0}^2}\left(1 - \frac{n_3^2}{n_1^2}\right)^{\frac{3}{2}}R. \tag{13}$$

The solid curves in Figs. 4 and 5 are graphs of the function

$$\alpha_c R\left(1 - \frac{n_3^2}{n_1^2}\right)^{\frac{1}{2}}$$

(which is proportional to the attenuation per radian) as a function of a/A using $\Re$ as a parameter. In Fig. 4, we further assume that

$$\frac{n_1}{n_3} = 1 + \Delta$$
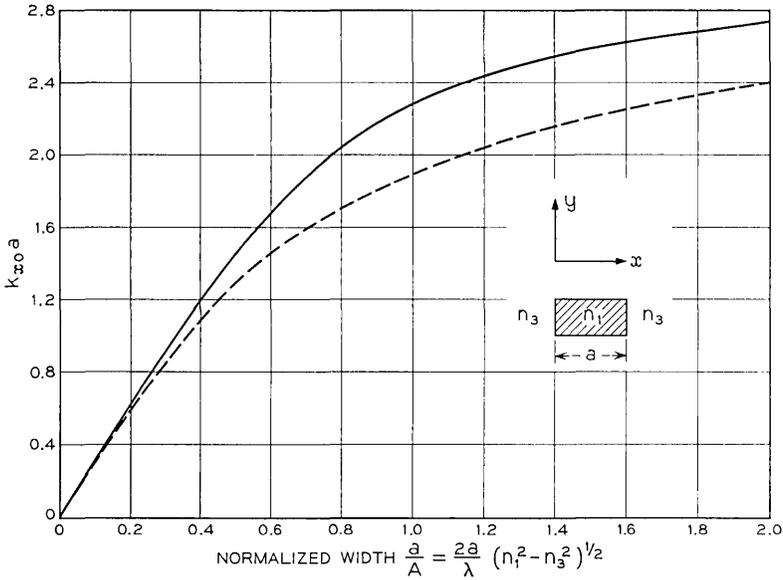
Fig. 3—Guide's electrical width. Solid line is for E^z_{11} mode with $n_1/n_3 = 1.5$; dashed line is for E^z_{11} mode with n_1/n_3 arbitrary, and for E^x_{11} mode with $n_1 \cong n_3$.

and

$$\Delta \ll 1;$$

in Fig. 5,

$$\frac{n_1}{n_3} = 1.5.$$

In the same figures each dashed line is a curve of constant conversion loss c. Since the calculations are valid for $c \ll 1$, we believe the solid curves are reliable to the left of the dotted curve $c = 0.3$ and grow progressively in error to the right of it.

To extend the use of this graph to arbitrarily large values of a/A, we calculate the loss per radian, equation (63), when $a/A \gg 1$ and $c \cong 1$. It is

$$\alpha_c R = \frac{n_3^2}{n_1^2} \left[1 - \left(\frac{n_3}{n_1}\right)^2 \right]^{-\frac{1}{2}} \exp\left\{ -\frac{\Re}{3} \left[1 - \left(\frac{9\pi}{2\Re}\right)^{\frac{2}{3}} + \frac{4n_3^2}{n_1^2\Re} \right]^{\frac{3}{2}} \right\}; \quad (14)$$

the dotted lines in Figs. 4 and 5 represent this loss. The reader can smoothly extend the solid curves to the right of the dashed line, $c = 0.3$, so that they become asymptotic to the dotted lines. Thus, the whole range of guide width a from 0 to R has been covered.
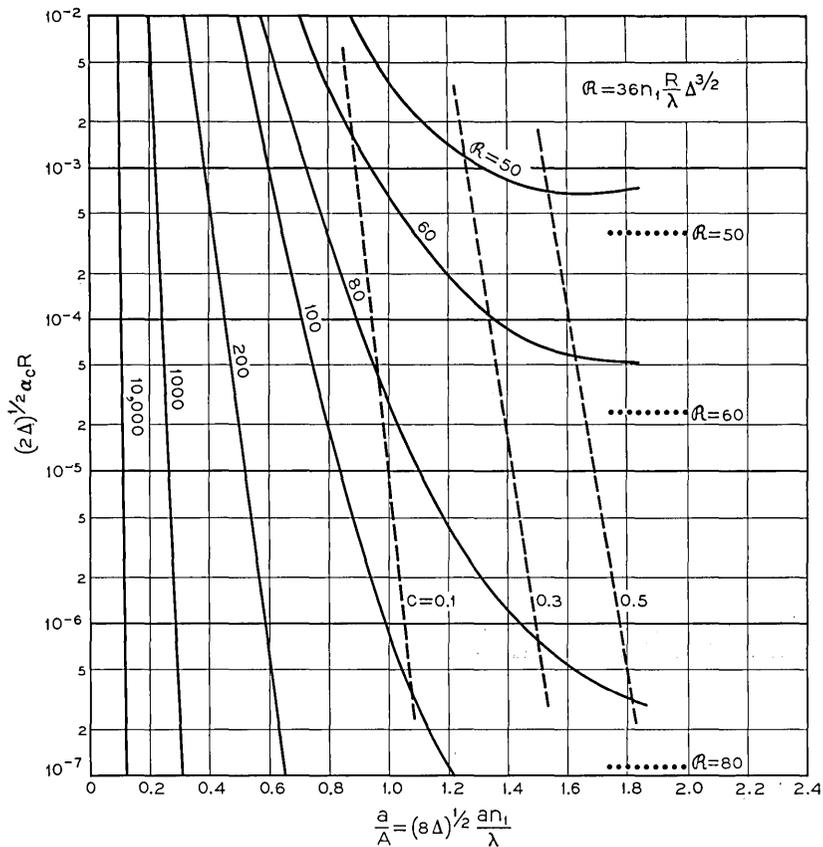
Fig. 4 — Attenuation per radian for $E_{11}{}^x$ and $E_{11}{}^y$ modes if $n_1/n_3 = 1 + \Delta$ and $\Delta \ll 1$.

To understand why these curves of constant R become asymptotic for $a/A \gg 1$, we have drawn in Fig. 6a a curved guide with a certain R; its width a is very large compared with A. Also the amplitudes of the field components E_x and H_y are plotted as functions of x and y.

Along x the field inside the guide behaves virtually as the Bessel function $J_\nu[k_1(R + x)]$ where ν is a very large number and outside of the guide decays exponentially. This guide has some radiation loss per radian.

Now, suppose that we start shrinking a without changing R. Since the field at $x = -a$ is very small, the radiation loss remains constant until a is made so short that the field at $x = 0$ and $x = -a$ are com-

parable (Fig. 6b). The field inside the guide varies almost sinusoidally, while outside decays exponentially and the attenuation per radian increases. If a is reduced even further (Fig. 6c) most of the power travels outside of the guide, and the loss increases even more. The field configuration along y is practically the same in the three cases (Fig. 6).

For resonant loops, such as the filter in Fig. 1, the intrinsic Q resulting from curvature radiation is more interesting than the attenuation α_c. They are related by the expression

$$Q_c = \frac{k_{z0}}{2\alpha_c}. \tag{15}$$
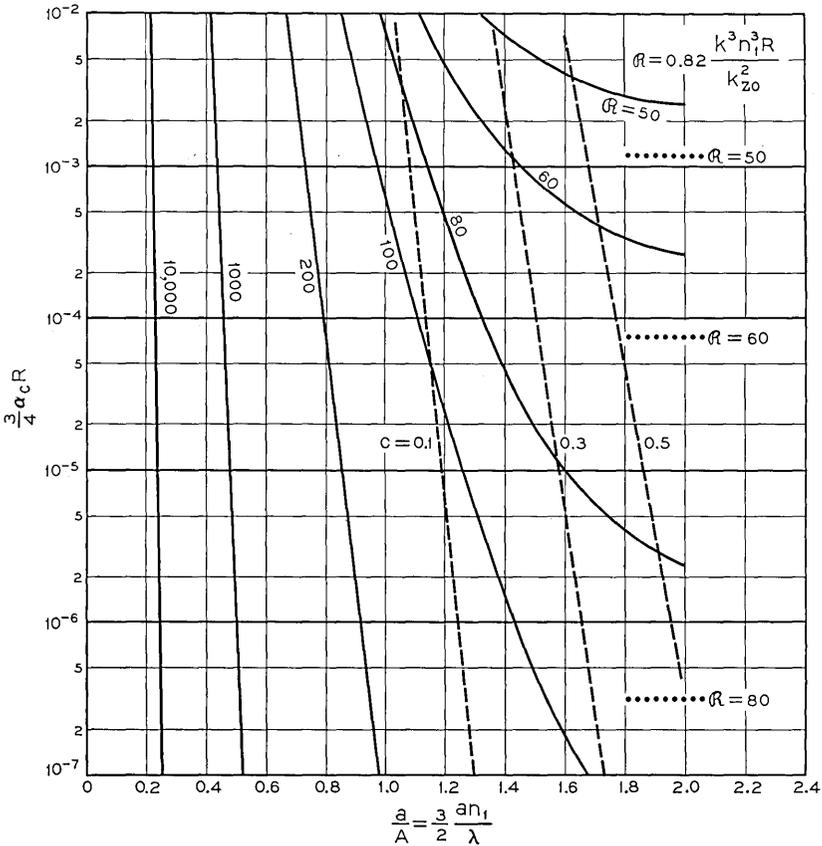
Fig. 5 — Attenuation per radian for E_{11}^x mode if $n_1/n_3 = 1.5$.
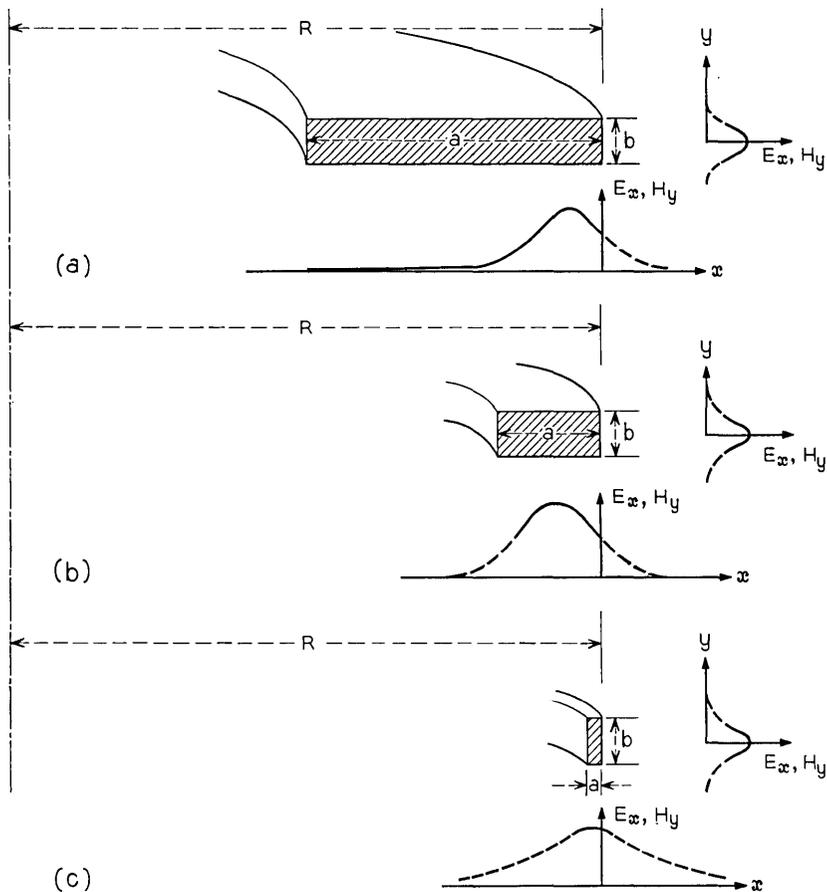
Fig. 6 — Field distribution as a function of guide width a with (a) $a/A \gg 1$, (b) $a/A \cong 1$, and (c) $a/A \ll 1$.

This function is plotted in Fig. 7, assuming

$$\frac{n_1}{n_3} = 1 + \Delta$$

and

$$\Delta \ll 1$$

and in Fig 8, assuming

$$\frac{n_1}{n_3} = 1.5,$$

using as before the normalized guide width a/A as variable and $\mathcal{R}$ as parameter. As in Figs. 4 and 5, the reader can easily match the solid and dotted curves. Further discussion of these curves is reserved for Section III.

The field components in media 2, 3, 4, and 5 decay almost exponentially away from the guiding rod, and the distances η_2, η_4, ξ_3, and ξ_5 over which the fields decrease by $1/e$ are
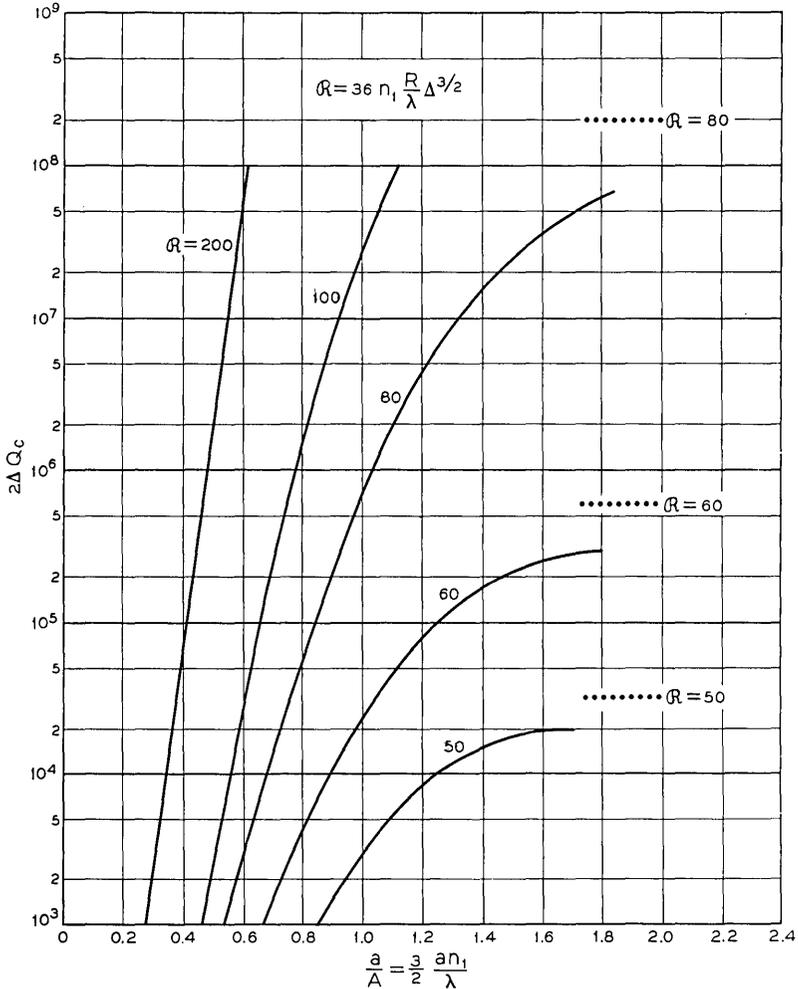
Fig. 7 — Intrinsic Q for E_{11}^z and E_{11}^y modes if $n_1/n_3 = 1 + \triangle$ and $\triangle \ll 1$.

$$\eta_{\underset{4}{2}} = \frac{1}{\left| k_{\underset{4}{y2}} \right|} = \frac{1}{\left(k_1^2 - k_2^2 - k_y^2 \right)^{\frac{1}{2}}} , \tag{16}$$

$$\xi_3 = \xi_5 = \frac{1}{\left| k_{x3} \right|} = \frac{1}{\left(k_1^2 - k_3^2 - \left| k_x^2 \right| \right)^{\frac{1}{2}}}. \tag{17}$$

2.2 E_{11}^y Mode

We now consider the E_{11}^y mode. The main components are E_y and H_x; they are qualitatively quite similar to components of the E_{11}^x mode, rotated $90°$.

The propagation constant k_z is still given by equation (2); but now k_y
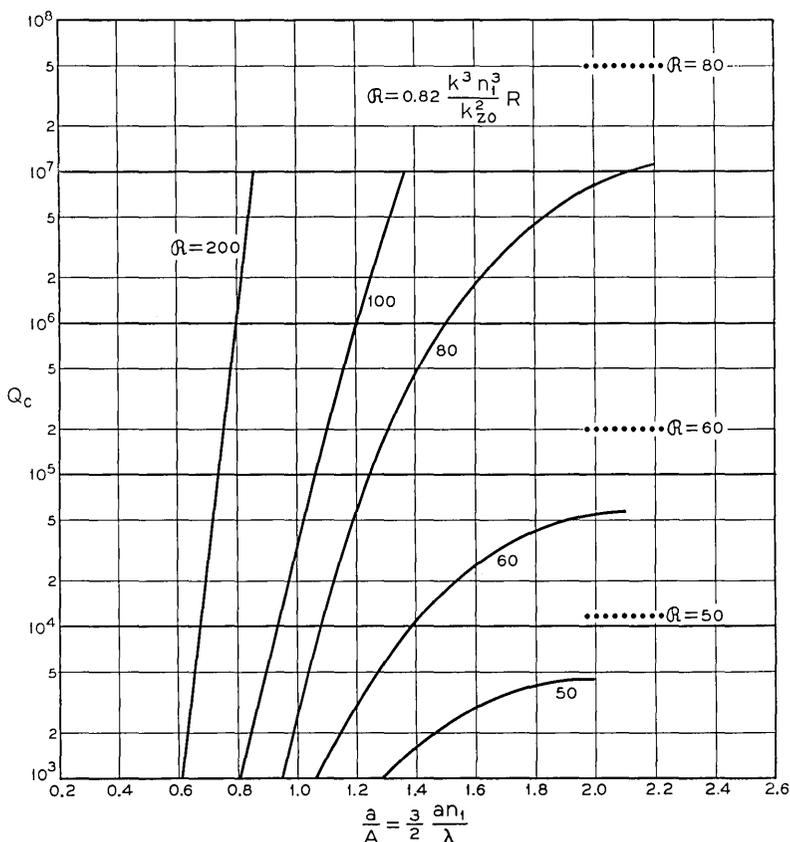
Fig. 8 — Intrinsic Q for E_{11}^x mode if $n_1/n_3 = 1.5$.

is the solution of

$$k_y b = \pi - \tan^{-1} \frac{n_2^2}{n_1^2} \left[\left(\frac{\pi}{k_y A_2} \right)^2 - 1 \right]^{-\frac{1}{2}} - \tan^{-1} \frac{n_4^2}{n_1^2} \left[\left(\frac{\pi}{k_y A_4} \right)^2 - 1 \right]^{-\frac{1}{2}}.$$

(18)

The equivalent formula of any of those between equation (7) and (17) can be derived from that formula by substituting the ratio of refractive indexes by unity, but leaving them unchanged wherever they are subtracted from unity. For example, equation (11) becomes

$$\alpha_c R = \frac{1}{2} \left[1 - \left(\frac{n_3}{n_1} \right)^2 \right]^{-\frac{1}{2}} (k_{x0}a)^2 \left(\frac{A}{\pi a} \right)^3 \left[1 - \left(\frac{k_{x0}A}{\pi} \right)^2 \right]^{\frac{1}{2}}$$

$$\cdot \frac{\mathfrak{R} \exp \left\{ -\frac{\mathfrak{R}}{3} \left[1 - \left(\frac{k_{x0}A}{\pi} \right)^2 \left(1 + \frac{2c}{ak_{x0}} \right)^2 \right]^{\frac{3}{2}} \right\}}{1 - 2 \left(1 - \frac{n_3^2}{n_1^2} \right) \left(\frac{k_{x0}A}{\pi} \right)^2 + 2 \frac{A}{a} \left[1 - \left(\frac{k_{x0}A}{\pi} \right)^2 \right]^{-\frac{1}{2}}},$$

(19)

while c and $\mathfrak{R}$ given by equations (12) and (13) remain unchanged.

Figure 9 is a graph of the function $\alpha_c R [1 - (n_3/n_1)^2]^{\frac{1}{2}}$, valid for any ratio n_1/n_3. In particular, for $n_1/n_3 = 1 + \Delta$ and $\Delta \ll 1$, equations (19) and (11) become the same, and consequently these curves coincide with those in Fig. 4. This means that for $n_1 \cong n_3$, the E_{11}^x and E_{11}^y modes have the same loss.

Figure 10 is a graph of the intrinsic Q of a loop operating in the E_{11}^y mode which can be derived from equations (15) and (19). As before, in a resonant loop with $n_1/n_3 = 1 + \Delta$ and $\Delta \ll 1$, the E_{11}^x or E_{11}^y modes have the same Q's.

III. DISCUSSION AND EXAMPLES

The attenuation per radian of any dielectric guide of rectangular cross section and the Q_c resulting from curvature are strongly dependent on the radius of curvature. With the help of equation (17), the attenuation per radian equation (11) can be written

$$\alpha_c R = MR \exp \left(-\frac{1}{6\pi^2} \frac{\lambda_z^2 R}{|\xi_3|^3} \right),$$

(20)

where M is independent of R, λ_z is the guided wavelength along z, and ξ_3 is the length over which the field in medium 3 decays by $1/e$. According to Fig. 11, the function

$$R \exp \left(-\frac{1}{6\pi^2} \frac{\lambda_z^2 R}{|\xi_3|^3} \right)$$

becomes negligibly small, and consequently the attenuation per radian becomes negligibly small when

$$R > \frac{24\pi^2 \, |\, \xi_3 \,|^3}{\lambda_z^2}. \tag{21}$$

This simple criterion is developed further in Ref. 15.

We are interested, though, in a more detailed description of transmission through a bent dielectric guide. Given a guide with a certain radius of curvature (that is, given R and a/A), in general the loss per radian of the E_{11}^x mode is much larger than that of the E_{11}^y mode (compare, for example, Figs. 5 and 9 for $n_1/n_3 = 1.5$). That difference becomes negligible if $n_1/n_3 - 1 \ll 1$.
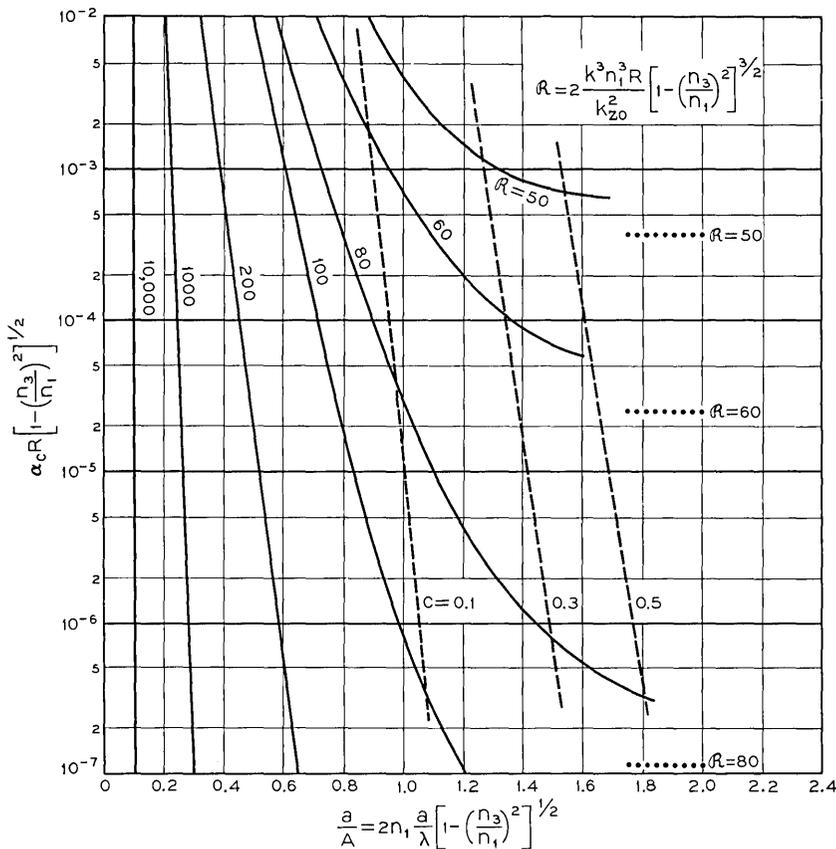
Fig. 9 — Attenuation per radian for E_{11}^y mode and $n_1/n_3 > 1$.
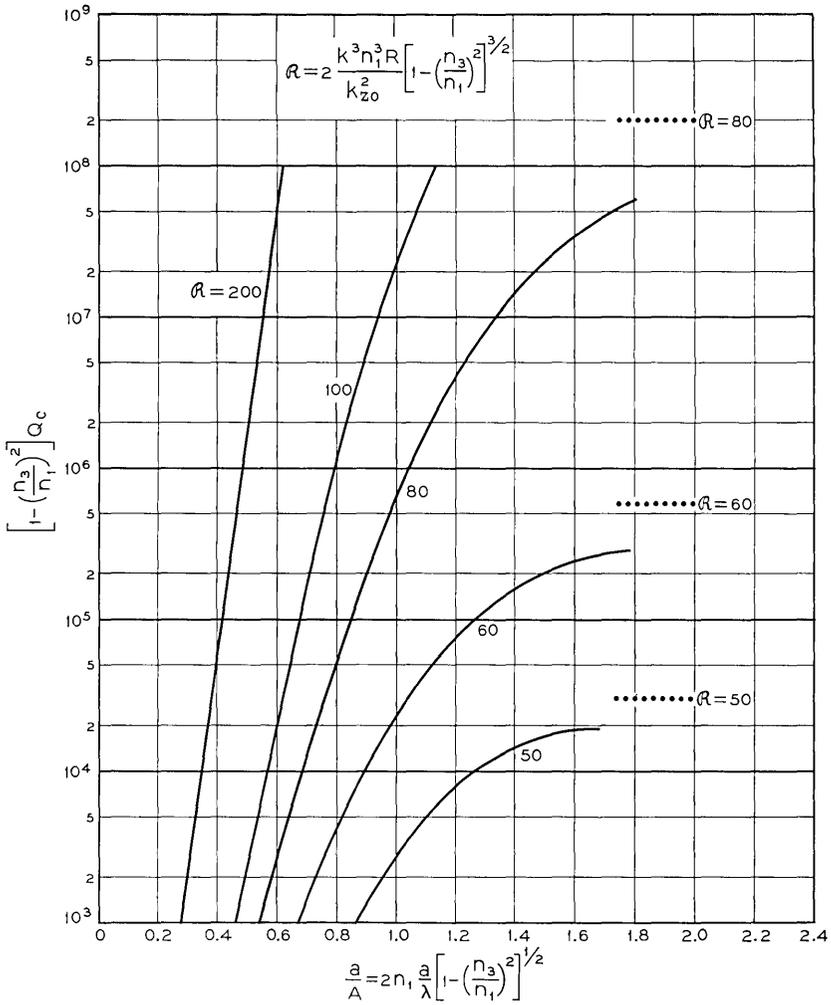
Fig. 10 — Intrinsic Q for $E_{11}{}^y$ mode and $n_1/n_3 > 1$.

Let us consider separately the three types of guide: thin, medium and large.

3.1 Thin or Low Loss Guides*

In thin guides the width a is so small that

_____

* Low loss for straight guide.

$$\frac{a}{A} = \frac{2a(n_1^2 - n_3^2)^{\frac{1}{2}}}{\lambda} \ll 1. \tag{22}$$

The height b of the guide must be large so that only a little part of the power travels in the shaded areas of Fig. 2. Assuming that the guiding rod dielectric is lossy, its refractive index is

$$n_1 = n\left(1 + \frac{i\alpha}{kn}\right), \tag{23}$$

where n is real and α is the attenuation constant of a plane wave in that medium.

Substituting equations (22) and (23) in equations (2), (11), and (12), we obtain

$$k_z = k_{z0} + i\alpha_s + i\alpha_c . \tag{24}$$

The first term

$$k_{z0} = (k_3^2 - k_y^2)^{\frac{1}{2}} \begin{cases} \left[1 + \dfrac{1}{8}\left[k_3a\left(1 - \dfrac{n_3^2}{n^2}\right)\right]\right]^2 & \text{for } E_{11}^x \text{ mode} \\[4mm] \left[1 + \dfrac{1}{8}\left[k_3a\left(\dfrac{n^2}{n_3^2} - 1\right)\right]\right]^2 & \text{for } E_{11}^y \text{ mode} \end{cases} \tag{25}$$

is the phase constant. Since most of the power travels in the external medium, its value for either mode is close to kn_3. The conversion loss term c is negligible.

The imaginary part of equation (24) is the attenuation constant, and is made of two terms. The first term

$$\alpha_s = \frac{\alpha}{2} nn_3k^2a^2\left(\frac{n^2}{n_3^2} - 1\right) \begin{cases} \left(\dfrac{n_3}{n}\right)^6 & \text{for } E_{11}^x \text{ mode} \\[3mm] 1 & \text{for } E_{11}^y \text{ mode} \end{cases} \tag{26}$$
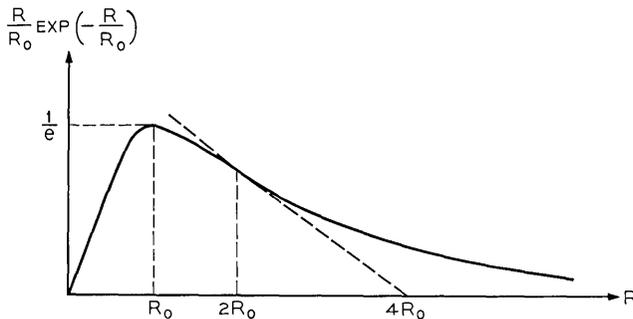
Fig. 11 — Plot of $R/R_0 \exp(-R/R_0)$ and tangent at inflection point.

is the attenuation that each mode would have if the guide were straight.[16] The second term

$$\alpha_c = \frac{k_3^3 a^2}{8}\left(\frac{n^2}{n_3^2} - 1\right)^2$$

$$\cdot \begin{cases} \left(\frac{n_3}{n}\right)^4 \exp\left\{-\frac{k_3^4 a^3 R}{12}\left(\frac{n_3}{n}\right)^6\left(\frac{n^2}{n_3^2} - 1\right)^3\left[1 - \frac{1}{2}\left(\frac{k_y}{k_3}\right)^2\right]\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(for } E_{11}^x \text{ mode)} \\ \exp\left\{-\frac{k_3^4 a^3 R}{12}\left(\frac{n^2}{n_3^2} - 1\right)^3\left[1 - \frac{1}{2}\left(\frac{k_y}{k_3}\right)^2\right]\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(for } E_{11}^y \text{ mode)} \end{cases} \tag{27}$$

is the attenuation resulting from the radiation introduced by the curvature. The E_{11}^y mode is more tightly bound to the guiding rod and consequently has more straight loss and less curvature loss than the E_{11}^x mode.

From equations (26) and (27), the radius of curvature R_d that doubles the straight guide loss is

$$R_d = \frac{12}{k_3}\left[\frac{\alpha n}{2\alpha_s n_3\left(\frac{n^2}{n_3^2} - 1\right)}\right]^{\frac{3}{2}}\left[1 - \frac{1}{2}\left(\frac{k_y}{k_3}\right)^2\right]^{-1}$$

$$\cdot \begin{cases} \left(\frac{n_3}{n}\right)^3 \log\left[\frac{kn}{4\alpha}\left(\frac{n^2}{n_3^2} - 1\right)\right] & \text{(for } E_{11}^x \text{ mode).} \\ \log\left[\frac{kn_3^2}{4\alpha n}\left(\frac{n^2}{n_3^2} - 1\right)\right] & \text{(for } E_{11}^y \text{ mode).} \end{cases} \tag{28}$$

Example 1: Consider a thin ribbon guide made of glass surrounded by air and assume that $n = 1.5$, $n_3 = 1$, $\alpha = 0.1$ nepers per m, and $b = \infty$. From equations (26) and (28) we calculate the values in Table I.

It is doubly advantageous to use the E_{11}^x mode rather than the E_{11}^y because (i) the thickness required for equal radiation loss and straight guide loss is roughly $(n/n_3)^3$ times larger, and (ii) R_d is about $(n/n_3)^3$ times smaller.

If the height b of the ribbon is finite, k_y/kn_3 is no longer zero and the radii are, according to equation (28), $[1 - \frac{1}{2}(k_y/k_3)^2]^{-1}$ times longer than those in Table I.

3.2 *Medium Size Guide for Integrated Optical Circuitry*

It is likely that guides for integrated optical circuitry will be possible to fabricate only with $n_1 \cong n_3$. The radiation loss per radian and the Q_c of

TABLE I—VALUES CALCULATED FROM EQUATIONS (26) AND (28)

α_s (nepers/m)	$E_{11}{}^y$ Mode		$E_{11}{}^z$ Mode	
	$\dfrac{a}{\lambda}$	$\dfrac{R_d}{\lambda}$	$\dfrac{a}{\lambda}$	$\dfrac{R_d}{\lambda}$
0.01	0.05	1.9×10^3	0.17	6.3×10^2
0.001	0.016	6.2×10^4	0.055	2×10^4
0.0001	0.005	2×10^6	0.017	6.5×10^5

loops made with these guides can be obtained from Figs. 4 and 7, considering abscissas around $a/A = 1$. For both modes, E_{11}^y and E_{11}^z, most of the power travels within the guiding rod.*

In general, the losses are very sensitive to the radius of curvature. They are also sensitive to the guide's width to the left of the dashed curve $c = 0.5$, but fairly insensitive to the right of it.

Example 2: Let us design a guide:

(*i*) The attenuation per radian resulting from radiation loss is

$$\alpha_c R = 0.01 \text{ nepers} = 0.087 \text{ dB.}$$

(*ii*) Its width a is the maximum compatible with single mode guidance in the infinitely high slab, that is

$$\frac{a}{A} = \frac{2a}{\lambda} (n_1^2 - n_3^2)^{\frac{1}{2}} = 1.$$

(*iii*) We assume $b = \infty$ and $n_3 = n_1(1 - \Delta)$, where $\Delta \ll 1$ and $n_1 = 1.5$.

From Fig. 4 we derive the guide dimensions for different values of Δ:

Δ	$\dfrac{a}{\lambda}$	$\dfrac{R}{\lambda}$
0.1	0.745	30
0.01	2.36	1,060
0.001	7.45	37,000

Unless Δ is 0.01 or larger, the radius of curvature R becomes uncomfortably large for integrated optical circuitry. Furthermore, if b is finite, k_y is no longer zero, and the radii become $[1 - (k_y/k_3)^2]^{-1}$ times larger than those in the table above.

---

* This is not true if $b/B_2 \ll 1$. Then k_{z0} must be calculated from equation (8).

Example 3:   We design a resonant loop (Fig. 1) such that its Q_c resulting from radiation is equal to the Q resulting from transmission loss in typical glass $(n_1 = 1.5, \alpha = 0.1$ neper/m at $\lambda = 1\mu)$; that is,

$$Q = Q_c = 5 \times 10^7.$$

Furthermore, let us assume as in Example 2 that $a/A = 1$, $n_3 = n_1(1 - \Delta)$, and $b = \infty$. With the help of Fig. 7 we derive

Δ	$\dfrac{a}{\lambda}$	$\dfrac{R}{\lambda}$
0.1	0.745	57
0.01	2.36	1,550
0.001	7.45	42,000

Again, unless Δ is larger than 0.01, the radius of curvature becomes unwieldily large for integrated optical circuitry.

Instead of using a loop as the resonant circuit of Fig. 1, it is possible to make $a = R$, and the loop becomes a pillbox (Fig. 12). This structure may be simpler to fabricate. For this case, also from Fig. 4, using the refractive indices of the previous example, we obtain

Δ	$\dfrac{R}{\lambda}$
0.1	42
0.01	1,170
0.001	32,000

The pillbox resonator requires a 30 percent shorter radius than the ring resonator. As before, if b is finite, the radii are $[1 - (k_y/k_3)^2]^{-1}$ times longer than those in the last two tables.
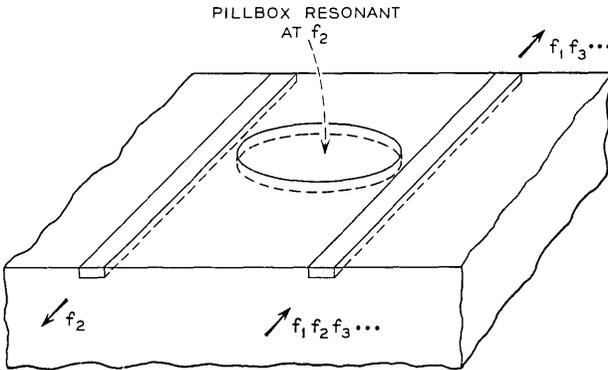
Fig. 12 — Channel dropping filter (pillbox type).

3.3 *Large Guides for Fiber Optics*

The large guide is multimode, $a/A \gg 1$, and the radius for small mode conversion is derived from equations (11) and (12), making $k_{x0}a = \pi$ and $k_{z0} = 2\pi n_1/\lambda$. Then

$$c = \pi n_1^2 \frac{a^3}{\lambda^2 R}.$$

For a power conversion $c^2 = 0.01$, and $n_1 = 1.5$, we have

$\dfrac{a}{\lambda}$	$\dfrac{R}{\lambda}$
5	8,900
10	71,000

The conversion loss is many orders of magnitude larger than the loss radiated by the fundamental mode because of the curvature. Radiation loss of higher order modes can be found in equations (51) and (63).

In general, clad fibers are of circular cross section; consequently our calculations do not strictly apply. Nevertheless, a guide of circular cross section and another of equal area but square cross section must have quite comparable attenuation per radian unless mode degeneracy occurs, but this is quite unlikely.

Though we have been talking throughout of light guides, it is obvious that all the calculations are equally applicable to microwave guides.

IV. CONCLUSIONS

Relations between radiation losses resulting from curvature, geometry, and electric characteristics of the bent dielectric guide are summarized in Figs. 4, 5, and 7 through 10 and they are discussed and exemplified in Section III.

The main qualitative results are that for a given radius of curvature R, the radiation loss can be reduced

(*i*) by increasing the difference between the refractive index n_1 of the guide and those of the media toward the outside, n_3, and inside, n_5, of the curved guide axis (Fig. 2) ;

(*ii*) by increasing the guide width a. Nevertheless, once a is bigger than

$$\left(\frac{R\lambda^2}{\pi n_1^2}\right)^{\frac{1}{3}},$$

(where λ is the free space wavelength), there is little reduction of the loss;

(iii) by choosing the height of the guide large enough to confine the fields as much as possible within the guide in the direction normal to the plane of curvature.

In general, the radiation losses are small if

$$R > \frac{24\pi^2 \mid \xi_3 \mid^3}{\lambda^2},$$

where ξ_3 is the length over which the field decays by $1/e$ in medium 3 (Fig. 2).

Thin ribbons of glass, surrounded by air and oriented as in Fig. 6c, operate better with the electric field perpendicular to the ribbon's plane. Choosing the thickness $a = 0.055\lambda$, the attenuation of the straight guide is 1 percent of the attenuation in glass, and the radius of curvature which doubles that low attenuation is $20,000\lambda$.

The dielectric guide for integrated optical circuitry seems suitable to negotiate bends and to make resonant loops of small radii of curvature and small radiation losses. For example, for

$$n_1 = 1.5$$

$$a = \frac{\lambda}{2n_1\left(1 - \frac{n_3^2}{n_1^2}\right)^{\frac{1}{2}}} \quad \text{(single mode guide)}$$

a 1 percent attenuation (0.087 dB) resulting from radiation in a length of guide equal to R is achieved with the following values

$1 - \dfrac{n_3}{n_1}$	$\dfrac{a}{\lambda}$	$\dfrac{R}{\lambda}$
0.1	0.745	30
0.01	2.36	1060
0.001	7.45	37000

The smaller $n_1 - n_3$, the larger the radius of curvature. For $\lambda = 0.63\mu$, if one wants to keep R below 1 mm, the difference between the internal and external refractive indices must be larger than 0.01.

Large cross section dielectric guides capable of supporting many modes are far more sensitive to mode conversions than to radiation losses. For the fundamental mode, the power conversion loss at the junction between a straight and a curved section of a multimode guide is

$$c^2 = \left(\pi n_1^2 \frac{a^3}{\lambda^2 R}\right)^2 .$$

For $n_1 = 1.5$, $a = 6.3\mu$, and $\lambda = 0.63\mu$, the radius of curvature R that produces a power conversion c^2 of 0.01 is 45 mm. The radiation loss in a length of guide equal to R is many orders of magnitude below 0.01.

APPENDIX

Field Analysis of the Curved Guide

Figure 2 shows the geometry and dielectric distribution of the curved guide. In this appendix two families of modes are found, E_{pq}^x and E_{pq}^y; each is studied separately.

A.1 E_{pq}^x Modes: Polarization Along x

The field components in each region should be written as integral expressions, but, as discussed in Section II, the power propagating through the shaded areas is neglected, and the field matching is performed only along the sides of region 1. Consequently, those field components do not need to be so general. As a matter of fact, the simplest field components in the mth of the five areas are[16]

$$H_{xm} = \frac{1}{k_m^2 - k_{ym}^2} \frac{\partial^2 H_{ym}}{\partial x\, \partial y} ,$$

$$H_{ym} = e^{-i\nu\theta + i\omega t}$$

$$\cdot \begin{cases} M_1 J_\nu[(k_1^2 - k_{\nu1}^2)^{\frac{1}{2}}(R + x) + \psi_1] \cos(k_{\nu1}y + \Omega_1) & \text{for } m = 1 \\ M_2 J_\nu\left[\left[k_2^2 - k_{\nu2}^2\right]^{\frac{1}{2}}(R + x) + \psi_2\right] \exp\left[\mp ik_{\nu2}y\right] & \text{for } m = \begin{matrix}2\\4\end{matrix} \\ M_3 H_\nu^{(2)}[(k_3^2 - k_{\nu3}^2)^{\frac{1}{2}}(R + x)] \cos(k_{\nu3}y + \Omega_3) & \text{for } m = 3 \\ M_5 J_\nu[(k_5^2 - k_{\nu5}^2)^{\frac{1}{2}}(R + x)] \cos(k_{\nu1}y + \Omega_5) & \text{for } m = 5 \end{cases} ,$$

$$H_{zm} = \frac{i}{k_m^2 - k_{ym}^2} \frac{\nu}{R + x} \frac{\partial H_{ym}}{\partial y} ,$$

$$E_{xm} = -\frac{\omega\mu}{k_m^2 - k_{ym}^2} \frac{\nu}{(R + x)} H_{ym} ,$$

$$E_{ym} = 0,$$

$$E_{zm} = \frac{-i\omega\mu}{k_m^2 - k_{ym}^2} \frac{\partial H_{ym}}{\partial x} , \tag{29}$$

in which M_m is the amplitude of the field in the mth medium; ψ_m and Ω_m are constants that locate the field maxima in region m; ω is the angular frequency; ϵn_m^2 and μ, the permittivity and permeability of each medium, are related by $k_m^2 = k^2 n_m^2 = \omega^2 \epsilon \mu n_m^2$; k_{ym} is the propagation constant along y in medium m; and J_ν and $H_\nu^{(2)}$ are Bessel and Hankel functions, respectively.

Strictly speaking, the H_y component in media 1, 2, and 4 should be written as a sum of Bessel functions of the first and second kind, but later on they are approximated by circular functions; therefore, we do not make any mistake using only the Bessel function of the first kind with an arbitrary phase constant in the argument.

We consider only guide geometries for which the guide wavelengths measured in the x and y directions in medium 1 are large compared with the wavelength measured in the z direction. This means that (i)

$$\frac{\partial H_{ym}}{\partial x} \ll \frac{\nu}{R} , \tag{30}$$

and, as a consequence, the field component H_{x1} is very small compared with H_z and is neglected; (ii) the propagating modes are basically of the TEM type.

In order to match the remaining components along the boundaries of medium 1, the field components in media 1, 2, and 4 must have the same dependence along x, while the field components in media 1, 3, and 5 must have the same dependence along y. Therefore

$$k_{y1} = k_{y3} = k_{y5} = k_y , \tag{31}$$

$$k_1^2 - k_y^2 = k_2^2 - k_{y2}^2 = k_4^2 - k_{y4}^2 , \tag{32}$$

$$\psi_1 = \psi_2 = \psi_4 = \psi, \quad \text{and} \quad \Omega_1 = \Omega_3 = \Omega_5 = \Omega. \tag{33}$$

Furthermore, the field matching yields the following four equations from which two characteristic equations will be derived

$$\tan\left(k_y \frac{b}{2} + \Omega\right) = i\frac{k_{y2}}{k_y} , \quad \tan\left(k_y \frac{b}{2} - \Omega\right) = i\frac{k_{y4}}{k_y} \tag{34}$$

$$\frac{J_\nu(\rho_{13})}{J_\nu'(\rho_{13})} = \frac{\rho_3}{\rho_{13}} \frac{H_\nu^{(2)}(\rho_3)}{H_\nu^{(2)'}(\rho_3)} , \quad \text{and} \quad \frac{J_\nu(\rho_{15})}{J_\nu'(\rho_{15})} = \frac{\rho_5}{\rho_{15}} \frac{J_\nu(\rho_5)}{J_\nu'(\rho_5)} \tag{35}$$

where

$$\left. \begin{array}{ll} \rho_{13} = R(k_1^2 - k_y^2)^{\frac{1}{2}} + \psi, & \rho_{15} = (R - a)(k_1^2 - k_y^2)^{\frac{1}{2}} + \psi \\ \rho_3 = R(k_3^2 - k_y^2)^{\frac{1}{2}}, & \text{and} \quad \rho_5 = (R - a)(k_5^2 - k_y^2)^{\frac{1}{2}}. \end{array} \right\} \tag{36}$$

Similar to what happens with the straight guide, equations (34) and (35) are the boundary conditions of two independent problems far simpler than the one depicted in Fig. 2. Thus, for a dielectric slab infinite in the x and z directions and with dimensions and refractive indices as depicted in Fig. 13a, the boundary conditions for modes with no E_y component coincide with equation (34). Similarly, for a bent slab infinite in the y direction as shown in Fig. 13b, the boundary conditions for modes with a negligible H_x component coincide with equation (35).

The elimination of Ω between the two expressions of equation (34) yields the characteristic equation for the plane slab[10]

$$k_y b = q\pi - \tan^{-1} \frac{1}{\left[\left(\dfrac{\pi}{A_2 k_y}\right)^2 - 1\right]^{\frac{1}{2}}} - \tan^{-1} \frac{1}{\left[\left(\dfrac{\pi}{A_4 k_y}\right)^2 - 1\right]^{\frac{1}{2}}} , \quad (37)$$

in which

$$A_{\substack{2\\4}} = \frac{\lambda}{2\left(n_1^2 - n_{\substack{2\\4}}^2\right)^{\frac{1}{2}}} ; \quad (38)$$

the $\tan^{-1}$ functions are to be taken in the first quadrant, and the arbitrary integer q is the order of the mode, that is, the number of extrema of each field component within the guiding rod in the y direction.

The transcendental equation (37) has an approximate closed form solution already found in Ref. 10

$$k_y \cong \frac{q\pi}{b} \left(1 + \frac{A_2 + A_4}{\pi b} + \cdots\right)^{-1} , \quad (39)$$
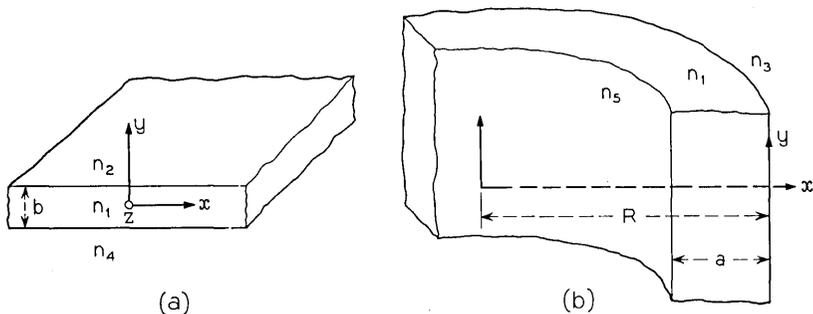
Fig. 13 — Guiding dielectric slabs.

which is valid only when b is so large that

$$\frac{A_2 + A_4}{\pi b} \ll 1 \tag{40}$$

and consequently the parenthesis is close to unity.

The field components in media 2 and 4 decay exponentially by $1/e$ in lengths η_2 and η_4, which are deduced from equation (32) to be

$$\eta_{2 \atop 4} = \frac{1}{\left| k_{y2 \atop 4} \right|} = \frac{1}{\left(k_1^2 - k_{2 \atop 4}^2 - k_y^2 \right)^{\frac{1}{2}}}. \tag{41}$$

Let us consider the solution of the characteristic equation of the bent slab (Fig. 13b). For guided modes, both the arguments and the order of the Bessel and Hankel functions involved in equation (35) are large compared with unity, and consequently they can be replaced by their Watson's first term approximations,[17]

$$
\left.
\begin{aligned}
J_\nu(\rho) &= \left[\frac{2}{\pi(\rho^2 - \nu^2)^{\frac{1}{2}}} \right]^{\frac{1}{2}}
\begin{cases}
\dfrac{1}{2} \exp\left[-\dfrac{(\nu^2 - \rho^2)^{\frac{3}{2}}}{3\nu^2} \right] & \text{for } \nu > \rho \\[2ex]
\sin\left[\dfrac{(\rho^2 - \nu^2)^{\frac{3}{2}}}{3\nu^2} + \dfrac{\pi}{4} \right] & \text{for } \rho > \nu
\end{cases} \\[4ex]
Y_\nu(\rho) &= -\left[\frac{2}{\pi(\rho^2 - \nu^2)^{\frac{1}{2}}} \right]^{\frac{1}{2}}
\begin{cases}
\exp\left[\dfrac{(\nu^2 - \rho^2)^{\frac{3}{2}}}{3\nu^2} \right] & \text{for } \nu > \rho \\[2ex]
\cos\left[\dfrac{(\rho^2 - \nu^2)^{\frac{3}{2}}}{3\nu^2} + \dfrac{\pi}{4} \right] & \text{for } \rho > \nu.
\end{cases}
\end{aligned}
\right\} \tag{42}
$$

These expressions are valid if

$$\frac{\nu^2}{(\rho^2 - \nu^2)^{\frac{3}{2}}} \ll 1. \tag{43}$$

Introducing these approximations for the Bessel functions in both equations (35) and eliminating ψ between them, we obtain the characteristic equation for the bent slab

$$\frac{1}{3\nu^2} \left[(\rho_{13}^2 - \nu^2)^{\frac{3}{2}} - (\rho_{15}^2 - \nu^2)^{\frac{3}{2}} \right]$$

$$= p\pi - \tan^{-1} \left(\frac{n_3^2}{n_1^2} \left[\frac{\rho_{13}^2 - \nu^2}{\nu^2 - \rho_3^2} \right]^{\frac{1}{2}} \left\{ 1 + i \exp\left[-\frac{2}{3} \frac{(\nu^2 - \rho_3^2)^{\frac{3}{2}}}{\nu^2} \right] \right\} \right)$$

$$- \tan^{-1} \frac{n_5^2}{n_1^2} \left(\frac{\rho_{15}^2 - \nu^2}{\nu^2 - \rho_5^2} \right)^{\frac{1}{2}}, \tag{44}$$

in which p is an arbitrary integer bigger than zero which determines the order of the mode in the x direction, and the $\tan^{-1}$ functions are to be taken in the first quadrant.

Let us rewrite this equation substituting ρ_3, ρ_5, ρ_{13}, and ρ_{15} by the values given in equation (36); furthermore, let

$$A_{\substack{3 \\ 5}} = \frac{\lambda}{2\left(n_1^2 - n_{\substack{3 \\ 5}}^2\right)^{\frac{1}{2}}}, \tag{45}$$

$$\nu = k_z R \tag{46}$$

and

$$k_z = (k_1^2 - k_y^2 - k_x^2)^{\frac{1}{2}}. \tag{47}$$

Because of these two last definitions, k_z, k_x, and k_y are the axial and the transverse propagation constants at $x = 0$. The characteristic equation (44) then becomes

$$\frac{R k_x^3}{3 k_z^2}\left[1 - \left(1 - \frac{2a k_z^2}{k_z^2 R}\right)^{\frac{3}{2}}\right]$$

$$= p\pi - \tan^{-1}\frac{n_3^2}{n_1^2}\frac{1 + i\,\exp\left\{-\frac{2}{3}\frac{\pi^3 R}{k_z^2 A_3^3}\left[1 - \left(\frac{k_z A_3}{\pi}\right)^2\right]^{\frac{3}{2}}\right\}}{\left[\left(\frac{\pi}{k_x A_3}\right)^2 - 1\right]^{\frac{1}{2}}}$$

$$- \tan^{-1}\frac{n_5^2}{n_1^2}\left[\frac{\left[1 - \frac{a}{R}\right]^2 (k_1^2 - k_y^2) - k_z^2}{k_z^2 - \left(1 - \frac{a}{R}\right)^2 (k_5^2 - k_y^2)}\right]^{\frac{1}{2}}. \tag{48}$$

To solve this equation for k_x we expand the left side and the second $\tan^{-1}$ in powers of $1/R$ and the first $\tan^{-1}$ in powers of the exponential. Assuming R is large and keeping the first term of each perturbation calculation, the solution of equation (48) is

$$k_x = k_{x0}\left(1 + \frac{2c}{a k_{x0}} - i\,\frac{k_{z0}\alpha_c}{k_{x0}^2}\right), \tag{49}$$

where

$$c = \frac{1}{2 k_{x0} a}\left(\frac{\pi a}{A_3}\right)^3 \frac{1}{\mathfrak{R}}\frac{1 + 2F_5}{1 + F_3 + F_5} \tag{50}$$

and

$$\alpha_c = \frac{k_{x0}^2}{k_{z0}} \left[1 - \left(\frac{k_{x0} A_3}{\pi} \right)^2 \right] F_3 \frac{\exp \left\{ -\frac{\Re}{3} \left[1 - \left(\frac{k_{x0} A_3}{\pi} \right)^2 \left(1 + \frac{2c}{ak_{x0}} \right)^2 \right]^{\frac{3}{2}} \right\}}{1 + F_3 + F_5},$$

(51)

in which

$$F_3 = \left[\frac{n_3}{n_1} \right]^2 \frac{\frac{A_3}{5}}{\pi a \left[1 - \left[\frac{k_{x0} A_3}{\pi} \right]^2 \right]^{\frac{1}{2}}} \frac{1}{1 - \left[1 - \frac{n_3^2}{n_1^2} \right] \left[\frac{k_{x0} A_3}{\pi} \right]^2}, \quad (52)$$

$$\Re = \frac{2\pi^3 R}{k_{z0}^2 A_3^3} = 2(n_1^2 - n_3^2)^{\frac{3}{2}} \frac{k^3 R}{k_{z0}^2}, \tag{53}$$

$$k_{z0} = (k_1^2 - k_y^2 - k_{x0}^2)^{\frac{1}{2}}, \tag{54}$$

and k_{x0} is the solution of the equation

$$k_{x0} a = p\pi - \tan^{-1} \frac{n_3^2}{n_1^2} \frac{1}{\left[\left(\frac{\pi}{k_{x0} A_3} \right)^2 - 1 \right]^{\frac{1}{2}}} - \tan^{-1} \frac{n_5^2}{n_1^2} \frac{1}{\left[\left(\frac{\pi}{k_{x0} A_5} \right)^2 - 1 \right]^{\frac{1}{2}}}.$$

(55)

This is the physical interpretation of equation (49): the transverse propagation constant k_x measured at $x = 0$ is made of three terms. The first term, k_{x0}, is the transverse propagation constant of the guide without curvature; the second and third terms are perturbations related to the change of field profile and radiation introduced by the curvature. It is easy to find that c^2 is the mode conversion loss that would exist at a junction between a straight guide and a curved one, and α_c is the attenuation coefficient of the curved guide.

The field components in media 3 and 5 decay almost exponentially away from the guide. The length ξ_3, over which the intensity in medium 3 decays by $1/e$, is derived as in equation (41) to be

$$\xi_3 = \frac{1}{|k_{x3}|} = \frac{1}{(k_1^2 - k_3^2 - |k_x^2|)^{\frac{1}{2}}} \tag{56}$$

and only approximately

$$\xi_5 = \frac{1}{|k_{x5}|} = \frac{1}{(k_1^2 - k_5^2 - |k_x^2|)^{\frac{1}{2}}}. \tag{57}$$

All these equations have been derived under the assumption that inequality (43) is satisfied; this means that the field configuration of the curved guide is very close to that of the straight guide. In other words, $c \ll 1$. For a given R, if one chooses the width a of the guide large enough, these inequalities are not satisfied, the previous results are no longer applicable, and a new solution is needed. We proceed to find it.

Let us assume as a limiting case that in Fig. 2

$$a = R. \tag{58}$$

The characteristic equation derived from the first equation of (35), making $\psi = 0$, is

$$\frac{(\rho_{13}^2 - \nu^2)^{\frac{3}{2}}}{3\nu^2} = (p - \tfrac{1}{4})\pi - \tan^{-1} \frac{n_3^2}{n_1^2} \left(\frac{\rho_{13}^2 - \nu^2}{\nu^2 - \rho_3^2}\right)^{\frac{1}{2}}$$
$$\cdot \left\{1 + i \exp\left[-\frac{2}{3} \frac{(\nu^2 - \rho_3^2)^{\frac{3}{2}}}{\nu^2}\right]\right\}. \tag{59}$$

Following similar steps to those taken to solve equation (44), we substitute ρ_{13}, ρ_3, and ν by the values given in equations (36) and (46); we obtain

$$\frac{R(k_x')^3}{3(k_z')^2} = (p - \tfrac{1}{4})\pi$$
$$- \tan^{-1} \frac{n_3^2}{n_1^2} \cdot \frac{1 + i \exp\left\{-\frac{2}{3} \frac{\pi^3 R}{(k_z')^2 A_3^3}\left[1 - \left(\frac{k_x' A_3}{\pi}\right)^2\right]^{\frac{3}{2}}\right\}}{\left[\left(\frac{\pi}{k_x' A_3}\right)^2 - 1\right]^{\frac{1}{2}}} \tag{60}$$

The primes distinguish the symbols from those used previously.

To solve this equation we notice that for small losses it must be that

$$\frac{k_x' A_3}{\pi} \ll 1. \tag{61}$$

Therefore, the $\tan^{-1}$ can be replaced by its argument and the approximate solution of equation (60) is

$$k_x' = k_{x0}'\left[1 - i \frac{k_{z0}'\alpha_c}{(k_{x0}')^2}\right], \tag{62}$$

where

$$\alpha_c = \frac{n_3^2}{n_1^2} \frac{k_{z0}'}{kR(n_1^2 - n_3^2)^{\frac{1}{2}}}$$

$$\cdot \exp\left(-\frac{\Re'}{3}\left\{1 - \left[\frac{6\pi(p - \frac{1}{4})}{\Re'}\right]^{\frac{2}{3}}\left[1 - \frac{2}{3}\frac{n_3^2}{n_1^2}\left(\frac{6}{\pi^2(p - \frac{1}{4})^2\Re'}\right)^{\frac{1}{3}}\right]\right\}^{\frac{3}{2}}\right),$$

$$(63)$$

$$k_{z0}' = [k_1^2 - k_y^2 - (k_{x0}')^2]^{\frac{1}{2}},$$

$$(64)$$

$$k_{x0}' = \frac{\pi}{A_3}\left[\frac{6\pi(p - \frac{1}{4})}{\Re'}\right]^{\frac{1}{3}}\left\{1 - \frac{1}{3}\frac{n_3^2}{n_1^2}\left[\frac{6}{\pi^2(p - \frac{1}{4})^2\Re'}\right]^{\frac{1}{3}}\right\},$$

$$(65)$$

and

$$\Re' = \frac{2\pi^3 R}{(k_{z0}')^2 A_3^3} = 2(n_1^2 - n_3^2)^{\frac{3}{2}}\frac{k^3 R}{(k_{z0}')^2}.$$

$$(66)$$

The field components outside the guide decay to $1/e$ in a length

$$\xi_3' = \frac{1}{|k_{x3}'|} = \frac{1}{[k_1^2 - k_3^2 - (k_{x0}')^2]^{\frac{1}{2}}}.$$

$$(67)$$

A.2 E_{pq}^y *Modes: Polarization Along y*

The field components and propagation constants can be derived from those in Section A.1 by changing E into H, μ into $-\epsilon$, and vice versa. Except for their polarizations, the E_{pq}^x and E_{pq}^y modes are very similar.

The formulas equivalent to equations (37) and (41) are

$$k_y''b = q\pi - \tan^{-1}\frac{n_2^2}{n_1^2}\frac{1}{\left[\left(\frac{\pi}{A_2 k_y''}\right)^2 - 1\right]^{\frac{1}{2}}} - \tan^{-1}\frac{n_4^2}{n_1^2}\frac{1}{\left[\left(\frac{\pi}{A_4 k_y''}\right)^2 - 1\right]^{\frac{1}{2}}}$$

$$(68)$$

$$\eta_2'' = \frac{1}{|k_{y2}''|}\frac{1}{\left(k_1^2 - k_2^2 - (k_y'')^2\right)^{\frac{1}{2}}}.$$

$$(69)$$

The double prime distinguish these symbols from those used before.

The equivalent formula to any of those between equation (45) and (67) can be derived from that formula by substituting the ratio of refractive indexes by unity, but leaving the differences between squares of indexes unchanged. For example, the formula equivalent to equation (52) for E_{pq}^y modes is

$$F_{3}'' = \frac{\frac{A_3}{5}}{\pi a\left[1 - \left[\frac{k_{x0}'' A_3}{\pi}\right]^2\right]^{\frac{1}{2}}}\frac{1}{1 - \left[1 - \frac{n_3^2}{n_1^2}\right]\left[\frac{k_{x0} A_3}{\pi}\right]^2}.$$

$$(70)$$

REFERENCES

1. Miller, S. E., U. S. Patent 3434774, applied for February 2, 1965 granted March 25, 1969.
2. Karbowiak, A. E., "New Type of Waveguide for Light and Infrared Waves," Elec. Letters, *1*, No. 2 (April 1965), p. 47.
3. Wolff, P. A., unpublished work.
4. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., this issue, pp. 2059–2069.
5. Kaplan, R. A., "Optical Waveguide of Macroscopic Dimension in Single-Mode Operation," Proc. IEEE, *51*, No. 8 (August 1963), p. 1144.
6. Schineller, E. R., "Summary of the Development of Optical Waveguides and Components," Wheeler Laboratories, Report #1471, April 1967.
7. Kapany, N. S., *Fiber Optics,* New York: Academic Press, 1967.
8. Stratton, J. A., *Electromagnetic Theory,* New York: McGraw-Hill, 1941, pp. 524–527.
9. Snitzer, E., "Cylindrical Dielectric Waveguide Modes," J. Opt. Soc. Amer., *51*, No. 5 (May 1961), pp. 491–498.
10. Marcatili, E. A. J., "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," B.S.T.J., this issue, pp. 2071–2102.
11. Schlosser, W., and Unger, H. G., "Partially Filled Waveguides and Surface Waveguides of Rectangular Cross-Section," *Advances in Microwaves,* New York: Academic Press, 1966, pp. 319–387.
12. Bracey, M. F., Cullen, A. L., Gillespie, E. F. F., and Staniforth, J. A., "Surface-Wave Research in Sheffield," IRE Trans. Antennas and Propagation, *AP7* (December 1959), pp. S219–S225.
13. Goell, J. E., "A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides," B.S.T.J., this issue, pp. 2133–2160.
14. Jones, A. L., "Coupling of Optical Fibers and Scattering in Fibers," J. Opt. Soc. Amer., *55*, No. 3 (March 1965), pp. 261–271.
15. Marcatili, E. A. J., and Miller, S. E., "Improved Relations Describing Directional Control in Electromagnetic Wave Guidance," B.S.T.J., this issue, pp. 2161–2188.
16. Stratton, J. A., *Electromagnetic Theory,* New York: McGraw-Hill, 1941, pp. 361.
17. Magnus, W., Oberhettinger, F., and Soni, R. P., *Formulas and Theorems for the Special Functions of Mathematical Physics,* New York: Springer-Verlag, 1966, p. 144.

A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides

By J. E. GOELL

This paper describes a computer analysis of the propagating modes of a rectangular dielectric waveguide. The analysis is based on an expansion of the electromagnetic field in terms of a series of circular harmonics, that is, Bessel and modified Bessel functions multiplied by trigonometric functions. The electric and magnetic fields inside the waveguide core are matched to those outside the core at appropriate points on the boundary to yield equations which are then solved on a computer for the propagation constants and field configurations of the various modes.

The paper presents the results of the computations in the form of curves of the propagation constants and as computer generated mode patterns. The propagation curves are presented in a form which makes them refractive-index independent as long as the difference of the index of the core and the surrounding medium is small, the case which applies to integrated optics. In addition to those for small index difference, it also gives results for larger index differences such as might be encountered for microwave applications.

I. INTRODUCTION

It is anticipated that dielectric waveguides will be used as the fundamental building blocks of integrated optical circuits. These waveguides can serve not only as a transmission medium to confine and direct optical signals, but also as the basis for circuits such as filters and directional couplers.[1] Thus, it is important to have a thorough knowledge of the properties of their modes.

Circular dielectric waveguides have received considerable attention because circular geometry is commonly used in fiber optics.[2-5] In many integrated optics applications it is expected that waveguides will consist of a rectangular, or near rectangular, dielectric core embedded in a dielectric medium of slightly lower refractive index. The modes

for this geometry are more difficult to analyze than those of the metallic rectangular waveguide because of the nature of the boundary.

Marcatili, using approximations based on the assumption that most of the power flow is confined to the waveguide core, has derived in closed form the properties of a rectangular dielectric waveguide.[6] In his solution, fields with sinusoidal variation in the core are matched to exponentially decaying fields in the external medium. In each region only a single mode is used. The results of this method are obtained in a relatively simple form for numerical evaluation.

The properties of the principal mode of the rectangular dielectric waveguide have been studied by Schlosser and Unger using a high-speed digital computer.[7] In their approach the transverse plane was divided into regions, as shown in Fig. 1, and rectangular coordinate solutions assumed in each of the regions. The longitudinal propagation constant was then adjusted so that a field match could be achieved at discrete points along the boundary. This method gives results which, theoretically, are valid over a wider range than Marcatili's, but with a significant increase in computational difficulty. One shortcoming of the method is that for a given mode, as the wavelength increases the field extent increases, so, in the limit it becomes increasingly difficult to match the fields along the boundaries between regions [1] and [2] and between regions [2] and [3].

A variational approach has been undertaken by Shaw and others.[8] They assume a test solution with two or three variable parameters in the core. From this test solution, the fields outside the core are then derived and the parameters are varied to achieve a consistent
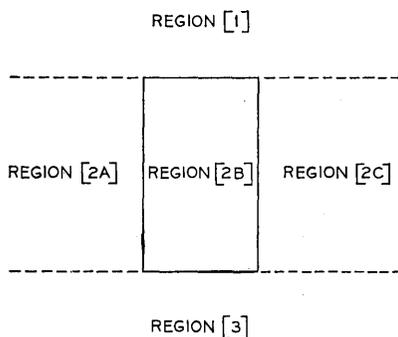
Fig. 1 — Matching boundaries for rectangular mode analysis.

solution. This approach, like that of Schlosser, requires involved computations. Also, it has the disadvantage that the test function must be assumed in advance. In addition, some of his preliminary results do not show the proper behavior for the limiting cases (waveguide dimensions which are very large or very small compared with the wavelength).

In the present analysis the radial variation of the longitudinal electric and magnetic fields of the modes are represented by a sum of Bessel functions inside the waveguide core and by a sum of modified Bessel functions outside the waveguide core. Solutions are found by matching the fields along the perimeter of the core. Thus, the matching boundary is not a function of the waveguide parameters, so the computational complexity does not increase with wavelength.

Section II discusses the underlying theory of the circular-harmonic analysis of rectangular dielectric waveguides. This is followed by a description of computational techniques and special graphical methods of presentation used. Section III is divided into three parts, the first describing the accuracy of the computations, the second describing field patterns, and the third presenting propagation curves.

II. DERIVATION OF EQUATIONS

The waveguide considered here consists of a rectangular core of dielectric constant, ϵ_1, surrounded by an infinite medium of dielectric constant, ϵ_0. Both media are assumed to be isotropic, and have the permeability of free space, μ_0. Figure 2 shows the coordinate systems (rectangular and cylindrical) and rod dimension used in this paper. The direction of propagation is in the $+z$ direction (towards the observer).

In cylindrical coordinates the field solutions of Maxwell's equations take the form of Bessel functions and modified Bessel functions multiplied by trigometric functions, and their derivatives. In order for propagation to take place in the z direction, the field solutions must be Bessel functions in the core and modified Bessel functions outside. Since Bessel functions of the second kind have a pole at the origin and modified Bessel functions of the first kind a pole at infinity, the radial variation of the fields is assumed to be a sum of Bessel functions of the first kind and their derivatives inside the core and a sum of modified Bessel functions and their derivatives outside the core.

In cylindrical coordinates, the z components of the electric and magnetic fields are given by

$$E_{z1} = \sum_{n=0}^{\infty} a_n J_n(hr) \sin (n\theta + \varphi_n) \exp [i(k_z z - \omega t)] \qquad (1a)$$

and

$$H_{z1} = \sum_{n=0}^{\infty} b_n J_n(hr) \sin (n\theta + \psi_n) \exp [i(k_z z - \omega t)] \qquad (1b)$$

inside the core, and by

$$E_{z0} = \sum_{n=0}^{\infty} c_n K_n(pr) \sin (n\theta + \varphi_n) \exp [i(k_z z - \omega t)] \qquad (1c)$$

and

$$H_{z0} = \sum_{n=0}^{\infty} d_n K_n(pr) \sin (n\theta + \psi_n) \exp [i(k_z z - \omega t)] \qquad (1d)$$

outside the core, where ω is the radian frequency and k_z the longitudinal propagation constant. The transverse propagation constants are given by

$$h = (k_1^2 - k_z^2)^{\frac{1}{2}} \qquad (2a)$$

and

$$p = (k_z^2 - k_0^2)^{\frac{1}{2}} \qquad (2b)$$

where $k_1 = \omega(\mu_0 \epsilon_1)^{\frac{1}{2}}$ and $k_0 = \omega(\mu_0 \epsilon_0)^{\frac{1}{2}}$. The terms J_n and K_n are the nth order Bessel functions and modified Bessel functions, respectively, and ψ_n and φ_n are arbitrary phase angles.

The transverse components of the fields are given by[9]

$$E_r = \frac{ik_z}{k^2 - k_z^2} \left[\frac{\partial E_z}{\partial r} + \left(\frac{\mu_0 \omega}{k_z r} \right) \frac{\partial H_z}{\partial \theta} \right] \qquad (3a)$$
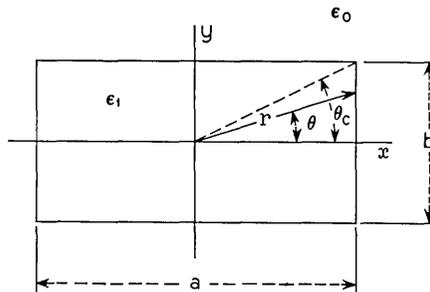
Fig. 2 — Dimensions and coordinate system.

$$E_\theta = \frac{ik_z}{k^2 - k_z^2} \left[\frac{1}{r} \frac{\partial E_z}{\partial \theta} - \left(\frac{\mu_0 \omega}{k_z} \right) \frac{\partial H_z}{\partial r} \right] \tag{3b}$$

$$H_r = \frac{ik_z}{k^2 - k_z^2} \left[-\left(\frac{k^2}{\mu_0 \omega k_z r} \right) \frac{\partial E_z}{\partial \theta} + \frac{\partial H_z}{\partial r} \right] \tag{3c}$$

$$H_\theta = \frac{ik_z}{k^2 - k_z^2} \left[\left(\frac{k^2}{\mu_0 \omega k_z} \right) \frac{\partial E_z}{\partial r} + \frac{1}{r} \frac{\partial H_z}{\partial \theta} \right], \tag{3d}$$

where k can be either k_1 or k_0.

Finally, the component of the electric field tangent to the rectangular core is given by

$$E_t = \pm (E_r \sin \theta + E_\theta \cos \theta) \quad \begin{aligned} -\theta_c < \theta < \theta_c \\ \pi - \theta_c < \theta < \pi + \theta_c \end{aligned} \tag{4a}$$

or

$$E_t = \pm (-E_r \cos \theta + E_\theta \sin \theta), \quad \begin{aligned} \theta_c < \theta < \pi - \theta_c \\ \pi + \theta_c < \theta < -\theta_c \end{aligned} \tag{4b}$$

where θ_c is the angle which a radial line to the corner in the first quadrant makes with the x axis. Similar expressions exist for the tangential magnetic field.

2.1 *Effects of Symmetry*

Since the waveguide is symmetrical about the x axis the fields must be either symmetric or antisymmetric about this axis. This is true because the structure is invarient under 180° rotations and therefore the field patterns must be invarient under a 180° rotation, except for sign. From this and the fact that $\partial/\partial \theta$ appears in each of equations (3), it is evident that two types of modes must exist, the first type with $\varphi_n = 0$ and $\psi_n = \pi/2$ and the second type with $\varphi_n = \pi/2$ and $\psi_n = \pi$.

Similarly, the field functions must also be symmetric or antisymmetric about the y axis. Suppose, for example, E_{z0} exhibits a sinusoidal angular dependence about $\theta = (E_{z0}$ is odd about the x axis). Then, letting $\alpha = \theta - \pi/2$, equation (1c) can be put in the form

$$E_{z0} = \sum_{n=0}^{\infty} c_n K_n(pr)(\sin n\alpha \cos n\pi/2 + \cos n\alpha \sin n\pi/2). \tag{5}$$

For E_{z0} to be purely symmetric about $\alpha = 0$ (the y axis), all n must be odd; for E_{z0} to be antisymmetric about $\alpha = 0$ all n must be even.

Since similar results apply for cosinusoidal variation of E_{z0} about $\theta = 0$, and all other field functions as well, any given mode must consist of either even harmonics or odd harmonics.

From the preceeding analysis it is evident that if the matching points are selected symmetrically about both the x and y axes, then, except possibly for sign, every point will have an equivalent point in each quadrant. Therefore, the field matching need only be performed in one quadrant. Thus, the use of the symmetry of the structure not only reduces the number of constants required to calculate the properties of a given mode by a factor of four, it also decreases the number of points to achieve a given degree of accuracy by the same factor.

2.2 Selection of Matching Points

As mentioned in Section 2.1, the matching point locations should be symmetrical about the x and y axes. For the odd harmonic cases, the points used to compute the results to be presented in Section III were $\theta_m = (m - 1/2)\pi/2N$; $m = 1, \cdots, N$, where N was the number of space harmonics.

The choice of points for the even harmonic cases was more complicated since simultaneous existence of an $n = 0$ harmonic for both the TE and TM circular modes is inconsistent with the waveguide symmetries. Thus, if the maximum n for both the TE and TM solutions are equal, the total number of coefficients to be found will be $4N - 2$ rather than $4N$ as in the previous case.

The method of choosing points for the even harmonic modes used for the computation of the results of Section III was to pick the points for the field components with even symmetry about $\theta = 0$ to be $\theta_m = (m - 1/2)\pi/2N$; $m = 1, 2, \cdots, N$, and for the field components with odd symmetry about $\theta = 0$ to be $\theta_m = (m - N - 1/2)\pi/2(N - 1)$; $m = N + 1, N + 2, \cdots, (2N - 1)$ for cases with unity aspect ratio, $(a/b = 1)$. For aspect ratios other than unity, all points were chosen according to the first formula, except that the first and last points for the odd z component were omitted.

2.3 Formulation of Matrix Elements

The coefficients of equation (1) were found by matching the tangential electric and magnetic fields along the boundary of the waveguide core. Since each type of field consists of both longitudinal and transverse components, four types of matching equations exist.

To facilitate computer analysis the matching equations were put in

matrix form. The matching equations in matrix form for the longitudinal field components are

$$E^{LA}A = E^{LC}C \tag{6a}$$

for the electric field and

$$H^{LB}B = H^{LD}D \tag{6b}$$

for the magnetic field. For the transverse fields the matrix matching equations are given by

$$E^{TA}A + E^{TB}B = E^{TC}C + E^{TD}D \tag{6c}$$

for the electric field and

$$H^{TA}A + H^{TB}B = H^{TC}C + H^{TD}D \tag{6d}$$

for the magnetic field. The A, B, C, and D matrices are N element column matrices of the a_n, b_n, c_n, and d_n mode coefficients, respectively. The elements of the $m \times n$ matrices E^{LA}, E^{LC}, H^{LB}, H^{LD}, E^{TA}, E^{TB}, E^{TC}, E^{TD}, H^{TA}, H^{TB}, H^{TC}, and H^{TD} are given by

$$e_{mn}^{LA} = JS, \tag{7a}$$

$$e_{mn}^{LC} = KS, \tag{7b}$$

$$h_{mn}^{LB} = JC, \tag{7c}$$

$$h_{mn}^{LD} = KC, \tag{7d}$$

$$e_{mn}^{TA} = -k_z(\mathbf{J}'SR + \mathbf{J}CT), \tag{7e}$$

$$e_{mn}^{TB} = k_0 Z_0(\mathbf{J}SR + \mathbf{J}'CT), \tag{7f}$$

$$e_{mn}^{TC} = k_z(\mathbf{K}'SR + \mathbf{K}CT), \tag{7g}$$

$$e_{mn}^{TD} = -k_0 Z_0(\mathbf{K}SR + \mathbf{K}'CT), \tag{7h}$$

$$h_{mn}^{TA} = \epsilon_r k_0(\mathbf{J}CR - \mathbf{J}'ST)/Z_0 , \tag{7i}$$

$$h_{mn}^{TB} = -k_z(\mathbf{J}'CR - \mathbf{J}ST), \tag{7j}$$

$$h_{mn}^{TC} = -k_0(\mathbf{K}CR - \mathbf{K}'ST)/Z_0 , \tag{7k}$$

$$h_{mn}^{TD} = k_z(\mathbf{K}'CR - \mathbf{K}ST), \tag{7l}$$

where

$$Z_0 = (\mu_0/\epsilon_0)^{\frac{1}{2}},$$

$$\epsilon_r = \epsilon_1/\epsilon_0 ,$$

$$S = \sin(n\theta_m + \varphi) \Big\rbrace \quad \text{or} \quad \begin{array}{l} \varphi = 0 \\ \\ \varphi = \pi/2 \end{array},$$

$$C = \cos(n\theta_m + \varphi)$$

$$J = J_n(hr_m), \qquad K = K_n(pr_m),$$

$$J' = J'_n(hr_m), \qquad K' = K'_n(pr_m),$$

$$\mathbf{J} = \frac{nJ_n(hr_m)}{h^2 r_m}, \qquad \mathbf{K} = \frac{nK_n(pr_m)}{p^2 r_m},$$

$$\mathbf{J'} = \frac{J'_n(hr_m)}{h}, \qquad \mathbf{K'} = \frac{K'_n(pr_m)}{p},$$

and

$$\left. \begin{array}{l} R = \sin\theta_m \\ \\ T = \cos\theta_m \\ \\ r_m = (a/2)\cos\theta_m \end{array} \right\rbrace \theta < \theta_c, \qquad \left. \begin{array}{l} R = -\cos\theta_m \\ \\ T = \sin\theta_m \\ \\ r_m = (b/2)\sin\theta_m \end{array} \right\rbrace \theta > \theta_c.$$

For $\theta = \theta_c$, the boundary at the corner was assumed to be perpendicular to the radial line connecting it to the origin, so for this case $R = \cos(\theta_m + \pi/4)$, $T = \cos(\theta_m - \pi/4)$, and $r_m = (a^2 + b^2)^{\frac{1}{2}}/4$.

2.4 Mode Designation

Unlike metallic waveguides, the field patterns of dielectric waveguides are sensitive to refractive index difference, wavelength, and aspect ratio. This complicates the problem of finding a reasonably descriptive mode designation scheme.

For rectangular metallic waveguides, the accepted approach is to designate the modes as TE (or H) and TM (or E), and to specify the number of field maxima in the x and y directions with a double subscript. When there is no variation the subscript 0 is used.

Since the rectangular dielectric waveguide modes are neither pure TE nor pure TM, a different scheme must be used. The scheme adopted is based on the fact that in the limit, for large aspect ratio, short wavelength, and small refractive index difference, the transverse electric field is primarily parallel to one of the transverse axes. Modes are designated as E^y_{mn} if in the limit their electric field is parallel to the y axis and as E^x_{mn} if in the limit their electric field is parallel to the x axis. The m and n subscript are used to designate the number of maxima in the x and y directions, respectively.[†]

[†] This scheme agrees with that used by Marcatili in Ref. 6.

2.5 *Electric and Magnetic Field Function Differences*

For a hollow metallic waveguide where pure TE and TM modes can exist, it is evident from equation (3) that E_r and H_θ have similar transverse variations as do E_θ and H_r, so that the impedance is independent of position. Furthermore, the transverse electric and magnetic fields are perpendicular and the power flow, Re $\{E \times H^*\}$, does not change sign anywhere across the waveguide.

By examination of equation (3), it is clear that for the mixed modes of the dielectric waveguide, the field functions are not similar and the impedance is a function of position. In order for the transverse fields E_t and H_t to be perpendicular,

$$E_t \cdot H_t = E_r H_r + E_\theta H_\theta = 0. \tag{8}$$

Now, from equation (3)

$$E_t \cdot H_t = \frac{k_z^2 - k^2}{k_z^2} \left(\frac{\partial H_z}{\partial r} \frac{\partial E_z}{\partial r} + \frac{1}{r^2} \frac{\partial H_z}{\partial \theta} \frac{\partial E_z}{\partial \theta} \right). \tag{9}$$

Thus, E_t and H_t are not necessarily perpendicular. Finally, since the transverse variations of E_t and H_t are not the same, the electric field and magnetic field can change sign at different points, which results in negative power flow.[†]

Three special cases exist where the electric and magnetic fields, and the impedance, have the same positional dependence, and where the power flow does not change sign across the waveguide:

(*i*) in one of the regions if the propagation constant is approximately equal to the bulk propagation constant of that region, that is, if $k \approx k_1$ or $k \approx k_0$,

(*ii*) everywhere in the limit for small refractive index difference, since case *i* will then hold in both regions, and

(*iii*) everywhere for circular symmetry of both the structure and the modes.

2.6 *Normalization*

The arguments of the Bessel and modified Bessel functions are given by $hr = (k_1^2 - k_z^2)^{\frac{1}{2}}r$ and $pr = (k_z^2 - k_0^2)^{\frac{1}{2}}r$, respectively. The first argument can be put in the form

$$hr = [k_1^2 - k_0^2 - p^2]^{\frac{1}{2}}r. \tag{10}$$

[†] This unusual property has also been observed for helices.[10] Presumably, if loss were included there would be a radial component of power to feed the reverse flow, and the lossless case can be thought of as the limit of the lossy case.

Letting

$$\mathcal{P}^2 = \frac{(k_z/k_0)^2 - 1}{n_r^2 - 1}, \tag{11}$$

and

$$\mathcal{R} = rk_0(n_r^2 - 1)^{\frac{1}{2}}, \tag{12}$$

where

$$n_r = (k_1/k_0)^{\frac{1}{2}} \tag{13}$$

is the index of refraction of the core relative to the outer medium, gives

$$pr = \mathcal{P}\mathcal{R} \tag{14}$$

and

$$hr = \mathcal{R}(1 - \mathcal{P}^2)^{\frac{1}{2}}. \tag{15}$$

The curves of the propagation constant given in Section III are drawn in terms of $\mathcal{P}^2$ and $\mathcal{B}$, where

$$\mathcal{B} = \frac{2b}{\lambda_0}(n_r^2 - 1)^{\frac{1}{2}} \tag{16}$$

and $\lambda_0 = 2\pi/k_0$. Since $\mathcal{R}$ is proportional to $1/(n_r^2 - 1)^{\frac{1}{2}}$ and $\mathcal{P}$ and $\mathcal{B}$ are proportional to $(n_r^2 - 1)^{\frac{1}{2}}$, the use of $\mathcal{P}^2$ and $\mathcal{B}$ as plotting variables eliminates the explicit dependence of the Bessel and modified Bessel function arguments on the refractive indices of the media.

Examination of the matching equations, equations (6), reveals that ϵ_r appears in the H^{TA} term. However, since ϵ_r appears as a multiplicative factor in H^{TA}, for sufficiently small values the normalized propagation constant, $\mathcal{P}^2$, is independent of ϵ_r.

The normalized propagation constant, $\mathcal{P}^2$, has two additional properties which make its use convenient. First, its range of variation is on the interval $(0, 1)$. Second, for $n_r \approx 1$,

$$\mathcal{P}^2 \approx \frac{k_z/k_0 - 1}{\Delta n_r}, \tag{17}$$

where $\Delta n_r = n_r - 1$; so for small n_r, $\mathcal{P}^2$ is proportional to $k_z - k_0$. The latter property is the reason that $\mathcal{P}^2$ rather than $\mathcal{P}$ was used as a plotting variable.

2.7 *Method of Computation*

2.7.1 *Propagation Constant*

Equation (6) yields $4N$ simultaneous homogeneous linear equations for the a_n, b_n, c_n, and d_n for the odd modes and $4N$-2 equations for

the even modes, using the matching points previously described. The equations can be combined to form a single matrix equation

$$[Q][T] = 0, \tag{18}$$

where

$$Q = \begin{bmatrix} E^{LA} & 0 & -E^{LC} & 0 \\ 0 & H^{LB} & 0 & -H^{LD} \\ E^{TA} & E^{TB} & -E^{TC} & -E^{TD} \\ H^{TA} & H^{TB} & -H^{TC} & -H^{TD} \end{bmatrix}$$

and the column matrix

$$[T] = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}.$$

All of the quantities in the matrices $[Q]$ and $[T]$ are themselves matrices as defined by equations (1), (6), and (7).

In order for a nontrivial solution to equation (18) to exist

$$\text{Det } [Q] = 0. \tag{19}$$

The normalized propagation constant, $\mathcal{P}^2$, was found by substituting test values into equation (19). First, values of $\mathcal{P}^2$ evenly distributed in the interval $(0, 1)$ were substituted to crudely locate the roots. Then, Newton's method was used to find the roots to the desired accuracy.[11] Generally, one Newton approximation was used to find $\mathcal{P}^2$ for the propagation curves and about ten Newton's approximations when $\mathcal{P}^2$ was to be used to calculate field plots.

Both the simple method of triangulation[12] and the more complicated Gauss pivotal condensation method[13] were used to evaluate the determinant, the former for almost all cases and the latter for a few cases when roundoff error was apparent because the value of the determinant was not a smooth function of $\mathcal{P}^2$. In all cases double precision arithmetic was used. For five space harmonics, about 0.1 second of IBM 360/65 computing time was required for each value of $\mathcal{P}^2$ to evaluate the determinant using the triangulation method.

Due to the wide dynamic range of the coefficients, steps had to be taken to prevent underflow and overflow of the computer and to re-

duce the effects of roundoff. Multiplying a row or column of the matrix by a finite constant is equivalent to multiplying the determinant by that constant. Thus, any row or column of the determinant can be multiplied by a positive function without shifting its zeroes.

A detailed theory giving the "best functions" can be derived. However, since a "brute force" method was used, the more sophisticated method, which was not used because it would have required a substantial increase in the complexity of the program logic, is not discussed. It was found that multiplying the Bessel function terms by $h^2 d/|J_n(hb)|$ and the modified Bessel function terms by $p^2 d/k_n(pb)$, where d is the average of the waveguide dimensions, kept the variation of the terms "under control." A further simplification was made by setting Z_0 to unity, which does not shift the zeroes of the determinant because if the H_t rows are multiplied by Z_0, then if Z_0 appears in a column, it will appear in a similar manner in every element of the column.

2.7.2 Mode Configurations

The electric and magnetic fields were calculated for representative cases from equation (3). To find the a_n, b_n, c_n, and d_n coefficients, k_z was first found from equation (19). Its value was then substituted into equation (18). By setting one of the elements of the T column matrix to unity, all of the other elements were then found by standard matrix techniques.[13]

Several approaches were used to obtain information that could be used to derive the field patterns. These included computation of the field components along radial cuts of the waveguide cross section, computer generated isoclines giving the direction of the electric field, and computer generated mode pictures.

The isoclines and pictures were drawn using a simulated Stromberg Carlson SC-4020 cathode ray tube plotter, which is capable of generating points and lines on a 1024×1024 grid.† A single quadrant was used for the isoclines and intensity picture since the results for all quadrants are identical except for orientation. In general, the dimensions were scaled so that the long dimension of the rectangular waveguide core extended over 80 percent of the displayed width. All figures were plotted at the points $(20m, 20n)$, where m and n take on all integer values from 0 to 49.

Isocline drawings were made by drawing a line at each of the coordinate points parallel to the electric field at that point (all lines

---

† An SC-4060 plotter was used to simulate the SC-4020 plotter to take advantage of previously existing programs.

had the same length). The isocline drawings were used as working tools to derive the field line drawings in Section III.

In order to draw pictures of mode patterns, the power density was calculated at each of the points to be plotted. The square root of the power density was then normalized to the square root of the peak power density and quantized into 21 levels. About each point in the picture, a portion of the figure shown in Fig. 3 was then plotted, starting at 1 and going to the point corresponding to the appropriate quantized level (except at the points where the quantized power was zero where no plotting was done). Since the size of the cathode ray tube spot is approximately equal to the line spacing in the figure, the plotted figures are filled in. Therefore, the light passed by these figures is approximately equal to the power density to be represented. For small index difference, since the power density is proportional to the square of the transverse electric field, the dynamic range of the pictures (in terms of the electric field) is 400.

Starting with the single quadrant pictures, complete pictures were generated by making quadruple exposures of the microfilm. In general, about 30 to 60 seconds of IBM 360/65 computing time were required for each picture.

III. RESULTS OF COMPUTATION

This section gives the computed results. Section 3.1 discusses accuracy. This is followed by a discussion of field plots and mode
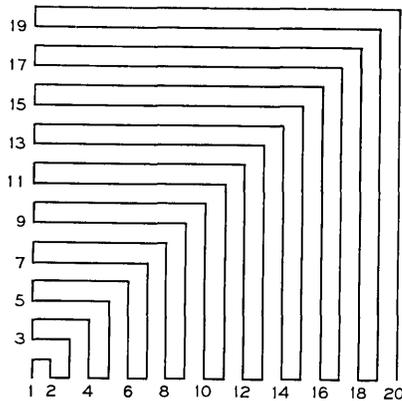
Fig. 3 — Intensity picture figure.

TABLE I—SAMPLE ACCURACY RESULTS

Number of Harmonics Used	$\mathcal{P}^2$			
	$a/b = 1$	$a/b = 2$	$a/b = 3$	$a/b = 4$
3	0.714	0.811	0.820	0.828
4	0.713	0.811	0.820	0.819
5	0.715	0.808	0.819	0.813
6	0.714	0.808	0.822	0.820
7	0.715	0.808	0.820	0.813
8	0.715	0.807	0.820	0.814
9	0.715	0.807	0.823	0.815
Variation	0.2%	0.4%	0.4%	1.5%

pictures in Section 3.2. Finally, curves of the propagation constant for a variety of conditions are presented in Section 3.3.

3.1 *Accuracy*

Numerous test runs were made in order to obtain an estimate of the accuracy of the computed results. The results of several of these runs are given in Table I for the first mode with $\mathcal{B} = 2$. The numbers at the bottom of the table represent the total variation for a given aspect ratio taken as a percentage of the full range possible (one).

For small aspect ratios, it is clear that the convergence is very rapid. However, for larger aspect ratios the convergence is not as good. For example, the variation for an aspect ratio of four is 1.5 percent (taken as a percentage of the full range of variation). For this case, from the table and from the limit for infinite aspect ratio[14] which is an upper bound for $\mathcal{P}^2$, it appears the error is about 3 percent. This error is achieved with a relatively small number of harmonics and can only be improved by using a prohibitively large number of harmonics on a computer which carries more significant digits than the one which was available for this study. However, since solutions exist for an infinite aspect ratio, the decrease in accuracy for the large aspect ratio of the circular-harmonic method is not a serious problem.

Computations similar to those for Table I were performed to obtain an estimate of the upper bound of the accuracy of the cases presented in Section 3.3. From these calculations, it is believed that all of the data to be presented in the following sections is accurate to 1 percent, except for the results of calculations using even harmonics for aspect ratios other than unity which are believed to be accurate to better than 2 percent. In general, accuracy decreases as the mode order increases, although not monotonically.

The results of the circular-harmonic analysis and of Marcatili's analysis agree.[6] In the regions where his method and the circular-harmonic method are both theoretically valid, the agreement is well within the tolerances given above. To avoid duplication, the reader is directed to his curves for a comparison.

The effect of the number of harmonics used in the field patterns is of some interest. This question has not been explored in great detail; however, a few comparisons of intensity pictures for different numbers of circular harmonics were made. In general, it was found that five harmonics were sufficient to give a good representation of the modes that this paper presents. An example of this is given in Fig. 4, comparing the E_{11}^y mode intensity patterns for five and nine harmonics. For the results which follow, five circular harmonics were used.

3.2 *Mode Configurations*

Figure 5 shows intensity pictures for the first six modes for unity aspect ratio, $\circledS = 3$, and an index difference of 0.01. Figure 6 gives similar data for an aspect ratio of two and $\circledS = 2$. For both, the plots are arranged in ascending order of cutoff frequency. All of the pictures are for E_{mn}^y modes. These pictures are virtually indistinguishable from the corresponding E_{mn}^x modes so both sets are not presented. In general, for small index differences the E_{mn}^y and E_{mn}^x can be considered to be near duals, that is, to have identical field patterns except that the electric and magnetic fields are interchanged.

The field distribution patterns for the modes of Figs. 5 and 6 are more complicated than those for the rectangular metallic waveguide
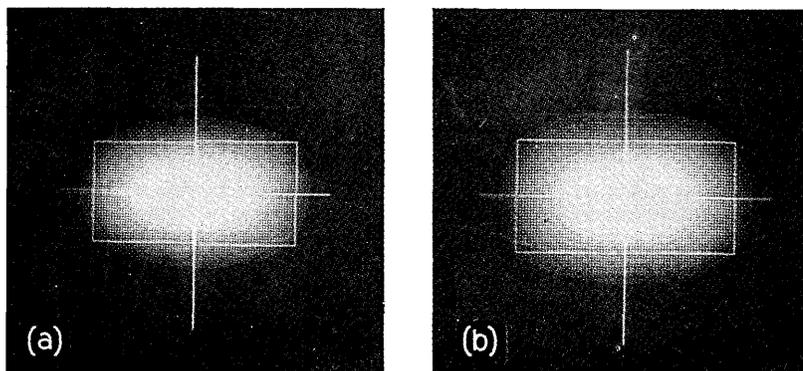
Fig. 4 — Intensity for the E_{11}^y mode for $a/b = 2$, $\circledS = 2$, and $\Delta n_r = .01$: (a) for five harmonics and (b) for nine harmonics.
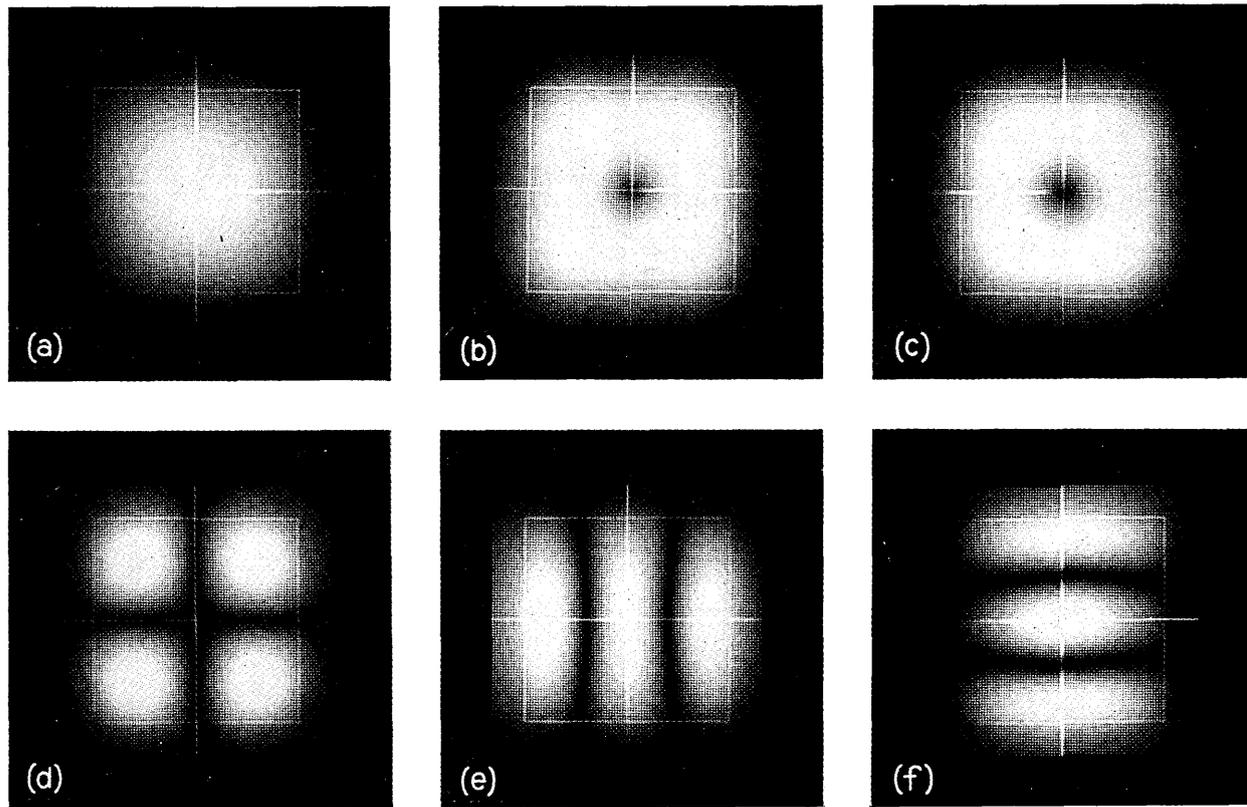
Fig. 5 — Intensity for some E_{mn}^y modes with unity aspect ratio, $\mathfrak{B} = 3$, and $\Delta n_r = 0.01$: (a) E_{11}^y, (b) E_{21}^y, (c) E_{12}^y, (d) E_{22}^y, (e) E_{31}^y, and (f) E_{13}^y.
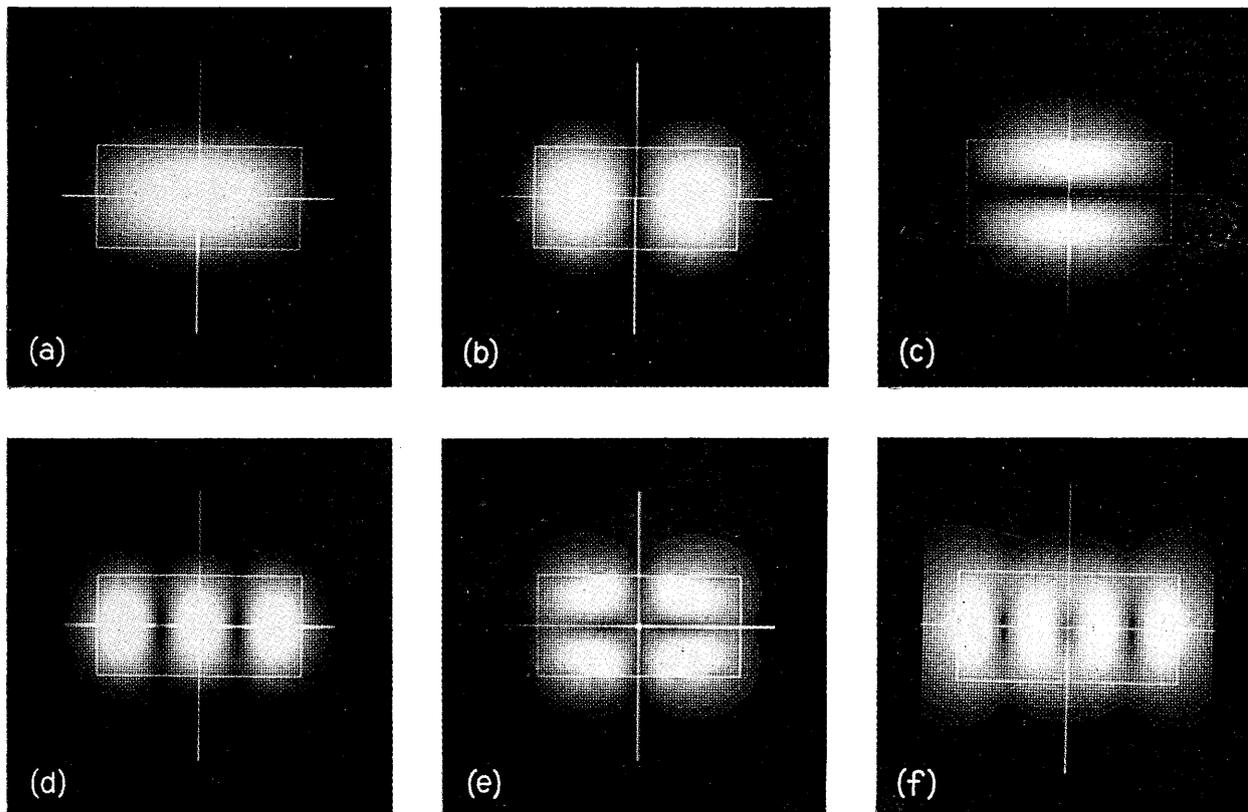
Fig. 6 — Intensity picture for some E^y_{mn} modes with $a/b = 2$, $\mathfrak{B} = 2$, and $\Delta n_r = 0.01$: (a) E^y_{11}, (b) E^y_{21}, (c) E^y_{12}, (d) E^y_{31}, (e) E^y_{22}, and (f) E^y_{41}.
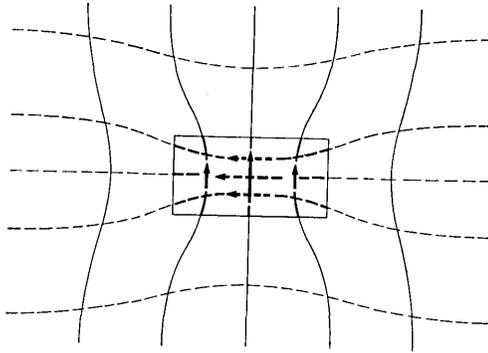
Fig. 7— Field configuration of the E_{11}^y mode.

since they extend beyond the waveguide boundary and, in general, their shape is dependent on waveguide parameters other than shape. The E_{11}^x and E_{11}^y modes have the simplest field patterns. Figure 7 shows the electric and magnetic field orientations for the E_{11}^y mode. In this figure and the following ones, there are heavy lines in the regions of high field intensity and light lines in regions of low field intensity. Only E_{mn}^y modes are shown since the E_{mn}^x modes can be obtained by interchanging the electric and magnetic field vectors.

Figure 8 shows the field lines for the E_{21}^y and E_{12}^y modes for a large aspect ratio. (For $a/b \rightarrow \infty$ the fields have the appearance of rectangular metallic waveguide modes.) However, as the aspect ratio approaches unity, the E_{12}^y and E_{21}^x modes and the E_{21}^y and E_{12}^y modes couple and shift to the patterns shown in Fig. 9. Most of the change takes place with the aspect ratio close to unity.

Figures 10, 11, and 12 show the field configurations for the E_{22}^y mode, the E_{31}^y mode, and the E_{13}^y mode, respectively. The field patterns of these modes do not change drastically with the aspect ratios.

Figure 13a shows an intensity picture of the E_{32}^y mode and Figure

(a)   (b)
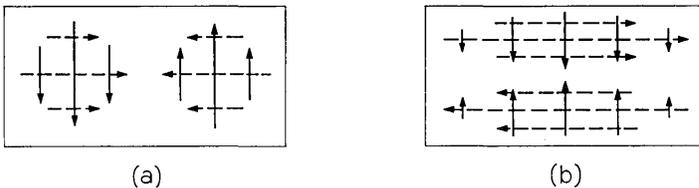
Fig. 8—Field configurations for the (a) E_{21}^y and (b) E_{12}^y modes far from cutoff.
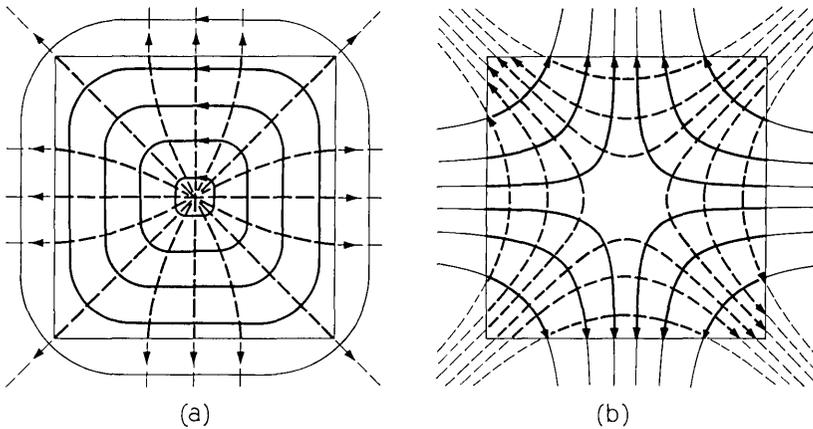
Fig. 9 — Field configurations for the square (a) E_{21}^y and (b) E_{12}^y modes.

13b its field pattern for unity aspect ratio. The field pattern inside the core is similar to a sum of the TE_{23} and TE_{32} of metallic waveguide, shown in Fig. 13c and d, respectively. Figure 13a demonstrates that the circular-harmonic analysis can generate complex field patterns with a relatively small number of harmonics.

Figures 14 and 15 show the variation of the intensity distribution with $\mathscr{P}^2$ for the E_{11}^y and E_{21}^y modes, respectively. As one would expect, for small values of $\mathscr{P}^2$ the radial extent of both modes increases very rapidly as $\mathscr{P}^2$ decreases. It is of significance, however, that most of the energy is contained within the waveguide core, even for relatively small values of $\mathscr{P}^2$ and Δn. Thus, Marcatili's assumption that very little energy propagates in the region of the corners is valid over a wide range.

3.3 Propagation Curves

In all cases of computed propagation curves, the normalized waveguide height $\mathscr{B}$, as given in equation (11), is plotted on the horizontal
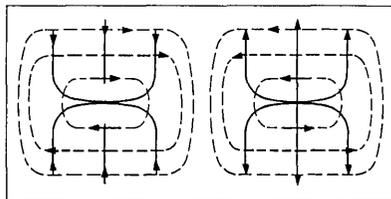
Fig. 10 — Field configuration of the E_{22}^y mode.
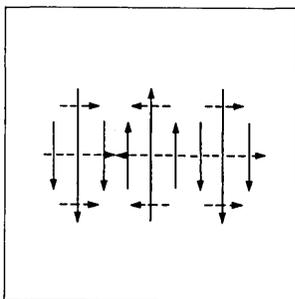
Fig. 11 — Field configuration of the E_{31}^y mode.

axis and the normalized propagation constant, $\mathcal{P}^2$, given in equation (16), along the vertical axis.

Figure 16 shows the case of vanishing index difference for an aspect ratio of one. The first 16 modes are shown. For this case the following six degenerate groups exist

$$E_{11}^y \, , \, E_{11}^x$$

$$E_{12}^y \, , \, E_{12}^x \, , \, E_{21}^y \, , \, E_{21}^x$$

$$E_{31}^y \, , \, E_{13}^x$$

$$E_{31}^x \, , \, E_{13}^y$$

$$E_{22}^x \, , \, E_{22}^y$$

$$E_{32}^y \, , \, E_{23}^x \, , \, E_{23}^y \, , \, E_{23}^x \, .$$

In addition, the E_{31}^y and the E_{31}^x modes are almost degenerate except
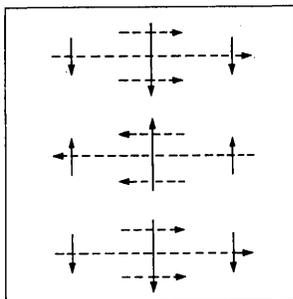
Fig. 12 — Field configuration of the E_{13}^y mode.
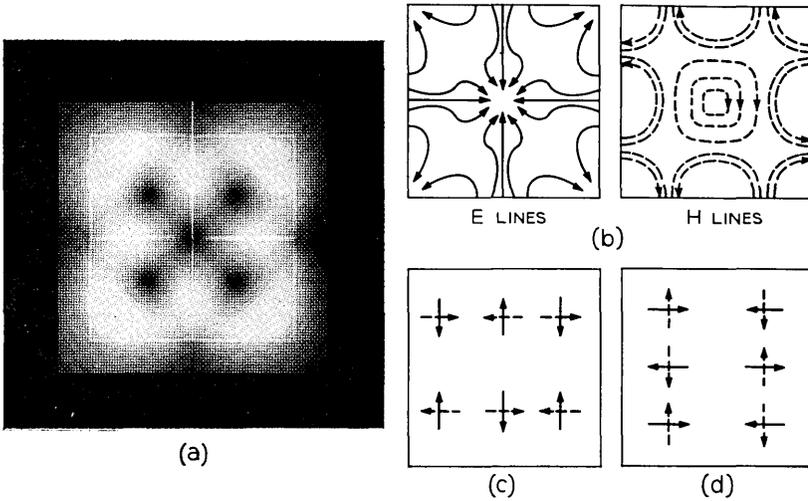
Fig. 13 — The E_{32}^y mode for unity aspect ratio: (a) intensity, (b) field configuration, (c) TE_{32}, and (d) TE_{23}.

near cutoff. The splitting of these modes can be accounted for by the differences of the field patterns shown in Fig. 11 and 12. Since the E_{31}^x mode reversals occur along the direction of the electric field lines, the electric field for this mode must have a larger longitudinal field component than for the E_{31}^y mode.

All degeneracies, except the $E_{mn}^y - E_{mn}^x$, are broken by a change in the aspect ratio as demonstrated in Fig. 17, which is drawn for the first 12 modes of a waveguide of aspect ratio 2. One interesting feature of this curve is the mode crossing of the E_{31}^y and E_{12}^y modes. Crossings of this type, which cannot occur in metallic waveguides, are possible because the field functions are frequency dependent. Qualitatively, it can be explained by noting that field reversals must take place in the core, therefore constraining the central lobe of the E_{31}^y more than any of the E_{12}^y mode lobes as cutoff is approached. Far from cutoff, however, all fields are well constrained and the E_{31}^y mode has a larger propagation constant than the E_{12}^y mode, as it does for the similar metallic waveguide mode with an aspect ratio of 2.

The effect of finite index difference on the modes can be observed by comparing Fig. 16, which is computed for unity aspect ratio and a vanishing index difference, with Fig. 18, which is computed for unity aspect ratio and a 0.5 index difference. The curves for modes whose
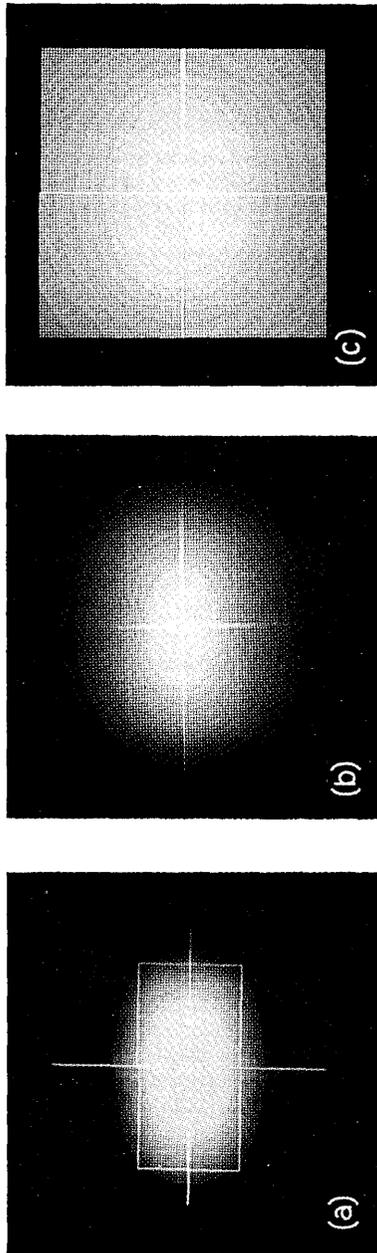
Fig. 14 — Intensity pictures of the E_{11}^y mode for (a) $\mathcal{G}^2 = 0.81$, (b) $\mathcal{G}^2 = 0.50$, and (c) $\mathcal{G}^2 = 0.02$.
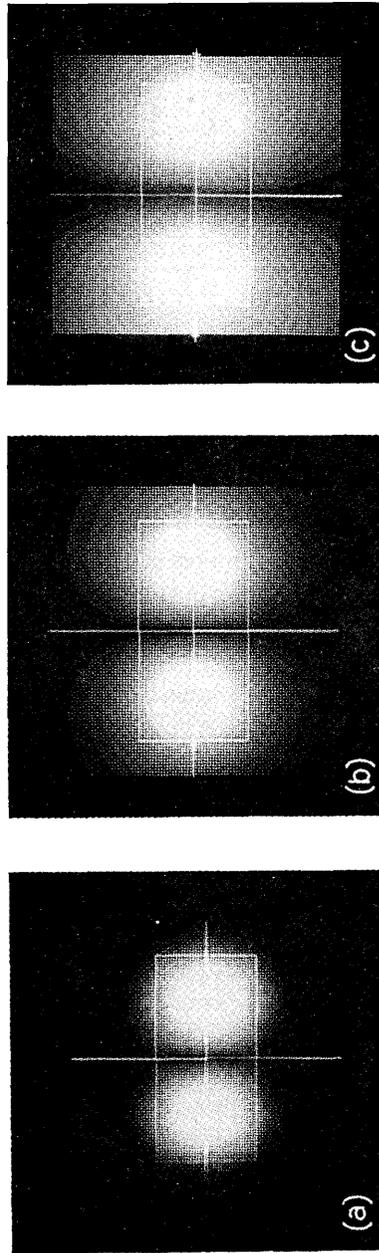
Fig. 15 — Intensity pictures of the E_{21}^y mode for (a) $\mathcal{O}^2 = 0.76$, (b) $\mathcal{O}^2 = 0.31$, and $\mathcal{O}^2 = 0.04$.
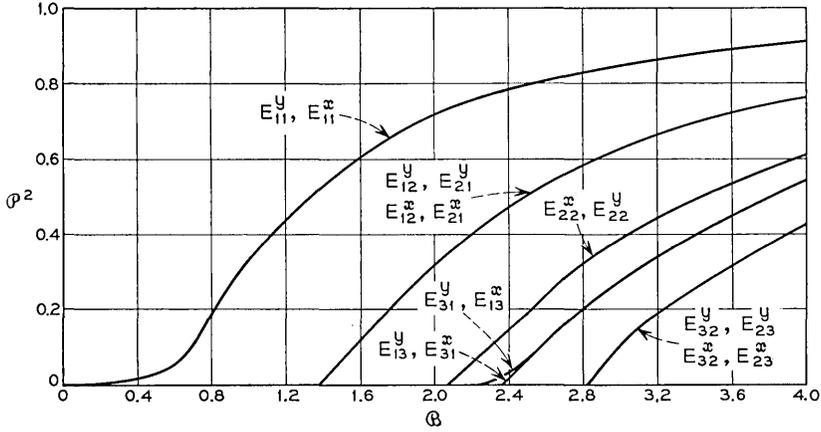
Fig. 16 — Propagation curves for the first 16 modes for unity aspect ratio and $\Delta n_r \to 0$.

field lines reverse direction across the origin are no longer degenerate, but those whose field lines do not reverse still are degenerate. For all degeneracies to be split, there must exist a finite index difference as well as an aspect ratio other than unity. Figure 19 illustrates one such case.

The effect of index difference on the degenerate principal modes for unity aspect ratio is examined in Fig. 20. The curve shows both a low and high index difference limit. In the range of interest for optical
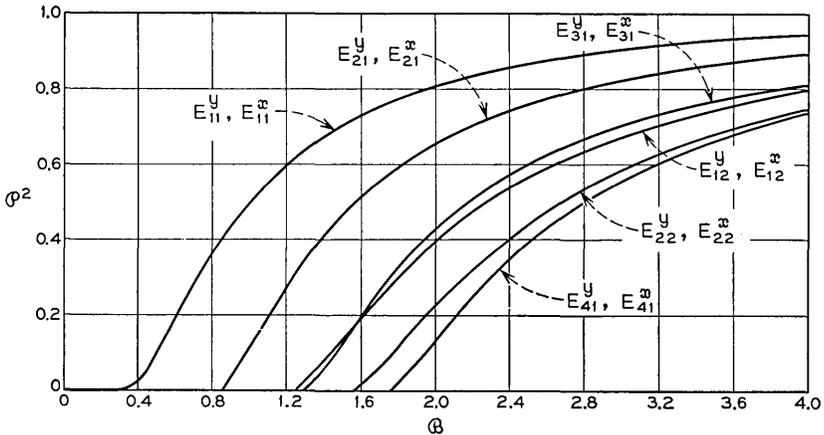
Fig. 17 — Propagation curves for the first 12 modes for $a/b = 2$ and $\Delta n_r \to 0$.
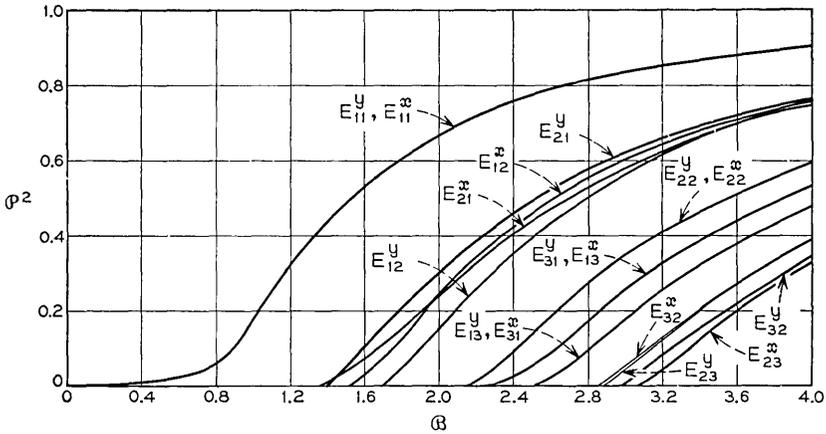
Fig. 18 — Propagation curves for the first 16 modes for unity aspect ratio and $\Delta n_r = 0.5$.

circuits $(0 - 0.1)$ the vanishing difference curve is an excellent approximation. The greatest changes occur in the $0.1 - 10$ range, which is the range of interest for some microwave problems.

Figure 21 presents the computed results for the effect of index changes on the principal modes for an aspect ratio of 2. The effect is much stronger on the E_{11}^y mode than the E_{11}^x mode. In fact, the effect on the E_{11}^x mode is comparatively small, except near cutoff.

The effect of aspect ratio on the principal modes is demonstrated for
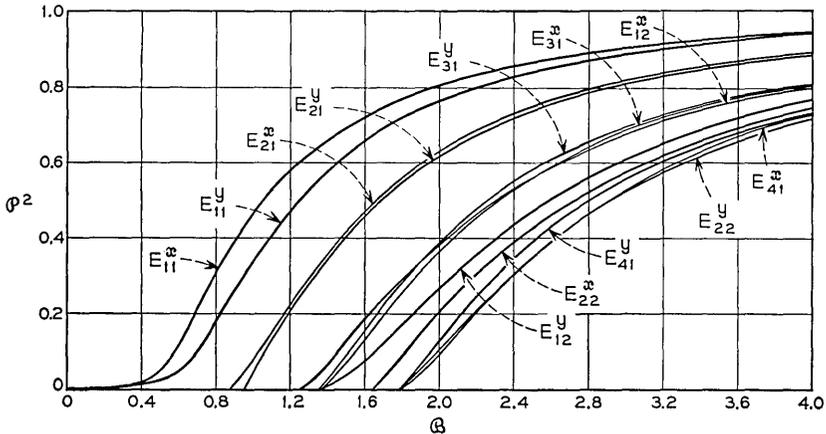
Fig. 19 — Propagation curves for the first 12 modes for $a/b = 2$ and $\Delta n_r = 0.5$.
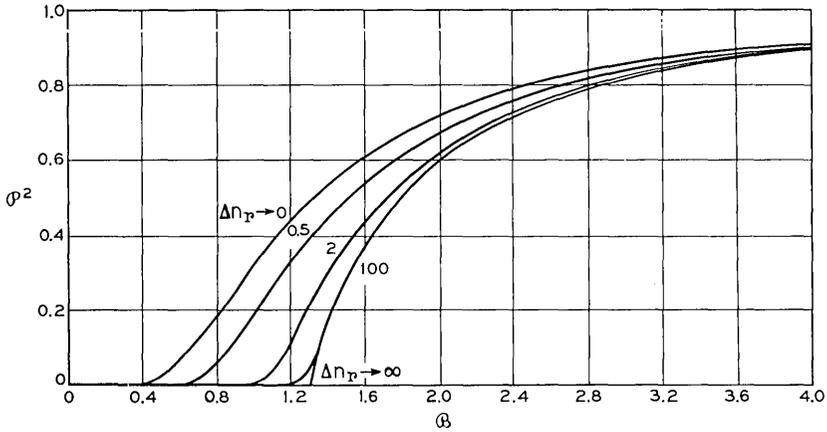
Fig. 20 — E_{11}^y and E_{11}^x mode propagation curves for several values of Δn_r with unity aspect ratio.

vanishing index difference in Fig. 22. The curve for infinite aspect ratio was obtained from the exact analysis of the slab case.[14]

IV. CONCLUSIONS

The results of the computations show that the circular harmonic method for analyzing rectangular dielectric waveguides gives excel-
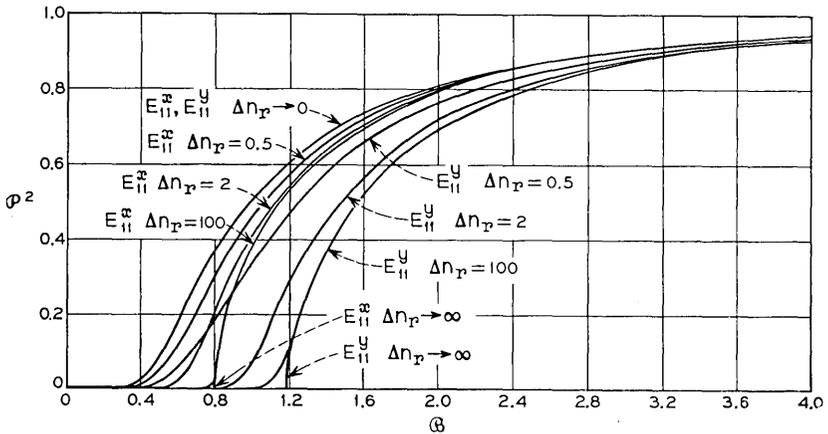
Fig. 21 — E_{11}^y and E_{11}^x mode propagation curves for several values of Δn_r with $a/b = 2$.
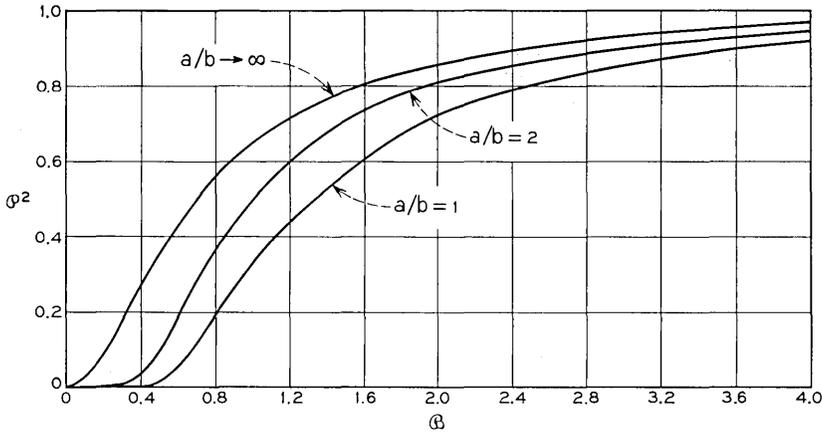
Fig. 22 — E_{11}^y and E_{11}^x mode propagation curves for several values of a/b with $\Delta n_r \to 0$.

lent results for waveguides of moderate aspect ratio. The convergence of the computed results was rapid and the results are in agreement with those of Marcatili's in the regions where his approximations apply. Furthermore, the results compare very well with Schlosser's curves for the principal mode.

Comparison of the results presented here with Marcatili's show that the two methods give values of the normalized propagation constant, $\mathcal{P}^2$, which are within a few percent for $\mathcal{P}^2 > 0.5$. Thus for $\mathcal{P}^2$ in this range his method is to be preferred since the calculations required are much simpler. However, for $\mathcal{P}^2 < 0.5$, and when it is desired to differentiate between modes for some of the near degenerate cases, another method must be used.

The circular harmonic analysis is attractive for small $\mathcal{P}^2$ because of the nature of the matching boundary. For large refractive index difference and moderate $\mathcal{P}^2$ both the method presented here and the one presented by Scholosser can be used.

V. ACKNOWLEDGMENTS

REFERENCES

1. Miller, S. E., "Integrated Optics: An introduction," this issue, pp. 2059–2069.
2. Kapany, N. S., *Fiber Optics*, New York: Academic Press, 1967, pp. 36–80.
3. Bracey, M. F., Cullen, A. L., Gillespie, E. F. F., and Staniforth, J. A., "Surface Wave Research in Sheffield," I.R.E. Trans. Antennas and Propagation, *AP-7*, No. 10 (December 1959), pp. 219–225.
4. Snitzer, E., "Cylindrical Dielectric Waveguide Modes," J. Opt. Soc. of Amer., *51*, No. 5 (May 1961), pp. 491–498.
5. Snitzer, E., and Osterberg, H., "Observed Dielectric Waveguide Modes in the Visible Spectrum," J. Opt. Soc. of Amer., *51*, No. 5 (May 1961), pp. 491–505.
6. Marcatili, E. A. J., "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," this issue, pp. 2071–2102.
7. Schlosser, W., and Unger, H. G., "Partially Filled Waveguides and Surface Waveguides of Rectangular Cross-Section," *Advances in Microwaves*, New York: Academic Press, 1966, pp. 319–387.
8. Shaw, C. B., French, B. T., and Warner, C. III, "Further Research on Optical Transmission Lines," Sci. Rep. No. 2, Contract AF449 (638)-1504 AD 625 501, Autonetics Report No. C7-929/501, pp. 13–44.
9. Stratton, J. A., "Electromagnetic Theory," New York: McGraw-Hill, 1941, p. 361.
10. Laxpati, S. R., and Mittra, R., "Energy Considerations in Open and Closed Waveguides," IEEE Trans. Antennas and Propagation, *AP-13*, No. 6 (November 1965), pp. 883–890.
11. Hamming, R. W., *Numerical Analysis for Scientists and Engineers*, New York: McGraw-Hill, 1962, pp. 81–82.
12. Freed, B. H., "Algorithm 41," Revision Evaluation of Determinant Comm. ACM, *6*, No. 9 (September 1963), p. 520.
13. "System/360 Scientific Subroutine Package," IBM, White Plains, N. Y., H20-0205-2, pp. 179–182.
14. Collin, R. E., "Field Theory of Guided Waves," New York: McGraw-Hill, 1960, pp. 480–495.

Improved Relations Describing Directional Control in Electromagnetic Wave Guidance

By E. A. J. MARCATILI and S. E. MILLER

(Manuscript received January 22, 1969)

The direction-changing capability of electromagnetic waveguides may be limited not only by mode conversion but also by radiation if the transverse field extends indefinitely into a freely propagating region. This paper gives new, more accurate expressions for the permitted bending radius with respect to mode conversion, using coupled-wave theory to categorize the wide variety of transmission media possible. This paper also makes a suggestion for estimating the permitted bending radius when radiation is a limitation. In single-mode "open" waveguides that have transverse fields extending indefinitely into a freely propagating region (such as a dielectric waveguide), the permitted bending radius is limited by radiation effects, whereas in either the open or completely shielded multimode waveguides, the permitted bending radius is usually limited by mode conversion.

I. INTRODUCTION

It is useful to be able to characterize the direction-changing capability of electromagnetic waveguides without detailed knowledge of the waveguiding structure. The first work in this area was reported by Miller in 1964.[1] A direction-determining parameter R_{min} was defined

$$R_{min} = \frac{a^3}{4\lambda^2} \tag{1}$$

in which R_{min} is a bend radius, a is the full transverse width of the field distribution, and λ is the wavelength in the medium in which the waveguide is embedded.* For bend radii longer than R_{min}, Ref. 1 indicates that wave propagation is virtually as in a straight guide; at radii less than R_{min} something drastic happens. Just what changes

---

* Notice that we have redefined a here; in Ref. 1 the full transverse width of the field distribution was $2a$.

occur in a straight guide depends on the nature of the medium in detail; for hollow conducting guides the change is large mode conversion and for beam transmission in a sequence of infinitely wide lenses the change is also mode conversion appearing as a wide oscillation of the beam about the nominal axis of propagation.

Following similar lines of thought, a parameter

$$\delta_{\max} = \frac{\lambda}{a} \tag{2}$$

is given to describe the transition region between essentially normal wave propagation and the region of drastic changes for abrupt angular changes in direction.[1] The only restriction on these order of magnitude direction-determining parameters given in Ref. 1 is the exclusion of degeneracy between the used mode and some other mode coupled by the direction change. It is well known that such a degeneracy results in complete loss of signal for certain lengths of bent guide regardless of the bending radius, and that removal of the degeneracy by dissipative or reactive means can in principle make the bend loss as small as desired.[2-4]

In recent studies of bend losses in dielectric waveguides, Marcatili found a serious disagreement between the implications of equation (1) and the bend losses predicted by analysis of the particular waveguiding structure.[5] For an "open" waveguide—that is, one in which the transverse field decays exponentially in a transverse plane but extends to great distances—he found that the bend radius required for tolerable losses was much larger than given by equation (1) and it followed a different law with relation to a and λ when only one mode could propagate.

It is now clear that two components of bend loss must be considered: the dissipative loss (resulting from either radiation or coupling to a high-loss undesired mode) for the normal mode of the bend region characterized by an attenuation coefficient α_r, and the mode conversion loss P_c for the straight-guide mode on entering and leaving the curved region. If mode transformers were used at the ends of the curved region (impractical for occasional bends in most transmission situations), the mode conversion loss would be zero and any bend R would be acceptable from that criterion.

Equation (1) relates to the mode conversion loss; it fails to give a correct estimate when dissipative loss is important. The permitted bend radius R must be assessed with respect to dissipative loss as

well as mode conversion loss; Section II gives relations which make this possible. Improved forms of equations (1) and (2) have also been derived which explicitly relate the maximum conversion loss to the bending radius for the generalized electromagnetic waveguide. The added quantitative factor should provide greater usefulness since the improved relations not only identify the transition region between virtually straight-guide behavior and violent changes, but also give detail about the transition. Section III gives these results and the appendices give the derivations.

II. RADIATION FROM CURVED OPEN WAVEGUIDES

Figure 1 shows a representation of an open waveguide. The shaded wave-guiding region has an effective index of refraction larger than that of the surrounding region, resulting in a transverse field distribution for the guided mode $F(x)$ which decays exponentially but remains finite. To derive a generalized expression for radiation loss as a function of bending radius R, we visualize this as a two-dimensional guide with an isotropic surrounding region capable of supporting a free-space radiating wave. We note that at some transverse distance x_r the maintenance of a pure guided mode with equiphase fronts on
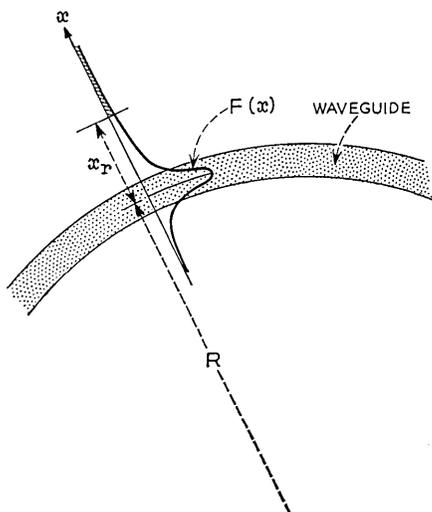
Fig. 1 — A two-dimensional open waveguide.

radial planes requires energy propagating at the speed of light, and for $x > x_r$ a pure guided mode implies energy propagating at greater than the velocity of light. This is true at some value of x_r for any finite bend radius R, since $F(x)$ extends indefinitely in the x direction. We postulate that the transverse field distribution $F(x)$ is virtually the same in the curved region as in a straight guide for large R. The fraction of the energy in the guided mode at $x > x_r$ is assumed to be lost to radiation; this loss is taken to occur in a longitudinal distance equal to the collimated-beam length associated with the field $F(x)$. All these assumptions imply that any mode propagating along the curved open guide radiates. This is indeed the case for the modes in the curved dielectric guide analyzed in Ref. 5.

As developed in Appendix A, the attenuation coefficient for the normal mode of the bend region is

$$\alpha_r = \frac{1}{2Z_c} \frac{\mathcal{E}_l}{\mathcal{E}_T}, \tag{3}$$

where

$$\mathcal{E}_l = \int_{x_r}^{\infty} F^2(x) \, dx, \tag{4}$$

$$\mathcal{E}_T = \int_{-\infty}^{\infty} F^2(x) \, dx, \tag{5}$$

$$Z_c = \frac{a^2}{2\lambda_s}, \tag{6}$$

$$x_r = \frac{(k_z - k_s)}{k_s} R, \tag{7}$$

k_z = longitudinal phase constant for the guided mode,
k_s = $2\pi/\lambda_s$ phase constant for a plane wave in the surrounding region, and
a = effective width of the transverse field $F(x)$.

Applying this formulation to a curved two-dimensional dielectric-slab waveguide of width t gives the following. From solutions of Maxwell's equations in a straight guide

$$F(x) = \cos k_x x \quad \text{for} \quad -\frac{t}{2} \leq x \leq \frac{t}{2}, \tag{8}$$

$$F(x) = \cos\left(\frac{k_x t}{2}\right) e^{\frac{-\left(|x| - \frac{t}{2}\right)}{\xi}} \quad \text{for} \quad |x| \geq \frac{t}{2}. \tag{9}$$

The resulting expressions for z_c, $\mathcal{E}_l$, and $\mathcal{E}_T$ are

$$\mathcal{E}_l = \frac{\xi}{2} \cos^2 \left(\frac{k_x t}{2}\right) e^{\dfrac{-2\left(x_r - \dfrac{t}{2}\right)}{\xi}}, \tag{10}$$

$$\mathcal{E}_T = \frac{t}{2} + \frac{1}{2k_x} \sin k_x t + \xi \cos^2 \left(\frac{k_x t}{2}\right), \tag{11}$$

$$z_c = \frac{\left[t + 2\xi \cos \left(\dfrac{k_x t}{2}\right)\right]^2}{2\lambda_s}. \tag{12}$$

These expressions, when put into equation (3), yield a radiation attenuation coefficient of the form*

$$\alpha_r = c_1 \exp(-c_2 R), \tag{13}$$

where c_1 and c_2 are independent of R. As Table I illustrates, in several cases of interest c_1 and c_2 are very large numbers (calculated for $\lambda = 0.6328$ μm). Case 1 corresponds to a thin glass sheet surrounded by air; cases 2 and 3 correspond to 1 percent and 0.1 percent index differences between the guide and the surrounding region, a possible guide of interest for miniature laser-beam circuitry.[6] Because c_1 and c_2 are so large, reasonable values of α_r occur only within a narrow range of bend radius R. Figure 2 illustrates α_r versus R for case 2. We can define a transition radius R_t as that value of R which gives $\alpha_r = 1$ neper per meter:

$$R_t = \frac{1}{c_2} \log c_1 \tag{14}$$

in which c_1 and c_2 are the constants of equation (13) found by evaluating equation (3). Because of the exponential nature of α_r versus R, radii smaller than R_t give excessive losses and radii slightly larger than R_t give negligibly small losses. We may therefore use R_t as an index of this transition for radiation losses analogous to the $R_{\min}$ of equation (1) for mode conversion losses.

Notice the size of x_r, the transverse distance to where wave propagation at the velocity of light is required. For cases 1, 2, and 3, x_r has the values 1.0, 3.9, and 16.5 μm, respectively, for $\alpha_r = 1$ neper per meter. Wave propagation at the velocity of light occurs quite close to the center of the guide, well within the bending radius.

---

* This paper uses mks units in all formulas.

TABLE I — VALUES FOR c_1 AND c_2

Case	Waveguide index of refraction	Slab width t (μm)	Sur-rounding index of refraction	c_1 (nepers per meter)	c_2 (meters^{-1})	R for $\alpha_r = 1$ neper/m
1	1.5	0.198	1.0	2.57×10^6	3.47×10^6	$4.25 \ \mu m$
2	1.5	1.04	1.485	1.04×10^5	1.46×10^4	0.79 mm
3	1.5	1.18	1.4985	5.4×10^3	81.4	0.106 m

In Appendix A the results using equation (3) are compared with the more exact values of α_r obtained from Maxwell's equations directly.[5] For a given α_r equation (3) yields a value of R about 0.6 times that obtained from Ref. 5. Moreover, Ref. 5 shows that, as the slab width t increases, the radiation loss does not decline indefinitely; the normal mode transverse field reshapes itself in the bend to increase $F(x)$ in the x_r region. However, the mode conversion loss usually becomes important at those values of t and for incidental bends (that is, without mode matching transformers) the mode conversion loss is limiting rather than radiation loss.

Another approach, which yields an expression for the radiation loss of the curved guide in terms of constants of the straight guide, consists
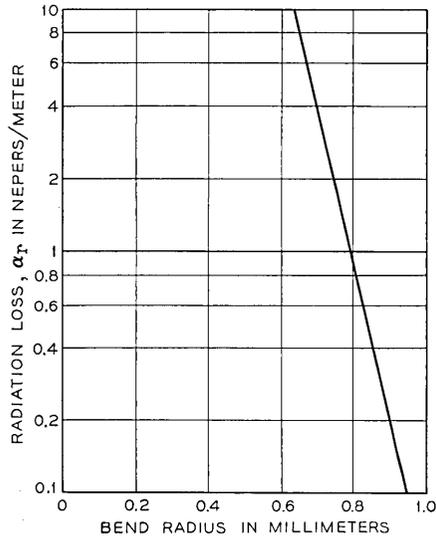
Fig. 2 — Radiation loss versus bend radius for a two-dimensional dielectric waveguide; case 2 of Table I.

of noticing that the boundary value problem, which can be solved exactly by matching the radial impedances at each interface, can also be solved approximately if the radius of curvature R is so large that the field components of the curved guide differ only slightly from those in the straight guide.[5] Then, all the impedances can be replaced by those of the straight guide except that on the external interface of the bend which, according to Ref. 5, must be multipled by

$$1 + i \exp\left(-\tfrac{2}{3}R \frac{k_{xe}^3}{k_z^2}\right).$$

In this expression k_{xe} and k_z are the propagation constants in the x and z directions in the external medium of the straight guide. The attenuation constant of the curved guide results

$$\alpha_r = k_{xe} \exp\left(-\tfrac{2}{3}R \frac{k_{xe}^3}{k_z^2}\right) \frac{\partial k_z}{\partial k_{xe}}. \tag{15}$$

This expression should give greater accuracy in general and does so in the case of the slab waveguide used in this section. It also shows that waveguides which present imaginary radial impedances have no radiation loss.

III. MODE CONVERSION LOSSES IN CURVED OPEN OR BOUNDED WAVEGUIDES

3.1 *General Formulation of Tilt Relation*

When a pure mode of a straight multimode waveguide enters and leaves a curved region, it generally suffers mode conversion loss. Coupled-mode theory has been applied to calculate these losses as a function of bend radius and to devise lower loss bend structures.[3,4,7,8] In these previous contributions, direct solution of Maxwell's equations is used to find which of the straight-guide modes are coupled in the bend, and for these important modes to find the transfer coupling coefficients and the associated differences in propagation constants which are needed in the coupled wave solution.

We present here a generalized use of coupled wave theory which gives an improvement on equations (1) and (2) in predicting approximate values of tolerable bend radius without direct solution for the transfer coupling coefficients or the phase constants. We do not imply that this provides accuracy comparable to a direct solution. It does yield an approximate answer to show where further work to get more accuracy is of interest.

The first approximation is used to derive the transfer coupling coefficient from the self-coupling coefficient. Consider a tilt (illustrated in Fig. 3) for a hollow metallic rectangular waveguide. The self-coupling in the tilt from the incident mode to the same mode beyond the tilt, of angle δ, is[9]

$$
| c_{sc} | = \left| \frac{\int_0^w \int_0^b \left[\left(\frac{\partial F}{\partial x} \right)^2 + \left(\frac{\partial F}{\partial y} \right)^2 \right] \exp \left(i \frac{2\pi}{\lambda_z} \delta x \right) dx \, dy}{\int_0^w \int_0^b \left[\left(\frac{\partial F}{\partial x} \right)^2 + \left(\frac{\partial F}{\partial y} \right)^2 \right] dx \, dy} \right| ,
\qquad (16)
$$

in which λ_z is the guided wavelength along z.

The function F is the axial field component which, for hollow metallic rectangular waveguides, is either $\sin \pi p \, x/w \, \sin \pi q \, y/b$ for TM_{pq} modes or $\cos \pi p x/w \, \cos \pi q \, y/b$ for TE_{pq} modes.

For small tilt angles $| c_{sc} |$ is of the form

$$
| c_{sc} | = 1 - \Delta,
\qquad (17)
$$

where $\Delta \ll 1$; Δ corresponds to the energy lost from the input mode at the tilt, whether by reflection or transmission into a single or into many modes. We now assume the incident mode to be well above cutoff so that reflection effects are small; that is, $w/\lambda > 1$ and preferably $w/\lambda \gg 1$. We further assume that all the lost energy at the tilt goes into a single undesired mode. For such a transfer

$$
| c_{sc} | = (1 - | c_t^2 |)^{\frac{1}{2}} \approx 1 - \tfrac{1}{2} | c_t |^2,
\qquad (18)
$$

where c_t is the transfer coupling coefficient. We then combine equations (17) and (18) to obtain the transfer coupling coefficient

$$
| c_t | = (2\Delta)^{\frac{1}{2}},
\qquad (19)
$$
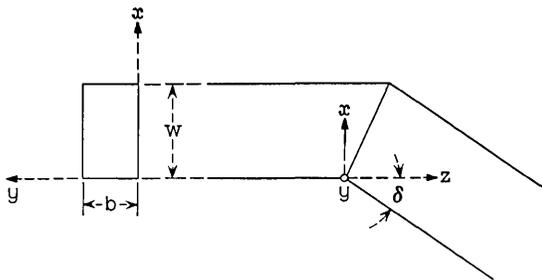
Fig. 3 — Tilt in hollow metallic rectangular waveguide.

and the fraction of the input power that is converted is

$$P_t = 2\Delta. \tag{20}$$

Carrying out the integration of equation (16) for the rectangular metallic waveguide, assuming $\delta w/\lambda_z \ll 1$, gives

$$P_t = B\left(\frac{\delta w}{\lambda_z}\right)^2. \tag{21}$$

Appendix C shows that for the lowest order TE mode TE_{10}, B is 5.28. For other modes, B ranges between 5.28 and 1.28; we somewhat arbitrarily select the geometric mean of these values to approximate P_t for any mode. Then,

$$P_t = 2.6\left(\frac{\delta w}{\lambda_z}\right)^2, \tag{22}$$

$$c_t = 1.61\left(\frac{\delta w}{\lambda_z}\right), \tag{23}$$

$$\delta = 0.62\,\frac{\lambda_z}{w}\,(P_t)^{\frac{1}{2}}, \tag{24}$$

which we have derived under the restrictions

$$\frac{w}{\lambda} \gg 1, \qquad \frac{\delta w}{\lambda} \ll 1.$$

Equation (24) is an improved form of equation (2). It shows the approximate tilt angle permitted versus fractional power converted. Derived for hollow metallic waveguide of width w, the "field" width is also w which is equivalent to a in equation (2); since we required the modes to be far from cutoff, $\lambda_z \cong \lambda$; however, we note that the converted power P_t is smaller in fact than indicated by using $\lambda_z = \lambda$ since the guided wavelength λ_z is greater than λ.

3.2 Formulation of Bend Coupling Coefficient

Using a limiting process, described in Section 2.3.2 of Ref. 10, the tilt conversion coefficient can be converted to a continuous bend conversion coefficient. Consider a sequence of straight guide sections, each of length l and connected making a tilt angle δ (Fig. 4). Let us assume that a mode entering in this guide couples at each tilt mostly to itself and lightly to one single spurious mode travelling in the forward direction. The tilt amplitude coupling coefficient is given by equation (23). The coupling per unit length is $|c_t/l|$; letting l and δ
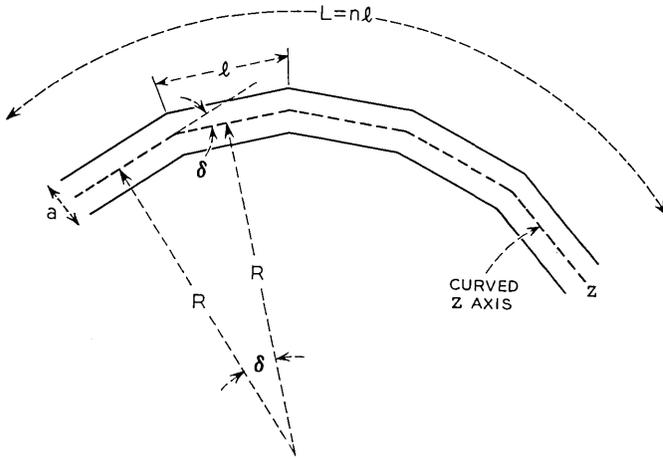
Fig. 4 — Waveguide bend made of a series of straight segments.

go to zero simultaneously in such a way that $l/\delta = R$, the bend amplitude coupling coefficient c_B is:

$$| c_B | = 1.61 \frac{w}{\lambda_z R}. \tag{25}$$

3.3 Coupled Wave Interaction

We are now prepared to discuss the effect of bends in producing mode conversion using coupled-wave theory. In this approach the signal amplitude E_1 is related to the undesired mode amplitude E_2 by the equations

$$\frac{dE_1}{dz} = -\Gamma_1 E_1 + k E_2 , \tag{26}$$

$$\frac{dE_2}{dz} = -\Gamma_2 E_2 + k E_1 , \tag{27}$$

in which

$\Gamma_1 = \alpha_1 + i\beta_1 = $ propagation constant of signal wave,
$\Gamma_2 = \alpha_2 + i\beta_2 = $ propagation constant of undesired wave, and
$k = $ transfer coupling coefficient.

These equations have been solved and the resulting wave interactions discussed in many papers.[3,4,8,10,11] Appendix B gives a few of the expressions relevant to this disscussion; we will draw from these. We

assume a boundary condition, $E_1 = 1.0$ and $E_2 = 0$ at $z = 0$, throughout. The effects of mode coupling depend importantly on $(\Gamma_1 - \Gamma_2)$ and k. In finding expressions which improve on equation (1) we break the discussion of a generalized waveguide down into a series of cases which are classified by the relation between the coupling coefficient k and $(\Gamma_1 - \Gamma_2)$.

3.4 Gradual Bends in Low-Loss Waveguides

We categorize the case of gradual bends in low-loss waveguides by

$$| k^2 | \ll (\beta_1 - \beta_2)^2, \tag{28}$$

$$(\alpha_1 - \alpha_2)^2 \ll (\beta_1 - \beta_2)^2, \tag{29}$$

$$\alpha_2 L \ll 1, \tag{30}$$

where L is the length of the bend.

This is the most likely case to be encountered in waveguides intended for low-loss transmission. The special case of degeneracy, $\beta_1 = \beta_2$, is treated in Section 3.6; degeneracy is not likely to occur accidentally since it is a very critical condition. Because β is very large compared with α in typical cases, equation (29) can be satisfied with relatively small changes from the degenerate condition, and the present case can be considered achievable except under very special circumstances.

With small α's, k is pure imaginary, $k = ic$; a value such as given by equation (25) applies. With equation (30) valid, the signal loss oscillates along the bend between zero and a maximum value

$$P_e = \left(\frac{2c}{\beta_1 - \beta_2} \right)^2. \tag{31}$$

To complete our derivation we need $(\beta_1 - \beta_2)$, which should be the difference between the phase constants of the modes coupled in the bend. We have not determined in our generalized waveguide case just which modes are coupled. We use as an approximation the rectangular metallic waveguide case of Fig. 3, and calculate the $\Delta\beta$ for the pth and $(p \pm 1)$ mode; again requiring the modes to be far from cutoff, we find

$$\Delta\beta = (\beta_1 - \beta_2) \simeq \frac{(2p \pm 1)\pi}{4} \frac{\lambda}{w^2}. \tag{32}$$

Combining equations (31), (32), and (25) with $c = | c_B |$ and solving for R yields

$$R = \frac{4.1}{(2p \pm 1)(P_c)^{\frac{1}{2}}} \frac{w^3}{\lambda^2}. \tag{33}$$

For the $p = 1$ mode only, the $(+)$ sign in $(2p \pm 1)$ applies; but for higher order modes either sign is applicable and the $(-)$ sign will be controlling. As a further rough approximation we may drop the ± 1 term, yielding

$$R = \frac{2.05}{p(P_c)^{\frac{1}{2}}} \frac{w^3}{\lambda^2}. \tag{34}$$

Equation (34) has the same general form as equation (1) but gives added accuracy by showing the quantitative influence of mode index and fractional conversion loss permitted.

3.5 Gradual Bends in Lossy Waveguides

Here we keep equations (28) and (29) but address the case where the undesired mode coupled to has high loss over the length L of the bend:

$$\alpha_2 L \gg 1. \tag{35}$$

Now, the true situation is very complex. The coupling coefficient k is complex and may have real and imaginary components that are equal. Energy conservation between c_{sc} and c_t, which was implied by equation (18), is not justified. Experience with helix waveguide for TE_{01}^0 waves shows, however, that the modulus of the helix coupling coefficient is comparable to that for a copper tube; therefore, we use equation (25) for the $| k |$ and proceed as before.

As the result of equation (35) the oscillations in the conversion loss are damped out and the conversion loss has the form of a simple exponential; that is, the normal mode of the curved region is set up with an attenuation coefficient $(\alpha_B + \alpha_1)$, where the extra loss resulting from the bend is

$$\alpha_B = \text{real} \left[\frac{k^2}{(\Gamma_1 - \Gamma_2)} \right]. \tag{36}$$

Using equation (25) with $| c_B | = | k |$, this becomes

$$\alpha_B = \frac{4.21}{(2p \pm 1)^2} \frac{(\alpha_2 - \alpha_1) w^6}{R^2 \lambda^4}. \tag{37}$$

This resembles a radiation loss in that it grows with length L, whereas in Section 3.4 the oscillatory loss peak was independent of L.

We can rearrange equation (37) to show the permitted bend radius R,

(again dropping the ± 1):

$$R = \frac{1.05}{p} \left[\frac{(\alpha_2 - \alpha_1)}{\alpha_B} \right]^{\frac{1}{2}} \frac{w^3}{\lambda^2}. \tag{38}$$

Here α_B may be regarded as a design criterion selected to meet the requirements of a particular use, analogous to P_c above; as such α_B may be independent of λ or may have some λ dependency.

Expression (38) has a character markedly different from equation (1). Since α_2 and α_1 are dependent on guide size and wavelength the a^3/λ^2 dependence given by equation (1) is not valid when coupling takes place to a very lossy mode.

3.6 *Bends in a Waveguide with Low-Loss Degenerate Coupled Modes*

When the modes coupled in the bend are degenerate, whether by design or misfortune, a far more stringent requirement on R develops. In this case

$$\beta_1 \equiv \beta_2 . \tag{39}$$

Because attenuation coefficients are small in many typical cases, it is relatively easy to obtain coupling coefficients that are larger, that is,

$$| c_B |^2 \gg | \alpha_2 - \alpha_1 |^2. \tag{40}$$

Then the signal wave output of a bend of length L is

$$| E_1 | = | \cos c_B L | \tag{41}$$

or, using the value of equation (25) for c_B,

$$| E_1 | = \left| \cos \left(\frac{1.61 \, wL}{\lambda_z R} \right) \right|. \tag{42}$$

The signal loss is infinite when the argument of the cosine is an odd multiple of $\pi/2$, and the corresponding bend radius R_∞ or bend length L_∞ are

$$R_\infty = \frac{1.02 \, wL}{m \lambda_z} \tag{43}$$

$$\left. \right\} \quad \text{for} \quad m = 1, 3, 5.$$

$$L_\infty = 0.98 \, m \, \frac{\lambda_z R}{w} \tag{44}$$

For small fractional power losses P_c, equation (42) may be approximated by the first term of the expansion; the resulting permitted

bend radius is

$$R = \frac{1.61}{(P_c)^{\frac{1}{2}}} \frac{wL}{\lambda}. \tag{45}$$

When $(\beta_1 - \beta_2)$ is nonzero, the signal transmission oscillates between unity and a minimum of

$$|E_1|_{\min} = \frac{\left|\dfrac{\beta_1 - \beta_2}{2c}\right|}{\left[\left(\dfrac{\beta_1 - \beta_2}{2c}\right)^2 + 1\right]^{\frac{1}{2}}} \tag{46}$$

which merges with equation (30) and the case considered in Section 3.4.

3.7 Bends in Waveguides with High-Loss Degenerate Coupled Modes

When the phase constants of the modes coupled in the bend are degenerate—that is, equation (37) holds—but the undesired mode is very lossy

$$|\alpha_2 - \alpha_1|^2 \gg |c_B|^2. \tag{47}$$

Then Appendix B shows that we again have normal-mode propagation in the bend region (as in Section 3.5) with an attenuation constant $(\alpha_1 + \alpha_B)$ where

$$\alpha_B = \frac{c_B^2}{\alpha_2 - \alpha_1}. \tag{48}$$

Using equation (25), this yields a bend radius:

$$R = \frac{1.61}{[\alpha_B(\alpha_2 - \alpha_1)]^{\frac{1}{2}}} \frac{w}{\lambda}. \tag{49}$$

This corresponds to very long bend radii in order to have equation (47) valid. Just as in equation (38), α_B of equation (49) is a discretionary design parameter.

IV. COMPARISON WITH KNOWN DIRECT SOLUTIONS

The principal usefulness of the preceding approximate relations for permissible tilt and bend radius is in new unstudied situations, where direct solutions are not available. However, we compare here the approximations with known direct solutions in order to gauge the accuracy to be expected.

4.1 *Tilt in a Sequence of Cylindrical Lenses: (Two-dimensional Problem)*

The input mode is gaussian, its spot size is w_o, and the transverse field distribution is $\exp[-(x/w_o)^2]$. The normalized power coupled to other modes at the tilt ($\delta \ll 1$) is[12]

$$P_2 = 1 - \left\{ \frac{\left| \int_{-\infty}^{\infty} \exp\left[-2\left(\frac{x}{w_o}\right)^2 - i\frac{2\pi}{\lambda}\delta x\right]dx \right|}{\int_{-\infty}^{\infty} \exp\left[-2\left(\frac{x}{w_o}\right)^2\right]dx} \right\} \cong \left(\frac{\pi\,\delta w_o}{\lambda}\right)^2. \qquad (50)$$

To compare this exact result with our approximate one, equation (22), we must define the width a of the beam. Somewhat arbitrarily we choose

$$a = 2w_o ; \qquad (51)$$

thus 95 percent of the power is traveling within the width a.

Substituting this value in equation (50) we obtain

$$P_2 = 2.5\left(\frac{\delta a}{\lambda}\right)^2. \qquad (52)$$

This compares to equation (21) with $p = 1$ and $w = a$,

$$P_t = 2.6\left(\frac{\delta a}{\lambda}\right)^2. \qquad (53)$$

Considering that equation (53) came from rectangular metallic waveguide and equation (52) from an open lens waveguide, the correspondence seems excellent.

4.2 *Tilt in a Cylindrical Metallic Waveguide Propagating* TE_{01}^0

For TE_{01}^0 at a tilt, important coupling is known to occur to three modes:[2,10]

Mode pair	Tilt coupling coefficient	
$TE_{01}^0 - TE_{11}^0$	$0.585\,\dfrac{a\delta}{\lambda}$	(54)
$TE_{01}^0 - TE_{12}^0$	$0.98\,\dfrac{a\delta}{\lambda}$	(55)
$TE_{01}^0 - TM_{11}$	$0.58\,\dfrac{a\delta}{\lambda}$	(56)

where a is the diameter of the round guide and is the full width of

the transverse field. This corresponds to equation (23) with $w = a$ and $p = 2$ (two extrema in the transverse field),

$$c_t = 1.61 \frac{a\delta}{\lambda}. \tag{57}$$

In the real case, the converted power is the sum of three conversions using the above three coupling coefficients; since the three components vary with a different period versus λ, or distance along the guide after the tilt, the actual mode conversion is a complicated function. We might take the root-sum-square combination of equations (54) through (56) to compare with equation (57), leading to

$$\mathrm{TE}_{01}^0 \ c_{t(rss)} \cong 1.65 \frac{a\delta}{\lambda}. \tag{58}$$

The converted power loss is $|c_t|^2$, so we see that equation (57) gives a correct order of magnitude indication, but it lacks significant detail.

4.3 Bends in Cylindrical Metallic Waveguide Propagating TE_{01}^0

The above discussion for tilt coupling coefficient applies directly to bend coupling coefficient in empty round guides, noting the interrelation

$$|c_B| = \frac{|c_t|}{R\delta}. \tag{59}$$

However, the maximum conversion loss in the bend is also controlled by the quantity $(\beta_1 - \beta_2)$ as given in equation (31). For the three important modes, the values are

Mode	$|\beta_1 - \beta_2|$	
$\mathrm{TE}_{01}^0 - \mathrm{TE}_{11}^0$	$3.6 \dfrac{\lambda}{a^2}$	(60)
$\mathrm{TE}_{01}^0 - \mathrm{TE}_{12}^0$	$4.4 \dfrac{\lambda}{a^2}$	(61)
$\mathrm{TE}_{01}^0 - \mathrm{TM}_{11}^0$	0	(62)

where a is again the guide diameter. These are to be compared with equation (32) with $w = a$ and $p = 2$,

$$|\beta_1 - \beta_2| = 3.9 \frac{\lambda}{a^2}. \tag{63}$$

The approximation (63) agrees well with the values for the $TE_{01}^0 - TE_{11}^0$ and $TE_{01}^0 - TE_{12}^0$ from expressions (60) and (61). However expression (62) shows that empty round guide has a degeneracy, which controls its behavior.[2] The permitted bend radius is controlled by the $TE_{01}^0 - TM_{11}^0$ interaction. Exact theory shows the bend length to the first extinction of signal is[2]

$$L_\infty = 2.7 \frac{R\lambda}{a} , \tag{64}$$

which is to be compared with equation (44) with $w = a$ and $m = 1$,

$$L_\infty = 0.98 \frac{R\lambda}{a}. \tag{65}$$

Here the agreement is again quite good. The permitted bend radius for P_c fractional power loss, from exact theory is

$$R = \frac{0.58}{(P_c)^{\frac{1}{2}}} \frac{aL}{\lambda} , \tag{66}$$

and the approximation from equation (45) is

$$R = \frac{1.61}{(P_c)^{\frac{1}{2}}} \frac{aL}{\lambda}. \tag{67}$$

In practical use of round guides for TE_{01}, however, the bare pipe is modified to eliminate the degeneracy. Intentionally making the empty guide elliptical is one way;[3] it takes only 1.7 percent diameter difference to make $(\beta_1 - \beta_2)^2 = 10(\alpha_2 - \alpha_1)^2$, making the relations of Section 2.4 valid. A more symmetrical modification is to add a thin dielectric lining; with a polyethelyne lining only 0.010 inches thick in a 2 inch inner diameter guide, the $(\beta_1 - \beta_2)$ for $TE_{01}^0 - TM_{11}^0$ is about 60 percent of that given above for $TE_{01}^0 - TE_{12}^0$.[12] This also yields $(\beta_1 - \beta_2)^2 \gg (\alpha_2 - \alpha_1)^2$ for all modes. Interestingly, exact theory shows that the lining drops the $TE_{01}^0 - TE_{11}^0$ bend coupling coefficient by an order of magnitude.[12,13] Thus only two small mode conversions occur in the bend of lined waveguide. Taking the simple sum of these conversion losses yields, from this "exact" treatment,

$$P_c = 0.098 \frac{a^6}{R^2\lambda^4}. \tag{68}$$

The exact radius relation is then

$$R = \frac{0.31}{(P_c)^{\frac{1}{2}}} \frac{a^3}{\lambda^2}. \tag{69}$$

This is to be compared with equation (33) with $w = a$ and $p = 2$,

$$R = \frac{1.02}{(P_c)^{\frac{1}{2}}} \frac{a^3 \cdot}{\lambda^2} \tag{70}$$

Considering the complexity of the true situation the estimate provided by equation (70) is good.

4.4 Helix Waveguide for TE_{01}^0

The helix waveguide for TE_{01}^0 is a very special structure designed to maximize the attenuation to the undesired modes.[14,15] This waveguide is unusual in presenting very large $(\alpha_2 - \alpha_1)$. The bend coupling coefficients k of equations (26) and (27) are no longer pure imaginary as they were in the simple metallic tube. For example, the complex nature of the helix coupling coefficients are shown for comparison with those of a metallic tube; we set $k = c' + jc''$, as shown in Table II. The helix values correspond to a longitudinal wall impedance of 196 ohms with a capacitive angle of 5°, both guides at $\lambda = 5.4$ mm and a guide diameter of 5.08 cm.

The attenuation coefficient of the normal mode of the bend region is

$$\alpha_1 + \sum_{n}^{n} \text{Real}\left[\frac{k_n^2}{(\Gamma_1 - \Gamma_n)}\right] \tag{71}$$

where the summation represents the contributions of the three modes above. Using the helix waveguide coupling values of Table II, the conversion loss contributions are given in Table III. Note that the contributions of the TE_{12} and TM_{11} modes are of opposite sign; experiment agrees well with this theory.[16] An approximate degeneracy exists between TM_{11} and TE_{12} in the helix waveguide.

When such direct computations were made over a range of numerical conditions in the 30 to 100 GHz region on helix waveguides varying in diameter from 0.25 inch to 3 inches, it was found that the mode conversion contribution to the bend-region normal-mode at-

TABLE II—HELIX WAVEGUIDE COUPLING VALUES

Mode	Solid Metallic Tube		Helix Waveguide	
	$c'R$	$c''R$	$c'R$	$c''R$
TE_{11}	0	5.5	−0.16	6.86
TM_{11}	0	5.46	−8.03	−5.71
TE_{12}	0	9.21	−3.76	11.88

TABLE III—CONVERSION LOSS IN HELIX WAVEGUIDE

Mode	Real $\dfrac{R^2 k_n^2}{\Gamma_1 - \Gamma_n}$
TE$_{11}$	0.713
TM$_{11}$	8.79
TE$_{12}$	−8.05
$\Sigma =$	1.55

tenuation coefficient is approximately

$$\alpha_B = 0.009 \frac{a^3}{R^2 \lambda^{2.7}}, \tag{72}$$

which yields a permitted bend relation from direct solution of the helix problem:

$$R = \frac{0.095}{(\alpha_B)^{\frac{1}{2}}} \frac{a^{1.5}}{\lambda^{1.35}}. \tag{73}$$

The corresponding approximate relation from Section 3.5 is equation (38) with $w = a$ and $p = 2$,

$$R = 0.52 \left(\frac{\alpha_2 - \alpha_1}{\alpha_B} \right)^{\frac{1}{2}} \frac{a^3}{\lambda^2}. \tag{74}$$

To compare functional dependence on a and λ, we need to know how $(\alpha_2 - \alpha_1)^{\frac{1}{2}}$ [which is $(\alpha_2)^{\frac{1}{2}}$] varies with a and λ in the helix waveguide. Unfortunately this is not readily available although it was implicitly used in the work which yielded equation (72). However, a single numerical point is known: at $a = 5.08$ cm and $\lambda = 5.4$ mm, $\alpha_2 = 1.4$ nepers per meter for TM$_{11}$, which will control the guide behavior in equation (74). With these numbers equation (73) yields

$$R_{\text{exact}} \cong \frac{1.12}{(\alpha_B)^{\frac{1}{2}}} \tag{75}$$

whereas equation (74) yields

$$R_{\text{approx}} = \frac{2.76}{(\alpha_B)^{\frac{1}{2}}}. \tag{76}$$

The approximation is only off a factor of about two, which is remarkable and may be fortuitous. We suggest that equations (38) and (74) be considered provisional until proven or disproven by additional work.

4.5 *Curved Beam Guide*

Let us consider a curved beam guide made of a sequence of confocal lenses propagating the fundamental gaussian mode. The radius of curvature R_e, the wavelength λ, and the beam size w are found, with the help of equation (50), to be related to the maximum power conversion P_c by

$$R_e = \frac{\pi^2 w_o^3}{\lambda^2 (P_c)^{\frac{1}{3}}}. \qquad (77)$$

As in a previous example, the width of the guide containing 95 per cent of the power in the wanted mode is $a = 2w_o$; therefore,

$$R_e = \frac{1.23}{(P_c)^{\frac{1}{3}}} \frac{a^3}{\lambda^2}. \qquad (78)$$

This exact result compares with the approximate value from equation (33) with $w = a$ and $p = 1$,

$$R = \frac{1.36}{(P_c)^{\frac{1}{3}}} \frac{a^3}{\lambda^2}. \qquad (79)$$

Considering that the exact value relates to an open lens waveguide and the approximate one relates to a hollow metallic rectangular waveguide, the agreement is excellent.

V. DISCUSSION AND CONCLUSION

The direction-changing capability of electromagnetic waveguides may be limited by (*i*) radiation, if the guided field extends into an open freely propagating region, and (*ii*) mode conversion. Radiation is the limitation for single-mode open guides that have transverse fields extending indefinitely into a freely propagating region. An estimate of permitted bending radius may be made by using equations (15) or (3) and the knowledge of the field for the straight guide. For a straight guide transverse field decaying exponentially [$\exp(-x/\xi)$], the radiation attenuation coefficient in a bend of radius R was found to be of the form

$$\alpha_R = c_1 \exp(-c_2 R), \qquad (13)$$

where c_1 and c_2 are large constants. As a result, α_R is large for

$$R < \frac{1}{c_2} \log c_1 \qquad (14)$$

and small for R greater than that value.

When the guide supports higher order modes, mode conversion loss tends to be the controlling factor. In Section III formulas are developed for permissible bend radius R versus transverse field width a, the guided wavelength λ_z, and fractional power P_c lost to other modes. Numerous possible cases are treated, depending on the relation between the mode coupling coefficients k, the signal mode propagation coefficient $\Gamma_1 = \alpha_1 + i\beta_1$, and the propagation coefficient of the mode coupled to, in the bend $\Gamma_2 = \alpha_2 + i\beta_2$. A case which should be very common is one of small or moderate losses and gradual bends:

$$| k^2 | \ll (\beta_1 - \beta_2)^2, \qquad\qquad (28)$$

$$(\alpha_1 - \alpha_2)^2 \ll (\beta_1 - \beta_2)^2, \qquad\qquad (29)$$

$$\alpha_2 L \ll 1, \qquad\qquad (30)$$

where L is the length of the bend. Then an approximation for the bend radius permitted is

$$R = \frac{4.1}{(2p \pm 1)(P_c)^{\frac{1}{3}}} \frac{a^3}{\lambda^2}, \qquad\qquad (33)$$

and for the permitted abrupt tilt angle δ

$$\delta = 0.62 \, (P_t)^{\frac{1}{2}} \frac{\lambda_z}{a}, \qquad\qquad (24)$$

in which p is the number of extrema in the transverse field distribution. Examples are given in Sections 4.1 through 4.4 which show that known theory for several hollow metallic and open lens waveguides agree well with these expressions.

One must use caution in applying these expressions to new waveguides where the modes coupled in the bend are not known and, more importantly, where the phase constant differences are not known. If by design or misfortune a degeneracy exists between modes coupled by the bend, $\beta_1 = \beta_2$, a radically more severe restriction on bend R occurs. Sections 3.6 and 3.7 discuss this situation. However, since β's are large compared with typical α's, it usually is possible to avoid these restrictive conditions and justify equations (28) and (29) by small modifications of the guiding structure.

If the mode coupled to is very lossy, so that $\alpha_2 L \gg 1$, equation (33) does not hold. Section 3.5 and equation (38) relate to this case. We cite one example in Section 5.4 which supports equation (38); but more experience with coupling to lossy modes is needed.

APPENDIX A

Supplement to Section II

We note that the maintenance of equiphase differences on $F(x)$ for all x, but on radial planes differing by $\Delta\varphi$ (Fig. 1), requires

$$k_z R \ \Delta\varphi \geqq k_s(R + x_r) \ \Delta\varphi, \tag{80}$$

where k_s is the phase constant for a plane wave in the region surrounding the waveguide. For the equal sign in equation (80) a plane wave in the x_r region is traveling at the velocity of light and equation (80) yields

$$x_r = \frac{(k_z - k_s)}{k_s} R. \tag{81}$$

The energy traveling at $x > x_r$ is presumed lost to radiation, since to remain guided would imply energy traveling at greater than the velocity of light. The fraction of the total energy in the cross section at $x > x_r$ is $\mathcal{E}_l/\mathcal{E}_T$, where $\mathcal{E}_l$ and $\mathcal{E}_T$ are given by equations (4) and (5). How rapidly, as a function of distance along the direction of propagation, does energy flow out from the main energy packet to this region at $x > x_r$? For a wave in an infinite uniform medium the energy remains collimated for a distance

$$z_c = \frac{a^2}{2\lambda_s}, \tag{82}$$

where a is the transverse field width and λ_s is the wavelength in that medium. It may be expected that an approximate distance z_c would be required for energy to flow out from the guided field of the same width a. Noting a power decay rate $e^{-2\alpha z} \approx 1 - 2\alpha z$, the fractional power loss becomes

$$\frac{\mathcal{E}_l}{\mathcal{E}_T} = 2\alpha_r z_c \tag{83}$$

or

$$\alpha_r = \frac{1}{2z_c} \frac{\mathcal{E}_l}{\mathcal{E}_T}. \tag{84}$$

Numerical Evaluations of a Specific Case

The potential usefulness of equation (3) is in estimating radiation losses of curved open waveguides for which the straight-guide fields are known, but for which a solution in the curved coordinate system is not

known. Here we compare the results of using equation (3) with the results of a direct solution, to obtain an indication of the accuracy that might be expected in other cases. The case is defined by equations (8) and (9), which lead to equations (10), (11), and (12) for $\mathcal{E}_l$, $\mathcal{E}_T$, and z_c.

We provide additional expressions needed in the numerical calculations: from known theory[5,16]

$$\frac{1}{\xi} = [k^2(n_1^2 - n_3^2) - k_x^2]^{\frac{1}{2}} \tag{85}$$

where k is the free space wave number, n_1 is the index of refraction of the dielectric slab, and n_3 is in the index of the surrounding region. The quantity k_x may be obtained graphically as a function of t/A and is reproduced here in Fig. 5, from Ref. 5. The quantity A is the value of t at which the second propagating mode appears,

$$A = \frac{\pi}{k(n_1^2 - n_3^2)^{\frac{1}{2}}} \frac{\lambda}{2(n_1^2 - n_3^2)^{\frac{1}{2}}}. \tag{86}$$
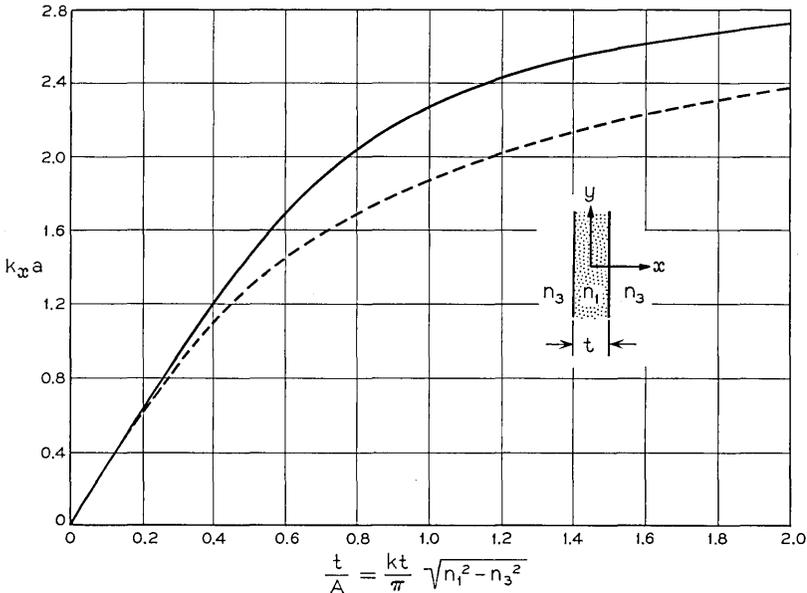
Fig. 5 — Normalized transverse wave number $k_x a$ versus normalized thickness t/A for a two-dimensional dielectric waveguide. ———— fundamental mode polarized perpendicular to the dielectric sheet and $n_3 = n_1/1.5$; — — — — fundamental mode polarized perpendicular to the sheet and $n_1/n_3 - 1 \ll 1$, or fundamental mode polarized parallel to the sheet and n_3 arbitrary.

Also known are [5,9]

$$k_z^2 = n_1^2 k^2 - k_x^2 , \tag{87}$$

$$k_z^2 = n_3^2 k^2 + \frac{1}{\xi^2} , \tag{88}$$

and in the region of considerable interest where

$$k_x \ll k n_1 , \tag{89}$$

the approximations

$$k_z = k n_1 - \frac{1}{2} \frac{k_x^2}{k n_1} \tag{90}$$

$$k_z - k n_3 \equiv k_z - k_s = k(n_1 - n_3) - \frac{1}{2} \frac{k_x^2}{k n_1} \tag{91}$$

are valid. Using the above relations, one can calculate α_r, given t, λ, n_1, and n_3.

Table IV lists the principal parameters and a comparison with more exact theory for several cases. In Table IV the first five columns define the waveguide; c_1 and c_2 are values from equation (13), found in turn by evaluating equations (3) through (7). The table also lists the radiation attenuation coefficient α_r, the estimate of the associated bend radius R, the value of R from Ref. 5, and the ratio. The estimate from equation (3) is consistently lower than the true required R (in the approximate ratio 0.6) for a wide range of index differences $(n_1 - n_3)$ and bend radii R.

The table also lists the transverse distance x_r at which the velocity of light condition occurs. It is interesting that it is so close to the waveguide.

Additional support for the approximate calculation based on equation (3) comes from an additional case. It is readily verified from exact theory that the case 1 condition, $n_1 = 1.5$ and $n_3 = 1.0$, yields different radiation losses for the two polarizations of wave if the thickness t is fixed. However, if t is adjusted to give the same external field decay constant ξ of equation (9), then the radiation losses are the same for the two polarizations of wave.

APPENDIX B

Solutions of the Coupled-Wave Equations (26) and (27)

If one assumes that the coupling coefficient k in equation (26) and (27) is pure imaginary, $k = ic$, one can express the fractional power

TABLE IV—TABULATION OF IMPORTANT PARAMETERS IN CURVED DIELECTRIC WAVEGUIDES

Case	t (10^{-6} m)	$\dfrac{t}{A}$	n_1	n_3	c_1	c_2	R (meter)	α_r (neper/m)	x_r (10^{-6} m)	R from Ref. 5 for same α_r (meter)	Ratio = $\dfrac{\text{col (8)}}{\text{col (11)}}$
1*	0.198	0.7	1.5	1.0	2.57 × 10^6	3.47 × 10^6	3.54 × 10^{-6}	11.6	0.846	5.49 × 10^{-6}	0.645
							6.15 × 10^{-6}	1.34 × 10^{-3}	1.47	11 × 10^6	0.656
2†	0.372	0.25	1.5	1.485	0.46 × 10^5	2.570	4.17 × 10^{-3}	1.0	6.5	6.89 × 10^{-3}	0.606
3†	1.04	0.7	1.5	1.485	1.037 × 10^5	1.46 × 10^4	0.807 × 10^{-3}	0.776	4.0	1.18 × 10^{-3}	0.68
							1.43 × 10^{-3}	0.895 × 10^{-4}	7.09	2.37 × 10^{-3}	0.60
4†	1.79	1.20	1.5	1.485	1.46 × 10^5	2.55 × 10^4	0.355 × 10^{-3}	16.9	2.55	0.593 ± 10^{-3}	0.6
							0.442 × 10^{-3}	1.89	3.18	0.711 × 10^{-3}	0.62
							0.6 × 10^{-3}	0.0321	4.31	0.948 × 10^{-3}	0.63
							0.938 × 10^{-3}	5.97 × 10^{-4}	6.75	1.18 × 10^{-3}	0.79
5†	2.38	1.6	1.5	1.485	2.18 × 10^5	3.04 × 10^4	0.336 × 10^{-3}	7.9	2.7	0.593 × 10^{-3}	0.566
							0.423 × 10^{-3}	0.568	3.4	0.711 × 10^{-3}	0.59
							0.585 × 10^{-3}	3.97 × 10^{-3}	4.7	0.948 × 10^{-3}	0.617
6†	1.18	0.25	1.5	1.4985	0.543 × 10^4	81.4	0.106	1.0	16.5	0.18	0.59

* The electric field is parallel to the dielectric slab.
† Applies for either polarization.

P_c converted out of the signal (that is case 1) mode as

$$P_c = 1 - \frac{\exp\,[(\alpha_1 - \alpha_2)z]}{4} \left| \left[1 - \frac{1}{(1 + \kappa^2)^{\frac{1}{2}}} \right] \exp \left[i\left(1 + \frac{1}{\kappa^2}\right)^{\frac{1}{2}} cz \right] \right.$$

$$\left. + \left[1 + \frac{1}{(1 + \kappa^2)^{\frac{1}{2}}} \right] \exp \left[-i\left(1 + \frac{1}{\kappa^2}\right)^{\frac{1}{2}} cz \right] \right|^2, \tag{92}$$

where

$$\kappa = \frac{i2c}{\Gamma_1 - \Gamma_2} \tag{93}$$

and

$$\Gamma_1 = \alpha_1 + i\beta_1 . \tag{94}$$
$$\quad {\scriptstyle 2} \quad {\scriptstyle 2} \quad {\scriptstyle 2}$$

In these formulas, Γ_1 and Γ_2 are the propagation constants of the wanted and spurious modes, respectively; in general, they are complex and their real parts, α_1 and α_2, are the attenuation constants; their imaginary parts, β_1 and β_2, are the phase constants. We bear in mind that $k = ic$ has only been proven valid in lossless waveguides, and for one case of coupling to a lossy mode (helix waveguide) k is complex.

Another useful expression is for the signal wave amplitude E_1 when the coupling k is small compared with $(\Gamma_1 - \Gamma_2)$, or more specifically,

$$| 4k^2 | \ll (\Gamma_1 - \Gamma_2)^2 \tag{95}$$

and

$$| k^2 | \ll | \Gamma_2(\Gamma_1 - \Gamma_2) | . \tag{96}$$

Then we may write

$$E_1 \cong \exp\,(-\Gamma_1 z)\left\{\left[1 - \frac{k^2}{(\Gamma_1 - \Gamma_2)^2} \right] \exp \left[-\frac{k^2 z}{(\Gamma_1 - \Gamma_2)} \right] \right.$$

$$\left. + \frac{k^2}{(\Gamma_1 - \Gamma_2)^2} \exp\,[(\Gamma_1 - \Gamma_2)z]\right\}. \tag{97}$$

The first term corresponds to the low-loss normal mode of the coupled region, and the second term to the high-loss mode (we assume $\alpha_2 > \alpha_1$).

For Section 3.4, it is valid to take $k = ic$; equation (92) yields a conversion loss of

$$P_c = \left(\frac{2c}{\beta_1 - \beta_2}\right)^2 \sin^2 \left[\frac{(\beta_1 - \beta_2)z}{2} \right]. \tag{98}$$

For Section 3.5 we use equation (97), keeping a complex k; note that for $\alpha_2 L \gg 1$, only the first term remains significant and the propagation constant of the normal mode is

$$\Gamma_1 + \frac{k^2}{(\Gamma_1 - \Gamma_2)}. \tag{99}$$

This yields equation (36) for α_B, the added attenuation resulting from the bend.

For Section 3.7 we again use equation (97); the first term predominates with the assumption

$$\frac{k^2}{(\Gamma_1 - \Gamma_2)} \ll 1 \tag{100}$$

and equation (99) yields equation (48).

For Section 3.6, the case of low-loss modes degenerately coupled, equation (92) yields

$$P_c \cong 1 - \exp\left[(\alpha_1 - \alpha_2)z\right] \left| \cos\left(cz + \frac{i}{\kappa}\right) \right|^2. \tag{101}$$

It is also well known that the signal amplitude is given by[2-4]

$$|E_1| = |\cos cz|, \tag{102}$$

the undesired mode amplitude by

$$|E_2| = |\sin cz| \tag{103}$$

and the fractional conversion loss P_c by

$$P_c = \sin^2 cz. \tag{104}$$

APPENDIX C

Supplementary Information Concerning the
Derivation of Equation (22)

Carrying out the integration of equation (16) for the rectangular metallic waveguide as outlined in Section 3.1 yields a conversion loss resulting from the tilt of

$$P_t = B\left(\frac{\delta w}{\lambda_z}\right)^2, \tag{105}$$

where

$$B = \frac{\pi^2}{3}\left[1 \pm \frac{\left(\frac{p}{w}\right)^2 - \left(\frac{q}{b}\right)^2}{\left(\frac{p}{w}\right)^2 + \left(\frac{q}{b}\right)^2}\frac{6}{\pi^2 p^2}\right].$$ (106)

The $+$ or $-$ sign corresponds to the TE_{pq} or TM_{pq} modes, respectively. For the lowest order TE mode, $p = 1$ and $q = 0$, B becomes 5.28. For the TE or TM mode with $p = 1$ and $q = 1$, B ranges from 5.28 to 1.28 as the dimensions of the guide vary between $w \ll b$ and $w \gg b$. The limits on B are 5.28 and 1.28 for any p or q. We somewhat arbitrarily chose a value $(5.28 \times 1.28)^{\frac{1}{2}} = 2.6$ to represent all modes simultaneously.

REFERENCES

1. Miller, S. E., "Directional Control in Light-Wave Guidance," B.S.T.J., *43* No. 4 (July 1964), pp. 1727–1739.
2. Jouguet, M., "Effects of the Curvature on the Propagation of Electromagnetic Waves in Guides of Circular Cross Section," Cables and Transmission (Paris), *1*, No. 2 (July 1947), pp. 133–153.
3. Miller, S. E., "Notes on Methods of Transmitting the Circular Electric Wave Around Bends," Proc. I.R.E., *40*, No. 5 (September 1952), pp. 1104–1113.
4. Miller, S. E., "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 3 (May 1954), pp. 661–719.
5. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., this issue, pp. 2103–2132.
6. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., this issue, pp. 2059–2069.
7. Schelkunoff, S. A., "Conversion of Maxwell's Equations into Generalized Telegraphist's Equations," B.S.T.J., *34*, No. 5 (September 1955), pp. 995–1043.
8. Unger, H. G., "Normal Mode Bends for Circular Electric Waves," B.S.T.J., *36*, No. 5 (September 1957), pp. 1292–1307.
9. Collin, R. E., *Field Theory of Guided Waves,* New York: McGraw Hill, 1960.
10. Rowe, H. E., and Warters, W. D., "Transmission in Multi-Mode Waveguide with Random Imperfections," B.S.T.J., *41*, No. 3 (May 1962), pp. 1031–1170.
11. Miller, S. E., "On Solutions for Two Waves with Periodic Coupling," B.S.T.J., *47*, No. 8 (October 1968), pp. 1801–1822.
12. Marcatili, E. A. J., "Effects of Redirectors, Refocusers, and Mode Filters on Light Transmission through Aberrated and Misaligned Lenses," B.S.T.J., *46*, No. 8 (October 1967), pp. 1733–1752.
13. Unger, H. G., "Lined Waveguide," B.S.T.J., *41*, No. 2 (March 1962), pp. 745–768.
14. Unger, H. G., "Normal Modes and Mode Conversion in Helix Waveguide," B.S.T.J., *40*, No. 1 (January 1961), p. 255; also "Helix Waveguide Theory and Application," B.S.T.J., *37*, No. 6 (November 1958) pp. 1599–1647.
15. Morgan, S. P., and Young, J. A., "Helix Waveguide," B.S.T.J, *35*, No. 6 (November 1956), pp. 1347–1384.
16. Young, D. T., "Measured TE_{01} Attenuation in Helix Waveguide with Controlled Straightness Deviations," B.S.T.J., *44*, No. 2 (February 1965), pp. 273–282.

Some Theory and Applications of Periodically Coupled Waves

By STEWART E. MILLER

(Manuscript received February 6, 1969)

Parallel-traveling waves can interact with complete power transfer even though they have different phase constants, provided that the coupling is periodic. This paper outlines some possible applications of this phenomenon, including mode transforming devices, frequency-selective filters in the microwave and laser wavelength regions, and parametric amplifiers or converters. This paper also gives some coupled-wave equations for interactions in a nonlinear medium and a generalization of the Tien conditions for parametric wave interaction.

I. INTRODUCTION

In a previous paper it was shown that two parallel-traveling coupled waves can interact with complete power interchange even though they have different phase constants.[1] This is accomplished by introducing a variation in coupling in the direction of wave propagation. The ideal coupling variation is a pure phase variation whose period exactly matches the beat period between the uncoupled waves, however, it was also shown in that paper that a simple periodic magnitude variation of the coupling can also yield complete power interchange between waves having different phase constants.

In this paper we outline some of the possible applications of periodic coupling. Complete power exchange between two modes of a single hollow metallic waveguide is illustrated. In two dielectric or hollow metallic waveguides, or in a combination of them, complete power exchange (or a desired fractional exchange) can be arranged. Frequency selective filters in the above structures can be obtained or broadband interactions can be chosen by suitable design. The periodic coupling phenomenon can be applied in lumped element parametric devices by modulating the pump waveform periodically; we give the

resulting conditions that the signal frequency, idler frequency, pump frequency, and modulation frequency must fulfill.

Finally, in distributed parametric devices the periodic coupling principle can be used to advantage; spatial variation of the coupling gives a modified phase-matching relation that may render useful long lengths (with guided waves or unguided waves) of materials not useful with previous vectorial phase matching relations; time modulation of the pump introduces new frequency relations of possible use in modulators or frequency translators. The frequency range in which such applications may be useful extends from the laser region to the lowest frequency at which distributed coupled-wave interactions are convenient.

Section II presents some theory needed to understand the device illustrations. In Appendices A and B and in the discussion of parametric devices, we develop some coupled-wave equations to facilitate analysis of nonlinear circuits with generalized time- and space- dependent couplings. This paper is a survey of potential applications and is intended as a stimulus for further work. Complete design relations and experimental verification are not included.

II. GENERAL THEORY

We deal with devices or situations in which two waves of amplitude E_1 and E_2 are coupled according to

$$\frac{d}{dz} E_1(z) = -\gamma_1 E_1 + c_{21}(z)E_2 \tag{1}$$

$$\frac{d}{dz} E_2(z) = -\gamma_2 E_2 + c_{12}(z)E_1 \tag{2}$$

in which γ_1 and γ_2 are the complex propagation constants and c_{12} and c_{21} are coupling functions. In a previous paper we showed that the coupling distributions summarized in Table I lead to wave interactions virtually the same as those which are familiar for c_{21} and c_{12} independent of z, provided that transformations for coupling magnitude c_* and differential phase constant $\Delta\beta_*$ are appropriately defined. For $E_1 = 1.0$ and $E_2 = 0$ at $z = 0$ the solutions for equations (1) and (2) are

$$E_1(z) = \exp\left(-\gamma_1 z\right)[A \exp\left(r_1 z\right) + B \exp\left(r_2 z\right)] \tag{3}$$

$$E_2(z) = \frac{\exp\left(-\gamma_1 z\right)}{2\sqrt{}} [\exp\left(r_1 z\right) - \exp\left(r_2 z\right)] \tag{4}$$

TABLE I — VALUES OF c_* AND $\Delta\beta_*$ FOR VARIOUS
PERIODIC COUPLING FUNCTIONS

Coupling Type	Coupling Definition	c_*	$\Delta\beta_*$
1	$c_{12} = c_{21} = jc$	c	$\Delta\beta = \beta_1 - \beta_2$
2	$c_{12} = jc \exp\left(-j\dfrac{2\pi z}{\lambda_m}\right)$ $c_{21} = jc \exp\left(j\dfrac{2\pi z}{\lambda_m}\right)$	c	$\Delta\beta - \dfrac{2\pi}{\lambda_m}$
3	$c_{12} = c_{21} = jc \sin\left(\dfrac{2\pi z}{\lambda_m}\right)$	$\dfrac{c}{2}$	$\Delta\beta - 2\pi/\lambda_m$
4	symmetrical square wave $c_{12} = c_{21} = jc$ for $n\lambda_m$ $\qquad < z < \lambda_m(n + \tfrac{1}{2})$ $c_{12} = c_{21} = -jc$ for $(n + \tfrac{1}{2})\lambda_m$ $\qquad < z < (n+1)\lambda_m \quad n = 0, 1, 2, \cdots$	$\dfrac{2}{\pi}c$	$\Delta\beta - \dfrac{2\pi}{\lambda_m}\left[1 - \left(\dfrac{c\lambda_m}{\pi}\right)^2\right]^{\frac{1}{2}}$
5	raised square wave $c_{12} = c_{21} = j2c$ for $n\lambda_m$ $\qquad\qquad\qquad < z < \lambda_m(n + \tfrac{1}{2})$ $\Big\}$ $c_{12} = c_{21} = 0$ for $(n + \tfrac{1}{2})\lambda_m < z$ $\Big]$ $\qquad\qquad < (n + 1)\,\lambda_m \ n = 0, 1, 2, \cdots$	$\dfrac{2}{\pi}c$ c	$\Delta\beta - \dfrac{2\pi}{\lambda_m}\left[1 - \left(\dfrac{c\lambda_m}{\pi}\right)^2\right]^{\frac{1}{2}}$ $\Delta\beta$

in which

$$A = \frac{1}{2} - \frac{1}{2}\frac{\left(\dfrac{\Delta\beta_*}{2c_*}\right) - i\left(\dfrac{\Delta\alpha}{2c_*}\right)}{\sqrt{}} \tag{5}$$

$$B = \frac{1}{2} + \frac{1}{2}\frac{\left(\dfrac{\Delta\beta_*}{2c_*}\right) - i\left(\dfrac{\Delta\alpha}{2c_*}\right)}{\sqrt{}} \tag{6}$$

$$r_{\frac{1}{2}} = \frac{\Delta\gamma_*}{2} \pm ic_*\sqrt{} \tag{7}$$

$$\sqrt{} = \left[1 + \left(\frac{\Delta\beta_*}{2c_*}\right)^2 - \left(\frac{\Delta\alpha}{2c_*}\right)^2 - i2\left(\frac{\Delta\beta_*}{2c_*}\right)\left(\frac{\Delta\alpha}{2c_*}\right)\right]^{\frac{1}{2}} \tag{8}$$

$$\Delta\gamma_* = (\alpha_1 - \alpha_2) + i\Delta\beta_*$$

$$\gamma_1 = \alpha_1 + i\beta_1$$

$$\gamma_2 = \alpha_2 + i\beta_2 \tag{9}$$

$$\gamma_1 - \gamma_2 = \Delta\alpha + i\Delta\beta.$$

In Table I we define the quantities c_* and $\Delta\beta_*$; λ_m is the wavelength of the coupling variation as defined in the second column of Table I.

In Table I, type 1 coupling is the familiar uniform coupling, independent of z. For negligible attenuation and for $\Delta\beta = 0$ the wave energy is exchanged cyclically between the two waves according to

$$E_1 = \cos (cz) \tag{10}$$

$$E_2 = i \sin (cz); \tag{11}$$

and for other values of $\Delta\gamma$ limited wave interactions occur. This has been described previously.[2]

In Table I, type 2 coupling corresponds to the exact transformations given for c_* and $\Delta\beta_*$; the other type couplings correspond to the approximate values given for c_* and $\Delta\beta_*$. For coupling types 1 and 2, equations (3) and (4) give exactly the coupled-wave amplitudes; for coupling types 3 and 4, equations (3) and (4) give the coupled wave amplitudes exactly at z equal to a multiple of $\lambda_m/2$, and may be in error by no more than about $0.2c\lambda_m/\pi$ at other values of z. The error may be slightly larger for coupling type 5, but is negligible for small $c\lambda_m$.

Figure 1 shows the initial buildup of the wave amplitude E_2 for coupling types 4 and 5. At $z = \lambda_m/2$ further extension of uniform coupling would result in added components to E_2 at such a phase as to diminish E_2. By reversing the sign of the type 4 coupling, the added components in the region $0.5\ \lambda_m < z < \lambda_m$ cause an increase in E_2. By reducing the magnitude of the type 5 coupling to zero at $\lambda_m = 0.5$, no components are added to E_2 in the region $0.5\ \lambda_m < z < \lambda_m$. At $z = \lambda_m$ the cycle repeats. In this way the amplitude variation in coupling versus z causes an average in-phase transfer of energy. The same behavior exists for an arbitrary amplitude variation of coupling $c(z)$; the fundamental Fourier component may be taken as the type 3 coupling and the resulting wave interaction calculated. The result is accurate provided that $c_p\lambda_m \ll 1$, where c_p is the peak of the coupling waveform.

III. FREQUENCY SENSITIVITY

In many coupled-wave devices the objective is to transfer all of the power from one wave to the other, and frequency sensitivity may be desirable (as in channel-selecting filters of a communication system) or may be undesirable. We show the magnitude of this frequency sensitivity.

Consider first two dielectric waveguides where most of the energy

travels in the central dielectrics designated n_1 and n_2 (indexes of refraction) in Fig. 2. Periodic coupling is induced by the dielectric sheets labeled n_3, corresponding to type 5 coupling in Table I. Then, approximately

$$\Delta\beta = \frac{2\pi}{\lambda} (n_1 - n_2) \tag{12}$$

in which λ is free space wavelength. We assume the complete transfer condition, which is

$$c_* L = \frac{\pi}{2} \tag{13}$$

with L being the length of the coupling region. Also let

$$L = N\lambda_m \tag{14}$$

with

$$\lambda_m = \frac{2\pi}{\Delta\beta_o} \tag{15}$$

and $\Delta\beta_o$ defined as $\Delta\beta$ at the midband frequency $f = f_o$. Now $\Delta\beta_*$ as a
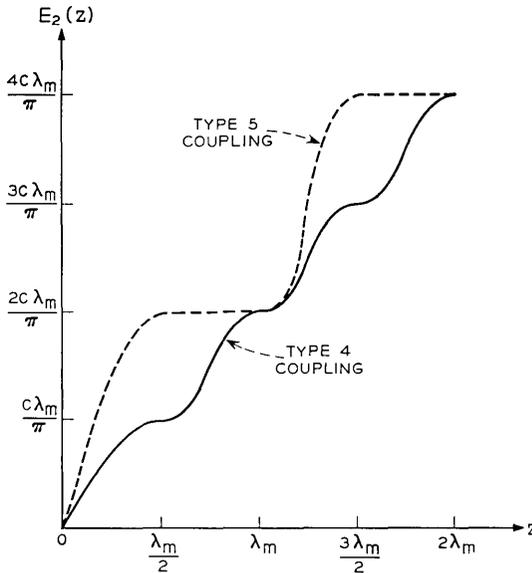
Fig. 1 — Transferred wave amplitude E_2 versus length of coupling region for type 4 and type 5 coupling (see Table I).

function of frequency is

$$\Delta\beta_*(f) = \Delta\beta(f) - \Delta\beta(f_o). \tag{16}$$

Expressing the frequency as a deviation from f_o

$$f = (1 + \delta)f_o, \tag{17}$$

we find

$$\Delta\beta_* = \frac{2\pi \delta}{\lambda_o}(n_1 - n_2) \tag{18}$$

with λ_o equal to λ at $f = f_o$. Using equations (18), (13), and (14) and assuming the typical case of negligible dependence of c_* on frequency, we find

$$\frac{\Delta\beta_*(f)}{c_*} = 4 \delta N. \tag{19}$$

This ratio uniquely determines the frequency sensitivity of the wave interaction, according to

$$|E_2| = \frac{1}{\left[\left(\frac{\Delta\beta_*}{2c_*}\right)^2 + 1\right]^{\frac{1}{2}}} \left| \sin\left\{\left[\left(\frac{\Delta\beta_*}{2c_*}\right)^2 + 1\right]^{\frac{1}{2}} c_* L\right\} \right| \tag{20}$$
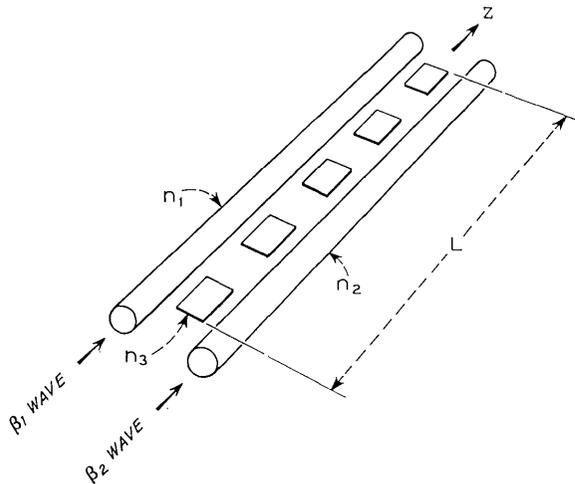
Fig. 2 — Dielectric waveguides (having indices of refraction n_1 and n_2) with periodic coupling.
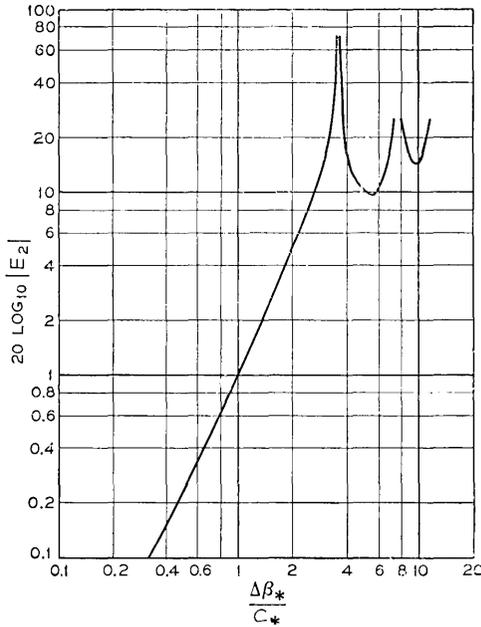
Fig. 3 — Transferred wave amplitude E_2 versus $\Delta\beta_*/c_*$, the frequency dependent parameters, for $c_*L = \pi/2$.

which follows from equation (4) with $\Delta\alpha = 0$. With complete transfer conditions $c_*L = \pi/2$ and with λ_m chosen to make $\Delta\beta_* = 0$ at $f = f_o$, equation (20) becomes unity at $f = f_o$ and falls off as $\Delta\beta_*(f)$ differs from zero, that is, as δ differs from zero in equation (17). Figure 3 shows E_2 versus $\Delta\beta_*/c_*$ for $c_*L = \pi/2$; values for this graph can be calculated from equation (20)[†]. Using these results and equation (19) we find the bandwidth properties of the periodically coupled wave interaction on dielectric waveguides. A few examples are listed in Table II. The first three rows illustrate broadband coupling; as long as N (the number of coupling periods in the total coupling length L) is five or less, very little variation from the complete transfer condition occurs. The fourth row illustrates that intentional frequency selectivity can be induced by using a large N; the 0.2 percent band at $N = 865$ yields $\Delta\beta_*/c_* = 3.46$, the location of the first null in Fig. 3. Structures analogous to Fig. 2 but actually fabricated in a solid sheet continuum are under consideration for laser beam circuitry. If a 20Å bandwidth to the first nulls is desired

[†] For $\Delta\beta_*/c_* < 2$, $20 \log_{10} | E_2 | \cong -1.1 \, |\Delta\beta_*/c_*|^2$ dB.

TABLE II — BANDWIDTH PROPERTIES OF PERIODICALLY
COUPLED WAVES ON DIELECTRIC WAVEGUIDES

Percentage Bandwidth $(200\,\delta)$	N	$20 \log |E_2|$ Band Edge Loss (dB)
10	1	-0.04
10	3	-0.36
10	5	-1.1
0.2	865	$-\infty$

at $10{,}000\,\text{Å}$ midband and if $(n_1 - n_2) = 0.1$, we find $\lambda_m = 10\ \mu\text{m}$ and the coupling length $L = 8.65$ mm. Frequency selectivity obtained in this way does not require low heat loss in the circuit; as long as the two waves have the same attenuation coefficient, loss does not limit the filter selectivity.

For waves in an infinite medium or in other types of waveguides, equation (20) remains valid but relations other than (19) must be found to describe the way $\Delta\beta_* / c_*$ varies with frequency. For waves in hollow metallic tubes the results are very similar to those for waves on dielectric rods. We show this with two illustrative examples as follows.

In any hollow metallic waveguide the phase constant of a mode is given by

$$\beta = \frac{2\pi}{\lambda} [1 - \mu^2]^{\frac{1}{2}} \tag{21}$$

where

λ = free space wavelength,
$\mu = f_c / f$,
f_c = cutoff frequency for the particular mode, and
f = operating frequency.

By defining $\lambda_o = \lambda$ at $f = f_o$

$$\mu_{10} = \mu \text{ for wave 1 at } f = f_o$$
$$\mu_1 = \mu \text{ for wave 1 at } f = f_o(1 + \delta);$$

and using similar definitions (not written out) for wave number 2, we find

$$\Delta\beta_*(f) = \Delta\beta(f) - \Delta\beta(f_o)$$

$$= \frac{2\pi}{\lambda_o} (1 + \delta)[(1 - \mu_1^2)^{\frac{1}{2}} - (1 - \mu_2^2)^{\frac{1}{2}}]$$

$$- \frac{2\pi}{\lambda_o} [(1 - \mu_{10}^2)^{\frac{1}{2}} - (1 - \mu_{20}^2)^{\frac{1}{2}}]. \tag{22}$$

To develop a physical model, we take parameters typical of a 24,000 MHz $TE_{10}^{\square} - TE_{01}^{\circ}$ transducer similar to one described in connection

with Fig. 42 of Ref. 2. We keep the same coupling length $L = 0.417$ m for complete transfer of power, corresponding to $c_* = 3.76$ m^{-1}. We arbitrarily choose to explore the bandwidth when $\lambda_m = L/3 = 0.139$ m. We keep the same rectangular guide width, 0.340 inches, which at $f_o = 24{,}000$ MHz gives $\mu_{10} = 0.723$. This determines that $\mu_{20} = 0.625$; there is a round guide diameter of 0.96 inches (optionally a particular μ_{20} larger than 0.723 could have been selected to give the same $|\; \Delta 2(f_o)\; |$ and λ_m). We can now calculate $\Delta\beta_*(f)/c_*$ from equation (22), neglecting variations in c_* for this estimate. For a 10 percent frequency band, that is, $\delta = 0.05$, we find $\Delta\beta_*/c_* = 1.01$ and the loss $20 \log_{10} E_2 = 1.1$ dB. The case, $N = L/\lambda_m = 3$, thus yields a result very similar to that obtained for dielectrically guided waves using equations (19) and (20) and shows broadband interaction capability for waves in guided tubes provided N is not too large. Sections V and VI discuss some factors which may motivate one to use periodic coupling instead of constant coupling.

Consider a second example in hollow metallic guides to illustrate intentional frequency selectivity. Assume we need a filter with center frequency $f = 50$ GHz and a 3 dB bandwidth of 1000 MHz. Then $\delta = 0.01$ and from equation (20) or Fig. 3 we find $(\Delta\beta_*/c_*) \cong 1.6$. We keep one wave at $\mu_{10} = 0.723$ as before and choose $\mu_{20} = 0.91$. We can calculate $\Delta\beta_*$ from these choices using equation (22) which yields $\Delta\beta_* = 8.95$ m^{-1} at $f = 1.01 f_o$. At this frequency we need $(\Delta\beta_*/c_*) = 1.6$, so c_* needs to be 5.58 m^{-1} and complete transfer at f_o (that is, $c_* L = \pi/2$), requires $L = 0.28$ m. These are reasonable values physically; Section IV illustrates possible coupling and waveguide cross-sectional geometries. We now note that $N = L/\lambda_m$ for this case is 12.7. The same values of $\delta(0.01)$ and $N(12.7)$ for a dielectrically guided wave pair yield from equation (19) $\Delta\beta_*/c_* = 5.1$, indicating somewhat more selectivity in the dielectrically guided waves than in the hollow-tube guided waves, for the same number of coupling periods N.

IV. STRUCTURES FOR PASSIVE WAVE INTERACTIONS

We describe a few structures in which guided waves may be coupled periodically. The general diagram is given in Fig. 4. Most typically there is no input to wave 2 in this discussion although the transformations of Table I and equations (1) and (2) may be used to treat general inputs to the periodically coupled region. In some cases the two waves occupy the same space as discrete modes of a single structure. In other cases separate guiding structures for the two waves are provided.

In Ref. 1 a structure is described for hollow metallic waveguide
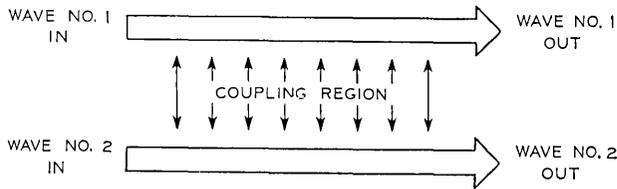
Fig. 4 — Two coupled waves; the dimension for the coupling region may be distance or time.

$TE_{10}^{\square} - TE_{01}^{\circ}$ coupling which closely approximates type 2 coupling and yields the simple transformation for $\Delta\beta_*$ of Table I without "harmonic" transformations for $\Delta\beta_*$. The harmonic transformations, discussed fully in Ref. 1, are characteristic of square-wave or sinusoidal coupling patterns and may yield appreciable wave interactions when $\Delta\beta_* \cong \Delta\beta p/\lambda_m$ with p an odd integer. The exponential type 2 coupling is thus a desirable one. However, because the harmonic interactions are weaker than the fundamental and may occur at greatly different frequencies, the square-wave and sinusoidal couplings are useful.

Figure 2 shows two dielectric waveguides periodically coupled with dielectric sheets yielding type 5 coupling of Table I. Its possible use as a frequency selective filter has already been referred to. Figure 5 shows the form it might take in laser circuitry where λ_m of 10 μm could be sought using photolithographic techniques; the substrate index n_* is to be less than n_1 and n_2 .[3]

Figures 6 and 7 illustrate the way two modes of a single hollow metallic waveguide can be coupled periodically to achieve complete or partial
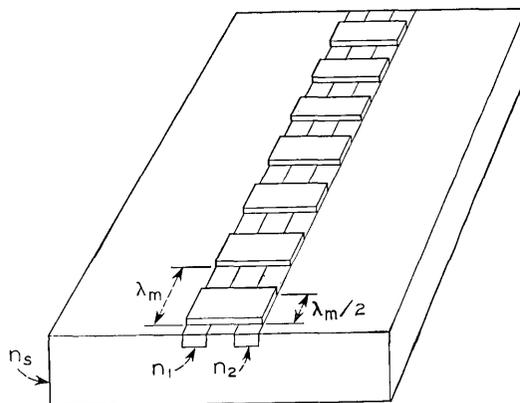
Fig. 5 — Periodically coupled dielectric waveguides.
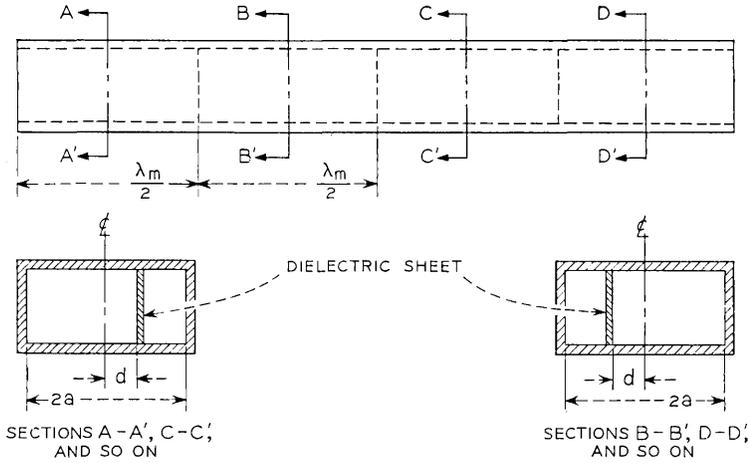
Fig. 6 — Periodic coupling structure for waves in a hollow, rectangular, metallic waveguide.

power interchange. In Fig. 6 the $TE_{10}^{\square}$ and $TE_{20}^{\square}$ modes are coupled by the dielectric sheet. The fields of these modes in a transverse plane are sketched in Fig. 8; a thin dielectric sheet introduces maximum coupling at a distance $d = 0.392a$, where the product of the two fields is a maximum. The coupling between the modes is reversed by moving the sheet to the opposite side of the guide centerline, as in section $B - B'$ of Fig. 6. A similar maximum coupling position can be found for the $TE_{11}^{\circ} - TE_{01}^{\circ}$ coupling, the fields for which are sketched in Fig. 9;
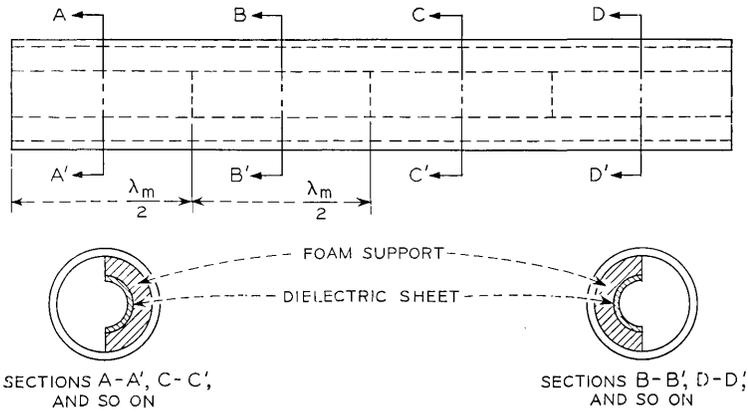
Fig. 7 — Periodic coupling structure for waves in a hollow, round, metallic waveguide.
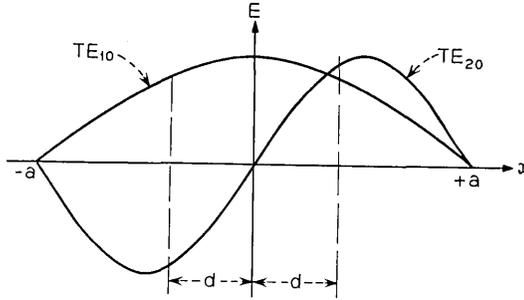
Fig. 8 — Transverse field distributions for $TE_{10}{}^{\square}$ and $TE_{20}{}^{\square}$.

Fig. 7 shows the structural form of coupler. In both Figs. 6 and 7 the length $\lambda_m/2$ is that at which the coupled modes develop π radians phase difference. This length is near that for π radians phase difference in an empty guide, which for Fig. 7 is approximately one diameter. (Specifically, in a $\frac{1}{3}$ inch-inside diameter guide at 54 GHz the half-beat wavelength for $TE_{11}{}^{\circ} - TE_{01}{}^{\circ}$ is about $\frac{1}{3}$ inch.) Structures of the type in Figs. 6 and 7 provide mode transformation without complicated and expensive shaping of the metallic walls.

Figures 10 and 11, which show the transverse cross sections of the guides, illustrate coupling between modes of different hollow metallic waveguides. Although $TE_{10}{}^{\square} - TE_{01}{}^{\circ}$ and $TE_{01}{}^{\circ} - TE_{01}{}^{\circ}$ couplings are indicated, any mode pair having common field components at the coupling aperture may be used. Figure 12 illustrates the type 5 coupling distribution, simulated by a series of discrete point couplings which should be spaced no more than about one-third guide wavelength. Either broadband power interchange or intentional frequency selectivity may be obtained.
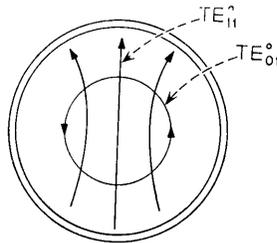
Fig. 9 — Transverse electric field lines for $TE_{11}{}^{\circ}$ and $TE_{01}{}^{\circ}$.
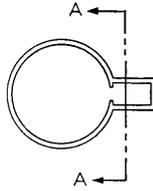
Fig. 10 — Transverse cross section for $TE_{10}^{\square} - TE_{01}^{\circ}$ coupling in hollow metallic waveguides.

V. LUMPED-ELEMENT PARAMETRIC DEVICES

Periodic coupling can be applied to lumped-element parametric devices; Figure 13 is a simplified version. The box labelled ω_1 is a filter presenting a short circuit at ω_1 and an open circuit at other frequencies; the filter box labelled ω_2 has similar characteristics

We assume a general time-varying capacitor

$$\mathcal{C}(t) = \mathcal{C}_0 + \mathcal{C}_p(t) \tag{23}$$

in which $\mathcal{C}_0$ is a constant. Appendix A shows that the normalized amplitudes representing the voltages and currents in the two resonant circuits can be described by the coupled-wave equations:

$$\frac{da_1}{dt} = j\omega_1 a_1 - \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_1 - a_1^*)}{\mathcal{C}_{11}} - \frac{(a_2 - a_2^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\} \tag{24}$$

$$\frac{da_1^*}{dt} = -j\omega_1 a_1^* + \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_1 - a_1^*)}{\mathcal{C}_{11}} - \frac{(a_2 - a_2^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\} \tag{25}$$

$$\frac{da_2}{dt} = j\omega_2 a_2 - \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_2 - a_2^*)}{\mathcal{C}_{22}} - \frac{(a_1 - a_1^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\} \tag{26}$$

$$\frac{da_2^*}{dt} = -j\omega_2 a_2^* + \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_2 - a_2^*)}{\mathcal{C}_{22}} - \frac{(a_1 - a_1^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\}. \tag{27}$$
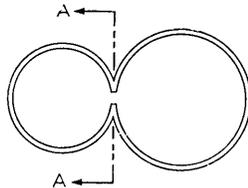
Fig. 11 — Transverse cross section for $TE_{01}^{\circ} - TE_{01}^{\circ}$ coupling in hollow metallic waveguides.
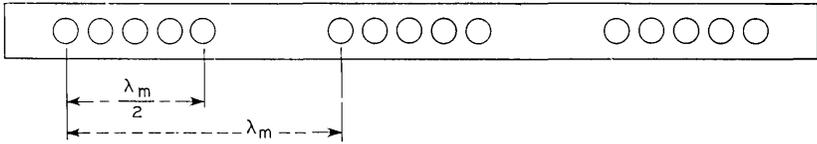
Fig. 12 — Section A-A' for Figs. 10 and 11.

For $\mathcal{C}_p = 0$ the solutions to equations (24) through (27) are of the form

$$a_1 = A_1 \exp (j\omega_1 t) \tag{28}$$

$$a_1^* = A_1^* \exp (-j\omega_1 t) \tag{29}$$

$$a_2 = A_2 \exp (j\omega_2 t) \tag{30}$$

$$a_2^* = A_2^* \exp (-j\omega_2 t). \tag{31}$$

We now specify a periodically varying capacitance component

$$\mathcal{C}_p = \Delta\mathcal{C} \cos (\omega_p t + \varphi) \cos \omega_c t \tag{32}$$

and we proceed to determine the coupling coefficients in equations (24) through (27) and to deduce the frequency interrelations governing the parametric interaction.

In equation (24) only the frequencies of the term in $d(\ \ )/dt$ at ω_1 result in large coupled-wave interaction; similarly in equations (25) through (27) only frequencies near $-\omega_1$ are important. Moreover, in equation (24) the term in $(a_1 - a_1^*)$ is a reaction of circuit 1 upon itself, which for small coupling is negligible; we drop terms of that type. With these criteria for selection of important terms we find that putting equation (32) in equations (24) through (27) leads to the following as the only significant wave interaction
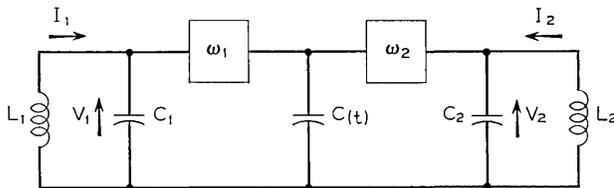
Fig. 13 — Lumped element parametric circuit.

$$\frac{da_1}{dt} = j\omega_1 a_1 - \frac{\Delta \mathcal{C}}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} \frac{d}{dt}$$

$$\cdot (a_2^* \{\exp [j(\omega_p t + \omega_c t + \varphi)] + \exp [j(\omega_p t - \omega_c t + \varphi)]\}) \qquad (33)$$

$$\frac{da_2^*}{dt} = -j\omega_2 a_2^* - \frac{\Delta \mathcal{C}}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} \frac{d}{dt}$$

$$\cdot (a_1 \{\exp [j(\omega_c t - \omega_p t - \varphi)] + \exp [j(-\omega_c t - \omega_p y - \varphi)]\}). \quad (34)$$

Noting that $dA_2^*/dt \ll (\omega_p \pm \omega_c)$ in our loose coupling approximation [A_2^* defined as in equation (31)] and similarly for dA_1/dt, we find equations (33) and (34) reduce to

$$\frac{da_1}{dt} = j\omega_1 a_1 + c_{121} a_2^* \exp [j(\omega_p + \omega_c)t] + c_{122} a_2^* \exp [j(\omega_p - \omega_c)t] \quad (35)$$

$$c_{121} = \frac{-\Delta \mathcal{C} \exp (j\varphi)}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} j(\omega_p - \omega_2 + \omega_c) \qquad (36)$$

$$c_{122} = \frac{-\Delta \mathcal{C} \exp (j\varphi)}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} j(\omega_p - \omega_2 - \omega_c) \qquad (37)$$

$$\frac{da_2^*}{dt} = -j\omega_2 a_2^* + c_{212} a_1 \exp [j(-\omega_p + \omega_c)t] + c_{211} a_1 \exp [j(-\omega_p - \omega_c)t]$$

$$\tag{38}$$

$$c_{211} = \frac{-\Delta \mathcal{C} \exp (-j\varphi)}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} j(-\omega_p - \omega_c + \omega_1) \qquad (39)$$

$$c_{212} = \frac{-\Delta \mathcal{C} \exp (-j\varphi)}{8[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}} j(-\omega_p + \omega_c + \omega_1). \qquad (40)$$

Note that

$$c_{212}^* = \frac{(\omega_p - \omega_c - \omega_1)}{(\omega_p - \omega_c - \omega_2)} c_{122} \qquad (41)$$

$$c_{211}^* = \frac{(\omega_p + \omega_c - \omega_1)}{(\omega_p + \omega_c - \omega_2)} c_{121} . \qquad (42)$$

Using relations (28) and (31) for a_1 and a_2^*, equations (35) and (38) reduce to

$$\frac{dA_1}{dt} = c_{121} A_2^* \exp [j(\omega_p + \omega_c - \omega_1 - \omega_2)t]$$

$$+ c_{122} A_2^* \exp [j(\omega_p - \omega_c - \omega_1 - \omega_2)t] \qquad (43)$$

$$\frac{dA_2^*}{dt} = c_{211}A_1 \exp\left[j(-\omega_p - \omega_c + \omega_1 + \omega_2)t\right]$$

$$+ c_{212}A_1 \exp\left[j(-\omega_p + \omega_c + \omega_1 + \omega_2)t\right]. \tag{44}$$

For simple exponential buildup of A_1 and A_2^* there are two possible frequency relations; one is

$$\omega_1 + \omega_2 = \omega_p + \omega_c \tag{45}$$

which reduces equations (43) and (44) to

$$\frac{dA_1}{dt} = c_{121}A_2^* + c_{122}A_2^* \exp\left(-j2\omega_c t\right) \tag{46}$$

$$\frac{dA_2^*}{dt} = c_{211}A_1 + c_{212}A_1 \exp\left(j2\omega_c t\right). \tag{47}$$

Here the $c_{121} - c_{211}$ terms are important; the other terms give a small cyclical variation on the exponential buildup.

The other important frequency relation is

$$\omega_1 + \omega_2 = \omega_p - \omega_c \tag{48}$$

which reduces equations (43) and (44) to

$$\frac{dA_1}{dt} = c_{121}A_2^* \exp\left(j2\omega_c t\right) + c_{122}A_2^* \tag{49}$$

$$\frac{dA_2^*}{dt} = c_{211}A_1 \exp\left(-j2\omega_c t\right) + c_{212}A_1 . \tag{50}$$

Here the $c_{122} - c_{212}$ terms are important; the other terms give a small cyclical variation on the exponential buildup.

Thus the effect of periodically varying the coupling in the lumped parametric circuit is to modify the frequency-relation requirement to equations (45) and (48). The result is eminently reasonable and perhaps superficially obvious. We can see this as follows: equation (32) can be rewritten

$$\mathfrak{C}_p = \frac{\Delta\mathfrak{C}}{2}\left\{\cos\left[(\omega_p + \omega_c)t + \varphi\right] + \cos\left[(\omega_p - \omega_c)t + \varphi\right]\right\}.$$

Suppose we assume a $\mathfrak{C}_p$ of

$$\mathfrak{C}_p = \frac{\Delta\mathfrak{C}}{2}\cos\left[(\omega_p + \omega_c)t + \varphi\right].$$

Then the previously known frequency condition for strong interac-

tion is[4]

$$\omega_1 + \omega_2 = \omega_p + \omega_c .$$

If instead we have

$$\mathcal{C}_p = \frac{\Delta \mathcal{C}}{2} \cos \left[(\omega_p - \omega_c)t + \varphi \right].$$

then the frequency condition for strong interaction is

$$\omega_1 + \omega_2 = \omega_p - \omega_c .$$

If we then assume linear superposition (unjustified in the nonlinear process) we could expect relations (45) and (48) for $\mathcal{C}_p$ of equation (32). The above analysis and associated discussion indicate the restrictions which must be met to achieve the desired result.

The periodic coupling variation need not be cosinusoidal as in (32). Instead, square wave or even low duty cycle pulse modulation of $\mathcal{C}_p$ again leads to equations (45) and (48), although care must be exercised to assure that pulse modulation of the pump properly reproduces the signal content in a parametric amplifier.

VI. DISTRIBUTED PARAMETRIC WAVE INTERACTIONS

Coupling in distributed parametric wave interactions can be periodic in two ways: (i) with respect to time at a particular point, and (ii) with respect to distance in the direction of propagation at a particular instant of time. We derive the constraints on propagation constants and on frequencies which result from such periodicity and then indicate some physical structures in which these wave interactions may prove useful.

Figure 14 shows a simplified model of a distributed transmission medium. The distributed capacitance is nonlinear and is a function of time as well as of the position z in the direction of propagation. A number of waves of frequencies ω_1, ω_2, and ω_p may propagate. The distributed inductance L_n is independent of current magnitude but may have
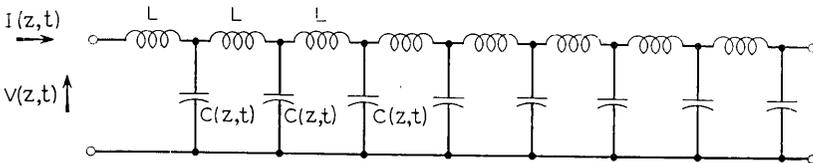
Fig. 14 — Distributed parametric circuit.

different values at different frequencies ω_n. In Appendix B the following coupled wave equations are derived for the normalized amplitudes of the traveling waves in Fig. 14

$$\frac{d}{dz} a_1 = -j\beta_1 a_1 - \frac{[z_{01}]^{\frac{1}{2}}}{2} \exp\,(-j\omega_1 t)\,\frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_1} \tag{51}$$

$$\frac{da_1^*}{dz} = j\beta_1 a_1^* - \frac{[z_{01}]^{\frac{1}{2}}}{2} \exp\,(j\omega_1 t)\,\frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_1} \tag{52}$$

$$\frac{da_2}{dt} = -j\beta_2 a_2 - \frac{[z_{02}]^{\frac{1}{2}}}{2} \exp\,(-j\omega_2 t)\,\frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_2} \tag{53}$$

$$\frac{da_2^*}{dt} = j\beta_2 a_2 - \frac{[z_{02}]^{\frac{1}{2}}}{2} \exp\,(j\omega_2 t)\,\frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_2} \tag{54}$$

in which we define, at frequency ω_n,

$$\mathcal{C}(z,\,t) = \mathcal{C}_{on} + \mathcal{C}_p(z,\,t) \tag{55}$$

$$z_{on} = \left[\frac{L_n}{\mathcal{C}_{on}}\right]^{\frac{1}{2}} \tag{56}$$

$$\beta_n = \omega_n[L_n\mathcal{C}_{on}]^{\frac{1}{2}}. \tag{57}$$

The time and space varying portion of $\mathcal{C}(z,\,t)$ is all contained within $\mathcal{C}_p(z,\,t)$, and $\mathcal{C}_{on}$ is dependent only on frequency.

Equations (51) through (54) may be used to explore the effects of any periodic coupling behavior. Because the normalized amplitudes a_1, a_1^*, a_2, and a_2^* are dependent on z only (according to equations 142, 143, 130, and 131), only the terms of the partial derivatives of (51) through (54), which yield zero time dependence of the coupling coefficient, result in appreciable coupled-wave interaction. This condition produces the frequency interrelations for parametric interaction. Similarly, only the terms of the partial derivatives of equations (51) through (54), which ultimately yield constant coupling between the traveling waves at all z, cause appreciable wave interaction; this condition produces the interrelations between the propagation constants (the β_n) necessary for parametric interaction. We proceed to apply this technique.

6.1 Traveling-Wave Pump with Spatial and Time Periodicity

We specify a function for the nonlinear distributed capacitance (type 3 coupling of Table I)

$$\mathcal{C}_p(z,\,t) = \frac{\Delta\mathcal{C}}{2} \cos\,\beta_c z\{\cos\,[(\omega + \omega_c)t - \beta_+ z] + \cos\,[(\omega - \omega_c)t - \beta_- z]\} \tag{58}$$

in which β_+ is the phase constant at frequency $(\omega + \omega_c)$ and β_- is the phase constant at $(\omega - \omega_c)$. This corresponds to driving the nonlinear medium with traveling wave at a modulated pump frequency $\cos \omega_c t$ in which ω_c is the modulation; the $\cos \beta_c z$ factor represents a spatial periodic variation in the coupling. Structures which produce spatially periodic parametric interactions are described later in this section.

We use equation (58) in equations (51) through (54) and select the terms which are capable of yielding a zero time dependence to the coupling terms. This shows a_1 and a_2^* to be the waves with significant coupling and the selected terms are

$$
\frac{da_1}{dz} = -j\beta_1 a_1 + c_{121} a_2^* \exp\left[-j(\beta_+ + \beta_c) + j(\omega + \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{121} a_2^* \exp\left[-j(\beta_+ - \beta_c) + j(\omega + \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{122} a_2^* \exp\left[-j(\beta_- + \beta_c) + j(\omega - \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{122} a_2^* \exp\left[-j(\beta_- - \beta_c) + j(\omega - \omega_c - \omega_1 - \omega_2)t\right] \tag{59}
$$

$$
\frac{da_2^*}{dz} = j\beta_2 a_2^* + c_{211} a_1 \exp\left[j(\beta_+ + \beta_c) - j(\omega + \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{211} a_1 \exp\left[j(\beta_+ - \beta_c) - j(\omega + \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{212} a_1 \exp\left[j(\beta_- + \beta_c) - j(\omega - \omega_c - \omega_1 - \omega_2)t\right]
$$

$$
+ c_{212} a_1 \exp\left[j(\beta_- - \beta_c) - j(\omega - \omega_c - \omega_1 - \omega_2)t\right] \tag{60}
$$

in which

$$
c_{121} = \frac{-\Delta \mathcal{C}}{16} j(\omega + \omega_c - \omega_2)(z_{01} z_{02})^{\frac{1}{2}} \tag{61}
$$

$$
c_{122} = \frac{-\Delta \mathcal{C}}{16} j(\omega - \omega_c - \omega_2)(z_{01} z_{02})^{\frac{1}{2}} \tag{62}
$$

$$
c_{211} = \frac{\Delta \mathcal{C}}{16} j(\omega + \omega_c - \omega_1)(z_{01} z_{02})^{\frac{1}{2}} \tag{63}
$$

$$
c_{212} = \frac{\Delta \mathcal{C}}{16} j(\omega - \omega_c - \omega_1)(z_{01} z_{02})^{\frac{1}{2}}. \tag{64}
$$

From equations (59) and (60) one sees that there are two frequency conditions which can yield large wave interactions. When

$$
\omega + \omega_c = \omega_1 + \omega_2 \tag{65}
$$

the c_{121} and c_{211} terms dominate and the other terms produce only minor

fluctuations. Also, when

$$\omega - \omega_c = \omega_1 + \omega_2 \tag{66}$$

the c_{122} and c_{212} terms dominate. When equation (65) is valid, $(\omega + \omega_c - \omega_2) = \omega_1$ and the coupling coefficients reduce to

$$c_{121} = \frac{-\Delta\mathcal{C}}{16} \, j\left(\frac{\omega_1}{\omega_2}\right)^{\frac{1}{2}} \left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}} \tag{67}$$

$$c_{211} = \frac{\Delta\mathcal{C}}{16} \, j\left(\frac{\omega_2}{\omega_1}\right)^{\frac{1}{2}} \left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}}. \tag{68}$$

Note that

$$c_{121} = \frac{\omega_1}{\omega_2} \, c_{211}^* \, . \tag{69}$$

When equation (66) is valid, the coupling coefficients of importance are c_{122} which reduces to equation (67) and c_{212} which reduces to equation (68), so that again

$$c_{122} = \frac{\omega_1}{\omega_2} \, c_{212}^* \, . \tag{70}$$

To find the necessary constraints on the phase constants we note that in the absence of coupling (that is, $\Delta\mathcal{C} = 0$) the solutions to equations (59) and (60) are of the form

$$a_1 = A_1 \exp\left(-j\beta_1 z\right) \tag{71}$$

$$a_2^* = A_2^* \exp\left(j\beta_2 z\right). \tag{72}$$

When equation (65) is valid, use of equations (71) and (72) in equations (59) and (60) reduces them to

$$\frac{dA_1}{dz} = c_{12}A_2^*\{\exp\left[-j(\beta_+ - \beta_c - \beta_1 - \beta_2)z\right]$$
$$+ \exp\left[-j(\beta_+ + \beta_c - \beta_1 - \beta_2)z\right]\} \tag{73}$$

$$\frac{dA_2^*}{dz} = c_{21}A_1$$

$$\cdot \{\exp\left[j(\beta_+ - \beta_c - \beta_1 - \beta_2)z\right] + \exp\left[j(\beta_+ + \beta_c - \beta_1 - \beta_2)z\right]\}. \tag{74}$$

We can now observe two conditions, either of which permit significant parametric wave interaction:

$$\beta_+ - \beta_c = \beta_1 + \beta_2 \tag{75}$$

$$\beta_+ + \beta_c = \beta_1 + \beta_2 . \tag{76}$$

Repeating the above procedure for equation (66) being valid instead of equation (65) yields two more permissible conditions at which in-phase wave interaction occurs at all z:

$$\beta_- - \beta_c = \beta_1 + \beta_2 \tag{77}$$

$$\beta_- + \beta_c = \beta_1 + \beta_2 . \tag{78}$$

When one of equations (75) through (78) is valid along with the corresponding frequency condition, equations (59) and (60) reduce to

$$\frac{dA_1}{dz} = c_{12}A_2^* \tag{79}$$

$$\frac{dA_2^*}{dz} = c_{21}A_1 . \tag{80}$$

These equations are satisfied by exponentials of the form

$$\exp\left[\pm (c_{12}c_{21})^{\frac{1}{2}}z\right].$$

When $(c_{12}c_{21})^{\frac{1}{2}}$ is pure real, growing and decaying waves are present and equations (67) and (68) meet this requirement. The parallel propagation of signal ω_1, idler ω_2, and pump ω results in gain, as is well known. Other configurations of signal, pump, and coupling peridocity can result in pure imaginary values of $(c_{12}c_{21})^{\frac{1}{2}}$ in which case a periodic interchange of power between waves is indicated.

The above discussion pertains to type 3 coupling of Table I, the difference between the sin and cos being negligible. For square wave coupling the physical model is often simpler to construct; we briefly consider this situation. In Fig. 15 we assume a region "a" in which the coupling is constant but the normal phase matching relations are not met, that is,

$$\beta_1 + \beta_2 \neq \beta.$$
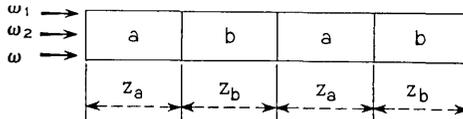
Fig. 15 — Model of a transmission medium with periodically varying properties.

Then the proper way to establish the periodic coupling is to make the length z_a such that the exponentials in equations (73) and (74) (with $\beta_c = 0$) become a half beat wavelength; for the specific case above there are two permissible choices,

$$(\beta_+ - \beta_1 - \beta_2)z_a = \pi \pm 2p\pi \tag{81}$$

$$(\beta_- - \beta_1 - \beta_2)z_a = \pi \pm 2p\pi \tag{82}$$

with the time modulation present; for cw pumping

$$(\beta - \beta_1 - \beta_2)z_a = \pi \pm 2p\pi \tag{83}$$

with p being any integer. Then, in the "b" region of Fig. 15, the coupling may be zero in which case we have type 5 coupling, or the coupling may be reversed compared with the "a" region, in which case we have type 4 coupling. In either case we require

$$(\beta - \beta_1 - \beta_2)z_b = \pi \pm 2p\pi. \tag{84}$$

The β's in the "a" and "b" regions need not be the same—the β's of equations (81) through (84) are to be those values characteristic of the waves' location. Earlier work has made use of some of these possibilities.[5,6]

Figure 15 shows square-wave coupling which, as discussed above, applies generally to passive wave interactions as well as to other parametric interactions. The conditions analogous to equations (81) through (84) follow from making the exponents in the appropriate coupled-wave equations, analogous to equations (73) and (74), equal to π or an odd multiple of π.

6.2 CW Traveling-Wave Pump with Simultaneous Modulation of the Entire Medium

A case related to that discussed in Section 6.1 is described by

$$\mathcal{C}_p(t) = \Delta\mathcal{C} \cos w_c t \cos (\omega t - \beta z). \tag{85}$$

Here the pump wave is a continuous wave and the entire array of variable capacitors is simultaneously modulated. This may occur when the modulating wave ω_c is brought into the nonlinear medium at right angle to z, or when ω_c is so small that the entire length of nonlinear medium is a lumped element in the ω_c circuit. Analysis similar to that in Section 6.1 shows that the frequency conditions are again given by equations (65) and (66), the coupling coefficients are twice those given by equations (67) and (68), and the phase constant condition is

$$\beta = \beta_1 + \beta_2 . \tag{86}$$

6.3 *Second-Harmonic Generation with Spatially Periodic Coupling*

The capacitance function for second-harmonic generation with spatially periodic coupling is

$$\mathcal{C}_p(t) = \Delta\mathcal{C} \cos \beta_c z \cos (\omega_1 t - \beta_1 z). \tag{87}$$

We look for coupling with $\omega_2 = 2\omega_1$ in equations (51) through (54) and find the interaction between a_1 and a_2. The coupling coefficients are

$$c_{12} = j \frac{\Delta\mathcal{C}}{4} \left(\frac{\omega_1}{\omega_2}\right)^{\frac{1}{2}} \left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}} \tag{88}$$

$$c_{21} = j \frac{\Delta\mathcal{C}}{4} \left(\frac{\omega_2}{\omega_1}\right)^{\frac{1}{2}} \left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}} \tag{89}$$

and the phase-constant requirement is

$$\beta_2 = 2\beta_1 \pm \beta_c . \tag{90}$$

In this case $(c_{12}c_{21})^{\frac{1}{2}}$ is pure imaginary, so the wave solutions, varying as

$$\exp [(c_{12}c_{21})^{\frac{1}{2}}z] \pm \exp [-(c_{12}c_{21})^{\frac{1}{2}}z],$$

represent a cyclical interchange of power between a_1 and a_2. However the mathematical model represented by equation (87) is not valid when a_1 diminishes appreciably because it no longer is the principal field on the variable capacitors as called for in equation (87).

If square-wave coupling is used in the configuration of Fig. 15, the phase constant and length relations are

$$(2\beta_{1a} - \beta_{2a})z_a = \pi \pm 2p\pi$$

$$(2\beta_{1b} - \beta_{2b})z_b = \pi \pm 2p\pi \tag{91}$$

with p being any integer including zero; the subscripts a or b on the β's denotes the region of Fig. 15 involved. As in the previous discussion of Fig. 15, a constant coupling in the "a" regions may be paired with either zero coupling or reversed coupling in the "b" regions to form types 4 or 5 coupling of Table I.

6.4 *Frequency Converter with Spatially Periodic Coupling*

Consider a medium driven nonlinear simultaneously at all z by a frequency ω_c according to

$$\mathcal{C}_p(z, t) = \Delta\mathcal{C} \cos \beta_c z \cos \omega_c t . \tag{92}$$

With waves of frequency ω_1 and ω_2 in the medium, from equations (51) through (54) we find that there is strong coupling between a_1 and a_2 at the frequency

$$\omega_c = \omega_1 - \omega_2 . \tag{93}$$

The phase constant condition is

$$\beta_1 - \beta_2 = \pm\beta_c \tag{94}$$

and the coupling coefficients are

$$c_{12} = j\,\frac{\Delta\mathcal{C}}{4}\left(\frac{\omega_1}{\omega_2}\right)^{\frac{1}{2}}\left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}} \tag{95}$$

$$c_{21} = j\,\frac{\Delta\mathcal{C}}{4}\left(\frac{\omega_2}{\omega_1}\right)^{\frac{1}{2}}\left(\frac{\beta_1\beta_2}{\mathcal{C}_{01}\mathcal{C}_{02}}\right)^{\frac{1}{2}}. \tag{96}$$

Since $(c_{12}c_{21})^{\frac{1}{2}}$ is pure imaginary there is a cyclical interchange of power between waves, and in this case the mathematical model is valid for complete interchange of power. If a wave at ω_1 is the input, the output will be solely a wave at ω_2 at a medium length z_t such that

$$\mid (c_{12}c_{21})^{\frac{1}{2}} \mid z_t = \frac{\pi}{2} \tag{97}$$

which yields

$$z_t = \frac{2\pi(\mathcal{C}_{01}\mathcal{C}_{02})^{\frac{1}{2}}}{\Delta\mathcal{C}(\beta_1\beta_2)^{\frac{1}{2}}}. \tag{98}$$

When square-wave coupling in a periodic structure of the form of Fig. 15 is used, the phase constant and length relations become

$$(\beta_{1a} - \beta_{2a})z_a = \pi \pm 2p\pi \tag{99}$$

$$(\beta_{1b} - \beta_{2b})z_b = \pi \pm 2p\pi \tag{100}$$

with p being any integer.

6.5 Structural Forms of Periodic Parametric Devices

We suggest here a few forms which periodic parametric devices might take. Figure 15 has already been referred to; it is apparent that the diagram is applicable to all of the preceding cases. The "b" region might simply be an index-matching oil without coupling effects. In other cases, it may be possible to achieve a reversal of the coupling.[6]

Figure 16 shows a centrosymmetric crystal such as (potassium tantalum niobate) with associated electrodes and potentials to achieve
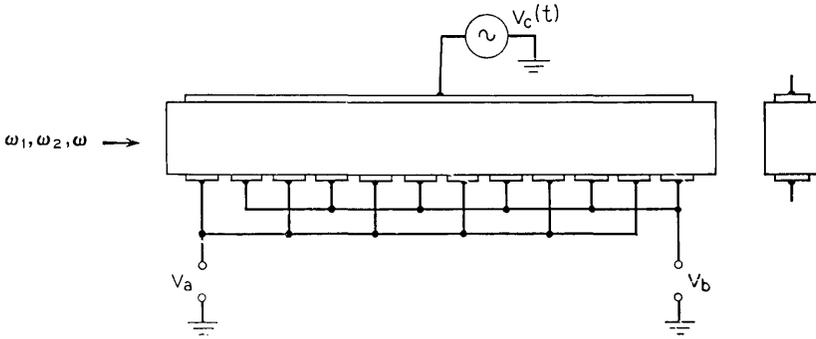
Fig. 16 — Model of a nonlinear crystal with wave coupling that is periodic both in time and space.

the periodic coupling. In such a crystal, a change in index of refraction is a parabolic function of the biasing field through the electro-optic effect. We have in mind laser wavelengths for the ω_1, ω_2, and ω waves. For V_a positive and V_b negative in Fig. 16, the slope of index versus RF field at frequency ω is positive in the "a" region and negative in the "b" region. Therefore, a spatial variation of coupling of the general form described in Section 6.1 is established; instead of the $\cos \beta_c z$ term in equation (58), a square-wave variation results from Fig. 16 with $V_c(t) = 0$ and dc biases of $V_a = +V$, $V_b = -V$.

With the addition of $V_c(t)$ in Fig. 16, a component $\cos \omega_c t$ as in equation (85) adds a simultaneous modulation of the medium, of the general form discussed in Section 6.2; to conform to equation (85) the voltages V_a and V_b should be made equal to zero. Second harmonic generation can be achieved using Fig. 16 with the ω wave omitted, $V_c(t) = 0$, $V_a = V$, and $V_b = -V$.

Frequency conversion of the type discussed in Section 6.4 might also be accomplished in the structure of Fig. 16. In this case the ω wave is omitted, the $\cos \omega_c t$ variation of equation (92) is produced by $V_c(t)$, and the biases V_a and V_b yield a square-wave spatial periodicity. Notice that V_b may be zero, approximating a type 5 coupling of Table I.

Figure 17 shows an alternate wave feeding arrangement for simultaneously modulating the entire nonlinear medium at a laser frequency rate. This could apply to Section 6.4 as well as to Section 6.2 with the addition of an ω wave parallel to the ω_1 and ω_2 waves.

In all cases, a guided wave may be used in the nonlinear medium by having a transverse index variation such as to produce a dielectric
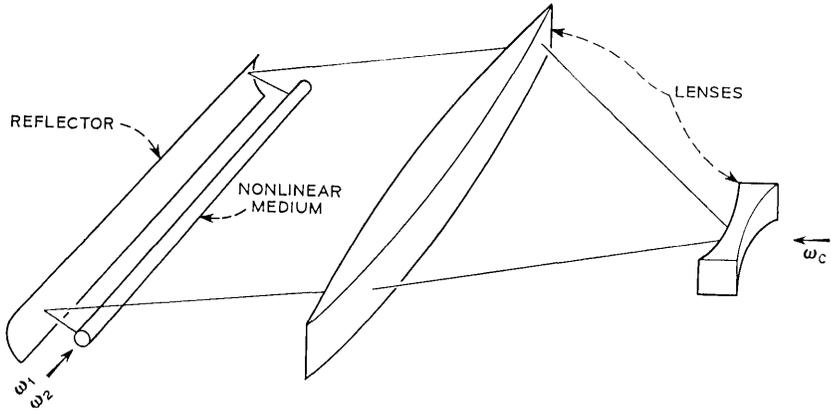
Fig. 17 — Parametric device with simultaneous modulation of the entire length of the nonlinear medium.

waveguide effect. This permits much longer regions of nonlinear interaction by holding the field within a small transverse area.

VII. CONCLUSION

We have outlined a wide variety of coupled-wave interactions in which a periodic variation in coupling may be used. The advantage in using periodic coupling rather then uniform coupling is frequently to achieve large power transfer between waves under conditions where uniform coupling will not do so—that is, where it is not possible for one reason or another to establish identical phase constants between the waves. Then by matching the periodicity of the coupling to the difference between the phase constants of the coupled waves, one can achieve nearly the same wave interactions as for matched phase constants and uniform coupling.

With frequency-selective filters, dispersion in the phase constants in combination with periodic coupling produces a desirable frequency-selective transfer of power. In the case of parmetric coupled-wave devices, periodic coupling requires a generalization of the Tien conditions which the frequencies and phase constants must meet.[7] These are outlined in Section VI.

VIII. ACKNOWLEDGMENT

The writer is indebted to D. H. Ring and Tingye Li for a careful reading of the manuscript.

APPENDIX A

Lumped Element Parametric Circuit

We now derive the coupled wave equations for the lumped circuit of Fig. 13 with a general time-varying capacitor[†]

$$\mathcal{C}(t) = \mathcal{C}_o + \mathcal{C}_p(t). \tag{101}$$

We define ω_1 and ω_2 by

$$\omega_1^2 L_1 \mathcal{C}_{11} = 1 \tag{102}$$

$$\omega_2^2 L_2 \mathcal{C}_{22} = 1 \tag{103}$$

$$\mathcal{C}_{11} = \mathcal{C}_1 + \mathcal{C}_o \tag{104}$$

$$\mathcal{C}_{22} = \mathcal{C}_2 + \mathcal{C}_o . \tag{105}$$

Then

$$\frac{dI_1}{dt} = -\frac{1}{L_1} V_1 \tag{106}$$

$$\frac{dI_2}{dt} = -\frac{1}{L_2} V_2 . \tag{107}$$

With the filter denoted by ω_1 in Fig. 13, a short circuit at ω_1 and an open circuit at other frequencies, and similarly for the filter ω_2

$$I_1 = \frac{d}{dt} \{[\mathcal{C}_1 + \mathcal{C}(t)V_1] - \mathcal{C}(t)V_2\} \tag{108}$$

$$I_2 = \frac{d}{dt} \{[\mathcal{C}_2 + \mathcal{C}(t)V_2] - \mathcal{C}(t)V_1\}. \tag{109}$$

Expanding equation (108)

$$I_1 = (\mathcal{C}_{11} + \mathcal{C}_p) \frac{dV_1}{dt} + V_1 \frac{d}{dt}(\mathcal{C}_{11} + \mathcal{C}_p)$$

$$- V_2 \frac{d}{dt}(\mathcal{C}_o + \mathcal{C}_p) - (\mathcal{C}_o + \mathcal{C}_p)\frac{dV_2}{dt}. \tag{110}$$

Rearranging terms,

$$\frac{dV_1}{dt} = \frac{I_1}{\mathcal{C}_{11}} - \frac{d}{dt}\left(\frac{\mathcal{C}_p V_1}{\mathcal{C}_{11}}\right) + \frac{V_2}{\mathcal{C}_{11}}\frac{d\mathcal{C}_p}{dt} + \frac{(\mathcal{C}_o + \mathcal{C}_p)}{\mathcal{C}_{11}}\frac{dV_2}{dt}. \tag{111}$$

[†] We follow the terminology of W. H. Louisell.[4]

Similarly,

$$\frac{dV_2}{dt} = \frac{I_2}{\mathcal{C}_{22}} - \frac{d}{dt}\left(\frac{\mathcal{C}_p V_2}{\mathcal{C}_{22}}\right) + \frac{V_1}{\mathcal{C}_{22}}\frac{d\mathcal{C}_p}{dt} + \left(\frac{\mathcal{C}_o + \mathcal{C}_p}{\mathcal{C}_{22}}\right)\frac{dV_1}{dt}. \tag{112}$$

As a result of the action of the ω_1 and ω_2 filters, V_1 contains only the frequency ω_1, and V_2 contains only the frequency ω_2. Hence dV_2/dt cannot contribute to dV_1/dt, and may be dropped in equation (111). Similarly, the last term of equation (112) may be dropped.

Multiplying the remainder of equation (111) by $j\omega_1\mathcal{C}_{11}$, adding equation (116), and multiplying each side by $(L_1)^{\frac{1}{2}}/2$, gives

$$\frac{(L_1)^{\frac{1}{2}}}{2}\left(\frac{dI_1}{dt} + j\omega_1\mathcal{C}_{11}\frac{dV_1}{dt}\right)$$

$$= \frac{(L_1)^{\frac{1}{2}}}{2}\left[-\frac{V_1}{L_1} + j\omega_1 I_1 - j\omega_1\mathcal{C}_{11}\frac{d}{dt}\left(\frac{\mathcal{C}_p V_1}{\mathcal{C}_{11}}\right) + j\omega_1 V_2\frac{d\mathcal{C}_p}{dt}\right]. \tag{113}$$

Using the normalized amplitudes a_1, a_2, and their complex conjugates

$$a_1 = \frac{(L_1)^{\frac{1}{2}}}{2}(I_1 + j\omega_1\mathcal{C}_{11}V_1) \tag{114}$$

$$a_1^* = \frac{(L_1)^{\frac{1}{2}}}{2}(I_1^* - j\omega_1\mathcal{C}_{11}V_1^*) \tag{115}$$

$$a_2 = \frac{(L_2)^{\frac{1}{2}}}{2}(I_2 + j\omega_2\mathcal{C}_{22}V_2) \tag{116}$$

$$a_2^* = \frac{(L_2)^{\frac{1}{2}}}{2}(I_2^* - j\omega_2\mathcal{C}_{22}V_2^*), \tag{117}$$

one may verify that equation (113) becomes

$$\frac{da_1}{dt} = j\omega_1 a_1 - \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_1 - a_1^*)}{\mathcal{C}_{11}} - \frac{(a_2 - a_2^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\}. \tag{118}$$

Using similar methods one can derive the other coupled wave equations

$$\frac{da_1^*}{dt} = -j\omega_1 a_1^* + \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_1 - a_1^*)}{\mathcal{C}_{11}} - \frac{(a_2 - a_2^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\} \tag{119}$$

$$\frac{da_2}{dt} = j\omega_2 a_2 - \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_2 - a_2^*)}{\mathcal{C}_{22}} - \frac{(a_1 - a_1^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\} \tag{120}$$

$$\frac{da_2^*}{dt} = -j\omega_2 a_2^* + \frac{d}{dt}\left\{\frac{\mathcal{C}_p}{2}\left[\frac{(a_2 - a_2^*)}{\mathcal{C}_{22}} - \frac{(a_1 - a_1^*)}{[\mathcal{C}_{11}\mathcal{C}_{22}]^{\frac{1}{2}}}\right]\right\}. \tag{121}$$

APPENDIX B

Distributed Parametric Medium

We now derive the coupled-wave equations for the distributed transmission medium of Fig. 14 with the general time- and space-varying distributed capacitance.

$$\mathcal{C}(z, t) = \mathcal{C}_{on} + \mathcal{C}_p(z, t) \tag{122}$$

where $\mathcal{C}_{on}$ is a constant relevant at angular frequency ω_n. Similarly the distributed inductance may have different values L_n at the various ω_n. From circuit theory

$$\frac{\partial V}{\partial z} = -L \frac{\partial I}{\partial t} \tag{123}$$

$$\frac{\partial I}{\partial z} = -\frac{\partial(V\mathcal{C})}{\partial t}. \tag{124}$$

Noting $\partial\mathcal{C}/\partial t = \partial\mathcal{C}_p/\partial t$, equation (124) becomes

$$\frac{\partial I}{\partial z} = -\mathcal{C}_{on} \frac{\partial V}{\partial t} - \frac{\partial}{\partial t}(V\mathcal{C}_p). \tag{125}$$

We define

$$Z_{01} = \left(\frac{L_1}{\mathcal{C}_{01}}\right)^{\frac{1}{2}} \tag{126}$$

$$z_{02} = \left(\frac{L_2}{\mathcal{C}_{02}}\right)^{\frac{1}{2}} \tag{127}$$

$$\beta_1 = \omega_1(L_1\mathcal{C}_{01})^{\frac{1}{2}} \tag{128}$$

$$\beta_2 = \omega_2(L_2\mathcal{C}_{02})^{\frac{1}{2}}. \tag{129}$$

Consider the case of propagating two waves in the medium of Fig. 14, one at ω_1 and one at ω_2. Then define

$$V(z, t) = V_1(z) \exp(j\omega_1 t) + V_2(z) \exp(j\omega_2 t) + V_1^*(z) \exp(-j\omega_1 t)$$
$$+ V_2^*(z) \exp(-j\omega_2 t) \tag{130}$$

$$I(z, t) = I_1(z) \exp(j\omega_1 t) + I_2(z) \exp(j\omega_2 t) + I_1^*(z) \exp(-j\omega_1 t)$$
$$+ I_2^*(z) \exp(-j\omega_2 t) \tag{131}$$

where the V_n and I_n are dependent only on z and the* denotes the complex conjugate. Then Equation (123) becomes

$$\exp{(j\omega_1 t)} \frac{dV_1}{dz} + \exp{(j\omega_2 t)} \frac{dV_2}{dz}$$

$$+ \cdots = -j\omega_1 L_1 \exp{(j\omega_1 t)} I_1 - j\omega_2 L_2 \exp{(j\omega_2 t)} I_2 + \cdots . \qquad (132)$$

Equating terms of equal frequency

$$\frac{dV_1}{dz} = -j\omega_1 L_1 I_1 \qquad (133)$$

$$\frac{dV_2}{dz} = -j\omega_2 L_2 I_2 \qquad (134)$$

$$\frac{dV_1^*}{dz} = j\omega_1 L_1 I_1^* \qquad (135)$$

$$\frac{dV_2^*}{dz} = j\omega_2 L_2 I_2^* . \qquad (136)$$

Using equations (130) and (131), equation (125) becomes

$$\exp{(j\omega_1 t)} \frac{dI_1}{dz} + \exp{(j\omega_2 t)} \frac{dI_2}{dz} + \cdots = -j\omega_1 \mathcal{C}_{01} V_1 \exp{(j\omega_1 t)}$$

$$- j\omega_2 \mathcal{C}_{02} V_2 \exp{(j\omega_2 t)} + \cdots - \frac{\partial(V\mathcal{C}_p)}{\partial t}. \qquad (137)$$

Equating terms of equal frequency yields

$$\frac{dI_1}{dz} = -j\omega_1 \mathcal{C}_{01} V_1 - \exp{(-j\omega_1 t)} \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_1} \qquad (138)$$

$$\frac{dI_1^*}{dz} = j\omega_1 \mathcal{C}_{01} V_1^* - \exp{(j\omega_1 t)} \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_1} \qquad (139)$$

$$\frac{dI_2}{dz} = -j\omega_2 \mathcal{C}_{02} V_2 - \exp{(-j\omega_2 t)} \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_2} \qquad (140)$$

$$\frac{dI_2^*}{dz} = j\omega_2 \mathcal{C}_{02} V_2^* - \exp{(j\omega_2 t)} \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_2} . \qquad (141)$$

The partial derivatives are to be evaluated in the vicinity of ω_1 for equation (138), $-\omega_1$ for equation (139), and so on. Considering only forward waves, we define a normalized wave amplitude

$$a_1(z) = \frac{V_1}{(z_{01})^{\frac{1}{2}}} = I_1(z_{01})^{\frac{1}{2}} \qquad (142)$$

$$a_1(z) = \frac{1}{2(z_0)^{\frac{1}{2}}} \{V_1 + z_{01} I_1\}. \qquad (143)$$

Using equations (143), (133), and (138),

$$\frac{da_1}{dz} = -\frac{1}{2(z_{01})^{\frac{1}{2}}} \left[j\omega_1 L_1 I_1 + z_{01} j\omega_1 \mathcal{C}_{01} V_1 + z_{01} \exp(-j\omega_1 t) \frac{\partial(V\mathcal{C}_p)}{\partial t} \right].$$

Using equations (143), (128), and (126),

$$-j\beta_1 a_1 = \frac{-j\omega_1 (L_1 \mathcal{C}_{01})^{\frac{1}{2}}}{2(z_{01})^{\frac{1}{2}}} (V_1 + z_{01} I_1) \tag{144}$$

$$= -\frac{1}{2(z_{01})^{\frac{1}{2}}} (j\omega_1 \mathcal{C}_{01} V_1 z_{01} + j\omega_1 L_1 I_1).$$

Hence

$$\frac{d}{dz} a_1 = -j\beta_1 a_1 - \frac{(z_{01})^{\frac{1}{2}}}{2} \exp(-j\omega_1 t) \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_1}. \tag{145}$$

Using similar substitutions one can show that

$$\frac{da_1^*}{dz} = j\beta_1 a_1^* - \frac{(z_{01})^{\frac{1}{2}}}{2} \exp(j\omega_1 t) \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_1} \tag{146}$$

$$\frac{da_2}{dt} = -j\beta_2 a_2 - \frac{(z_{02})^{\frac{1}{2}}}{2} \exp(-j\omega_2 t) \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{\omega_2} \tag{147}$$

$$\frac{da_2^*}{dt} = j\beta_2 a_2^* - \frac{(z_{02})^{\frac{1}{2}}}{2} \exp(j\omega_2 t) \frac{\partial(V\mathcal{C}_p)}{\partial t}\bigg|_{-\omega_2}. \tag{148}$$

With the mode amplitudes normalized as above, the square of the amplitudes represents the power carried by the mode.

REFERENCES

1. Miller, S. E., "On Solutions for Two Waves with Periodic Coupling," B.S.T.J., 47, No. 8 (October 1968), pp. 1801–1822.
2. Miller, S. E., "Coupled Wave Theory and Waveguide Applications," B.S.T.J., 333, No. 3 (May 1954), pp. 661–719.
3. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., this issue, pp. 2059–2069.
4. Louisell, W. H., Coupled Mode and Parametric Electronics, New York: John Wiley, 1960.
5. Bloembergen, N., "Apparatus for Converting Light Energy from One Frequency to Another", Patent 3,384,433, applied for July 9, 1962, issued May 21, 1968, see also Armstrong, J. A., Bloembergen, N., DuCuing, J., and Pershan, P. S., "Interactions between Light Waves in a Nonlinear Dielectric," Phys. Rev., 127, No. 6 (September 15, 1962), pp. 1918–1939.
6. Miller, Robert C., "Optical Harmonic Generation in Single Crystal BaTiO₃," Phys. Rev., 134, No. 5A (June 1964), p. A1313.
7. Tien, P. K., "Parametric Amplication and Frequency Mixing in Propagating Circuits," J. Applied Phys., 29, No. 9 (September 1958), p. 1347.

The Cutoff Region of a Rectangular
Waveguide with Losses,
Its Properties and Uses*

By L. U. KIBLER

(Manuscript received February 10, 1969)

The effect of the wall and the dielectric losses on the operation of a rectangular waveguide at frequencies in the cutoff region was investigated both theoretically and experimentally. A new measurement technique that permits determining the electrical properties of metals and dielectrics at microwave frequencies was developed from these investigations.

I. INTRODUCTION

Physical waveguides have walls with finite conductivity and enclose dielectric regions that have finite losses. The usual high conductivity metals and low-loss dielectrics have little effect on wave propagation at frequencies well above and below the cutoff frequency region.[†] These metallic and dielectric losses, however, have a pronounced effect in a small frequency region that includes the nominal cutoff frequency for a particular mode as determined for a lossless waveguide of the same geometry.[1] The purpose of this paper is the theoretical and experimental investigation of the properties of a physical waveguide operated at frequencies in this latter region.

The scope of this investigation is limited to a rectangular waveguide operated in the 8.2 to 12.4 GHz band of frequency (X band). The dominant mode of the lossless waveguide, the TE_{10} mode, serves as the initial model for an analysis of a similar mode configuration when losses are present. The analysis is divided into two parts: first, the waveguide is assumed to have two narrow walls with conductivity

σ_1, and the two broad walls with conductivity σ_2; and second, the same waveguide is analyzed with a lossy dielectric slab centered between the narrow walls of the waveguide.

The results of this analysis are examined experimentally using waveguide sections that have several different wall conductivities. Additional experiments were conducted with two types of lossy dielectrics. The results of these experiments demonstrate the effect of wall losses and dielectric losses on propagation in the cutoff frequency region of the waveguide.

The major use motivating this study of the cutoff properties of a lossy waveguide is that of determining the electrical constants of metals and dielectrics. The conductivity of three metals, and the dielectric constant and loss tangent of two dielectrics are determined experimentally using the cutoff properties of the waveguide. Copper, nickel, and a nichrome-copper composite were chosen. The effect of a dc magnetic field on the conductivity of copper and nickel were also investigated. The magnetic field produced no measurable effect on copper; however, the apparent conductivity of nickel decreased. A tentative explanation of this observation is advanced. Lucite and micarta were chosen for the dielectric experiments.

Experimental values of these physical constants are determined to within an accuracy of less than 2 percent. Where published values of these constants are available, they are found to agree within a few percent with the electrical values we obtained. The difference between published values and the values determined by this cutoff measurement technique reflect the use of certain approximations in the analysis and the experiment errors. These two sources of error are not separable, but it is evident that they are quite small.

This experimental technique provides a marked departure from the classical resonant cavity techniques used in the past.[2] The chief advantage of the waveguide cutoff measurement technique lies in the ability to measure the properties of metals accurately at microwave frequencies. The properties of dielectrics can also be measured although the accuracy of the cutoff technique is about the same as that of the classical resonant cavity techniques. There remains, however, the general advantage of having alternate measurement techniques which may be more convenient in some instances.

II. ANALYSIS

There are many analyses of the effects of losses in waveguides.[3-6] These efforts have been concerned with the effect of wall or dielectric

losses at frequencies well above or below the cutoff frequency of a particular mode. Barrow and Lender treated the effect of a finite wall conductivity on the propagation constant near the nominal cutoff of a circular waveguide.[7,8] Southworth noted that a decrease in the wall conductivity will decrease the frequency at which the cutoff region occurs.[9]

The classic method of treating wall losses and dielectric losses of a single mode in waveguides in the propagation region of the guide is to consider the power loss in the walls.[10] These approximate solutions are not valid in the vicinity of cutoff, since the lossless analysis on which they are based has a singularity at cutoff.

In order to accurately calculate the propagation constant of a waveguide with lossy walls and possibly containing a lossy dielectric, we must consider what field components must be present in the walls and in the dielectric. We will direct our attention to the TE$_{10}$ mode in the lossless waveguide as a starting point.

Refer to Fig. 1. The TE$_{10}$ mode in the lossless rectangular waveguide has the following field components; a y-directed electric field, a z-directed magnetic field and an x-directed magnetic field.

From these fields of the lossless waveguide, we can consider what other field components are necessary when the bounding walls have finite conductivity. At the side walls $x = 0$, $x = a$, there must be a y-directed electric field at the wall. On the top and bottom walls there must be an x-directed electric field and a z-directed electric field in the metal as a result of the finite currents in these directions. These fields must be supported by like directed fields in the dielectric region of the waveguide since the tangential electric and magnetic fields must be continuous across the dielectric-metal boundaries. Figure 1 shows the required field distributions.

In order to solve the field in the waveguide we must find a solution of Maxwell's equations for a possibly lossy dielectric surrounded by walls with finite conductivity. The finite conductive walls will be considered to be describable by their intrinsic impedance.

In a source free region, Maxwell's equations for a source free region with sinusoidal time dependence can be written for solution by vector potentials.[1] These vector potential equations became

$$\mathbf{E} = -\nabla \times \mathbf{F} - z\mathbf{A} + \frac{1}{y} \nabla(\nabla \cdot \mathbf{A}) \tag{1}$$

$$\mathbf{H} = \nabla \times \mathbf{A} - y\mathbf{F} + \frac{1}{z} \nabla(\nabla \cdot \mathbf{F}) \tag{2}$$
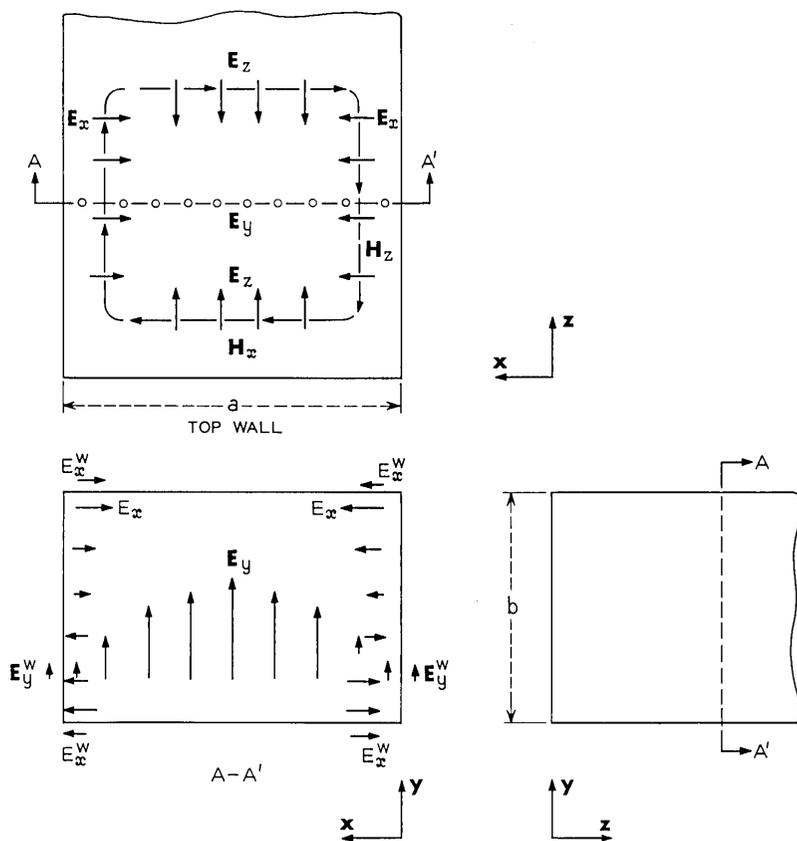
Fig. 1 — Field distribution of a lossy, waveguide.

where $\mathbf{A}$ = magnetic vector potential
         $\mathbf{F}$ = electric vector potential
     $y(\omega) = \sigma + j\omega(\epsilon' - j\epsilon'')$ (admittance per unit length)
     $z(\omega) = j\omega\mu$ (impedance per unit length)
         $j = (-1)^{\frac{1}{2}}$

and

     ϵ' = dielectric constant
     ϵ'' = dielectric loss factor
     σ = conductivity
     μ = permeability.

The choice of a y-directed complex magnetic vector potential of the form

$$\mathbf{A} = \mathbf{u}_y A_m \sin K_{x_o} x \cos K_{y_o} y \exp (-\gamma z) \tag{3}$$

where A_m is an arbitrary constant yields, after substitution in equations (1) and (2), the complex electric and magnetic field vectors in the dielectric region.

$$\mathbf{H}_x^o = A_m \gamma \sin K_{x_o} x \cos K_{y_o} y \exp (-\gamma z) \mathbf{u}_x \tag{4a}$$

$$\mathbf{H}_{z_o} = A_m K_{x_o} \cos K_{x_o} x \cos K_{y_o} y \exp (-\gamma z) \mathbf{u}_z \tag{4b}$$

$$\mathbf{E}_{x_o} = \frac{A_m K_{x_o} K_{y_o}}{y_o} \cos K_{x_o} x \sin K_{y_o} y \exp (-\gamma z) \mathbf{u}_x \tag{4c}$$

$$\mathbf{E}_{y_o} = \frac{A_m \sin K_{x_o} x \cos K_{y_o} y}{y_o} (k_o^2 - K_{y_o}^2) \exp (-\gamma z) \mathbf{u}_y \tag{4d}$$

$$\mathbf{E}_{z_o} = \frac{A_m K_{y_o} \gamma}{y_o} \sin K_{x_o} x \sin K_{y_o} y \mathbf{u}_z \tag{4e}$$

$$K_{x_o}^2 + K_{y_o}^2 - \gamma^2 = k_o^2 \tag{4f}$$

where

$$k_o^2 = -y_o z_o$$

and

γ = complex longitudinal propagation constant
K_{x_o}, K_{y_o} = complex transverse propagation constants.

In order to account for the properties of the metals that make up the waveguide walls, we define the surface impedance of a metal at microwave frequencies as

$$\left. \frac{E_y}{H_z} \right|_{\substack{x=a \\ x=o}} = Z(\text{side walls}) \equiv Z_s = R_s + jX_s$$

$$\left. \frac{E_x}{H_z} \right|_{\substack{y=b \\ y=o}} = Z(\text{top or bottom walls}) \equiv Z_T = R_T + jX_T . \tag{5}$$

In order to evaluate the surface impedance of the metal walls of the waveguide, we use the surface impedance for TEM waves in an unbounded lossy medium. This approximation is exact for a plane wave incident on a lossy metal. In the cutoff region the dominant mode fields can be approximately described by plane waves reflecting between the

side walls. Thus this definition of the surface impedance closely approximates the lossy waveguide in the cutoff region.

These wall impedances can be defined from $z(\omega)$ and $y(\omega)$ such that

$$Z_{S,T} = \text{Re} \left(\frac{z_{S,T}}{y_{S,T}}\right)^{\frac{1}{2}} + j \, IMAG \left(\frac{z_{S,T}}{y_{S,T}}\right)^{\frac{1}{2}}. \tag{6}$$

From (6) the intrinsic wall impedance for the conventional good conductors where $\sigma \gg \omega\epsilon$ can be obtained as

$$Z_S = \left(\frac{\omega\mu_s}{2\sigma_s}\right)^{\frac{1}{2}} (1 + j)$$

$$Z_T = \left(\frac{\omega\mu_T}{2\sigma_T}\right)^{\frac{1}{2}} (1 + j) \tag{7}$$

The determination of the propagation constants K_x, K_y, and γ results from application of the boundary conditions. These boundary conditions require continuity of the tangential E and H fields at each boundary. From equations (4) with conditions (5) and (7), we obtain

$$\frac{k_o^2 - K_{y_o}^2}{K_{x_o} y_o} \tan K_{x_o} a = \left(\frac{\omega\mu_s}{2\sigma_s}\right)^{\frac{1}{2}} (1 + j) \tag{8a}$$

$$\frac{K_{y_o}}{y_o} \tan K_{y_o} b = \left(\frac{\omega\mu_T}{2\sigma_T}\right)^{\frac{1}{2}} (1 + j). \tag{8b}$$

These equations (8) are transcendental and are solvable on a digital computer. Solution of (8b) for K_{y_o} allows the solution of (8a) for K_{x_o} for each frequency of interest. The z-directed propagation constant γ can be determined by substituting K_{x_o} and K_{y_o} into (4f).

It is evident that a set of curves for γ can be plotted for various values of $R_{S,T}$ and $X_{S,T}$. Thus from measured values of γ, the values of $R_{S,T}$ and $X_{S,T}$ can be determined, and hence the values of σ_s or σ_T.

The solutions represented by equations (8) can be used to determine the characteristics of the cutoff region of waveguides that have walls made of composite or coated metals. The intrinsic impedance of such conductors has been solved by Ramo and Whinnery.[11] The solution is given below.

$$Z_{COMP} = R_1(1 + j) \left[\frac{\sinh \tau_1 d + \dfrac{R_2}{R_1} \cosh \tau_1 d}{\cosh \tau_1 d + \dfrac{R_2}{R_1} \sinh \tau_1 d}\right] \tag{9}$$

where

$$d = \text{thickness of coating metal}$$

$$\tau_1 = (1 + j)(\pi f \mu_1 \sigma_1)^{\frac{1}{2}}$$

$$R_1 = \left(\frac{\pi f \mu_1}{\sigma_1}\right)^{\frac{1}{2}}$$

$$R_2 = \left(\frac{\pi f \mu_2}{\sigma_2}\right)^{\frac{1}{2}}$$

σ_1 = conductivity of coating metal

σ_2 = conductivity of coated metal

μ_1 = permeability of coating metal

μ_2 = permeability of coated metal.

Z_{COMP} can be substituted for either Z_S or Z_T in equations (8) depending on which walls of the waveguide are coated.

The most general solution of a waveguide in the cutoff region must include not only the effects of walls with finite conductivity but also the effect of a lossy dielectric. When the dielectric completely fills the interior of the waveguide the solutions just given can be used by inserting the complex dielectric constant defined above.

Because of a limited physical size of the available dielectrics or to accommodate certain measurement techniques discussed later, it may be necessary to use a thin slab of dielectric material that only partially fills the interior of the waveguide. Figure 2 is a sketch of such a dielectric slab waveguide.
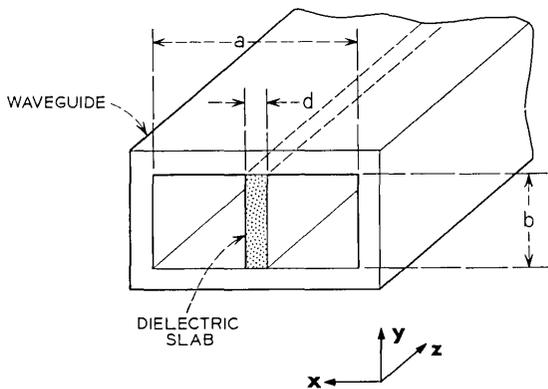
Fig. 2 — Dielectric slab waveguide.

The analysis of a lossy dielectric slab centered in a waveguide whose walls have finite conductivity proceeds from equations (1) and (2). The field solution if losses are assumed present only in the top and bottom walls can be obtained by choosing a y-directed complex electric vector potential

$$\mathbf{F} = \mathbf{u}_y \varphi. \tag{10}$$

A solution which satisfies the physical requirements of the dielectric slab waveguide dominant mode can be obtained by choosing the φ's as

$$\varphi_d = B_l \cos K_{x_d}\left(x - \frac{a}{2}\right) \cos K_{y_d} y \exp(-\gamma z) \quad \frac{a - d}{z} \leqq x \leqq \frac{a + d}{z}$$

$$\varphi_o = A_l \sin K_{x_o}(a - x) \cos K_{y_o} y \exp(-\gamma z) \quad \frac{a + d}{2} \leqq x \leqq a \tag{11}$$

$$\varphi_o' = A_l \sin K_{x_o} x \cos K_{y_o} y \exp(-\gamma z) \qquad 0 \leqq x \leqq \frac{a - d}{z}.$$

where A_l and B_l are arbitrary constants. The field components in the three regions of the dielectric slab waveguide are determined by substituting the electric vector potentials represented by (10) and (11) into the field equations (1) and (2).

In the dielectric region $(a - d)/2 \leqq x \leqq (a + d)/2$, the field components are

$$\mathbf{E}_{y_d} = \gamma B_l \cos K_{x_d}\left(x - \frac{a}{2}\right) \cos K_{y_d} y \exp(-\gamma z)\mathbf{u}_y$$

$$\mathbf{E}_{z_d} = -K_{y_d} B_l \cos K_{x_d}\left(x - \frac{a}{2}\right) \sin K_{y_d} y \exp(-\gamma z)\mathbf{u}_z$$

$$\mathbf{H}_{x_d} = \frac{(k_d^2 - K_{x_d}^2)}{z_d} B_l \cos K_{x_d}\left(x - \frac{a}{2}\right) \cos K_{y_d} y \exp(-\gamma z)\mathbf{u}_x \tag{12}$$

$$\mathbf{H}_{y_d} = \frac{B_l}{z_d} K_{x_d} K_{y_d} \sin K_{x_d}\left(x - \frac{a}{2}\right) \sin K_{y_d} y \exp(-\gamma z)\mathbf{u}_y$$

$$\mathbf{H}_{z_d} = \frac{B_l \gamma K_{x_d}}{z_d} \sin K_{x_d}\left(x - \frac{a}{2}\right) \cos K_{y_d} y \exp(-\gamma z)\mathbf{u}_z$$

where

γ = complex longitudinal propagation constant
$$\gamma^2 = K_{x_d}^2 + K_{y_d}^2 - k_d^2$$

K_{x_d} = x-directed propagation constant
K_{y_d} = y-directed propagation constant
k_d^2 = $\omega^2 \mu_d \epsilon_d$
z_d = $j\omega\mu_d$
a = guide width
d = width of dielectric slab
μ_d = dielectric permeability
ϵ_d = $\epsilon_d' - j\epsilon_d''$
ϵ_d' = dielectric permitivity
ϵ_d'' = dielectric loss factor.

In the region defined by $0 \leqq x \leqq (a - d)/2$ the field components become

$$\mathbf{E}_{y_o}' = \gamma A_l \sin K_{x_o}x \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_y$$

$$\mathbf{E}_{z_o}' = -K_{y_o}A_l \sin K_{x_o}x \sin K_{y_o}y \exp(-\gamma z)\mathbf{u}_z$$

$$\mathbf{H}_{x_o}' = \frac{(k_o^2 - K_{x_o}^2)}{z_o} A_l \sin K_{x_o}x \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_x$$

$$\mathbf{H}_{y_o}' = \frac{A_l}{z_o} K_{x_o} \cos K_{x_o}x \sin K_{y_o}y \exp(-\gamma z)\mathbf{u}_y$$ (13)

$$\mathbf{H}_{z_o}' = -\frac{A_l K_{x_o}\gamma}{z_o} \cos K_{x_o}x \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_z .$$

In the region defined by $[(a + d)/2] \leqq x \leqq a$ the field components are

$$\mathbf{E}_{y_o} = \gamma A_l \sin K_{x_o}(a - x) \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_y$$

$$\mathbf{E}_{z_o} = -K_{y_o}A_l \sin K_{x_o}(a - x) \sin K_{y_o}y \exp(-\gamma z)\mathbf{u}_z$$

$$\mathbf{H}_{x_o} = \frac{(k_o^2 - K_{x_o}^2)}{z_o} A_l \sin K_{x_o}(a - x) \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_x$$

$$\mathbf{H}_{y_o} = \frac{A_l K_{x_o}K_{y_o}}{z_o} \cos K_{x_o}(a - x) \sin K_{y_o}y \exp(-\gamma z)\mathbf{u}_y$$ (14)

$$\mathbf{H}_{z_o} = \frac{A_l K_{x_o}\gamma}{z_o} \cos K_{x_o}(a - x) \cos K_{y_o}y \exp(-\gamma z)\mathbf{u}_z$$

where

γ = complex longitudinal propagation constant
γ^2 = $k_{x_o}^2 + K_{y_o}^2 - k_o^2$
K_{x_o} = x-directed propagation constant
K_{y_o} = y-directed propagation constant

$$k_o^2 = \omega^2 \mu_o \epsilon_o$$
$$z_o = j\omega\mu_o$$
$$\mu_o = \text{permeability of vacuum}$$
$$\epsilon_o = \text{permitivity of vacuum.}$$

The total field solution of the dielectric slab waveguide is obtained by matching the tangential electric and magnetic fields at the dielectric-air boundaries. Matching these field quantities yields

$$\frac{K_{x_d}}{z_d} \tan K_{x_d} \frac{d}{2} = \frac{K_{x_o}}{z_o} \cot K_{x_o}\left(\frac{a - d}{2}\right) \tag{15a}$$

$$\frac{z_d K_{y_d}}{(k_d^2 - K_{x_d}^2)} \tan K_{y_d} b = Z_T \tag{15b}$$

$$\frac{z_o K_{y_o}}{(k_o^2 - K_{x_o}^2)} \tan K_{y_o} b = Z_T \tag{15c}$$

where Z_T is the surface impedance of the top and bottom walls of the waveguide.

Attempts to include the effects of the finite conductivity of the side walls in this solution were not successful. This failure stems from the lack of conformance of the boundary conditions and the coordinate surfaces. However, the fields in a waveguide operated in the cutoff region are approximately TEM waves in the transverse direction. We can use this fact to modify equations (15) to include the effects of the finite side wall conductivity. For this modifying solution we turn to a transverse resonance type of analysis.

Consider the dielectric slab waveguide in Fig. 2 at cutoff as a lossless parallel plate waveguide. The side wall of the waveguide is represented by its intrinsic admittance, y_s. The center of the guide is considered an open circuit. The admittance looking to the right, y_1, and to the left, y_2, of the dielectric-air boundary, is given by

$$y_1 = y_o \frac{y_s + jy_o \tan K_{x_o}\left(\dfrac{a - d}{2}\right)}{y_o + jy_s \tan K_{x_o}\left(\dfrac{a - d}{2}\right)}$$

$$y_2 = jy_d \tan K_{x_d} \frac{d}{2} \tag{16}$$

where

$$y_o = \left(\frac{\epsilon_o}{\mu_o}\right)^{\frac{1}{2}}$$

$$y_d = \left(\frac{\epsilon_d}{\mu_d}\right)^{\frac{1}{2}}$$

$$y_s = \frac{1}{z_s} = \frac{1}{\left(\frac{\omega\mu}{2\sigma_s}\right)^{\frac{1}{2}}(1 + j)}$$

$K_{x_d} = \omega_c(\mu_d\epsilon_d)^{\frac{1}{2}}$ at resonance

$K_{x_o} = \omega_c(\mu_o\epsilon_o)^{\frac{1}{2}}$ at resonance

ω_c = resonance or cutoff angular frequency.

The condition for resonance is then given by

$$y_1 = -y_2 . \tag{17}$$

Substitution of (16) into (17) yields

$$y_o \left[\frac{1 + jy_o z_s \tan K_{x_o}\left(\frac{a - d}{2}\right)}{y_o z_s + j \tan K_{x_o}\left(\frac{a - d}{2}\right)} \right] = -jy_d \tan K_{x_d} \frac{d}{2}. \tag{18}$$

Since $jy_o z_s$ will be a small value for practical wall metals, we can use the approximation

$$jy_o z_s = \alpha \approx \tan \alpha. \tag{19}$$

Equation (18) then becomes

$$-jy_o \left[\frac{1 + \tan \alpha \tan K_{x_o}\left(\frac{a - d}{2}\right)}{-\tan \alpha + \tan K_{x_o}\left(\frac{a - d}{2}\right)} \right] = -jy_d \tan K_{x_d} \frac{d}{2}. \tag{20}$$

The bracketed function on the left side of equation (20) is the expansion of $\cot(A\text{-}B)$; hence (20) becomes

$$y_o \cot \left[K_{x_o}\left(\frac{a - d}{2}\right) - jy_o z_s \right] = y_d \tan K_{x_d} \frac{d}{2}. \tag{21}$$

Equation (21) has the general trigonometric form of the field solution given in equation (15a). Substituting the values for y_o and y_d and multiplying and dividing by K_{x_o} and K_{x_d} yields when the common terms of μ_o, ϵ_o, ϵ_d and μ_d are cancelled and K_{x_o} and K_{x_d} are reintroduced and the

equation rearranged

$$\frac{K_{x_o}}{z_o} \cot \left[K_{x_o}\left(\frac{a - d}{2} \right) + \frac{K_{x_o}z_s}{z_o} \right] = \frac{K_{x_d}}{z_d} \tan K_{x_d} \frac{d}{2}. \tag{22}$$

Equation (22) is the same form as equation (15a) except for the modification of the argument of the cotangent function by the effect of the side wall impedance. Because of the evident similarity of equations (15a) and (22), we can interpret K_{x_o} as the transverse complex propagation constant that is valid not only at cutoff but over a range of frequencies extending on either side of cutoff.

The term z_s/z_o can be expanded by using the definition of the skin depth, δ

$$\delta = \left(\frac{2}{\omega\mu\sigma} \right)^{\frac{1}{2}} \tag{23}$$

to yield

$$\frac{z_s}{z_o} = \frac{(1 - j)\delta}{2}. \tag{24}$$

The argument of the cotangent term becomes

$$K_{x_o}\left[\frac{a - d}{2} + \frac{(1 - j)\delta}{2} \right]. \tag{25}$$

Argument (25) indicates that the width of the air region of the dielectric-slab waveguide is increased by an amount proportional to the skin depth. A similar result was obtained by Adler, Chu, and Fano, who analyzed the minimum of the standing wave pattern for a plane wave at an air-lossy metal interface.[12]

The equations which furnish the solution of a lossy dielectric slab centered in a waveguide with lossy walls in the region of cutoff are given by

$$\frac{K_{x_o}}{z_o} \cot K_{x_o}\left[\frac{(a - d)}{2} + \frac{(1 - j)\delta}{2} \right] = \frac{K_{x_d}}{z_d} \tan K_{x_d} \frac{d}{2} \tag{26a}$$

$$\frac{z_d K_{y_d} \tan K_{y_d}b}{[k_d^2 - K_{x_d}^2]} = Z_T \tag{26a}$$

$$\frac{z_o K_{y_o} \tan K_{y_o}b}{[k_o^2 - K_{x_o}^2]} = Z_T \tag{26c}$$

$$\gamma^2 = K_{x_o}^2 + K_{y_o}^2 - k_o^2 \tag{26d}$$

$$\gamma^2 = K_{x_d}^2 + K_{y_d}^2 - k_d^2 . \tag{26e}$$

These are complex transcendental equations and were solved by a digital computer.

Equations (26), which provide the solution for the propagation constant of a lossy waveguide with a centered lossy dielectric slab operated in the cutoff region, were derived using several approximations. These approximations are considered quite accurate for frequencies in the cutoff region. At frequencies outside the cutoff region, these approximations lead to an increasing error in the computation of the longitudinal propagation constant. Thus, the electrical properties of materials cannot be determined accurately for frequencies outside the cutoff region. The accuracy of the measurements in the cutoff region will be evident from the experimental results.

III. EXPERIMENTAL CIRCUITS

The analysis in Section II shows that the cutoff region of a waveguide with losses in the walls and dielectric is basically characterized by the complex longitudinal propagation constant. The other descriptive parameters such as impedance, admittance, scattering coefficients, and so on, depend on this propagation constant.

The propagation constant can be measured experimentally by determining the total attenuation and the total phase shift of a section of uniform waveguide whose length is known accurately. Accurate measurement of either attenuation or phase shift is difficult to obtain. However, differences in phase shift and in attenuation can be measured with great accuracy.

The experimental circuit was designed to measure differences in attenuation and phase shift. Figure 3 shows the basic experimental circuit. It is a microwave form of the usual low frequency comparison circuit. A common source supplies two paths. The path B is used as a reference path. Path A, the comparison path, has two separate test paths, A_1 and A_2, either of which may be chosen by proper positioning of the waveguide switches 1 and 2. The two main circuit paths are connected to a phase detector and to an amplitude detector by positioning switches 3 and 4.

These experiments could be conducted at any number of microwave frequencies. Available tables of the properties of metals and dielectrics show that these materials have a marked change in their dc properties in the X band of frequencies.[11,13] For this reason and the availability of accurate test equipment, the center of the X band of frequencies, about 9.5 GHz, was chosen for the design of the experimental circuit.
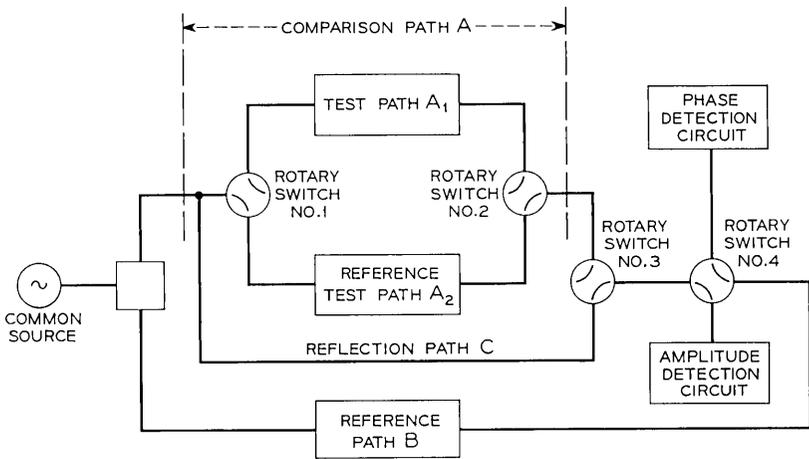
Fig. 3 — Simplified schematic diagram of experimental circuit.

The complete experimental circuit is shown schematically in Fig. 4. Standard commercial precision waveguide components were used throughout. Isolators were chosen to have voltage standing wave ratio's of less than 1.08 and isolation greater than 40 dB. The phase detection circuit was the kind described by Cohn.[14] It can measure phase differences to an accuracy of 0.05°. The amplitude detection circuit was used in conjunction with the tandem precision rotary vane attenuators (path A, Fig. 4). This combination was capable of measuring attenuation differences to an accuracy of 0.005 dB.

A precision rotary vane phase shifter was calibrated against the phase detection circuit and both were calibrated with selected lengths of precision X band waveguide. The phase shifter (path B, Fig. 4) and the phase detection circuit were used in combination for phase difference measurements.

Two types of waveguide test sections (see Fig. 5) were designed for use in these studies of the properties of a waveguide operated at cutoff. We designate these as a type A and a type B test section. Both types were electroformed of oxygen-free hard copper. Mandrels of the required dimensions for each test section were machined from aluminum, and polished to remove any roughness. The wall thickness of both types of test sections was a nominal $\%_{16}$ inch.

Each type A test section was electroformed in one piece. Standard X-band flanges, type UG-39/1, were soldered to each end of a test
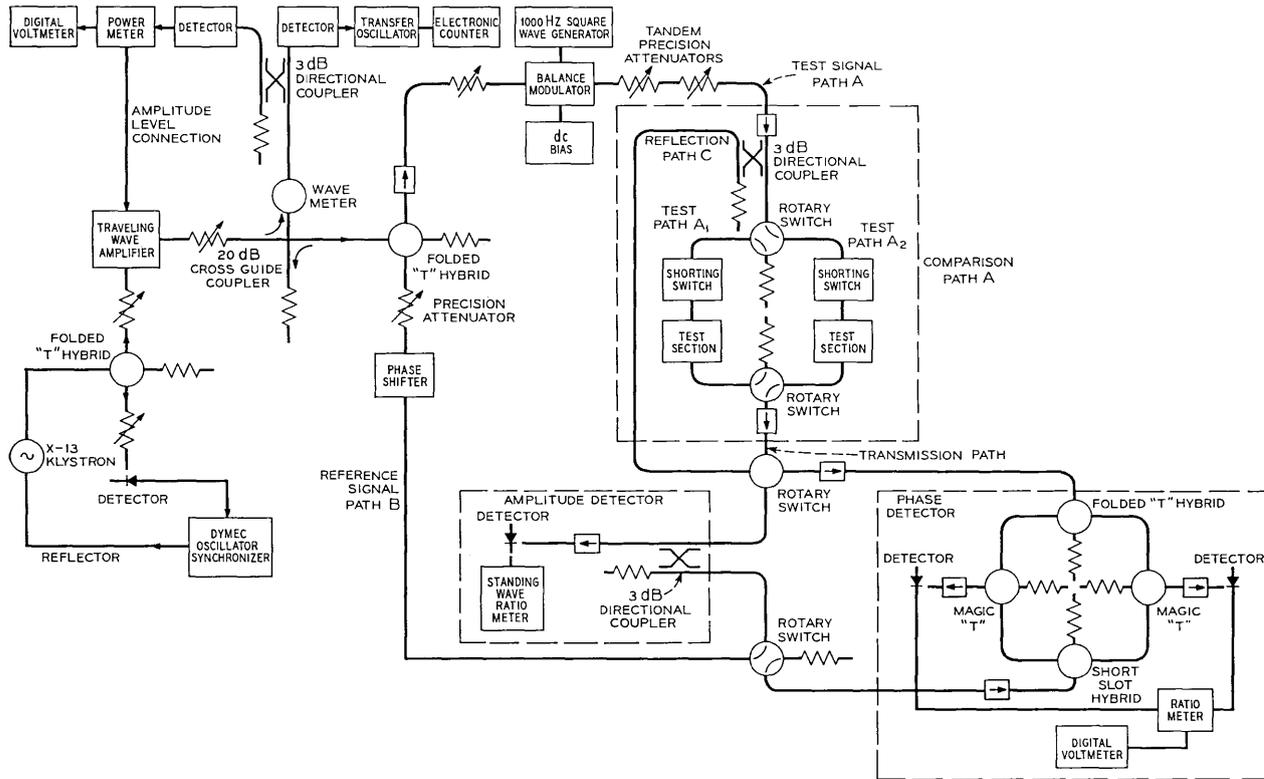
Fig. 4 — Experimental circuit for cutoff waveguide measurement.
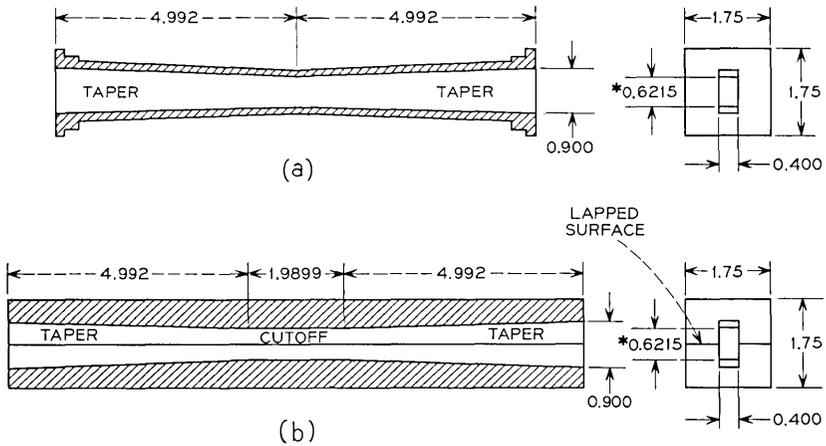
Fig. 5 — (a) Type A and (b) type B waveguide test sections. Dimensions are in inches. * The widths of the test sections were measured with gauge blocks to insure accuracy.

section. The flanges were machined and lapped to a smooth mating surface. Two alignment pins were inserted in each flange. Figure 6 shows a complete type A section.

The type B test sections were made in two halves (see Fig. 7). These halves were joined along the center of the broad faces of the wave-guide walls. Each half of the type B test section was mounted in a brass channel for rigid support. The joining faces of these brass mounting channels were polished to achieve the required width. Alignment pins assured accurate assembly of the two halves. The joining surfaces were machined and lapped for accurate mating. The halves were held
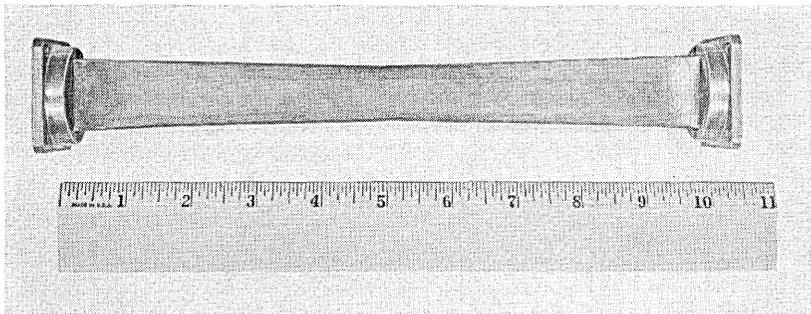
Fig. 6 — A complete type A waveguide test section.

together by 24 10-32 bolts. Threaded holes to mate standard X-band flanges were placed in each end of the type B test section. A dielectric slab is shown inserted in one half.

The type B test sections were used to examine the conductivity of various metals. These metals were placed on the walls of the cutoff portion of the type B sections by plating or evaporation. These test sections were made in halves for two reasons. First, it was possible to obtain uniform metal deposit on the three walls of the channel that results from making the section in halves. Uniform metal deposits on the interior walls of a closed test section was difficult if not impossible. Second, when metal is deposited on the walls of a waveguide, the interior dimensions are reduced by the metal thickness. At cut-
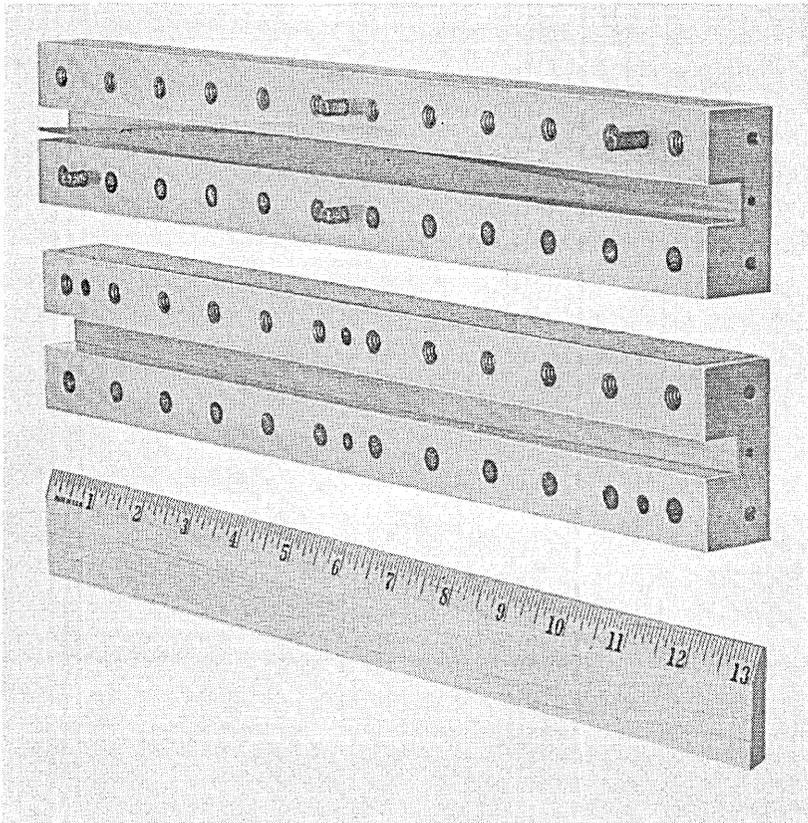
Fig. 7 — Type B waveguide test section with a dielectric slab in the top half.

off these small changes (0.0001 inch or less) in the waveguide height are insignificant. However, the same magnitude of dimension change in the width are very significant. By using two halves and by depositing the same thickness of metal on the joining surfaces as on the walls, there is an automatic compensation of the width change. The metal deposited on the narrow walls decreases the waveguide width, but the metal deposited on the joining surfaces increases the width on joining the halves by the same amount. Thus, the waveguide width was kept constant regardless of the thickness of the deposited metal.

IV. EXPERIMENTAL MEASUREMENT PROCEDURE

The general procedure for measuring the properties of a waveguide operated in the cutoff region is divided into three steps. We use Fig. 3 to help in the discussion of these steps. First, the phase and attenuation difference between the reference path (B in Fig. 3) and the reference test waveguide (A_2 in Fig. 3) are measured. This measurement includes both the transmission through, and reflection from (C in Fig. 3) the reference test waveguide. Second, the phase and attenuation difference between the reference path (B in Fig. 3) and the test waveguide section (A_1 in Fig. 3) are measured. As above, this measurement includes both the transmission through and the reflection from (path C) the test waveguide section. Third, the phase and attenuation difference between the reference test waveguide and the test waveguide are determined from the first two measurements.

The measurement of the effect of copper walls on the properties of a waveguide operated at cutoff required a copper type A waveguide test section and a copper type B waveguide test section. The type A section was placed in the position of the reference test waveguide (A_2 in Fig. 3); and the type B section in the position of the test waveguide section (A_1 in Fig. 3). The three part measurement procedure was followed.

These measurements yielded two results. The transmission measurements result in the differences in the total phase shift and the total attenuation of the cw signal transmitted through the type A and type B waveguide test sections. The reflection measurements yielded the difference in the total phase shift and the total attenuation of the cw signal reflected from the type A and type B sections. From Fig. 5 we see that the type A and type B test sections have identical tapers. These tapers were adjusted to be electrically equal. The total phase shift and attenuation differences thus became the phase shift and at-

tenuation of the transmission through and the reflection from the 2-inch-long cutoff section contained in the type B test section.

The measurement of the effect of the dielectric slabs of lucite and micarta on the properties of the waveguide at cutoff followed the same procedure. The dielectric slabs (see Fig. 8) were centered in the type A and type B copper test sections. Since the tapers at the ends of the dielectric slabs are identical, the result of the measurement is the phase shift and the attenuation of the transmission through, and the reflection from, the dielectric loaded cutoff section of the type B test section.

The measurement of the effect of the other metallic walls, nickel and nichrome-copper, required only type B test sections. A copper type B test section was placed in the reference test waveguide position (A_2 in Fig. 3). An identical second copper type B test section was placed in the test section position (A_1 in Fig. 3). The electrical difference between these two test sections was determined for use in correcting future measurements.

The type B test section (A_1 in Fig. 3) was removed and the metals, nickel or nichrome, were applied over the copper walls of the cutoff region. The type B section was then reinserted into test position A_1. The measurement steps just described were repeated. The results of these measurements after correction for the possible electrical difference yielded the phase shift and attenuation of the signal transmitted through and reflected from the 2-inch long cutoff section of the type B waveguide section. The properties of the waveguide with nickel or nichrome walls operated in the cutoff region are determined from these results.

(a)

(b)
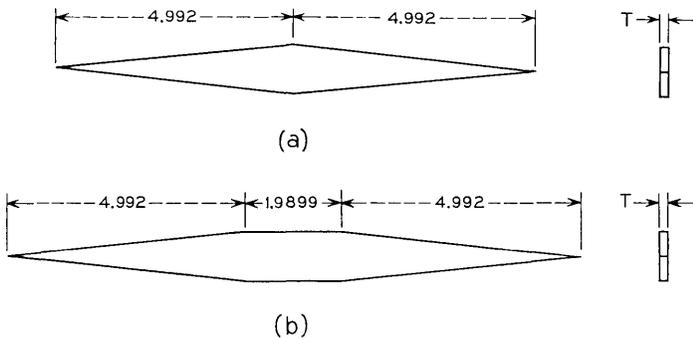
Fig. 8 — Dielectric slabs for (a) type A and (b) type B test sections. Dimensions are in inches.

These phase and attenuation measurements are used to calculate the complex propagation constant of a waveguide operated at frequencies in the cutoff region. These calculations are based on an analysis developed by Southworth (see pp. 57 and 58 of Ref. 9). He defines a voltage wave progressing down a finite length transmission line and suffering repeated reflections from mismatches at the input and output of the line. Southworth derives an expression for the steady state voltage at any point on the line. Using Southworth's notation we define V_t as the voltage transmitted to the output of the cutoff section and V_0 as the voltage reflected to the input of the cutoff section. The difference between these two "voltages" can be written using Southworth's results as

$$(V_t - V_0) = [\exp(-\gamma_0 l) - 1]. \tag{27}$$

The measurement of $(V_t - V_0)$ yields from (27), the experimental value of γ_0, the complex propagation of the cutoff region of the waveguide section.

The experimental measurements given by the phase shifter, the phase detection circuit, and the tandem attenuators were used to determine the value of $(V_t - V_0)$. Figures 3 and 4 should help in the following discussion. For the measurement of copper and dielectrics we have a type A test section in path A_2 and a type B test section in path A_1 of the comparison path A. We consider two voltage waves, E_0 and E_1. E_0 propagates in the comparison path A, and E_1 in the reference path B.

The attenuation and phase shift of test path A_2 from the input rotary switch to the center of the type A test section is defined as $A_0 \exp(j\Phi_0)$, and from the center of the type A test section to the output rotary switch as $B_0 \exp(j\Phi_1)$. The attenuation and phase shift of test path A_1 from the input rotary switch to the junction of the type B section taper and the cutoff section is defined as $A_0^1 \exp(j\Phi_0^1)$, and from the output of the cutoff section to the output rotary switch, $E_0^1 \exp(j\Phi_1^1)$. The voltage wave in the reference path B is defined as $E_1 \exp(j\theta_1)$.

Test path A_2 with a type A section inserted is connected to the comparison path. With the tandem attenuators set at an arbitrary value, the phase shifter is adjusted to provide a 45° phase difference between the comparison and reference paths. The outputs of the phase measuring circuit and the amplitude measuring circuit are proportional to

$$\Phi_0 + \Phi_1 = \theta_1 \pm 45° \tag{28a}$$

and

$$E_0 A_0 B_0 = M_0 = \text{SWR meter reading,} \qquad (28b)$$

respectively.

When the comparison path is connected to test path A_1 with a type B test section inserted, the outputs of the phase measuring circuit and the amplitude measuring circuit are proportional to $\Phi_0^1 + \Phi_1^1 + \measuredangle V_l$ and $E_0 A_0^1 B_0^1 \mid V_l \mid$, respectively.

The tandem attenuators and the phase shifter are adjusted to return the outputs of the amplitude measuring circuit and the phase measuring circuit to their values when test path A_2 was connected to the comparison path. This condition is expressed as

$$\Phi_0^1 + \Phi_1^1 + \measuredangle V_l = \theta_1 + \theta_p \pm 45° \qquad (29a)$$

$$T_1 E_0 A_0^1 B_0^1 \mid V_l \mid = M_0 \qquad (29b)$$

where θ_p is the change in the phase shifter and T_1, the change in the attenuators' reading.

The same analysis is applied to the voltage waves reflected from test paths A and B. The reflection from test path B is expressed as

$$2\Phi_0 = \theta_1 \pm 45° \qquad (30a)$$

$$E_0 2A_0 = M_1 . \qquad (30b)$$

The reflection from test path A is expressed as

$$2\Phi_0^1 + \measuredangle V_0 = \theta_1 + \theta_p^1 \pm 45° \qquad (31a)$$

$$T_2 E_0 2A_0^1 \mid V_0 \mid = M_1 \qquad (31b)$$

where θ_p^1 is the change in the phase shifter and T_2 is the change in the attenuators' reading.

Subtracting equation (28) from (29) yields

$$\Phi_0^1 - \Phi_0 + \Phi_1^1 - \Phi_1 + \measuredangle V_l = \theta_p \qquad (32a)$$

$$T_1 A_0^1 B_0^1 \mid V_l \mid = A_0 B_0 . \qquad (32b)$$

A calibration procedure determines the difference between Φ_0 and Φ_0^1, and Φ_1 and Φ_1^1, and the ratio of A_0^1/A_0 and B_0^1/B_0. With these measured differences, V_l is determined from the phase shifter change, θ_p in degrees, and the tandem attenuators' change, T_1 in dB.

Subtracting equation (30) from (31) yields

$$\measuredangle V_0 + 2\Phi_0^1 - 2\Phi_0 = \theta_p^1 \qquad (33a)$$

$$T_2 A_0^1 \mid V_0 \mid = A_0 . \tag{33b}$$

Again the calibration procedure furnished the values of $\Phi_0^1 - \Phi_0$ and A_0^1/A_0 . The value of V_0 was determined from the phase shifter change θ_p^1 and the tandem attenuators' change T_2 . The value of $V_l - V_0$ is determined by the phase shifter's and attenuators' change in reading. The value of γ_0 is calculated from these experimental measurements by equation (27).

This analysis is also applicable to measurements with type B test sections in both the reference test path and the test path as required for measurement of nickel and nichrome. We define the propagation constant for one type B test section cutoff region (copper) as γ_0 and for the second type B test section (nickel or nichrome) as γ_1 . We can then write for the nickel or nichrome section

$$(V_l - V_0)_A = [\exp(-\gamma_{1l}) - 1] \tag{34a}$$

and for the copper section

$$(V_l - V_0)_{cu} = [\exp(-\gamma_0 l) - 1]. \tag{34b}$$

The difference of these two equations yield

$$\begin{aligned}\Delta_{A-cu} &= (V_l - V_0)_A - (V_l - V_0)_{cu}\\ &= [\exp(-\gamma_1 l) - \exp(-\gamma_0 l)].\end{aligned} \tag{35}$$

This difference represents the measured difference between a copper type B test section and a type B test section with a nickel or nichrome cutoff section. The analysis used previously to describe the copper cutoff section measurements is applicable here to show that Δ_{A-cu} was evaluated by this measurement and γ_0 by the copper test section measurement; thus the value of γ_1 is determined.

Since nickel is a magnetic material, measurements were made on the nickel plated type B test section with a magnetic field applied. These measurements demonstrated qualitatively that this measuring technique could be used to detect the magnetic induced changes in conductivity of certain metals.

The magnetic field was obtained from a surplus horseshoe shaped magnetron magnet. The width of the pole pieces was approximately two inches and the gap approximately two and one quarter inches. The measured field between the poles was approximately 2100 gauss. The magnet was oriented with the type B test section to produce a magnetic field parallel to the electromagnetic field lines in the side walls of the waveguide. Because of its construction (horseshoe

shape), this magnet produced a nonuniform magnetic field in the wave-guide walls. Uniform magnetic field sources were not available; hence, these experiments were qualitative.

Measurements were made with nichrome to determine the effect on the waveguide cutoff properties of a lossy metal. A second type B test section was inserted in test path A_1 in place of the nickel plated test section. The reference type B test section remained in test path A_2. The phase shift difference and the attenuation difference between test path A_1 and test path A_2 was measured for both transmitted and reflected signals. This was done to establish a reference for the type B test section in test path A.

The type B test section was removed from test path A_1 and disassembled. The two halves of the test section were masked and nichrome was vacuum evaporated on the two narrow sidewalls to a thickness of 800Å. The two halves were reassembled and the nichrome type B test section reinserted in its original position in test path A_1.

The properties of a waveguide partially loaded with a dielectric was examined in the cutoff frequency region. Two types of dielectrics were used, one with low loss, lucite; and one with moderate loss, micarta.* In addition to demonstrating the effects of dielectric at cutoff, the dielectric constant and the loss tangent were determined from these measurements. Figure 8 is a diagram of the dielectric slabs.

V. EXPERIMENTAL RESULTS

In order to place the results of the experiments in the proper perspective, we note that the nominal cutoff frequency of a lossless wave-guide 0.62150-inch wide operated in the dominant mode is 9502.030 MHz.

5.1 *Metallic Test Sections*

The general effect of decreasing the conductivity of the waveguide walls for frequencies in and near the cutoff region can be seen in Fig. 9. The conductivity ranges from the dc value of oxygen-free hard electroformed copper, 5.8×10^5 mho/cm (reciprocal ohms per centimeter), to approximately one tenth the conductivity of copper. At a single frequency, as the conductivity is decreased, the imaginary part of the propagation constant, β, increases. The real part of the propagation constant, α, decreases with decreasing conductivity for fre-

---

* The micarta used was made of woven cotten impregnated with cresylic acid formaldehyde resin.
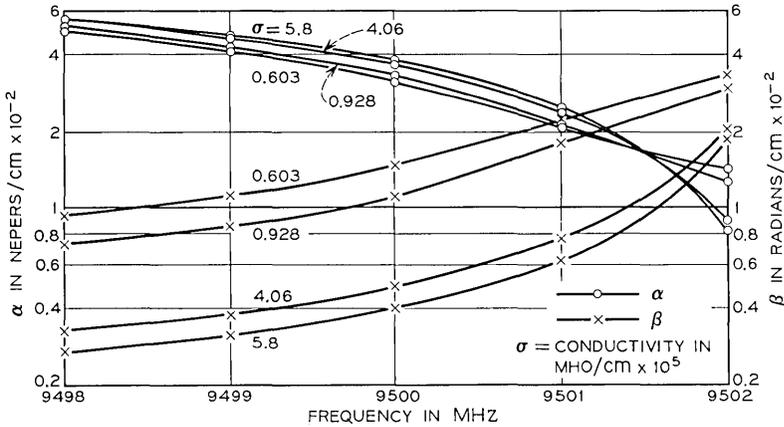
Fig. 9 — Propagation constant in cutoff region as a function of conductivity. $\gamma = a + j\beta$.

quencies just below cutoff, but α increases for frequencies just above cutoff.

It is interesting that in Fig. 9 there is one frequency in the cutoff region for each value of conductivity where the real part of the propagation constant (α) in nepers per centimeter equals the imaginary part of the propagation constant (β) in radians per centimeter. This frequency is properly defined as the cutoff frequency when the waveguide has walls with finite conductivity. Further examination of Fig. 9 shows that this defined cutoff frequency shifts to a lower frequency as the conductivity of the waveguide walls is decreased. We see that a decrease in conductivity by a factor of ten causes this defined cutoff frequency to shift by 850 KHz. Since microwave frequencies in this frequency region can now be measured to 1 KHz, the potential accuracy obtained by using this frequency shift for the measurement of the conductivity of metals is less than 1 per cent.

5.1.1 Copper Test Sections

The copper test section was measured at two different frequencies in the cutoff region. These frequencies, 9500.873 and 9497.960 MHz, were chosen to cover a region where Fig. 9 shows a maximum difference in α and β for the range of expected copper conductivity. The calculated values of γl (l in cm) for the first test frequency are plotted in Fig. 10 as a function of the conductivity (in mho/cm.) The value of γl was used instead of γ to make comparison with the measured values easier, since

the measurement involves the total length of the test section and since the experimental errors are in dB and degrees. Each figure contains two curves; one for the total attenuation in dB, and one for the total phase shift in degrees.

The experimental values are plotted as points marked α_M and β_M on the figure. The vertical broken lines with markings $\Delta\alpha$ and $\Delta\beta$ indicate the error limits in the experimental measurements. In Fig. 10, the error limits are smaller for the β_M measured value than for the α_M value. The results at the second frequency, although not plotted, support the plotted results.

The average value of the conductivity of copper based on the α_M
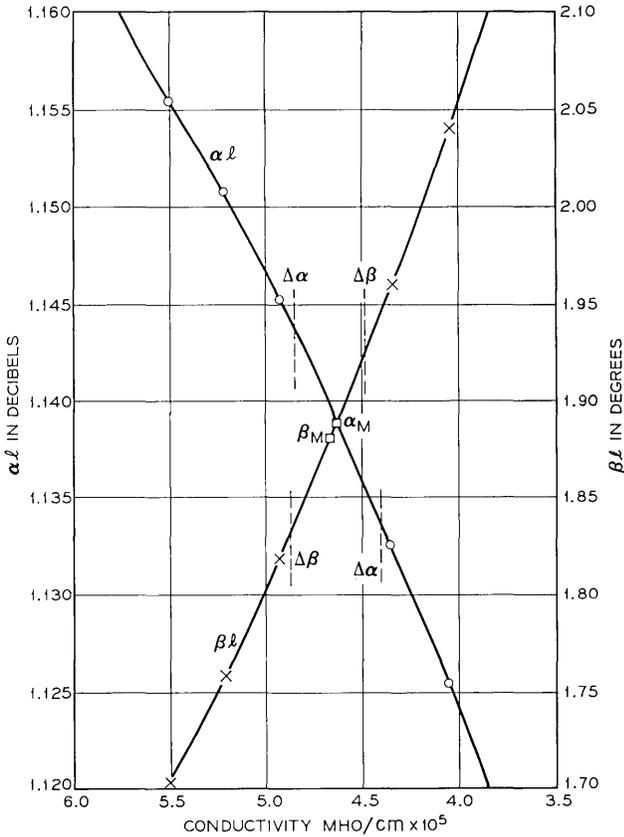
Fig. 10 — Copper experimental results at 9500.873 MHz yielding experimental value of conductivity. ○ calculated αl, × calculated βl, and □ measured values.

measurement at the two frequencies is 4.635×10^5 mho/cm. The average value of the β_M values is 4.685×10^5 mho/cm. The average of the mean of the maximum and minimum values of α_M and β_M at the two frequencies is 4.69×10^5 mho/cm. The best value for the conductivity of oxygen-free hard electro-formed copper at 9500 MHz is taken to be the average of the α_M and β_M averages, 4.66×10^5 mho/cm. The error in this value, based on the errors in the measured values is ± 1.5 percent. This value of the conductivity of copper is 80.3 percent of the dc value of copper. Previously reported values for the conductivity of copper in this frequency range, based on measurements of long lengths of waveguide operated in the propagation frequency region, vary between 85 and 78 percent of the dc conductivity.[15]

5.1.2 *Nickel Test Sections*

The nickel test section was made by electroplating a 0.001-inch-thick layer of commercial grade nickel on the four walls of the cutoff section of an electroformed copper test section. The wall conductivity of this copper test section was measured before plating. These results are not repeated since this measured conductivity agrees with the previously measured value of the conductivity of the copper within the error limits mentioned earlier.

The experimental measurements were made in the manner already described, at two test frequencies, 9500.873 and 9497.963 MHz. The calculated values of γl and the experimental points for the first test frequency is plotted in Fig. 11 as a function of conductivity. As was done for the copper measurements, two curves are plotted on the figure, one for the total real part of γl, and one for the total imaginary part of γl.

Figure 11 shows some interesting features of the cutoff region. The total attenuation of the nickel test section is less than that for the copper test section, indicating that the cutoff region has shifted to a lower frequency. Consistent with this shift in the cutoff region to lower frequencies is the increase in the total phase shift. However, as the conductivity is decreased below 6×10^4 mho/cm, the total attenuation increases. This result indicates that the loss resulting from the decreased conductivity is increasing faster than the cutoff region is shifting to a lower frequency by the decreasing conductivity. This difference leads to a net increase in the total loss of the cutoff section.

Only the experimental values for the first test frequency are plotted. The results at the second frequency support these results. The points
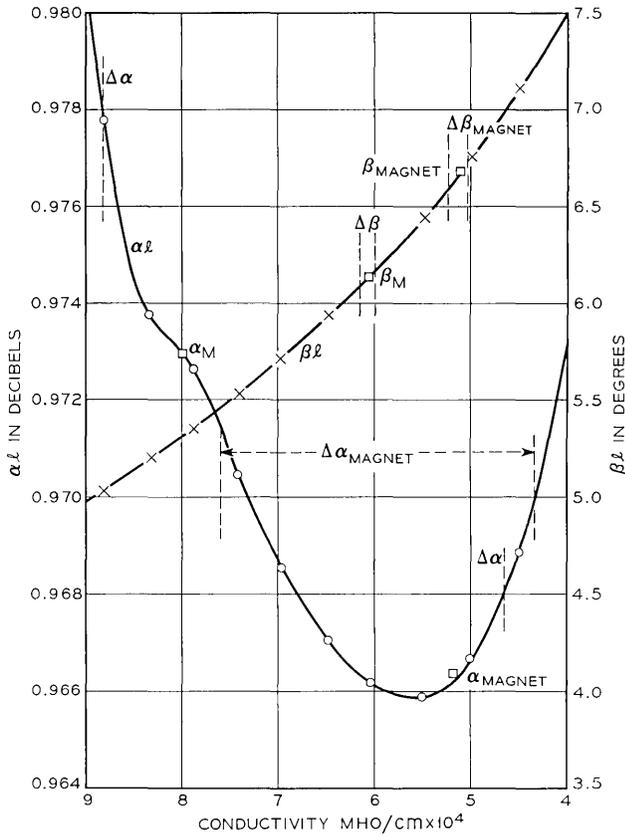
Fig. 11 — Nickel experimental results at 9500.873 MHz yielding experimental value of conductivity including effect of applied dc magnetic field. "Magnet" indicates results with magnetic field applied. ○ calculated αl, × calculated βl, and □ measured values.

labeled α_M and β_M are the experimental values. The vertical broken lines enclosing $\Delta\alpha$ and $\Delta\beta$ define the error limits in the experimental measurements. The error limits in Fig. 11 are small for the phase measurement, β_M, while the error limits for the attenuation measurement, α_M, are large; hence the measured β value yields the more accurate result. The figures for the copper and nickel test sections illustrate a feature of this experimental technique. At some frequency in the cutoff region both the α and β measurement may have the same accuracy, while at another frequency either the α or β measurement will yield a more accurate result. This feature, of course, depends on the errors in the measuring

equipment. It does allow one to choose test frequencies which will compensate for the errors in the measuring equipment.

The β_M measurements at the two test frequencies both yield a value of 6.10×10^4 mho/cm for the conductivity of commercial grade nickel plating. The α_M measurements yield conductivity values of 8.0×10^4 mho and 6.4×10^4 mho/cm. However, the error limits of the α_M measurements include the β_M measured values. In this case, one concludes that the most accurate measure of the conductivity of nickel is the β_M value. The maximum range of the measured value of conductivity based on the β_M measurement is 6.3×10^4 to 5.9×10^4 mho/cm or an error of 3.6 percent. The minimum range of the measured values is 6.0×10^4 to 6.15 mho/cm or an error of 1.7 percent.

The dc conductivity of nickel as given by various tables of the properties of metals[16] is 1.28×10^5 mho/cm. The experimentally determined value for the conductivity of nickel at 9500 GHz is 6.10×10^4 mho/cm or 47.6 per cent of the dc conductivity. The conductivity of copper at 9500 GHz was determined earlier in this paper to be 80.3 per cent of the dc value. Electroplated metals have been reported to be more porous than solid metals.[15] This increased porosity would account for the larger decrease in the conductivity of nickel compared with copper in these experiments.

The nickel plated test section was used for a second experiment. The test section was subjected to a magnetic field of 2100 gauss as discussed in Section IV. The actual field applied to the nickel was difficult to determine accurately because of the size of the Hall plate available to measure the field. It is estimated that a field of 500 gauss was applied to the nickel walls. This same field was applied to the copper test section before plating. No measurable effects were obtained.

The results of the experiment with the magnetic field are plotted in Fig. 11 as the points, α_{MAGNET} and β_{MAGNET}. It is evident from the location of these points on the calculated curves that the application of the magnetic field has caused an apparent decrease in the conductivity of nickel. The mean value of the conductivity resulting from the application of the magnetic field is 4.95×10^4 mho/cm.

The exact cause of this decrease in conductivity is not known. Since there was no effect of the magnetic field on the copper test section before plating, we can assume that the decrease in the conductivity of nickel resulted from the ferromagnetic properties of nickel. This effect can then be explained by assuming that the magnetic field increases the effective microwave permeability of nickel by the ratio of

6.10/4.95. The effect of the conductivity of the metal walls of a waveguide enter into the calculation of the propagation constant through the expression

$$z = \left(\frac{\omega\mu}{2\sigma}\right)^{\frac{1}{2}}.$$

Hence, the increase in the permeability, μ, causes the same result as a decrease in conductivity. It is well known that nickel is ferromagnetic at low frequencies. The ferromagnetic property is described by its permeability. Evidently, if this explanation is correct, nickel exhibits a small ferromagnetic effect at microwave frequencies.

5.1.3 *Nichrome-Copper Test Section*

The conductivity of the walls of one of the electroformed copper type B test sections was measured and found to agree with the original copper test section within the stated error limits. Nichrome was applied to the two narrow walls as described in Section IV, and an experiment was conducted at 9497.936 MHz. The calculated values and the experimental results are plotted in Fig. 12. This figure shows the effect of conductivity on the cutoff region not present in the previous results. The α and β curves have a somewhat unusual shape.

These new features are not too unusual when it is considered that we are dealing with the combination of a composite metal, nichrome over copper, on two walls of the waveguide, and a single metal, copper, on the remaining walls. The thickness of the evaporated nichrome is much less than the nichrome skin depth. The effect of this combination of metals is best understood by considering the intrinsic wall impedance defined in equation (9). The variation in this wall impedance with the change in the conductivity of the coating metal for a fixed coated metal is discussed by Ramo and Whinnery.[11]

The variation in the total attenuation as the nichrome conductivity is decreased results from two factors, the shift in the cutoff region and the increase in the skin depth of the nichrome. The rate at which these two factors change as the conductivity decreases governs the shape of the α and β curves. This effect can be explained simply by considering the α curve. The explanation of the shape of the β curve is more complicated and would involve repeating the analysis of Ramo and Whinnery. This is not necessary for our purposes.

As the nichrome conductivity is decreased, the skin depth of the nichrome is increased. The presence of the nichrome has less effect on
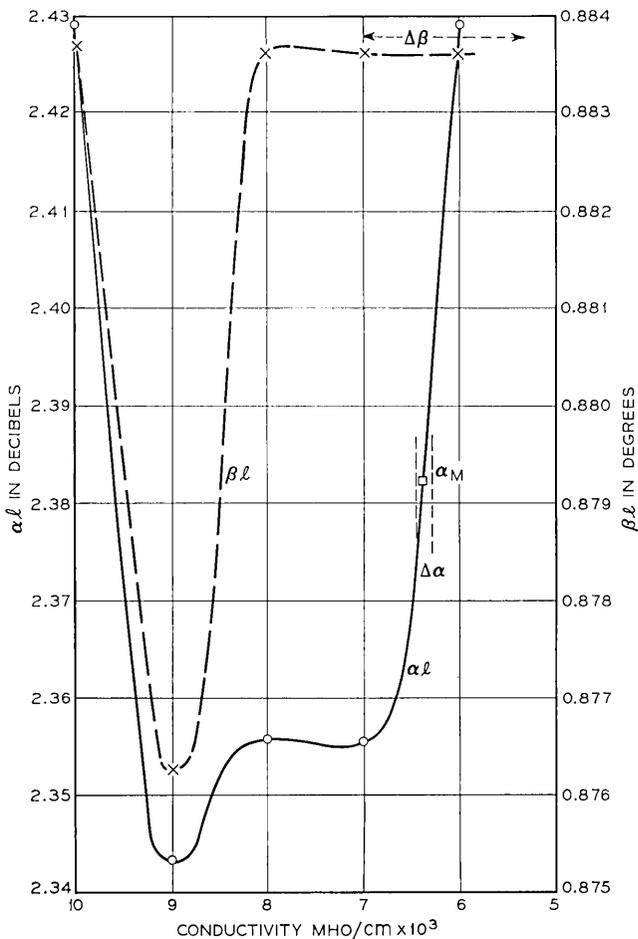
Fig. 12 — Nichrome-copper experimental results at 9497.936 MHz yielding experimental value of conductivity. ○ calculated αl, × calculated βl, and □ measured values.

the microwave currents and the total conductivity approaches that of copper. However, until the nichrome conductivity decreases sufficiently, the conductivity of the nichrome-copper combination is less than that of copper and hence causes a shift of the cutoff region to lower frequencies with an attendant decrease in the total attenuation. As the nichrome conductivity is decreased further, the effective conductivity approaches that of copper and the cutoff region shifts to

higher frequencies. The attenuation increases again approaching that of a copper test section.

The experimental value of α, α_M, is plotted in Fig. 12. The experimental value of β, β_M, lies on the flat portion of the β curve with error limits that cover the extent of the flat portion plotted. Thus the β measurement gives no accurate measure of the conductivity of nichrome. The experimental value of nichrome conductivity at 9497 MHz based on the α measurement is 6.4×10^3 mho/cm. The maximum error is 1.5 percent. The dc conductivity of nichrome is 10^4 mho/cm (See Ref. 16). The measured value of nichrome at 9497 MHz is 64 percent of the dc value.

5.2 Dielectric Loaded Cutoff Test Sections

It is well known that the insertion of a dielectric into a waveguide section causes an increase in the phase shift per unit length for frequencies above the cutoff region. The effect of lossy dielectrics placed in a waveguide section operated in the cutoff region is not well known.

Figure 13 shows the effect of a lossy dielectric in a waveguide over a frequency range covering the cutoff frequency region. The curves of the real and imaginary part of the propagation constant were plotted for several dielectric constants and loss tangents for a dielectric slab 0.059 inch thick inserted in a copper waveguide 0.62150 inch wide. The unloaded cutoff frequency of this waveguide is approximately 9500 MHz. The waveguide has a wall conductivity of 4.64×10^5 mho/cm.

Examination of these curves shows that increasing the dielectric constant for a constant loss tangent shifts the cutoff region to a lower frequency (α decreases, β increases). For a constant dielectric constant an increase in the loss tangent shifts the cutoff region to a higher frequency (α increases, β decreases). Although not shown in Fig. 13, a decrease in the wall conductivity of the waveguide shifts the set of curves to a lower frequency.

5.2.1 Lucite Dielectric

The copper test section used for the original measurement of the conductivity of copper was used for the dielectric experiments. The 0.056-inch thick slab of lucite was inserted in the copper test sections.

The experimental values of α and β were compared with a series of curves calculated for various combinations of dielectric constants, ϵ'/ϵ_0, and loss tangents, ϵ''/ϵ_0, for the two test frequencies, 8361.653
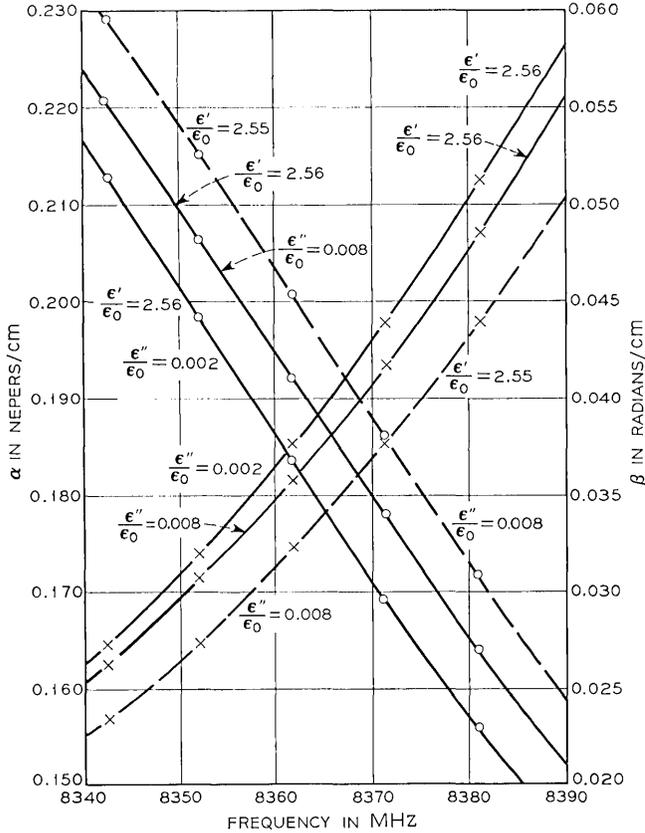
Fig. 13 — Propagation constant in cutoff region as a function of relative dielectric constant and loss tangent. ◯ calculated αl, and ✕ calculated βl.

and 8351.945 MHz. Fig. 14 was plotted for those values which agreed with the experimental results. The results for 8351.945 MHz are not presented because they give the same result as in Fig. 14. The experimental values, α_M and β_M, are plotted on the respective curves of these two figures. The vertical broken lines indicate the error limits of the measured values. The curve in Fig. 14 was plotted for $\epsilon'/\epsilon_0 = 2.55$ and a range of loss tangent values, ϵ''/ϵ_0.

The error in the value of the measured α_M can be seen to be much less than that of β_M. This is an example of a case discussed in Section 5.1.2 where one of the parts of the propagation constant can be measured with greater accuracy than the other at the chosen frequency in the cutoff

region. From the experimental results, the experimental value of the dielectric constant of lucite is 2.55. The error limits, although not shown on the curves are 2.56 and 2.545. The measured values of the loss tangent are 0.0065 and 0.0066 giving a mean value of 0.00655. The error limits at 8361.653 MHz are 0.0064 and 0.0066; and at 8351.945 MHz, 0.0065 and 0.0067. The maximum error in the mean value of the loss tangent is 4.5 percent, and the minimum error, 1.5 percent.

The measured values of β_M do not lie at the same values of loss tangent as those of α_M. However, the error limits of the experimental values of β_M enclose the error limits of the experimental values of α_M. The measured values of β_M, while not agreeing with those of α_M support the more accurate values of α_M.

The experimental values of the dielectric constant and the loss tangent
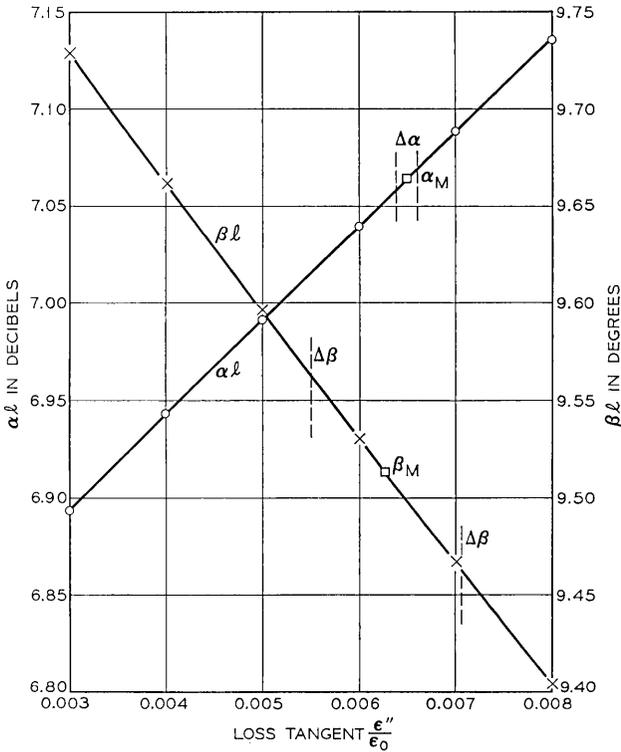
Fig. 14 — Lucite dielectric experimental results at 8361.653 MHz yielding experimental values of relative dielectric constant and loss tangent. ○ calculated αl, × calcuated βl, and □ measured values.

for the lucite dielectric at 8350 MHz can be taken as 2.55 and 0.00655, respectively. The values reported for lucite at 10 GHz are 2.59 and 0.006, respectively.[13]

5.2.2 Micarta Dielectric

The lucite dielectric slabs in the copper test sections were replaced with 0.031-inch thick micarta slabs for an experiment in which the total attenuation and the total phase shift for various values of the dielectric constant ϵ'/ϵ_0 and the loss tangent ϵ''/ϵ_0 were calculated. The results which satisfy the experimental results are plotted in Fig. 15 for the test frequency 8477.289 MHz. Other experiments were performed at 8455.512 MHz. These experiments, although not plotted, support the results of Fig. 15.
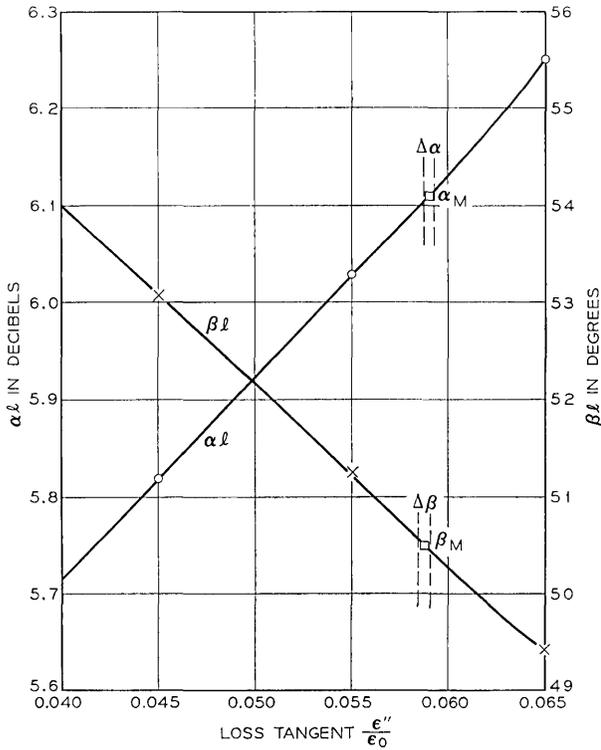
Fig. 15 — Micarta dielectric experimental results at 8477.289 MHz yielding experimental values of relative dielectric constant and loss tangent. ○ calculated αl, × calculated βl, and □ measured values.

In these experiments with micarta, the experimental results at 8477.289 MHz yielded a dielectric constant $\epsilon'/\epsilon_0 = 3.62$, and a loss tangent based on the α measurement of $\epsilon''/\epsilon_0 = 0.0575$, and for the β measurement, $\epsilon''/\epsilon_0 = 0.0580$. The results at 8455.512 MHz yielded a dielectric constant of 3.60 and a loss tangent for the α measurement of 0.585 and for the β measurement of 0.0585. The mean dielectric constant, determined from the measurement at the two test frequencies, is $\epsilon'/\epsilon_0 = 3.61$ with an error of ± 1.5 percent. The mean loss tangent determined from these measurements is $\epsilon''/\epsilon_0 = 0.058$ with an error of ± 1 percent.

Published tables of the properties of dielectric materials list dielectric constants ranging from 3.42 to 3.78 and loss tangents ranging from 0.05 to 0.08 for micarta at 10 GHz. The range of values stems from slightly different formulations used in the manufacture of micarta. Since the definite composition of our sample of micarta is not known, it is evident that our results are quite justified.

The results discussed in the preceding sections are summarized in Table I, which lists the metal or dielectric, the frequency of measurement, the measured values of the indicated electrical properties, and the value of these properties as determined by other measurement techniques.

VI. CONCLUSIONS

The effects of various metals and dielectrics on the properties of the cutoff region of a rectangular waveguide operated in the dominant mode have been investigated. It has been shown that a waveguide with walls of finite conductivity has a cutoff region instead of a singular cutoff frequency associated with a lossless waveguide. As the conductivity of the waveguide walls is reduced, the cutoff region is shifted to a lower frequency.

It is evident that the definition of the cutoff frequency for a lossless guide does not apply when losses are present. The definition of cutoff frequency should take into account the conductivity of the walls. The cutoff frequency for a given mode may be defined (for a given conductivity) as that frequency where the real part of the propagation constant in nepers per unit length is equal to the imaginary part of the propagation constant in radians per unit length. For the same physical dimensions, a waveguide operated in the dominant mode with walls of conductivity σ_1 would have a higher cutoff frequency than a waveguide with walls of conductivity σ_2 for $\sigma_1 > \sigma_2$.

TABLE I—MEASUREMENTS AT 72°F, 50 PERCENT
RELATIVE HUMIDITY

Metal	Measured		Published Values*	
	Conductivity (mho/cm)	Frequency (MHz)	Conductivity (mho/cm)	Frequency (MHz)
Copper (electroformed)	4.66×10^5	9500	5.8×10^5 4.64×10^5 3.15×10^5	0 (dc) 10,000 24,000
Nickel (commercial plated)	6.10×10^4	9500	1.28×10^5	0 (dc)
Nickel (with 800 gauss H field)	4.95×10^4	9500	None Available	
Nichrome (evaporated)	6.4×10^3	9500	1.0×10^4	0 (dc)

Dielectric	Measured			Published Values†		
	ϵ'/ϵ_0	ϵ''/ϵ_0	Frequency (MHz)	ϵ'/ϵ_0	ϵ''/ϵ_0	Frequency (MHz)
Lucite (sheet)	2.55	0.00655	8350	2.59	0.006	10,000
Micarta (sheet)	3.61	0.058	8460	3.62	0.057	10,000

* See Refs. 6, 9, 15, and 16.
† See Refs. 9 and 13.

The introduction of a lossy dielectric into a rectangular waveguide operated in the dominant mode with walls of finite conductivity has a pronounced effect on the cutoff frequency region. A lossless dielectric inserted into a lossless waveguide produces a singular cutoff frequency at a frequency lower than that of the waveguide alone. When the waveguide walls have finite conductivity and the dielectric has a finite loss tangent, there is a cutoff region rather than a singular frequency. In this cutoff frequency region, for a constant dielectric constant, an increase in the loss tangent causes the cutoff frequency region to shift to a higher frequency. For a constant loss tangent, an increase in the dielectric constant causes the cutoff region to shift to a lower frequency. For a constant dielectric constant and loss tangent, a decrease in the wall conductivity of the waveguide causes a shift of the cutoff region to a lower frequency.

In the general case of a waveguide with walls of finite conductivity and a dielectric with a finite loss tangent, the cutoff frequency may again be defined as that frequency at which the real part of the propagation constant in nepers per unit length is equal to the imaginary part

in radians per unit length. So defined, there are generally distinct cut-off frequencies for each combination of wall conductivity, dielectric constant, loss tangent, and waveguide dimensions.

Having discussed the effect of metals and dielectrics on the cutoff region of the dominant mode of a rectangular waveguide, we turn to uses of this waveguide phenomenon. The most prominent use of the cutoff region has been examined in detail; that of measuring the properties of metals and dielectrics at microwave frequencies.

The properties of three metals, copper, nickel, and nichrome, and two dielectrics, lucite and micarta, were measured using the effect of these materials on the cutoff region. The experimental values of the metal conductivities and the relative dielectric constant and loss tangent of the dielectrics are given in Section V. The accuracy of all measured values was about ±2 per cent, although some measurements were accurate to ±1 per cent.

There are little published data on the microwave conductivity of these metals at the frequencies used for the experiments. What data are available agrees with our results to within 5 per cent. The error limits of the published values were not given; hence it is not possible to check the accuracy of the experimental values in this way.

The decrease in the conductivity of nickel in the presence of a dc magnetic field demonstrates an effect not observed in the measurements of the other metals. The exact cause of this effect is not known. It is suggested that, since nickel is ferromagnetic, the magnetic field caused a small increase in the microwave permeability of nickel. The analysis of a lossy cutoff waveguide operated at cutoff depends on the intrinsic wall impedance. Within this approximation, it is evident that an assumed increase in permeability produces the same effect as the measured decrease in the conductivity of nickel.

There are published data for lucite at 10 GHz. The values obtained from the cutoff waveguide measurements agree within 2 per cent of these values. Interpolating between the published values to obtain values for 8.5 GHz brings the agreement to about 1 per cent. The exact chemical composition of the micarta dielectric was not known. There are a range of values given in *Tables of Dielectric Properties* for different micarta compositions.[13] These published values bracket the experimental results obtained from the cutoff measurements.

Lucite was chosen as one of the test materials in order to establish a known reference to determine the total error inherent in this analysis and measurement technique. The experimental results show that the measured value of the electrical properties of lucite agree to within

1 per cent of values determined by other techniques.[13] The analysis of a lossy dielectric slab centered in a lossy waveguide operated in the cutoff frequency region requires more approximations than the analysis of the empty lossy waveguide. Hence, we would expect the maximum error to be present in the measurement of the lucite dielectric. The small error for lucite, 1 per cent, is indicative of the accuracy of this technique for measuring electrical properties of metals and dielectrics.

REFERENCES

 1. Harrington, R. F., *Time Harmonic Electromagnetic Fields,* New York: Mc-Graw-Hill, 1961.
 2. Ginzton, E. L., *Microwave Measurements,* New York: McGraw-Hill, 1957.
 3. Karbowiak, A. E., "Theory of Imperfect Waveguides: The Effect of Wall Impedance," Proc. IEEE (London), *102,* part B. No. 5 (September 1955), pp. 698–707.
 4. Papadopoulos, V. M., "Propagation of Electromagnetic Waves in Cylindrical Waveguides with Imperfectly Conducting Walls," Quart. J. Mechanical and Applied Math. *VII,* pt. 3 (September 1954), pp. 326–334.
 5. Kahn, W. K., "Power Transmission Through General Uniform Waveguides," IRE Trans. Microwave Theory and Techniques, *MTT-10,* No. 5, (September 1962), pp. 328–331.
 6. Marcuvitz, N., *Waveguide Handbook,* vol. 10, Radiation Laboratory Series, New York: McGraw-Hill, 1951.
 7. Barrow, W. L., "Transmission of Electromagnetic Waves in Hollow Tubes of Metal," Proc. IRE, *24,* No. 10, (October 1936), pp. 1298–1329.
 8. Linder, E. G., "Attenuation of Electromagnetic Fields in Pipes Smaller than the Critical Size," Proc. IRE, *30,* No. 12 (December 1942), pp. 554–556.
 9. Southworth, G. C., *Principles and Applications of Waveguide Transmission,* Princeton, N. J., D. Van Nostrand Co., 1951.
10. Montgomery, C. G., Dicke, R. H., and Purcell, E. M., *Principles of Microwave Circuits,* vol. 8, Radiation Laboratory Series, New York: McGraw-Hill, 1948.
11. Ramo, S. and Whinnery, J. R., *Fields and Waves in Modern Radio,* New York: John Wiley and Sons, Inc., 1956.
12. Adler, R. B., Chu, L. J., and Fano, R. M., *Electromagnetic Energy Transmission and Radiation,* New York: John Wiley and Sons, Inc., 1965.
13. von Hipple, A., *Tables of Dielectric Materials,* Cambridge, Mass.: Laboratory for Insulation Research, Massachusetts Institute of Technology, April 1957.
14. Cohn, S. B., and Weinhouse, N. P., "An Automatic Microwave Phase-Measurement System," Microwave J., 7, No. 2 (February 1964), pp. 49–56.
15. E. Maxwell, "Conductivity of Metallic Surfaces," J. Appl. Phys. *18,* No. 7 (July 1947), pp. 629–638.
16. Hodgman, C. D., ed., *Handbook of Chemistry and Physics,* Cleveland, Ohio: Chemical Rubber Pub. Co., 1967.

Intermodal Coupling at the Junction
Between Straight and Curved Waveguides

By C. P. BATES

This paper analyses the coupling of electromagnetic modes at the junction between straight and continuously curved rectangular waveguides. The method of solution is based on an integral equation formulation, applicable for sharp as well as gradual bends. Such quantities as the average power transmitted or reflected into each of the various modes propagating in the straight and curved waveguide sections are readily obtained.

The article presents the results of representative calculations for the two types of waveguide bends. These include graphs of the energy distribution in the transmitted and reflected modes as a function of dimensionless ratios for a sharp bend; the range of values considered allows immediate application of the results to standard C-band waveguides. The gradual bend example uses parameters encountered in the waveguide connections to an antenna in a typical microwave relay network.

I. INTRODUCTION

In a microwave system for guiding electromagnetic waves, often there are bends formed by connecting straight and continuously curved rectangular waveguides (see Fig. 1). Precise numerical computations and extensive analytical investigations of the angular propagation constants for the various electromagnetic modes in the curved section alone have been published by Cochran and Pecina.[1] The propagation constants and modal fields which may exist in the straight sections alone are trivial. To understand propagation of electromagnetic waves through these waveguide bends, therefore, requires a complete comprehension of the intermodal coupling that takes place at the various junctions and discontinuities. This paper investigates the coupling that occurs where straight and continuously curved rectangular waveguides join.

This type of structure has been studied to some extent by others.

2259
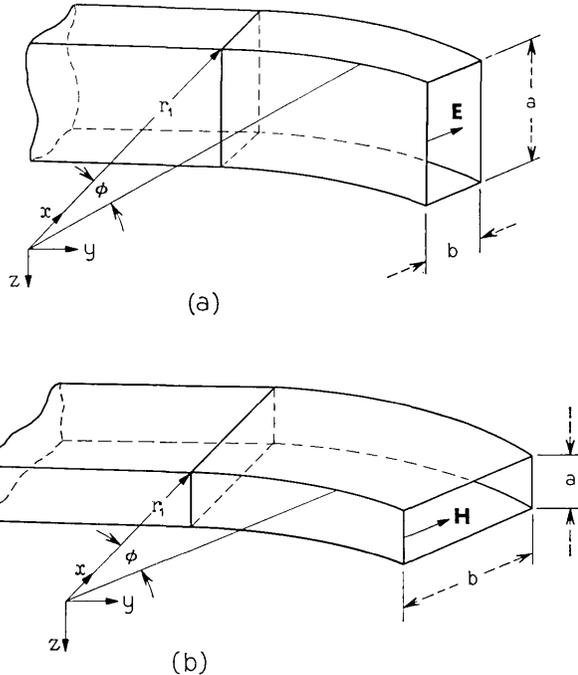
Fig. 1 — Waveguide bends formed by connecting straight and continuously curved rectangular waveguides. (a) E-plane bend. (b) H-plane bend.

There is an approach based on a matrix calculus formulation by Rice.[2] Using a perturbation method, Jouguet obtained expressions for the fields in the curved waveguide up to terms of second order, that is, to terms in $1/R^2$, where R denotes the radius of curvature of the axis of the curved guide.[3] He uses these approximate expressions to determine the intermodal coupling that results at the junction between the straight and curved waveguides for a particular polarization of the field. In contrast with Jouguet's approach, the analysis we use permits the waveguide bends to be as sharp as desired, while still including the gradual bend within the permissible range of parameters.

Our approach involves the solution of a boundary value problem formulated in terms of the appropriate modal expansions for the fields in the straight and curved waveguides (see Section 2.1). The modal functions and propagation constants in these waveguide sections consist of certain combinations of trigonometric and Bessel functions and the zeros of such combinations. Evaluation of the appropriate quan-

tities for the curved waveguide is one of the more difficult aspects of this problem and necessitates not only numerical methods for determining zeros and asymptotic expansions but also computer algorithms for the accurate evaluation of Bessel functions. Such algorithms have been recently developed and programmed at Bell Telephone Laboratories.[4]

With the modal expansions in hand, one can formulate an integral equation for the aperture field at the junction between the straight and curved waveguides. This equation, as discussed in Section 2.2, may be solved numerically by the method of moments to within a reasonable accuracy (error criteria are discussed in Section 4.1). A solution for the fields in the waveguides can then be easily obtained, and such quantities as the power reflected or transmitted into various modes at the junction may be evaluated.

Section 4.2 gives examples of the intermodal power coupling for both sharp and gradual bends. Section 4.2.1 presents the results for the sharp bend example as a function of certain dimensionless ratios; the range of values considered allows direct application of the results to standard C-band waveguides. The results clearly demonstrate that significant intermodal power coupling takes place; they also establish the exaggeration which occurs in the reflected powers near the cutoff frequencies of the individual modes. Section 4.2.2 gives the results for the gradual bend example and shows that reflections are negligible and hence only the forward coupling has significant levels for the gradual bend considered.

II. FORMULATION AND SOLUTION OF THE BOUNDARY VALUE PROBLEM

2.1 Fields in Straight and Continuously Curved Waveguides

An arbitrary electromagnetic field, which may exist in either the straight or continuously curved waveguide, may be expressed as a sum of the longitudinal electric (LE) and longitudinal magnetic (LM) modes appropriate to that section (for explicit details on such modal representations in a continuously curved waveguide see Cochran and Pecina and for the straight waveguide see Harrington[1,5]). The LE modes have an electric field transverse to the z-direction which means this field component lies in the longitudinal plane, while the LM modes have their magnetic field similarly positioned.

The explicit form of the transverse components of the LE model expansion, suppressing an $\exp(jwt)$ time convention, is given below

(for a *LE* field and straight waveguide):

$$^s\mathbf{E}^e = \sum_{m,n} A^{\pm}_{mn} \; ^s\mathbf{e}^e_{mn} \; \exp \; (\mp j\beta_{mn}y),$$

$$^s\mathbf{H}^e = \pm \sum_{m,n} A^{\pm}_{mn} \; ^s\mathbf{h}^e_{mn} \; \exp \; (\mp j\beta_{mn}y)$$

(1)

with

$$^s\mathbf{e}^e_{mn} = \varphi^s_m(x)\psi_n(z)\hat{x},$$

$$^s\mathbf{h}^e_{mn} = \frac{1}{j\beta_{mn}Z} \frac{d}{dx} \varphi^s_m(x) \frac{d}{dz} \psi_n(z)\hat{x} + \frac{h^2_n}{j\beta_{mn}Z} \varphi^s_m(x)\psi_n(z)\hat{z},$$

$$\varphi^s_m(s) = (\epsilon_m/b)^{\frac{1}{2}} \cos \; [m\pi/b(r_2 - x)],$$

$$m = 0, 1, 2, \cdots, \quad \epsilon_0 = 1, \quad \epsilon_m = 2, \quad m \geqq 1,$$

$$\psi_n(z) = (2/a)^{\frac{1}{2}} \sin \; (n\pi z/a), \quad n = 1, 2, \cdots,$$

$$\beta_{mn} = [h^2_n - (m\pi/b)^2]^{\frac{1}{2}} = -j[(m\pi/b)^2 - h^2_n]^{\frac{1}{2}}, \quad m = 0, 1, 2, \cdots,$$

$$h_n = [k^2 - (n\pi/a)^2]^{\frac{1}{2}}, \quad n = 1, 2, \cdots,$$

and

$$k = \omega/c, \quad Z = j\omega\mu, \quad r_2 = r_1 + b.$$

Here $^s\mathbf{E}^e$ is the transverse electric field intensity, $^s\mathbf{H}^e$ the transverse magnetic field intensity, ω the angular frequency, k the wave number, μ the permeability, and c the phase velocity of the medium filling the guide. The vector components which make up the field are given by the lower case letters. The $A^{\pm}_{mn}$ are the unknown expansion coefficients of the individual *LE* modes in the straight guide with the ($\pm$) indicating waves traveling either in the positive or negative y-direction (towards or away from the junction in the straight section of Fig. 1a). The propagation constant of a particular mode is β_{mn}, and it is either real or purely imaginary (providing the guide is filled with a lossless medium) thus indicative of either a traveling or evanescent mode.

In the curved waveguide, using polar coordinates (ρ, φ, z), one has (for the *LE* field and curved guide):

$$^c\mathbf{E}^e = \sum_{m,n} C^{\pm}_{mn} \; ^c\mathbf{e}^e_{mn} \; \exp \; (\mp j\nu_{mn}\varphi),$$

$$^c\mathbf{H}^e = \pm \sum_{m,n} C^{\pm}_{mn} \; ^c\mathbf{h}^e_{mn} \; \exp \; (\mp j\nu_{mn}\varphi)$$

(2)

with

$$^c\mathbf{e}^e_{mn} = \frac{1}{\rho} \varphi^c_{mn}(h_n\rho)\psi_n(z)\hat{\rho},$$

$$\mathbf{h}_{mn}^e = \frac{1}{j\nu_{mn}Z} \frac{d}{d\rho} \varphi_{mn}^c(h_n\rho) \frac{d}{dz} \psi_n(z)\hat\rho + \frac{h_n^2}{j\nu_{mn}Z} \varphi_{mn}^c(h_n\rho)\psi_n(z)\hat z,$$

$$\varphi_{mn}^c(h_n\rho) = C_{\nu_{mn}}(h_n\rho)/\| C_{\nu_{mn}}(h_n\rho) \|,$$

$$C_{\nu_{mn}}(h_n\rho) = J'_{\nu_{mn}}(h_nr_2)Y_{\nu_{mn}}(h_n\rho) - Y'_{\nu_{mn}}(h_nr_2)J_{\nu_{mn}}(h_n\rho),$$

and

$$\| C_{\nu_{mn}}(h_n\rho) \| = \left(\int_{r_1}^{r_2} C_\nu^2/\rho \, d\rho\right)^{\frac{1}{2}}.$$

In these expressions $J_\nu(x)$ and $Y_\nu(x)$ are the Bessel functions of the first and second kind respectively; the prime indicates differentiation with respect to the argument. The permissible propagation numbers ν_{mn}, in this case, are given by the implicit solutions of

$$\frac{d}{d\rho} C_{\nu_{mn}}(h_n\rho) \big|_{\rho=r_1} = 0, \qquad m = 0, 1, 2, \cdots,$$

and again they are either real or purely imaginary providing the guide is filled with a lossless medium.[6] Section III discusses the modal function, C_ν, in more detail. The $C_{mn}^\pm$ are the expansion coefficients of the individual modes with the ($\pm$) again designating the direction of mode travel.

In equations (1) and (2) the superscript e indicates that the particular vector is an LE component and the superscript s or c indicates that the vector or function is associated with the straight or curved sections. The subscripts m and n are the modal indices. In Section IV, where results are also given for an LM polarization, a superscript m designates such fields.

One may easily verify that these transverse LE field components, along with their longitudinal counterparts, satisfy Maxwell's equations in the appropriate regions and that the required boundary conditions, namely, zero tangential electric field and zero normal magnetic field on the waveguide walls, are met. Such representations are complete in that any arbitrary fields in the straight and curved waveguides which have their electric components confined to the longitudinal plane can be expanded in the form of equation (1) or (2), respectively.

Appropriate expressions may also be written for the transverse components of the LM modal expansions. They would also be complete in the sense that any arbitrary fields in the straight and curved waveguides which have their magnetic components confined to the longitudinal plane could be expanded in such a representation.

It can be shown, for the geometry indicated in Fig. 1, that an LE

source in either the straight or curved waveguide excites only an *LE* field, and conversely an *LM* source excites only an *LM* field. Hence a waveguide bend excited by an *LE* mode is usually referred to as an E-plane bend, in keeping with the fact that the *LE* source sets up only an *LE* field for which the electric field is confined to the longitudinal plane, that is, the plane of the bend. Figure 1a shows the typical waveguide geometry for an E-plane bend. Analogously, an H-plane bend is one for which the magnetic field is confined to the plane of the bend; this occurs when the source and hence the resulting fields are *LM*. Figure 1b shows typical geometry for this case.

Notice that the transverse vector components can be shown to satisfy

$$\iint {}^{s}\mathbf{e}_{mn}^{e} \cdot {}^{s}\mathbf{e}_{rs}^{e} \, dA = \delta_{mr} \, \delta_{ns}^{\dagger} , \qquad \iint {}^{c}\mathbf{e}_{mn}^{e} \cdot {}^{c}\mathbf{e}_{rs}^{e} \, \rho \, dA = \delta_{mr} \, \delta_{ns} . \qquad (3)$$

In equation (3) the integration is taken over the cross-sectional area of the appropriate waveguide interior. Such orthogonalities are a consequence of the differential equations and the boundary conditions satisfied by the scalar parts of the transverse vector components.

2.2 *Integral Equation Formulation and Solution*

As discussed in Section 2.1, the fields in the guides need only be expanded in a representation consistent with the given source. In the sequel, the unknown coefficients of the modal expansions are determined through an integral equation approach.

If there are *LE* modes incident on the junction in Fig. 1a in both the straight and curved guides, the continuity in the transverse electric and magnetic fields at the junction between the guides requires

$$S_{rs}^{s} \, {}^{s}\mathbf{e}_{rs}^{e} + \sum_{m,n} A_{mn}^{-} \, {}^{s}\mathbf{e}_{mn}^{e} = \sum_{m,n} C_{mn}^{+} \, {}^{c}\mathbf{e}_{mn}^{e} + S_{rs}^{c} \, {}^{c}\mathbf{e}_{rs}^{e} \qquad (4)$$

and

$$S_{rs}^{s} \, {}^{s}\mathbf{h}_{rs}^{e} - \sum_{m,n} A_{mn}^{-} \, {}^{s}\mathbf{h}_{mn}^{e} = \sum_{m,n} C_{mn}^{+} \, {}^{c}\mathbf{h}_{mn}^{e} - S_{rs}^{c} \, {}^{c}\mathbf{h}_{rs}^{e} . \qquad (5)$$

Here the source coefficients have been designated, for emphasis, by S_{rs}^{s} and S_{rs}^{c} for the straight and curved sections, respectively; they are assumed specified. The unknowns are the modal expansion coefficients A_{mn}^{-} and C_{mn}^{+}.

Each side of equation (4) is actually an expansion of the unknown aperture electric field $\mathbf{E}_{a}(x, z)$. Referring to the orthogonal properties of

† Kronecker delta.

the transverse vector components in Section 2.1, it follows that

$$S^s_{rs} \; \delta_{mr} \; \delta_{ns} + A^-_{mn} = \iint_{S_A} \mathbf{E}_a \cdot {}^s\mathbf{e}^e_{mn} \, dA \tag{6}$$

and

$$C^+_{mn} + S^c_{rs} \; \delta_{mr} \; \delta_{ns} = \iint_{S_A} \mathbf{E}_a \cdot {}^c\mathbf{e}^e_{mn} \, \rho dA \tag{7}$$

for $m = 0, 1, \ldots$ and $n = 1, 2, \ldots$ with the integration being performed over the aperture area.[†] Rearranging equation (5) and substituting the relationships (6) and (7) for the expansion coefficients results in:

$$2S^s_{rs} \; {}^s\mathbf{h}^e_{rs} + 2S^c_{rs} \; {}^c\mathbf{h}^e_{rs} = \iint_{S_A} \mathbf{E}_a(x', z') \cdot \bar{\mathbf{G}}(x, z; x', z') \, dA', \tag{8}$$

where the dyadic kernel is given by

$$\bar{\mathbf{G}}(x, z; x', z') = \sum_{m,n} [{}^s\mathbf{e}^e_{mn}(x', z') \; {}^s\mathbf{h}^e_{mn}(x, z) + x' \; {}^c\mathbf{e}^e_{mn}(x', z') \; {}^c\mathbf{h}^e_{mn}(x, z)]. \tag{9}$$

Notice that equation (8) is precisely in the form of a vector Fredholm integral equation of the first kind for the unknown aperture electric field.

A solution of this integral equation by the method of moments would proceed as follows.[7] Expanding the aperture electric field in terms of the modes of the straight waveguide gives

$$\mathbf{E}_a(x, z) = \sum_{m,n} a_{mn} \; {}^s\mathbf{e}^e_{mn} . \tag{10}$$

Substituting into equation (8) and interchanging summation and integration then requires

$$2S^s_{rs} \; {}^s\mathbf{h}^e_{rs} + 2S^c_{rs} \; {}^c\mathbf{h}^e_{rs} = \sum_{m,n} a_{mn}({}^s\mathbf{h}^e_{mn} + \sum_p b_{mpn} \; {}^c\mathbf{h}^e_{pn}), \tag{11}$$

with b_{mpn} defined by

$$b_{mpn} = \int_{r_1}^{r_2} \varphi^s_m(x)\varphi^c_{pn}(h_n x) \, dx. \tag{12}$$

Taking the inner product of equation (11) with ${}^s\mathbf{h}^e_{iq}$ finally leads, after some algebra, to

† The indices m and n of the modes are chosen such that in the limit of $r_1 \rightarrow \infty$ the mode in the curved guide with index numbers m and n is asymptotic to the mode in the straight guide with index numbers m and n.

$$\frac{2S_{rs}^{s}}{\beta_{rs}} \delta_{jr} \delta_{qs} + \frac{2S_{rs}^{c}}{\nu_{rs}} b_{jrs} \delta_{qs} = \sum_{m} a_{mq} \left(\frac{\delta_{jm}}{\beta_{mq}} + \sum_{p} \frac{b_{mpq}b_{jpq}}{\nu_{pq}} \right),$$

$$j = 0, 1, 2, \cdots , \quad q = 1, 2, \cdots \quad (13)$$

as the infinite set of algebraic equations to be solved for a_{mq}, the expansion coefficients of the aperture electric field.

As a first observation, we note from equation (13) that $a_{mq} \equiv 0$ if $q \neq s$, and hence the aperture field is actually given by

$$\mathbf{E}_a(x, z) = \sum_{m} a_{ms} \, {}^s\mathbf{e}_{ms}^{e}(x, z), \quad (14)$$

that is, the z-variation of the excited modes is the same as the z-variation of the source mode. At this point, therefore, the second subscript may be dropped without loss of generality by merely realizing it is the same as the second subscript of the exciting modes.

Equations (13) form an infinite set of equations for the infinite number of unknown expansion coefficients of the aperture field. A truncation is now made in order to solve for a_m by standard matrix methods, including sufficient terms in the field expansions in order to ensure reasonable accuracy (see Section 4.1).[†]

2.3 *Reflected and Transmitted Modes*

Let us assume here that the expansion coefficients for the aperture field have been obtained by solving equation (13). A relationship between the coefficients of the modes in the straight guide and the aperture field was given by equation (6). Substituting the expansion of the aperture field, equation (14), into this equation gives an expression for the coefficients of the modes in the straight guide as

$$A_m^- = a_m - \delta_{mr}S_r^s , \qquad m = 0, 1, 2, \cdots . \quad (15)$$

Likewise equation (7) yields, for the coefficients of the modes excited in the curved waveguide,

$$C_m^+ = \sum_{q} a_q b_{qm} - \delta_{mr}S_r^c , \qquad m = 0, 1, 2, \cdots . \quad (16)$$

These relations are deceptively simple in that much of the complex interplay between incident, reflected, and transmitted modes is hidden in the "assumed known" coefficients a_m and a_q.

The average power carried by the incident rth mode in the straight

---

† The solution of equation (13) also requires knowledge of the b_{mp} defined by equation (12). Their determination is at the crux of this method and their evaluation will be discussed in Section 3.

and curved guide may be determined:

$$PI_r^s = -\iint {}^sE_x^e \, {}^sH_z^{e*} \, ds = \frac{h^2}{j\beta_r Z^s *} \mid S_r^s \mid^2 \tag{17}$$

and

$$PI_r^c = \iint {}^cE_x^e \, {}^cH_z^{e*} ds = \frac{h^2}{j\nu_r Z^c*} \mid S_r^c \mid^{2 \ddagger}. \tag{18}$$

Here we assume that the incident rth mode is a propagating mode with real β_r and ν_r, and that ($*$) designates the complex conjugate of a quantity.

The average power coupled into the mth mode from the incident rth mode may be evaluated in a similar manner yielding for the straight guide

$$PC_{mr}^s = \frac{h^2}{j\beta_m Z^s *} \mid A_m^- \mid^2 \tag{19}$$

and for the curved guide

$$PC_{mr}^c = \frac{h^2}{j\nu_m Z^c *} \mid C_m^+ \mid^2. \tag{20}$$

The index m in equations (19) and (20) is anyone such that β_m or ν_m is real; that is, the mth mode must be a propagating one which carries energy away from the junction. There are, of course, only a finite number of such propagating modes for a particular operating frequency (see Cochran and Pecina[1]).

Equations (19) and (20) thus determine the power coupling, that is, the power excited in the mth propagating mode either transmitted or reflected when the rth mode is incident in either the straight or curved sections. Naturally, these quantities become of dominant importance as one moves away from the junction and the evanescent modal contributions die out. Section 4.2 gives some examples of the power coupling for both sharp and gradual bends.

A similar analysis can be performed for an LM excitation. Section IV presents the numerical results for this case.

III. PROPAGATION CONSTANTS AND MODAL FUNCTIONS

The modal functions for the continuously curved waveguide are de-

---

‡ At this stage we assume that the waveguides are filled with a lossless dielectric; hence, the total power is the sum of the power in each individual mode.

fined in terms of Bessel functions of the first and second kind in Section 2.1. Obviously they are solutions of Bessel's differential equation, which we write in the form

$$\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{d}{d\rho}C_{\nu mn}\right) + \left(h_n^2 - \frac{\nu_{mn}^2}{\rho^2}\right)C_{\nu mn} = 0; \tag{21}$$

moreover, for LE excitation, they are such that

$$\frac{d}{d\rho}C_{\nu mn}(h_n\rho)\mid_{\rho=r_1}^{r_2} = 0. \tag{22}$$

The boundary condition at $\rho = r_2$ is automatically satisfied by the particular choice of the cross-product Bessel functions in Section 2.1, whereas the boundary condition at $\rho = r_1$ determines the admissible angular propagation constants ν_{mn}.

The real angular propagation constants result in propagating modes and hence are the most important in this analysis. These are obtained for the sharp bend by a program of precise calculations of the real ν-solutions of the transcendental equation (22). The Bessel functions appearing in equation (22) were approximated with six-figure accuracy by the use of algorithms recently developed and programmed for a digital computer as discussed in Ref. 1.

There are other methods to determine the propagation constants of gradual bends. For instance, a large parameter expansion of the differential equation (21) can be made; that is, the modal functions and propagation constants may both be expanded in negative powers of r_2. The unknown coefficients of each series can then be determined by imposing the boundary conditions at $\rho = r_1$ and r_2. This approach has been used by Kislyuk, as well as others; Ref. 8 gives these results. Four terms in the expansion are all that are available, because higher order terms are extremely tedious to determine.

A comparison of the real values of ν evaluated from Kislyuk's results with the precise ν-zeros of equation (22) shows five digit agreement for gradual bends $(r_1/b > 12)$. In the final program, we chose to calculate all angular propagation constants by Kislyuk's equation for large (r_1/b), that is, 12 or greater.

In the sharp bend case the imaginary propagation constants cannot be obtained precisely, because there are no computer algorithms for the evaluation of Bessel functions for this range (imaginary orders). So other techniques must be used.

When the propagation constants lie on the negative imaginary axis,

that is, $\nu = -i\mu$ and μ is not close to zero, asymptotic expansions for the modal functions can be obtained. One approach is to approximate the Bessel functions in C_ν (the derivative with respect to $h_n r_2$ is not performed as yet) by the first term of the asymptotic series developed by Olver.[9] This yields an expression in terms of the familiar Airy functions Ai and Bi. When the Airy functions are approximated by the leading terms of their phase-amplitude expansions (see Abromowitz and Stegun[10]) and the derivative with respect to $h_n r_2$ is taken, the approximation of C_ν becomes

$$C_\nu \approx \frac{-2}{\pi\mu[(1 + \zeta^2\eta^2)(1 + \eta^2)]^{\frac{1}{4}}} \left\{ (1 + \eta^2)^{\frac{1}{4}}/\eta \cos \left[\mu(\omega(\eta) - \omega(\zeta\eta)) \right] \right.$$

$$\left. + \frac{\eta}{2\mu(1 + \eta^2)} \sin \left[\mu(\omega(\eta) - \omega(\zeta\eta)) \right] \right\}, \tag{23}$$

where

$$\eta = \frac{hr_2}{\mu}, \qquad \zeta = \frac{r}{r_2},$$

and

$$\omega(\eta) = \ln \left[\frac{1 + (1 + \eta^2)^{\frac{1}{2}}}{\eta} \right] - (1 + \eta^2)^{\frac{1}{2}}.$$

The imaginary propagation constants for the sharp bend are now determined by numerically finding the μ-zeros of the asymptotic expressions, equation (23).

The evaluations of the inner products, equation (12), required in Section 2.2 are performed numerically. When the propagation constants are real, we again use the computer algorithms for the evaluation of the Bessel functions for both the sharp and gradual bends. The latter evaluation was required because the evaluation of the modal functions by means of a large parameter series expansion as determined by Kislyuk's approach (really only three terms available) does not exhibit very good agreement with the precise evaluation of the modal function even in a region where the agreement between the two methods of determining the propagation constant is very good. When the propagation constants are imaginary the modal functions for the curved waveguide are evaluated by means of the approximate expression, equation (3), for both the sharp and gradual bends.

A similar analysis may be made for an LM polarization; Section IV presents the results of appropriate numerical calculations, as well as,

comments concerning the verification of the numerical solution and some representative examples.

IV. ERROR CRITERIA AND REPRESENTATIVE EXAMPLES

4.1 *Error Criteria*

As discussed in Section 2.2, the solution of the integral equation for the aperture field reduces to an infinite set of algebraic equations for its expansion coefficients. We make a truncation so that standard matrix techniques may be used to solve for these unknown coefficients. Sufficient terms must be included to obtain reasonable accuracy; including more terms than necessary wastes computing time.

One can verify that a particular truncation is adequate by determining how well the field solutions satisfy the continuity requirement at the aperture. The conservation of energy, which requires that the average power in all the propagating modes traveling away from the junction between the guides be equal to the average power in the propagating modes incident on the junction, is always satisfied (within roundoff error) by the solution obtained (that is, regardless of the number of modes used); therefore it cannot be used as an accuracy check. This redundancy in the conservation of power, which may be established by an analysis suggested by Amitay and Galindo,[11] is a consequence of the method of moments approach which has been used to solve the integral equation.

When the incident field is an *LE* mode, the aperture electric field is determined. From this field one can derive the modal coefficients and hence the magnetic fields in the straight and curved guides. These derived magnetic fields should be continuous at the aperture; therefore, a mean square error (MSE—refer to its application by Cole and others[12]), normalized with respect to the incident field, can be defined as

$$\text{MSE} = \frac{\iint (^s\mathbf{H}^e - {}^c\mathbf{H}^e) \cdot (^s\mathbf{H}^e - {}^c\mathbf{H}^e)^* \, ds}{\iint (^s\mathbf{H}^e_i + {}^c\mathbf{H}^e_i) \cdot (^s\mathbf{H}^e_i + {}^c\mathbf{H}^e_i)^* \, ds}.$$

The subscript i designates the incident exciting field; the terms in the numerator constitute the total fields, all evaluated at the junction between the guides. This mean square error is a meaningful measure of how well continuity in the aperture field is approached, and is, of course, a function of the number of modes used in expanding the fields.

It was found, in the examples of Section 4.2, that the mean square error could be maintained smaller than 10^{-5}. This corresponds to three to four digit agreement between the samples of the transverse components of the magnetic fields on both sides of the aperture. A similar mean square error may be defined for the LM case with corresponding error levels.

4.2 *Representative Examples*

Some of the following representative examples correspond to very sharp bends ($r_1/b \approx 1$); the others correspond to very gradual bends ($r_1/b \gg 1$).

4.2.1 *Sharp Bends*

Figures 2 through 5 and Table I give the results for the sharp bends. We give the power transmitted into the modes of the curved guide and reflected into the modes of the straight guide for an incident mode in the straight guide in terms of the dimensionless ratios b/λ, r_1/b, and a/b. Any structure with these ratio numbers has a coupling characteristic as displayed.

The incident modes used in Figs. 2 and 3 are from the set of $LM_{mo}^{\bullet}$ modes in the straight guide which have an electric field given by

$$^{\bullet}\mathbf{E}_{m0}^m = \frac{-B_{mo}^+ h_0^2}{j\beta_{mo} Y} (2/b)^{\frac{1}{2}} \sin [m\pi/b(r_2 - x)] \exp (-j\beta_{mo}y)\hat{z}.$$

That is, the incident fields are the familiar TE_{mo} modes of a uniform rectangular guide. The curved guides used here are referred to as H-plane bends since the magnetic field lies in the plane of the bend (see Section 2.1). In Figs. 4 and 5 the incident modes are from the $LE_{m1}^{\bullet}$ mode set in the straight guide. The curved guides there are referred to as E-plane bends since the electric field lies in the plane of the bend. The $LE_{01}^{\bullet}$ mode incident corresponds to the familiar TE_{01} mode with an electric field given by

$$^{e}\mathbf{E}_{01}^{\bullet} = A_{01}^+(2/ab)^{\frac{1}{2}} \sin (\pi z/a) \exp (-j\beta_{01}y)\hat{x}$$

whereas the $LE_{11}^{\bullet}$ mode incident is a combination of the TE_{11} and TM_{11} modes in a uniform rectangular guide. The coefficients B_{mo}^+ and A_{m1}^+ of the incident modes are chosen so that the incident power is unity.

The sharp bend results may be used, for example, to depict the operation of the standard C-band guide 0.872 by 1.872 inch for a frequency range from 3 to 18 GHz for the LM excitation and from 3 to 20 GHz for the LE excitation. For convenience we have superimposed

Fig. 2 — Intermodal coupling at H-plane bend — sharp bend case. (Incident mode = LM_{10}^s; incident power = 1.0; r_1/b = 1.068; a/b = 0.466.) Solid line is for a curved guide; dashed line is for a straight guide.
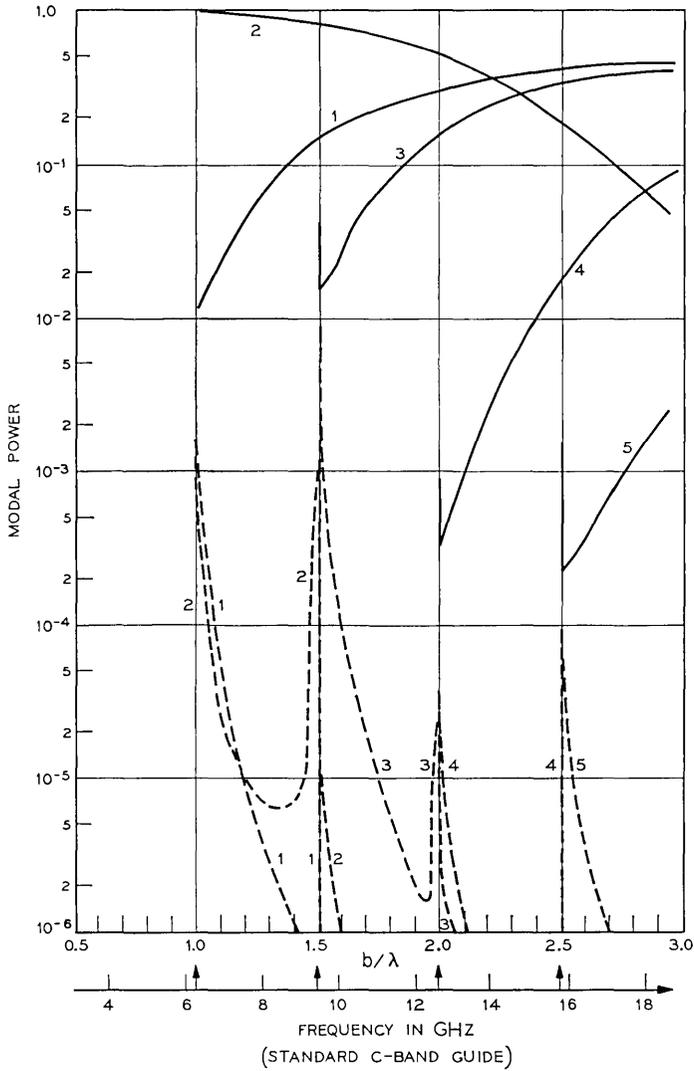
Fig. 3 — Intermodal coupling at H-plane bend — sharp bend case. (Incident mode = LM_{20}^s; incident power = 1.0; r_1/b = 1.068; a/b = 0.466.) Solid line is for a curved guide; dashed line is for a straight guide.
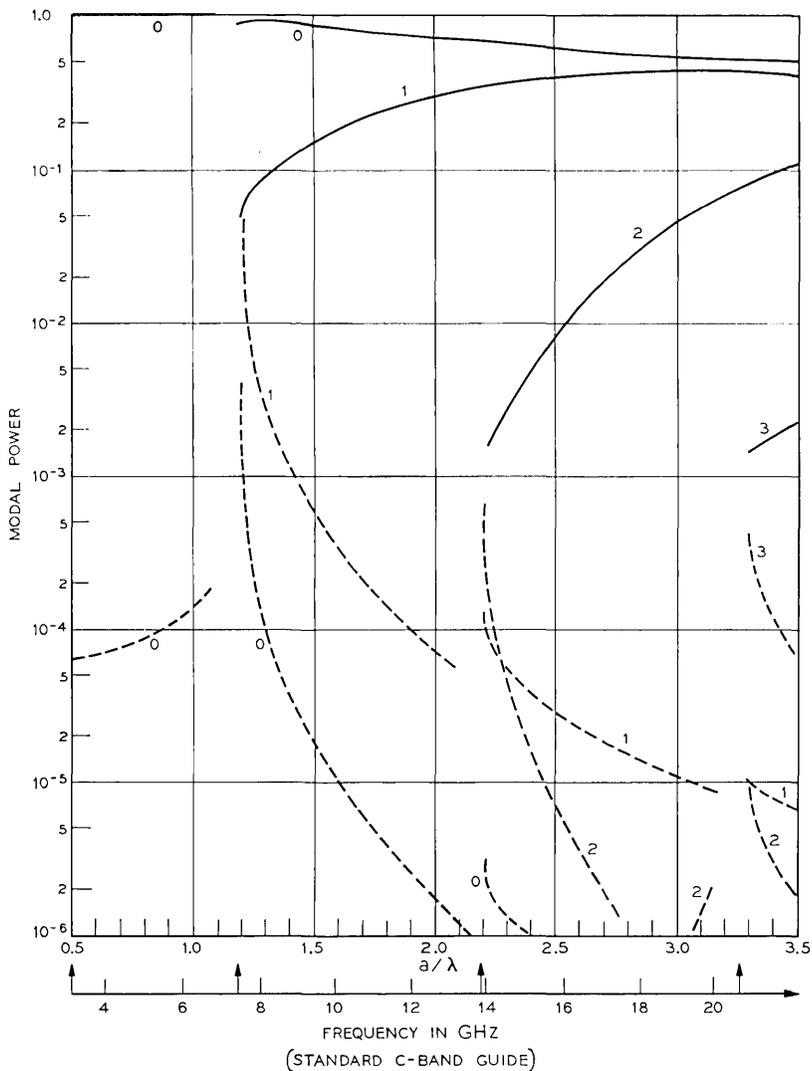
Fig. 4 — Intermodal coupling at E-plane bend — sharp bend case. (Incident mode = LE_{01}^s; incident power = 1.0; $r_1/b = 1.148$; $a/b = 2.15$.) Solid line is for a curved guide; dashed line is for a straight guide.
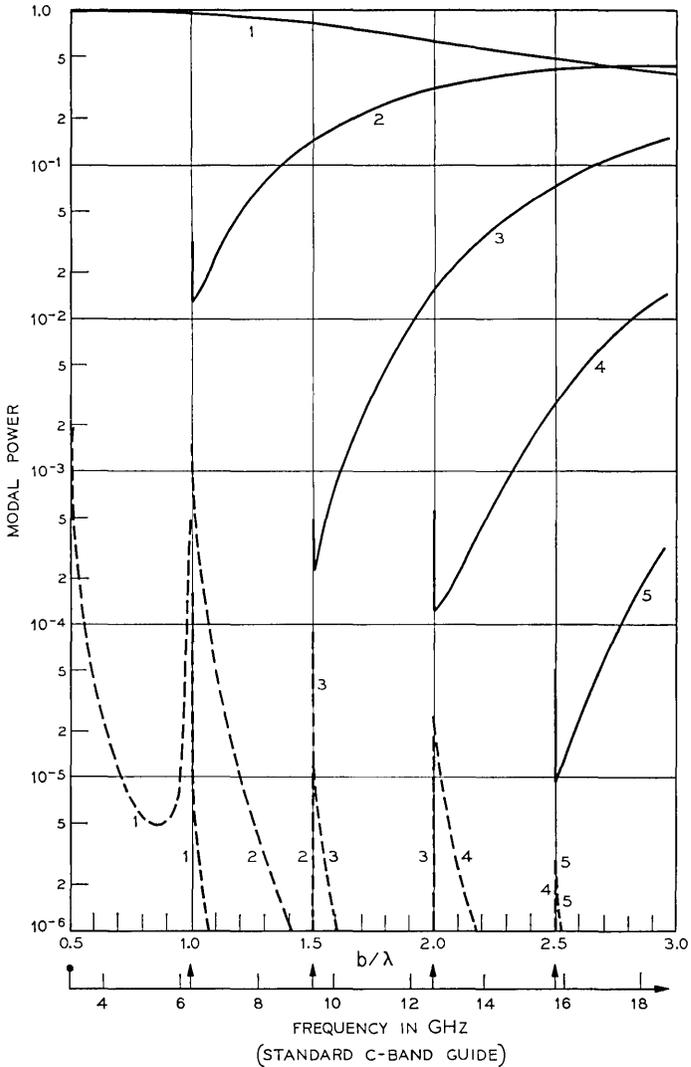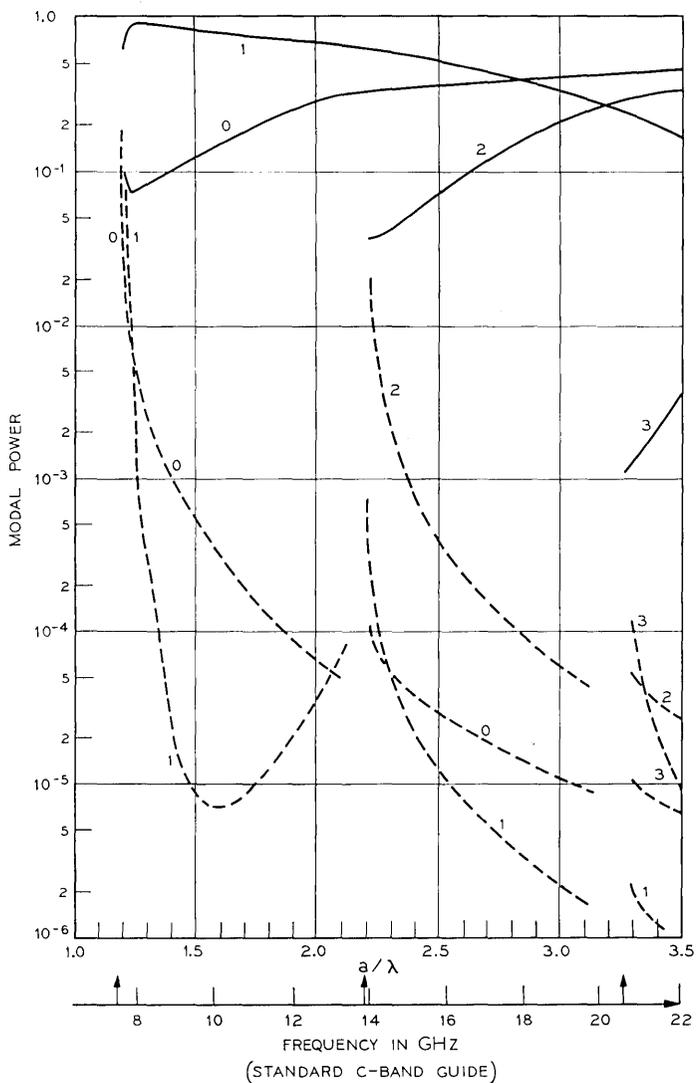
Fig. 5 — Intermodal coupling at E-plane bend — sharp bend case. (Incident mode = LE^s_{11}; incident power = 1.0; $r_1/b = 1.148$; $a/b = 2.15$.) Solid line is for a curved guide; dashed line is for a straight guide.

TABLE I—INTERMODAL COUPLING FOR LONGITUDINAL ELECTRIC
AND LONGITUDINAL MAGNETIC FIELDS FOR A SQUARE
GUIDE WITH $r_1/b = 1.068$

b/λ	Incident Mode Power	Reflected Mode Power	Transmitted Mode Power
1.19	$LM^s_{10} = 1.0$	$LM^s_{10} \approx 10^{-7}$	$LM^c_{10} = 0.952174$
		$LM^s_{20} = 0.000012$	$LM^c_{20} = 0.047815$
	$LE^s_{01} = 1.0$	$LE^s_{01} = 0.000001$	$LE^c_{01} = 0.618860$
		$LE^s_{11} = 0.000046$	$LE^c_{11} = 0.376074$
		$LE^s_{21} = 0.000022$	$LE^c_{21} = 0.004997$
1.79	$LM^s_{10} = 1.0$	$LM^s_{10} = 10^{-9}$	$LM^c_{10} = 0.736720$
		$LM^s_{20} = 0.000001$	$LM^c_{20} = 0.258699$
		$LM^s_{30} \approx 10^{-7}$	$LM^c_{30} = 0.004581$
	$LE^s_{01} = 1.0$	$LE^s_{01} \approx 10^{-7}$	$LE^c_{01} = 0.456751$
		$LE^s_{11} = 0.000005$	$LE^c_{11} = 0.384346$
		$LE^s_{21} = 0.000001$	$LE^c_{21} = 0.153564$
		$LE^s_{31} = 0.000027$	$LE^c_{31} = 0.005306$

another coordinate scale on Figs. 2 through 5 demonstrating the frequency of operation if the guide has these dimensions. The vertical arrows on this frequency scale indicate the cutoff frequencies of the modes in the straight guide. As these examples show, the frequency band covered corresponds to a situation where up to 5 modes can propagate. The overmoded operation demonstrates the possible coupling between modes. Also notice that for the H-plane bend (Figs. 2 and 3), b is greater than a and for the E-plane bend (Figs. 4 and 5) b is less than a.

The strong coupling between the modes for sharp bends is clearly demonstrated for H-plane bends in Figs. 2 and 3. In Fig. 2 we see that the LM^s_{10} mode incident in the straight guide can actually couple more energy into the LM^c_{20} than into the LM^c_{10} mode of the curved guide when b/λ is greater than 2.7. Conversely the LM^s_{20} mode incident in the straight guide can couple more energy into the LM^c_{10}, LM^c_{30}, and LM^c_{40} modes than into the LM^c_{20} mode of the curved guide for the appropriate b/λ, as Fig. 3 shows.

For an LM excitation it is possible to have a mode propagating in the curved waveguide while still cut off in the straight waveguide. This leads to a value of coupling into the curved guide mode which drops sharply as the corresponding mode begins to propagate in the straight guide but then increases with increasing frequency. Figures 2 and 3 show this at $b/\lambda = 1.0, 1.5, 2.0,$ and 2.5. The reflections are also exaggerated at the cutoff frequencies of the modes in the straight guide. In

contrast, the cutoff frequencies of the *LE* modes in the curved guide
are greater than or equal to those in the straight guide. This leads to
the possibility of having a mode propagating in the straight guide
while still cut off in the curved guide. Sharp jumps in reflections and
transmissions are thus also expected at such cutoff frequencies for this
situation; unfortunately, we are not able to examine them in detail.
Recall that the procedure outlined in Section III allows us only to find
the imaginary propagation constants in the curved waveguide if they
are not too small. Unfortunately the case just described violates this
restriction since the pertinent propagation constants in the curved
waveguide are imaginary with a magnitude infinitesimally close to
zero.

In Figs. 4 and 5 one can see that the power coupling at an E-plane
bend is also very strong. The reflections for this case are more pro-
nounced over a wider frequency band than in the H-plane case. Again
there are exaggerated reflections at the cutoff frequencies of the modes.
Notice that with both *LE* and *LM* polarizations, the forward coupling
is greatest into the modes adjacent to the one corresponding to the
incident mode.

It is valuable to compare the coupling for an E-plane bend with that
of a H-plane bend for equivalent problems. To this end, consider a
square guide with the ratio r_1/b set at 1.068 for each polarization and
the frequency of operation set at the same value for both cases. Table
I gives the coupling for this situation. (Notice that the number of *LE*
modes propagating is one more than the number of *LM* modes prop-
agating at the frequencies used.) From the results one sees that much
less energy is forward coupled into the mode corresponding to the in-
cident one when the fields are *LE* and that the total reflected energy is
greater in the *LE* case. This is not unexpected if one examines the dis-
continuity in the geometry encountered by the electric field intensity
for both cases.

4.2.2 *Gradual Bends*

For a very gradual bend situation we chose square waveguides with
$r_1/b = 250$. This very gradual bend simulates the curvatures encoun-
tered in the waveguide connection between receiver and antenna of the
Bell System TD-2 microwave relay system (one must realize though
that waveguides with circular cross sections are used there). Tables II
and III give some pertinent results. Again a frequency scale is super-
imposed, this time corresponding to a 2.4 inch guide. This value was
picked so that the fundamental mode in the straight guide with a square

cross section would have the same propagation constant as the fundamental mode of a straight guide with a circular cross section, 2.812 inches in diameter. This we felt, permits us to stimulate, at least qualitatively, the situation encountered with circular cross section guides in the TD-2 system.

The coupling in the reverse direction (reflected power) was at least 60 dB down for both *LE* and *LM* fields regardless of which mode was incident; hence these tables do not give them. The H-plane bend (Table II) forward couples power ($\approx$40 dB down) into modes adjacent to the one corresponding to the incident mode only at the higher frequencies. The E-plane bend (Table III) exhibits much larger forward coupled power into such modes ($\approx$30 dB down) at these same higher frequencies. The levels of the undesired forward coupling at the lower frequencies is much less ($\approx$50 dB down). The results suggest that with such gradual bends reverse coupling is totally insignificant and only forward coupling can have a meaningful effect.

All the results discussed in Section 4.2 have been based on the excitation from the straight guide side of the junction. The results for excitation from the curved guide side are of the same form, and hence, have not been given for the sake of brevity. However, forward coupling into the straight guide from the curved guide may be deduced from the data already presented by realizing that the power forward coupled into mode m of the straight guide from mode n in the curved

TABLE II—INTERMODAL COUPLING RESULTING FROM A LONGITUDINAL MAGNETIC MODE INCIDENT IN STRAIGHT GUIDE WITH $r_1/b = 250$

b/λ	f ($b = 2.4$ inches) GHz	Incident Mode	Excited Mode	Excited Mode Power Level (dB)
0.813	4	LM_{10}^s	LM_{10}^c	0.00
1.22	6	LM_{10}^s	LM_{10}^c	0.00
			LM_{20}^c	-55.84
		LM_{20}^s	LM_{10}^c	-55.84
			LM_{20}^c	0.00
2.24	11	LM_{10}^s	LM_{10}^c	0.00
			LM_{20}^c	-41.54
			LM_{30}^c	< -60
			LM_{40}^c	< -60
		LM_{20}^s	LM_{10}^c	-41.54
			LM_{20}^c	0.00
			LM_{30}^c	-47.59
			LM_{40}^c	< -60

TABLE III—INTERMODAL COUPLING RESULTING FROM A
LONGITUDINAL ELECTRIC MODE INCIDENT IN STRAIGHT
GUIDE WITH $r_1/b = 250$

b/λ	f ($b = 2.4$ inches) GHz	Incident Mode	Excited Mode	Excited Mode Power Level (dB)
0.813	4	LE_{01}^s	LE_{01}^c	0.00
			LE_{11}^c	−50.06
		LE_{11}^s	LE_{01}^c	−50.06
			LE_{11}^c	0.00
1.22	6	LE_{01}^s	LE_{01}^c	0.00
			LE_{11}^c	−39.39
			LE_{21}^c	< -60
		LE_{11}^s	LE_{01}^c	−39.39
			LE_{11}^c	0.00
			LE_{21}^c	−55.46
2.24	11	LE_{01}^s	LE_{01}^c	−0.01
			LE_{11}^c	−27.38
			LE_{21}^c	< -60
			LE_{31}^c	< -60
			LE_{41}^c	< -60
		LE_{11}^s	LE_{01}^c	−27.38
			LE_{11}^c	−0.01
			LE_{21}^c	−39.88
			LE_{31}^c	< -60
			LE_{41}^c	< -60

guide is the same as the power forward coupled into mode n of the curved guide from mode m in the straight guide (reciprocity).[†]

V. CONCLUSION

This paper has investigated the coupling of electromagnetic waves between straight and curved rectangular waveguides. Numerical results have been obtained by using a numerical method which leads to solutions applicable for sharp as well as gradual bends. Two representative examples have been given. One was a sharp bend and could be used to depict the coupling that takes place, say, in standard C-band guides. The other was a very gradual bend; this was used to obtain some insight into the coupling that occurs in the waveguide connections between the receiver and antenna in typical microwave networks.

The coupling discussed has been confined to a one junction struc-

---

[†] This modal reciprocity, although surprising at first glance, is a direct consequence of Maxwell's equations, the lossless character of the guides, and the orthogonalities between the modes as discussed in Section 2.1.

ture, that is, a straight to a curved guide or a curved to a straight guide. In any practical system, however, at least two junctions generally occur, that is, one encounters straight-curved-straight or curved-straight-curved connections. For very gradual bends it is merely necessary to account for the forward coupling at each junction since any reflections are negligible. Sharp bends, on the other hand, require one to account for multiple reflections; this appears to be most effectively handled by the scattering matrix approach.

VI. ACKNOWLEDGMENT

REFERENCES

 1. Cochran, J. Alan, and Pecina, Robert G., "Mode Propagation in Continuously Curved Waveguides," Radio Sci., *1* (new series), No. 6 (June 1966), pp. 679–695.
 2. Rice, S. O., "Reflections from Circular Bends in Rectangular Wave Guides—Matrix Theory," B.S.T.J., *27*, No. 2 (April 1948), pp. 305–349.
 3. Jouguet, M., "Les effets de la courbure et des discontinuitiés de courbure sur la propagation des ondes dans les guides à section rectangulaire," Câbles et Transmission, *1*, No. 1 (April 1947), pp. 39–60.
 4. Cochran, J. Alan, Nagel, C. M., and Alsberg, P. A., unpublished work.
 5. Harrington, R. F., *Time Harmonic Electromagnetic Fields*, New York: McGraw-Hill, 1961.
 6. Cochran, J. Alan, "Remarks on the Zeros of Cross-Product Bessel Functions," J. Soc. Ind. Appl. Math., *12*, No. 3 (September 1964), pp. 580–587.
 7. Harrington, R. F., "Matrix Methods for Field Problems," Proc. IEEE, *55*, No. 2 (February 1967), pp. 136–149.
 8. Kislyuk, M. Zh., "Waveguide Bend of Rectangular Cross-Section," Radio Eng., *16*, No. 4 (April 1961), pp. 1–8.
 9. Olver, F. W. J., "The Asymptotic Expansion of Bessel Functions of Large Order," Trans. Royal Soc. London, *247*, Series A, (1954), pp. 328–368.
10. Abromowitz, M., and Stegun, I., *Handbook of Mathematical Functions*, New York: Dover Publications, 1965.
11. Amitay, N., and Galindo, V., "On Energy Conservation and the Method of Moments in Waveguide Discontinuity and Scattering Problems," 1969 United States Nat. Committee of the Int. Sci. Radio Union Spring Meeting, April 21-24, 1969, Washington, D. C.
12. Cole, W. J., Nagelberg, E. R., and Nagel, C. M., "Iterative Solution of Waveguide Discontinuity Problems," B.S.T.J., *XLVI*, No. 3 (March 1966), pp. 649–672.

A Quadratic Conduction Gas Lens

By C. A. FRITSCH and D. J. PRAGER

(Manuscript received January 17, 1969)

A lens system for a periodic light-beam waveguide is proposed and analyzed in which gas is enclosed in a circular cylinder heated with a cos 2ϕ temperature distribution. We show that this temperature distribution may be produced by cutting a cylindrical hole in the center of a square block which has two opposite sides of equal temperature above the ambient temperature, and two sides of a lower temperature. Heat conduction across the gas produces an index of refraction variation which, in two orthogonal azimuthal planes, increases or decreases as the radius squared. The effect of thermal convection is analyzed by solving the governing equations as an expansion in powers of the Rayleigh number; the solution reveals that convection effects can be made negligible over a practical range of lens parameters. The major attributes of the lens system are that only temperature controls are required and the aberrations associated with thermal convection can be readily minimized.

I. INTRODUCTION

A gas lens system to transmit a light beam through a tube should have a favorable refractive index, negligible aberrations, and a simple construction. The favorable refractive index must be such that all light rays parallel to the tube axis, but of varying distances from that axis, converge at approximately the same point on the axis, the distance being called the focal length. Within the paraxial ray approximation it is easy to show that an r^2 variation of the refractive index has this property (see, for example, Refs. 1 and 2).

Berreman obtained a refractive index (which varied approximately as the square of the radius) by flowing a gas through a cold cylinder enclosing a warm helix aligned on the axis.[3] The interior of the helix has the desired refractive index. Marcuse and Miller simplified Berreman's lens by considering a cool gas flowing through a heated cylinder of uniform temperature (the Graetz problem).[1,2]

In order to reduce the distortion resulting from spherical aberrations, Berreman built a counterflow arrangement composed of two back to back tubular lenses.[4] Marcuse calculated the principal surfaces of a flow type lens noting that the one with the light beam parallel to the flow differs only slightly from that of the beam antiparallel to the flow.[2] He then numerically calculated the fate of a beam as it passes through a large number of flow lenses and compared the results with those with a counterflow arrangement.[5] This arrangement decreased the distortion. Kaiser later found that this configuration also lessens the asymmetric distortion due to thermal convection.[6]

The major drawback to the flow-type lens is the need for control of the flow. Gu performed a compressible flow analysis and found that, as a result of the wall friction, choking could occur for the optimal flow rate in a few hundred meters.[7] This could be overcome only by further complexities in the system.

A conduction-type lens was proposed by Suematsu, Iga, and Ito, in which they analyzed a configuration composed of hyperbolic, convex inward walls, two of which are at one temperature and the other opposing two at a lower temperature.[8] The concomitant temperature distribution varies as the square of the distance in the transverse direction. Then the refractive index bears the r^2 variation* in two orthogonal planes, being convergent in one and divergent in the other. This quadratic variation has two highly desirable characteristics. First, within the paraxial approximation, the focal length of every ray passing through a quadratic lens is independent of the radius, and hence the field reproduces itself after each period.[1] Second, Marcatili has shown that the eigenfunctions associated with a quadratic lens are Gaussian. Therefore a laser beam which is also Gaussian can be mode-matched to a waveguide consisting of quadratic lenses. This means that all the energy will remain in the launched mode; the only mode conversion that would take place is that resulting from higher order variation, that is, aberrations.

The advantage of the conduction lens is that only temperature controls are required since no gas flow is involved. However, thermal convection is present in this lens and although Suematsu, and others, observed a degradation of their lens at high temperature differences they did not analyze the thermal convection effects.

---

* For negligible pressure changes the refractive index is virtually only a function of the temperature; for small temperature variations the changes in refractive index are directly proportional to and of opposite sign from the temperature changes.

We will consider a quadratic conduction-type lens, which is formed by imposing a cos 2ϕ temperature distribution on the wall of a circular cylinder. For sufficiently small temperature variations the change in the refractive index is approximately quadratic and lensing action similar to that of Suematsu, and others, is obtained. The three central questions considered are: (i) What are the effects of thermal convection on the quadratic distribution? (ii) How does one readily obtain a cos 2ϕ wall temperature distribution? (iii) What are the optical properties of a waveguide consisting of these lenses?

We show that the cos 2ϕ distribution can be achieved very simply by boring a circular hole in a square block in which two opposite sides bear a higher temperature than ambient and the other two bear a lower temperature. If sections of the above lens are placed in tandem, each consecutive one rotated by 90 degrees, there then exists in one plane a series of alternating, convergent-divergent lenses. In the perpendicular plane this series is, so to speak, 180 degrees out of phase. We may then use Miller's[9] analysis of a sequence of alternating gradient lenses,* and determine criteria for the optical properties as a function of the parameters of the system.

We study the effect of thermal convection by using a straightforward perturbation analysis which is found to be in agreement with preliminary results of an experiment. We investigate the method of producing the wall temperature distribution by constructing an approximate solution which reveals how the wall temperature distribution can be established, as well as discuss the experimental program in progress and compare this lens and the other cited above.

II. ANALYSIS

2.1 *Analysis of Thermal Convection*

Consider a circular cylinder with the geometry given in Fig. 1. The governing equations for the steady motion of the gas within the cylinder are:

(i) continuity equation,

$$\nabla \cdot (\rho \mathbf{u}) = 0; \tag{1}$$

(ii) equation of motion,

$$\rho(\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \rho \mathbf{g} - \mu \nabla^2 \mathbf{u} = 0; \tag{2}$$

---

* Alternating gradient focusing in gas lens systems was first proposed by A. R. Hutson.[3]
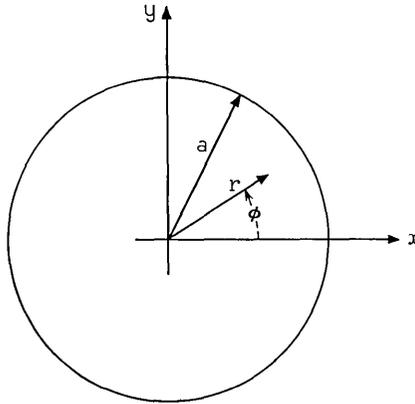
Fig. 1 — Geometry of the lens cylinder.

(*iii*) energy equation,

$$\rho \mathbf{u} \cdot \nabla (c_p T) \;=\; k \nabla^2 T + \mathbf{u} \cdot \nabla p + \mu \Phi. \tag{3}$$

Here,

ρ is the density,

$\mathbf{u}$ the velocity,

p the pressure,

$\mathbf{g}$ the gravitational acceleration,

μ the viscosity,

c_p the specific heat at constant pressure,

k the thermal conductivity, and

Φ the dissipation function (associated with the frictional work).

The boundary conditions at the cylinder surface are

$$T(a, \phi) \;=\; T_o\!\left(1 + \frac{\Delta T}{T_o} \cos 2\phi\right) \tag{4}$$

and

$$\mathbf{u}(a, \phi) \;=\; 0, \tag{5}$$

where ΔT is the maximum excursion about the average wall temperature, T_o.

At this point we use the Boussinesq approximation which consists of two elements; the density changes are significant only in the body force term, and these changes are a function of temperature only. The latter element amounts to neglecting the product of the isothermal

compressibility, κ, times the pressure change in comparison with the product of the volumetric expansivity, β, times the temperature change. In other words, for $\rho = \rho(p, T)$

$$\frac{d\rho}{\rho} = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial p}\right)_T dp + \frac{1}{\rho}\left(\frac{\partial\rho}{\partial T}\right)_p dT$$

$$= \kappa \, dp - \beta \, dT,$$

and the Boussinesq approximation requires the second term to be much larger than the first term but still small enough so that

$$\rho = \rho_0[1 - \beta(T - T_o)], \tag{6}$$

where the subscript denotes conditions at the center of the cylinder in the absence of fluid motion.

We nondimensionalize the variables in the hope that a perturbation scheme for a solution to our problem may be suggested. We define

$$\mathbf{U} = \frac{\mathbf{u}}{k/(\rho_o c_p a)} \,, \qquad \mathbf{x} = \mathbf{X}/a, \qquad \theta = \frac{T - T_o}{\Delta T}. \tag{7}$$

Since the density changes are considered important only in the body force term, equation (1) yields the incompressible continuity equation,

$$\nabla \cdot \mathbf{U} = 0. \tag{8}$$

The pressure term can be eliminated from the equation of motion by taking the curl of equation (2). The result of this operation leads us to define the velocity components in terms of the stream function ψ so that in cylindrical coordinates we have

$$U_r = \frac{1}{r}\frac{\partial\psi}{\partial\phi} \,, \qquad U_\phi = -\frac{\partial\psi}{\partial r}. \tag{9}$$

The continuity equation, (8), is identically satisfied, and the equation of motion becomes

$$\sigma\nabla^4\psi + \frac{1}{r}\left(\frac{\partial\psi}{\partial r}\frac{\partial}{\partial\phi} - \frac{\partial\psi}{\partial\phi}\frac{\partial}{\partial r}\right)\nabla^2\psi = \sigma\lambda\left(\cos\psi\frac{\partial}{\partial r} - \frac{\sin\phi}{r}\frac{\partial}{\partial\psi}\right)\theta, \tag{10}$$

where

$$\sigma = \frac{\mu c_p}{k} \,, \quad \text{the Prandtl number, and}$$

$$\lambda = \frac{\beta\,\Delta T\,g c_p \rho_0^2 a^3}{\mu k} \,, \quad \text{the Rayleigh number.}$$
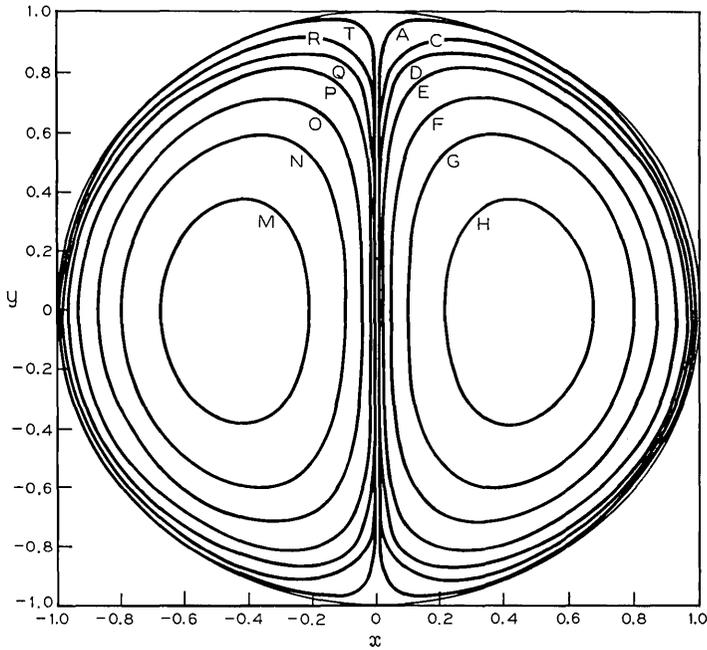
Fig. 2 — Contour plot of the first approximation for the stream function with values of $\psi^{(1)}$ on indicated contours:

A = 0.00001	F = 0.0005	M = −0.002	Q = −0.0001
C = 0.00005	G = 0.001	N = −0.001	R = −0.00005
D = 0.0001	H = 0.002	O = −0.0005	T = −0.00001
E = 0.0002		P = −0.0002	

If the velocities are sufficiently small then the viscous dissipation can be neglected and the energy equation, (3), in terms of the new variable becomes

$$\nabla^2 \theta - \frac{1}{r}\left(\frac{\partial \psi}{\partial r}\frac{\partial}{\partial \phi} - \frac{\partial \psi}{\partial \phi}\frac{\partial}{\partial r}\right)\theta = 0. \tag{11}$$

The boundary conditions, (4) and (5), become

$$\theta(1, \phi) = \cos 2\phi \tag{12}$$

and

$$\frac{\partial \psi}{\partial r}(1, \phi) = \frac{\partial \psi}{\partial \phi}(1, \phi) = 0. \tag{13}$$

In the case of a small Rayleigh number it is fruitful to seek a solu-

tion in powers of λ;

$$\psi = \lambda\psi^{(1)} + \lambda^2\psi^{(2)} + \cdots \tag{14}$$

and

$$\theta = \theta^{(0)} + \lambda\theta^{(1)} + \lambda^2\theta^{(2)} + \cdots . \tag{15}$$

This expansion is valid in the limit, $\lambda \to 0$, and an upper bound of λ for the validity of the expansion will be obtained subsequently.

When we insert equations (14) and (15) into (10) and (11), the coefficients of like powers of λ must individually be set equal to zero for the equations to hold as λ is varied. Beginning with the lowest order we obtain from equation (11)

$$\nabla^2\theta^{(0)} = 0. \tag{16}$$

The solution to the equation, with the boundary condition given by equation (12), is

$$\theta^{(0)}(r, \phi) = r^2 \cos 2\phi. \tag{17}$$

Next, from equation (10) the lowest order contribution to the stream function is obtained from

$$\nabla^4\psi^{(1)} = \cos\phi \frac{\partial\theta^{(0)}}{\partial r} - \frac{\sin\phi}{r} \frac{\partial\theta^{(0)}}{\partial\phi} \tag{18}$$

with the boundary conditions given by equation (13). Inserting equation (17) into equation (18) and expressing the biharmonic operator in cylindrical coordinates yield

$$\left(\frac{\partial^4}{\partial r^4} + \frac{2}{r}\frac{\partial^3}{\partial r^3} - \frac{1}{r^2}\frac{\partial^2}{\partial r^2} + \frac{1}{r^3}\frac{\partial}{\partial r}\right.$$

$$\left. - \frac{2}{r^3}\frac{\partial^3}{\partial\phi^2\,\partial r} + \frac{2}{r^2}\frac{\partial^4}{\partial\phi^2\,\partial r^2} + \frac{1}{r^4}\frac{\partial^4}{\partial\phi^4} + \frac{4}{r^4}\frac{\partial^2}{\partial\phi^2}\right)\psi^{(1)} = 2r \cos\phi. \tag{19}$$

The solution to this inhomogeneous biharmonic equation is

$$\psi^{(1)}(r, \phi) = \frac{1}{96} [r^4 - 2r^2 + 1]r \cos\phi. \tag{20}$$

Figure 2 is a contour plot of the stream function, equation (20).

Finally we wish to determine the perturbation on $\theta^{(0)}$. This will indicate the effect of thermal convection in distorting the lens and afford an estimation of the upper bound of the Rayleigh number. Again, from equation (11) we get
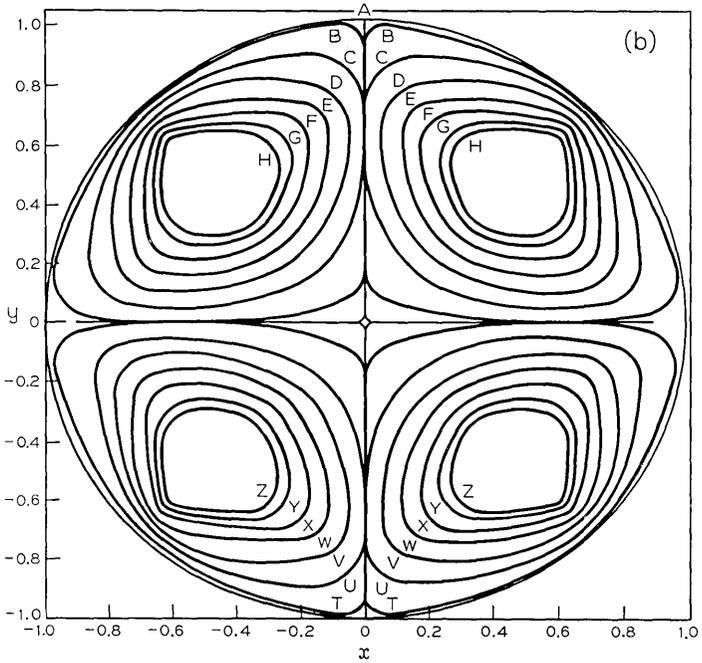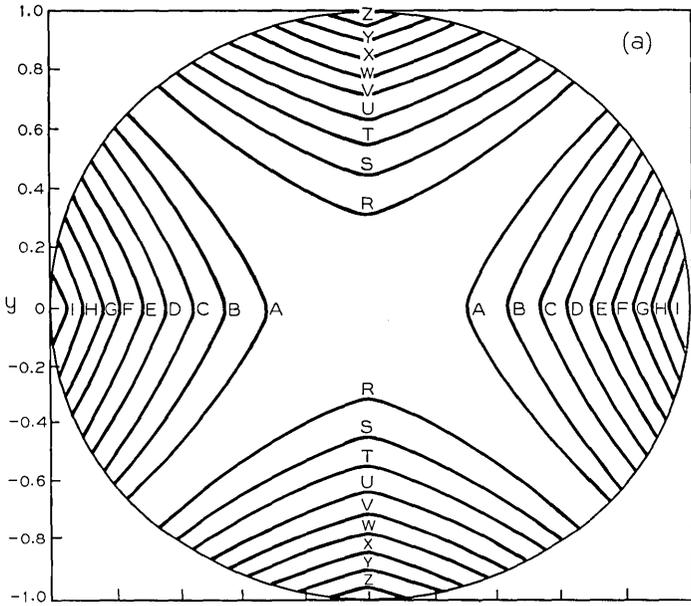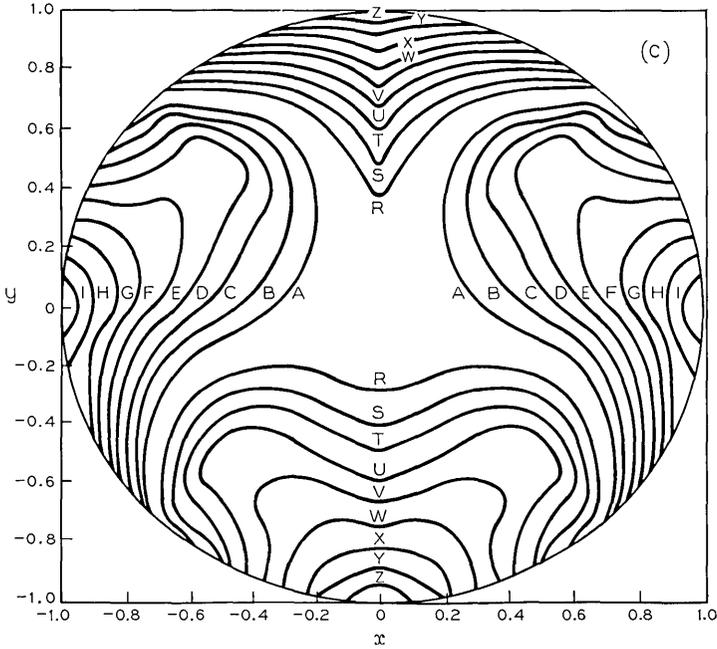
Fig. 3 — Contour plot of the (a) zeroth approximation for the temperature distribution with values of $\theta^{(0)}$ on indicated contours:

$$A = 0.1 = -R \qquad D = 0.4 = -U \qquad G = 0.7 = -X$$
$$B = 0.2 = -S \qquad E = 0.5 = -V \qquad H = 0.8 = -Y$$
$$C = 0.3 = -T \qquad F = 0.6 = -W \qquad I = 0.9 = -Z$$

(b) first perturbation for the temperature distribution with values of $\theta^{(1)}$ on indicated contours:

$$A = 0.0001 = -S \qquad C = 0.005 = -U \qquad E = 0.015 = -W \qquad G = 0.025 = -Y$$
$$B = 0.001 \;\; = -T \qquad D = 0.01 \;\; = -V \qquad F = 0.02 \;\;\; = -X \qquad H = 0.028 = -Z$$

(c) first approximation for the temperature distribution ($\lambda = 10^3$) with values of $\theta^{(0)} + \lambda\theta^{(1)}$ on indicated contours:

$$A = 0.1 = -R \qquad D = 0.4 = -U \qquad G = 0.7 = -X$$
$$B = 0.2 = -S \qquad E = 0.5 = -V \qquad H = 0.8 = -Y$$
$$C = 0.3 = -T \qquad F = 0.6 = -W \qquad I = 0.9 = -Z$$

$$\nabla^2 \theta^{(1)} = \frac{1}{r} \left(\frac{\partial \psi^{(1)}}{\partial \phi} \frac{\partial \theta^{(0)}}{\partial r} - \frac{\partial \psi^{(1)}}{\partial r} \frac{\partial \theta^{(0)}}{\partial \phi} \right) \qquad (21)$$

with the boundary condition

$$\theta^{(1)}(1, \phi) = 0. \qquad (22)$$

Inserting equation (17) and (10) into equation (21) yields

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{2}\frac{\partial^2}{\partial \phi^2}\right)\theta^{(1)} = -2(r^4 - 2r^2 + 1)r \sin \phi \cos 2\phi$$

$$+ 2(5r^4 - 6r^2 + 1)r \cos \phi \sin 2\phi. \qquad (23)$$

The solution which satisfies equation (22) is

$$\theta^{(1)}(r, \phi) = [f_1(1) - f_2(1)]\frac{r}{2}\sin \phi - [f_1(1) + f_2(1)]\frac{r^3}{2}\sin 3\phi$$

$$+ f_1(r) \sin \phi \cos 2\phi + f_2(r) \cos \phi \sin 2\phi, \qquad (24)$$

where

$$f_1(r) \equiv \frac{r^4}{96}\left(\frac{32}{379} - \frac{48}{2041}r^2\right)$$

and

$$f_2(r) \equiv \frac{r^3}{4(96)}\left(2r^4 - \frac{48(31)}{2041}r^3 - 4r^2 + \frac{32(11)}{379}r + 2\right).$$

A numerical calculation reveals that the maximum value of $\theta^{(1)}$ is approximately 3×10^{-4}. Since $\theta^{(0)}$ is bounded by unity, the expansion should be valid for Rayleigh numbers less than the order of 10^4. Figures 3a, b, and c show contour plots of $\theta^{(0)}$, $\theta^{(1)}$, and $\theta^{(0)} + \lambda\theta^{(1)}$, respectively. In Fig. 3c, $\lambda = 10^3$ to demonstrate the distortion possible.

Experiments are being conducted to verify the foregoing results and to better understand thermal convection in other circumstances. Figure 4 is a photograph of the streamlines made visible by the introduction of cigarette smoke into a circular cylinder having a $\cos 2\phi$ temperature distribution. The Rayleigh number is 575. Notice the resemblance between this pattern and the contour plot of the preceding analytical results (Fig. 2). The slight shift upward of the smoke streamlines can be attributed to higher order terms in θ and ψ. The steadiness of the observed flow supports our seeking time-independent solutions of the equations of motion.

2.2 Establishing the cos 2φ Wall Temperature Distribution

If one imposes a linear temperature distribution across a slab by heating one face and cooling the other, and then if one drills a cylindrical hole parallel to the faces of the slab, it is well known that a temperature distribution varying as $\cos \phi$ will appear on the wall of the cylindrical hole. Extending this to a square region with one pair of opposite faces heated and the other pair cooled one might presume

that a cos 2ϕ temperature distribution would appear on a cylindrical hole cut in the center of the square. To determine the degree of approximation of this presumption the heat conduction problem in a region bounded on the exterior by a square and on the interior by a circle is analyzed in the following paragraphs. Figure 5 shows the geometry of the problem.

The problem of a square with a hole in it cannot be solved exactly, as we show. An approximate solution could be sought in either cartesian or cylindrical coordinates. However, considering the problem in cylindrical coordinates allows one to compare the relative magnitude of the portion of the distribution, which varies with cos 2ϕ, to that associated with higher order terms. Secondly, the solution is more nearly
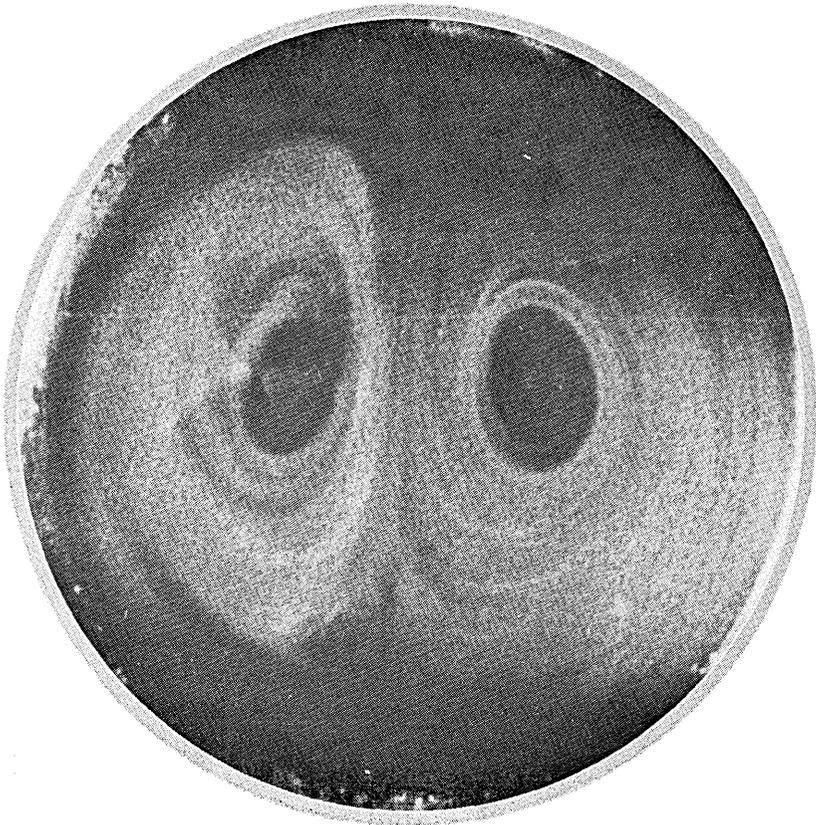
Fig. 4 — Convective motion illuminated by cigarette smoke.
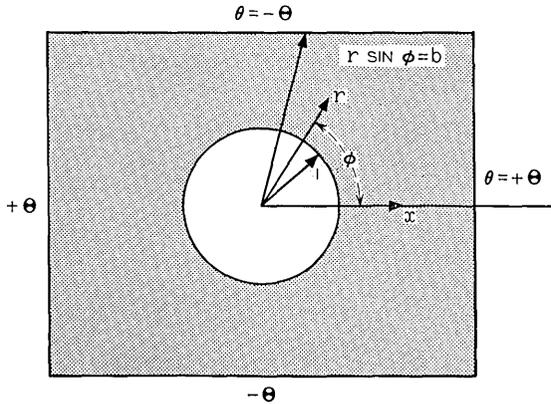
Fig. 5 — Geometry for conduction problem in solid cross section of gas lens.

exact on the cylindrical hole if the approximate solution is sought in cylindrical coordinates. Furthermore, since the heating arrangement has a certain amount of symmetry, only a sector $\pi/4 \leqq \phi \leqq \pi/2$ need be considered.

For steady two-dimensional conduction in a material having constant thermal conductivity the heat conduction equation becomes

$$\frac{\partial^2 \theta}{\partial r^2} + \frac{1}{r} \frac{\partial \theta}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \theta}{\partial \phi^2} = 0. \tag{25}$$

The boundary conditions are:

$$\text{at} \quad r \sin \phi = b, \qquad \theta = -\Theta; \tag{26}$$

$$\text{at} \qquad r = 1, \qquad \frac{\partial \theta}{\partial r} = 0; \tag{27}$$

$$\text{at} \qquad \phi = \frac{\pi}{4}, \qquad \theta = 0; \tag{28}$$

$$\text{at} \qquad \phi = \frac{\pi}{2}, \qquad \frac{\partial \theta}{\partial \phi} = 0. \tag{29}$$

Notice that $\theta = (T - T_o)/\Delta T$ as before, where T is the temperature excursion desired on the cylindrical hole. Consequently, $\Theta = (T_w - T_o)/\Delta T$ where T_w is the wall temperature. The insulated condition, (27), assumes that the heat lost to the gas in the hole is negligibly small compared with the heat conduction in the solid. This is reasonable as

long as $k_{\text{solid}} \gg k_{\text{gas}}$.* Condition (29) results from the symmetry about $\pi/2$. The radius r is normalized with respect to the cylinder radius as in the Section 2.1.

Assume a separable solution of the form

$$\theta = R(r)\Phi(\phi), \qquad (30)$$

so that

$$r^2 R'' + rR' - \alpha^2 R = 0, \qquad (31)$$

$$\Phi'' + \alpha^2 \Phi = 0. \qquad (32)$$

The solution of equation (31) and (32) is

$$\theta = A\left(r^\alpha + \frac{B}{r^\alpha}\right)(C \sin \alpha\phi + \cos \alpha\phi). \qquad (33)$$

The insulated condition (27) is satisfied if $B = 1$. To satisfy both conditions (28) and (29) simultaneously, $C = 0$ and

$$\alpha = 2n, \qquad n = 1, 3, 5, \cdots . \qquad (34)$$

Therefore,

$$\theta_n = A_n\left(r^{2n} + \frac{1}{r^{2n}}\right)(\cos 2n\phi), \qquad n = 1, 3, 5, \cdots , \qquad (35)$$

where A_n is determined to satisfy equation (26), that is,

$$-\Theta = \sum_{n=1,3,5,\cdots}^{\infty} A_n\left(\frac{b^{2n}}{\sin^{2n} \phi} + \sin^{2n} \phi\right) \cos 2n\phi. \qquad (36)$$

Because of equation (26) our problem in r is not a Sturm-Liouville system and we have no assurance that equation (36) will converge even if the A_n's could be determined in general. In what follows we determine the first few A_n's so that equation (36) is satisfied in two different senses as accurately as our needs dictate—collocation and minimization of the error in a least-squares sense.[10]

In the collocation method the error is made to vanish at, say, three particular points on the boundary $r \sin \phi = b$. This gives us three simultaneous equations through which A_1, A_3, and A_5 can be determined. For two different sets of collocation points, the corresponding coefficients are listed in Table I for the ratio of the side length to the

---

* The k for most plastics is a factor of 10 greater than that for air. For formed plastics $k_{\text{solid}} \approx k_{\text{gas}}$ and the behavior of θ at $r = 1$ can be assessed from the solution for conduction in a square with two sides at Θ and two sides at $-\Theta$.

TABLE I—COEFFICIENTS FOR APPROXIMATE SOLUTION BY COLLOCATION

| | Collocation points = 60°, 75°, and 90° | | |
	$b = 2$	$b = 4$	$b = 6$
a_1	1.03531	1.08889	1.09191
a_3	−0.11082	−0.10469	−0.10434
a_5	0.01083	0.01155	0.01159

| | Collocation points = 50°, 70°, and 90° | | |
	$b = 2$	$b = 4$	$b = 6$
a_1	1.04521	1.09927	1.10232
a_3	−0.14096	−0.13457	−0.13421
a_5	0.03046	0.03101	0.03104

cylinder diameter, $b = 2, 4, 6$. These coefficients are normalized with respect to Θ. Furthermore, some of the dependence on b is suppressed when the coefficients are defined as:

$$a_n = \frac{A_n b^{2n}}{\Theta}, \tag{37}$$

so that

$$\frac{\theta}{\Theta} \cong \sum_{n=1,3,5} \frac{a_n}{b^{2n}} \left(r^{2n} + \frac{1}{r^{2n}} \right) \cos 2n\phi, \qquad \frac{\pi}{4} \leq \phi \leq \frac{\pi}{2}. \tag{38}$$

Figure 6 contains a plot of $\theta(r = b/\sin \phi, x/b)$ using both sets of collocation points. This illustrates the degree of approximation entailed at the outer boundary where $-(\theta/\Theta)$ should equal unity over $0 \leq X/b < 1$.

The least-squares method requires that the mean square error over the boundary $r = b/\sin \phi$, $\pi/4 \leq \phi \leq \pi/2$, be as small as possible. Defining

$$\epsilon = \theta - (-\Theta); \tag{39}$$

we then wish to minimize

$$\int_\pi^\pi \epsilon^2 \, d\phi. \tag{40}$$

For convenience we take only the first two terms of equation (35) for θ and note that $r^{2n} \gg 1/r^{2n}$ close to the outer boundary so long as $b \geq 2$. Consequently,

$$\theta \cong A_1 r^2 \cos 2\phi + A_3 \cos 6\phi, \tag{41a}$$

and

$$\epsilon \cong A_1 b^2 \frac{\cos 2\phi}{\sin^2 \phi} + A_3 b^6 \frac{\cos 6\phi}{\sin^6 \phi}. \tag{41b}$$

Inserting equation (41b) into integral (40), performing the integration, and setting the derivative with respect to A_1 and A_3 equal to zero we find that:

$$\frac{\theta}{\Theta} \cong \frac{1.3118}{b^2}\left(r^2 + \frac{1}{r^2}\right)\cos 2\phi - \frac{0.1805}{b^6}\left(r^6 + \frac{1}{r^6}\right)\cos 6\phi. \tag{42}$$

Figure 6 also has a plot of equation (42) evaluated at $r = b/\sin \phi$. Apparently the collocation method yields a much closer approximation for heat conduction problems. (The square of the temperature has no particular physical meaning.)

Returning to the collocation solution (Table I) the temperature distribution on the cylindrical wall, for $b = 6$, is given as

$$\frac{\theta(1, \phi; 6)}{\Theta} \cong \frac{2a_1 \cos 2\phi}{6^2} + \frac{2a_3 \cos 6\phi}{6^6} + \frac{2a_5 \cos 10\phi}{6^{10}}$$
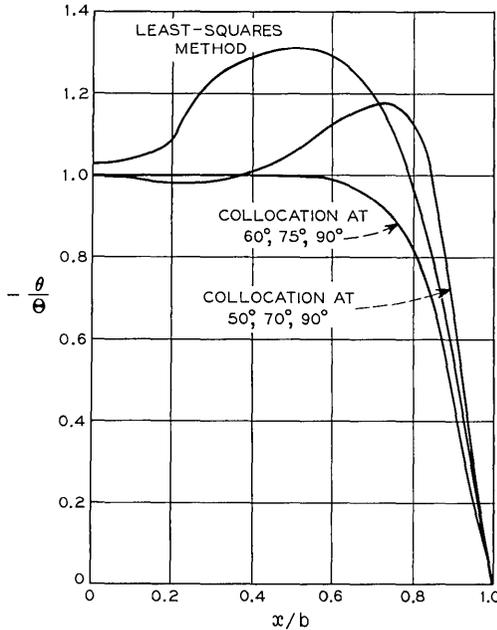
Fig. 6 — Comparison of approximation at outer boundary in conduction problem.

$$\cong 0.06066 \cos 2\phi - 0.0447 \times 10^{-4} \cos 6\phi \qquad (43)$$
$$+ 0.0383 \times 10^{-8} \cos 10\phi,$$

and we see that the distribution varies as $\cos 2\phi$ within one part in 10,000. For $b = 2$ the deviation is somewhat greater, being

$$\frac{\theta(1, \phi; 2)}{\Theta} = 0.5176 \cos 2\phi - 0.00346 \cos 6\phi + 0.02115 \times 10^{-3} \cos 10\phi.$$
$$(44)$$

Similar results can be obtained from equation (42).

Recall that this solution is for $k_{\text{solid}} \gg k_{\text{gas}}$. When $k_{\text{solid}} \approx k_{\text{gas}}$ the deviation of $\theta(1)$ from $\cos 2\phi$ can be evaluated from the analytical solution for a solid square two sides at Θ and two other sides at $-\Theta$, v.[11] In doing this we found that the deviations from $\cos 2\phi$ are of the same order as those cited above.

The power necessary to operate the lens can be readily found from integrating along the radial line at $\phi = \pi/4$. The heat flow rate $\dot{Q}$ through one sector is given by

$$\dot{Q} = -k \, \Delta T \int_{1}^{b} \frac{1}{r} \frac{\partial \theta}{\partial \phi} \, dr$$

$$\dot{Q} = \Delta T \sum_{n=1,3,5} k A_n \sin \frac{n\pi}{2} \left[\left(\frac{2}{\sqrt{2}} \right)^{2n} - \frac{(\sqrt{2}/2)^{2n}}{b^{4n}} \right] \qquad (45)$$

which is nearly independent of b. In terms of the collocation coefficients, where the b^{4n} term has been neglected,

$$\frac{\dot{Q}}{k\Theta \, \Delta T} \cong 2a_1 - 8a_3 + 32a_5 . \qquad (46)$$

For $b = 4$

$$\dot{Q} = 3.38 \, k\Theta \, \Delta T = 3.38 \, k(T_w - T_o). \qquad (47\text{a})$$

Since

$$\theta(1, \pi/2; 4) = -0.136\Theta = -0.136 \left(\frac{T_w - T_o}{\Delta T} \right) = 1, \qquad (47\text{b})$$

then for a ΔT of 1°C excursion $T_{\text{wall}} - T_o = 7.4°C$. For a gas lens whose solid portion is made of polystyrene ($k = 0.1$ W/m°C) the heat flow rate would be $Q = 2.5$ W/m for each sector; then the power requirement would be 10 W/m for 1°C ΔT across the lens. If a foamed polystyrene could be used the power requirement would be 3.5 W/m°C.

2.3 *Optical Properties of the Conduction Lens*

With the effect of thermal convection in mind we now consider how to evaluate the optical properties of a gas lens characterized by the lowest order temperature distribution (that is, $r^2 \cos 2\phi$). We wish to determine these properties as functions of ΔT, cylinder radius a, and lens section length L; we are constrained by the requirement of minimizing ΔT so that the distortion depicted in Fig. 3c shall be tolerable. In the following paragraphs we only write down the relevant equation; we do not establish explicit design criteria.

The system of lenses consists of a sequence of sections with each succeeding one rotated 90 degrees. Therefore, for any angle, ϕ (see Fig. 1), as one marches axially, the sections act alternately as divergent and convergent lenses. Since the temperature varies angularly, as well as radially, so does the refractive index; hence, in addition to the ray bending toward or away from the axis it will, in general, be twisted. However, at $\phi = 0$ and $\pi/2$ the refractive gradient has no angular gradient and, hence, rays originally in either of those planes remain there; they undergo convergent and divergent displacements alternately. All other rays have radial displacements intermediate to those at $\phi = 0, \pi/2$.

The trajectories of the rays in the $\phi = 0, \pi/2$ planes may be calculated analytically and turn out to be sinusoidal and exponential in the convergent and divergent sections, respectively.

Although a numerical solution must be used for the other trajectories, some qualitative observations may be made. In the neighborhood of $\phi = 0$ the angular component of the refractive index causes rays to be twisted away from that attitude, while near $\phi = \pi/2$ rays are restored to that angular position. Therefore, as one moves down a section the density of rays tends to increase near $\phi = \pi/2$ and to decrease near $\phi = 0$.

In order to obtain the intensity of the beam through a lens section, the Helmholtz-type equation with the appropriate refractive index must be solved. This was done by Marcatili[12] for an asymmetrical but convergent-type refractive index.* He established conditions for the stability of a lens system and calculated the focal length.

For our present purposes there is no need for a detailed solution of the field equations but rather for the ray displacement, stability criterion, and focal length. Toward this end Miller's[9] analysis of the

---

* Marcatili informed us that there is no basic reason why his analysis could not be extended to include divergent sections.

ray equation is applicable.† He obtained these quantities by solving the difference equations which govern the passage of the rays through the sequence of lenses. If we only consider the $\phi = 0$, $\pi/2$ planes, then the sections act as alternating convergent and divergent lenses, with the rays remaining in their original planes; we may then apply Miller's results.

Miller obtained the ray displacement after the nth convergent and mth divergent lens for an initially convergent and an initially divergent sequence. He also found the stability condition which keeps the ray trajectory bounded after an infinite number of lenses. This condition is

$$0 < \frac{L}{f} < 2. \tag{48}$$

Here, to serve as an example, we only display the expression for the ray displacement after the nth convergent lens for an initially convergent lens:

$$r_n = r_o k_1 \cos (n\,\delta - \phi_1) + r_0' L k_2 \sin n\,\delta \tag{49}$$

where r_o and r_0' are the initial displacement and slope, respectively, and

$$\delta = \cos^{-1}\left[1 - \frac{1}{2}\left(\frac{L}{f}\right)^2\right], \qquad k_1 = \frac{2}{1 - \dfrac{L}{2f}},$$

$$\phi_1 = |\cos^{-1} k_1^{-1}|, \quad \text{and} \quad k_2 = \frac{2 + \dfrac{L}{f}}{\sin \delta}. \tag{50}$$

Furthermore, we must stipulate that the ray does not intersect the cylinder wall, that is,

$$\frac{r_n}{a} < 1. \tag{51}$$

The relationship between the focal length and the refractive index may be obtained from Marcuse and Miller.[1] For a thin lens the focal length is given by*

---

† We are indebted to Marcatili for several clarifying remarks on this subject.
 * A thin lens is one in which the principal surface generated by rays incident from the left coincides with that surface constructed by rays incident from the right. Since there is no preferred direction with the conduction-type lens it is thin; the flow-type lenses cited in Section I may be approximately thin.

$$f = \tfrac{1}{2}\beta_o \frac{r^2}{\Delta\phi} , \tag{52}$$

where $\beta_o = 2\pi/\lambda$, λ is the wave length of the light, and $\Delta\phi$ is the difference of the phase of a ray incident a distance r from and parallel to the axis after traveling a distance L, compared with a ray on the axis traveling the same distance.

To calculate $\Delta\phi$ in terms of the refractive index, we invoke the paraxial approximation in which the rays are regarded as approximately parallel to the axis. Then the required phases are easy to calculate, that is,

$$\phi(r, z) \cong \beta_o \int_0^L n(r)\, dx = \beta_o n(r)L \tag{53}$$

and

$$\phi(0, z) = \beta_o n(0)L. \tag{54}$$

The refractive index at $\phi = 0$ is

$$n(r, 0) = 1 + (n_o - 1)\frac{T_o}{T(r, 0)} = 1 + \frac{n_o - 1}{1 + \dfrac{\Delta T}{T_o}\dfrac{r^2}{a^2}}$$

$$\cong n_o - (n_o - 1)\frac{\Delta T}{T_o}\frac{r^2}{a^2} , \tag{55}$$

where n_o is the refractive index at the axis at temperature T_o.

Hence,

$$\Delta\phi = \beta_o(n_o - 1)\frac{\Delta T}{T_o}\frac{r^2}{a^2} L \tag{56}$$

and the focal length is obtained from equation (52):

$$f = \frac{1}{2}\frac{a^2 T_o}{(n_o - 1)\,\Delta T L} \tag{57}$$

independent of r.

With the aid of equations (48), (49), (51), and (57) we may determine the focal length and ray displacement as a function of the lens section and radius and temperature excursion. For a complete discussion of the foregoing subject, see Ref. 9.

To establish precise design criteria the foregoing equations must be solved on a computer. However, for illustrative purposes and as one

aspect of the problem we shall make use of some of Miller's simplified expressions valid in certain limits.[9]

If we use equation (49) with the value of the section length to focal length ratio which yields the smallest value of the maximum ray displacement and, furthermore, insure that the rays do not intersect the wall, ΔT obtained is unacceptable for three major reasons (*i*) the power requirement is excessive, (*ii*) the moderate Rayleigh number will cause appreciable distortion, and (*iii*) the temperature excursion is sufficiently large so that section end effects may be significant.

In order to overcome these objections we now examine the case of weak focusing, that is, $2f/L \gg 1$. We consider the initial conditions such that

$$r_o \ll r_o'f. \tag{58}$$

(The opposite inequality for weak focusing yields a trivial design problem since it does not involve the focal length.) From Miller the maximum ray radius, r_{max}, is [9]

$$r_{max} = 2fr_0'. \tag{59}$$

To insure that the ray does not intersect the wall we have

$$\frac{a}{L} \geqq \frac{r_{max}}{L}. \tag{60}$$

Inserting equations (57) and (59) into equation (60) yields

$$1 \geqq \frac{a}{L} \frac{T_o}{(n_o - 1)\, \Delta T} r_o'. \tag{61}$$

As an example we use the following values, where air is the medium of the lensing action

$$T_o = 290°\text{K},$$

$$n_o - 1 = 0.295 \times 10^{-3},$$

and

$$\lambda \text{ (Rayleigh number)} = 9.15 \times 10^7\ \Delta T a^3$$

with ΔT in degrees Celsius and a in meters. In addition, let $r_0' = 10^{-4}$, $a = 3 \times 10^{-3}$m, and $L = 0.5$ m. Then from inequality (61) we obtain

$$\Delta T \geqq 0.59°\text{C}. \tag{62}$$

Hence, $\lambda = 2.9$. Consequently,

$$| \lambda \theta^{(1)} | < 10^{-3} \ll | \theta^{(0)} |_{\mathrm{max}} .$$

It should be borne in mind that with the above values of a and L, the weak focusing limit is satisfied for $\Delta T \lesssim 5\,°\mathrm{C}$. In addition, to satisfy inequality (58) with the foregoing values we must have

$$r_o \ll r_0' f = 1.5 \times 10^{-3} \; \mathrm{m} \tag{63}$$

which is easy to satisfy.

Using equation (47a), we obtain for the heat flow rate through one sector

$$\dot{Q} = 1.5 \; \mathrm{W/m}$$

which results in a power requirement of 6.0 W/m. The required exterior wall temperature is calculated from equation (47b) as $T_w = T_o + 4.4\,°\mathrm{C}$. Therefore, in the limit of weak focusing the temperature excursion is sufficiently small to make the lens system promising.

Considering the flow-type lens of Marcuse and Miller to have the same characteristics as the above conduction type lens, we calculate the power expended at optimum flow rate to be 1.14 W.[1] Hence, the lens proposed here requires somewhat more power for heating than those previously investigated. However, the flow-type lens also requires power to drive the gas.

Since the input beam will be more complicated then was assumed above, the foregoing calculation is very cursory. However, the reasonable magnitudes of a and L together with the small Rayleigh number lend encouragement to a more detailed analysis.

III. CONCLUSIONS AND RECOMMENDATIONS

The conduction-type lens proposed here is found to be feasible on the basis of negligible distortion resulting from thermal convection and reasonable power requirements to maintain the desired temperature distribution. Although the lens design illustrated was predicated on the weak focusing limit a wider range of parameters can be found by using Miller's complete expression.[9]

The effect of thermal convection was calculated from a two dimensional analysis, which is certainly valid away from the ends of the section since $a/L \ll 1$. For the temperature excursion required and the lens illustrated, the convection effect was found to be negligible. However, at the interface between the sections, the axial temperature gradients could be large depending on the spacing left between sections. Axial gradients were present in the experiments of Suematsu

and others for their hyperbolic shaped conduction-type lens system.[8] They found that no significant aberration existed as long as $\Delta T < 42/a^{1.34}$ (a in millimeters) so that the effects of the axial gradients must have been insignificant.

The analyses presented indicate that a system of conduction-type lenses might be practical for an alternating gradient light-beam waveguide. Such a system would require straight square rods with a cylindrical hole. Two sides of the rod would be heated while the other two would be held at a uniform and constant temperature. This could be done by attaching aluminum fins which project into a constant temperature heat sink to the cooled sides. Such a heat sink is available for buried systems since, at depths greater than about five feet, the surface temperature changes are virtually damped out. Therefore, cooing is not required.

The hole in the rod would be of the order of 6 mm in diameter and the exterior could be as small as 2.4 cm across a face. Larger hole dimensions could be used but, for the same size beam and lensing action, the temperature difference and power requirement would increase proportionately.

After only a preliminary design analysis, where the simplest of Miller's expressions have been used, parameters have been obtained in the weak focusing limit which yield a power consumption somewhat greater than but of the same order of magnitude as flow-type gas lenses.[9] Additional investigations are, of course, necessary. The distortion of a gaussian beam as it is launched through a lens system should be numerically calculated (similar to Marcuse's study for the flow-type lens.[5]) The effect of the axial gradients that will be present at the interface between two lens sections will have to be assessed through experimental measurements of the optical performance of such a lens system.

REFERENCES

1. Marcuse, D., and Miller, S. E., "Analysis of a Tubular Gas Lens," B.S.T.J., *43*, No. 4 (July 1964), pp. 1759–1787.
2. Marcuse, D., "Theory of a Thermal Gradient Gas Lens," IEEE Trans., *MTT-13*, No. 6 (November 1965), pp. 734–739.
3. Berreman, D. W., "A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases," B.S.T.J., *43*, No. 4 (July 1964), pp. 1469–1475.
4. Berreman, D. W., "Convective Gas Light Guides or Lens Trains for Optical Beam Transmission," J. Opt. Soc. Amer., *55*, No. 3 (March 1965), pp. 239–247.
5. Marcuse, D., "Deformation of Fields Propagating Through Gas Lenses," B.S.T.J., *45*, No. 8 (October 1966), pp. 1345–1358.

6. Kaiser, P., unpublished work.

7. Gu, A. L., unpublished work.

8. Suematsu, Y., Iga, K., and Ito, S., "A Light Beam Waveguide Using Hyperbolic-Type Gas Lenses," IEEE Trans., *MTT-14*, No. 12 (December 1966), pp. 657–667.

9. Miller, S. E., "Alternating-Gradient Focusing and Related Properties of Conventional Convergent Lens Focusing," B.S.T.J., *43*, No. 4 (July 1964), pp. 1741–1758.

10. Collatz, L., *The Numerical Treatment of Differential Equations,* New York: Springer-Verlag, 1966.

11. Carslaw, H. S., and Jaeger, J. C., *Conduction of Heat in Solids,* New York: Oxford University Press, 1959.

12. Marcatili, E. A. J., "Modes in a Sequence of Thick Astigmatic Lens-Like Focusers," B.S.T.J., *43*, No. 7 (Novemeber 1964), pp. 2887–2904.

Resonances in Waveguide Antennas with Dielectric Plugs

By C. P. WU

This paper discusses an analysis of the radiation from a parallel-plate waveguide to determine the effects of loading the waveguide with dielectric plugs near the aperture. We devote special attention to the situation in which the higher order modes, generated by the aperture discontinuity, propagate inside the dielectric plug but are evanescent in the unloaded waveguide region. We show that the dielectric plug may function as a resonant cavity for this type of wave mode. When one of these modes is at resonance, it is strongly excited by the incident wave; the presence of this resonance is manifested by the appearance of sharp spikes in the reflection coefficient either as a function of the frequency or the plug thickness. We also discuss the relation between the resonances in a single waveguide and in array configuration.

I. INTRODUCTION

The radiation from a parallel-plate waveguide with infinitesimally thin walls is one of the relatively few electromagnetic boundary value problems for which the Wiener-Hopf integral equation technique may be applied to obtain a closed form solution.[1] Unfortunately, this elegant mathematical technique quickly loses its usefulness even when rather minor modifications of the physical system are introduced, such as, for example, by allowing the waveguide to have finite wall thickness or loading the waveguide with a dielectric material.

The somewhat simpler problem of determining the radiation admittance of a waveguide terminated in an infinite conducting plane has been treated by several workers using the variational technique.[2,3] The field of the incident wave is used to approximate the true aperture field in these calculations. The results thus obtained appear adequate for engineering purposes. The implication is that the radiation admittance of an empty waveguide is rather insensitive to the approxima-

tion used for the aperture field distribution. There is no way, however, to ascertain without more elaborate calculations how well the aperture field is approximated by that of the incident wave.

The variational technique has also been widely used in a broad class of scattering problems. Although it seems that useful approximate answers are often obtainable even when rather crude approximations are used for the trial funnctions, there are numerous instances, notably in the area of phased arrays[4] and in problems involving dielectric material,[5] wherein it has been found that good approximations of the trial functions are necessary to obtain meaningful results. An important factor contributing to this knowledge undoubtedly is the widespread availability of high speed electronic computers, which have made it possible to perform elaborate computations hitherto regarded as too time-consuming and costly to be practical.

In this paper, we discuss the radiation properties of a waveguide which is loaded with dielectric plugs near the aperture and is terminated in an infinite conducting plane. A waveguide antenna has the advantage that it can be flush mounted. This feature makes it attractive for applications such as missile and aircraft antennas. Dielectric plugs, moreover, provide convenient covers to protect the antenna feed system against environmental influences. The introduction of dielectric material, however, makes it possible to excite the wave modes which have a surface wavelike field distribution within the waveguide because they propagate inside the dielectric plug but are evanescent in the empty waveguide region. (This excitation is caused by the aperture discontinuity.)

We show that because of the excitation of this type of wave mode, the antenna impedance (or the reflection coefficient) exhibits resonance characteristics versus both the frequency and the thickness of the dielectric plug. These resonances occur when the parameters are such that the impedances of a surface wavelike mode (or "ghost mode") satisfy a transverse resonance condition. The implication of this observation is that the dielectric plug acts like a resonance cavity for the surface wavelike modes. When the combination of the parameters is such as to permit one of these modes to resonate, the effect is to cause rapid variation in the radiation impedances (or reflection coefficient) which are manifested as sharp spikes.

The radiation patterns generally show smooth variations versus the angle of observation. Only when the parameters are in the close vicinity of a resonance such that the higher order mode is exceedingly strongly excited do pattern dips appear. Moreover, the dips are rather

broad and shallow. It is therefore necessary to exercise extreme care in order to detect resonances by examining the patterns alone.

An earlier analysis of phased arrays using the waveguide at hand as the radiation element has revealed that resonance characteristics also exist in both the reflection coefficients and the mutual coupling of the array.[6,7] These resonances are related in certain ways to those of the present problem. We briefly discuss the relationship with the view toward using a single waveguide for the detection of the resonances in an array configuration.

The boundary value problem is formulated in two ways, one in a pure integral equation with the tangential magnetic field as the unknown and the other in an integro-differential equation with the aperture electric field as the unknown. It appears that no known analytical method is available for solving either equation. It is possible, however, to use numerical technique to determine approximate but accurate solution from the latter equation. We discuss the method of obtaining solutions by the method of moments; we also point out certain salient features with regard to the formulation.

II. FORMULATION OF THE PROBLEM

Consider a parallel-plate waveguide, terminated in an infinite conducting plane as illustrated in Fig. 1. The waveguide is loaded with a dielectric plug (or window) near the aperture. We consider the system to be excited by the lowest TE mode incident upon the aperture from the waveguide side, and assume the fields to be invariant with respect to y. Under these conditions, it is easily shown that the scat-
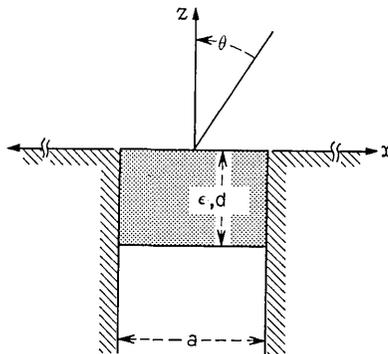
Fig. 1 — A flush mount parallel-plate waveguide with dielectric plug.

tered fields consist of TE modes alone. We determine the radiation characteristics of the antenna by using the integral equation approach.

2.1 *Integral Equations*

The problem may be formulated in terms of integral equations having as the unknown function either the tangential electric field or the tangential magnetic field in the plane $z = 0$. In order to do so, we must first introduce suitable representations for the tangential fields in the regions both inside and outside the waveguide. The application of boundary conditions using these field representations across the common $z = 0$ plane then leads to the desired integral equations. We derive first the equation with the tangential electric field as the unknown.

2.2 *Integro-Differential Equation for Aperture Electric Field*

By virtue of the equivalence principle,[8] the fields in $z \geq 0$ may be derived from an equivalent magnetic dipole $\mathbf{M} = \mathbf{E}_t \times \hat{z}$ situated above a perfectly conducting plane, where $\mathbf{E}_t$ denotes the tangential electric field which exists at the aperture and $\hat{z}$ is a unit vector in the z direction. According to the image theorem, these fields are equal to twice the fields produced by the same equivalent source in free space. Since $\mathbf{E}_t = \hat{y}E_y(x', 0)$, $\mathbf{M} = \hat{x}E_y(x', 0)$. The vector potential due to this source distribution may be determined easily to be

$$\mathbf{F} = \hat{x}\, \frac{-j}{2} \int_A H_0^{(2)}(kR)E_y(x', 0)\ dx', \tag{1}$$

where A denotes the waveguide aperture, $H_0^{(2)}(\mu)$ is the zeroth order Hankel function of the second kind, and $R = [(x - x')^2 + z^2]^{\frac{1}{2}}$. We use the time convention $\exp j\omega t$, which is suppressed for brevity.

The electromagnetic fields in $z \geq 0$ may be derived from $\mathbf{F}$ by

$$\mathbf{E} = -\nabla \times \mathbf{F},$$

$$\mathbf{H} = \frac{1}{j\omega\mu_0}\, [k^2\mathbf{F} + \nabla(\nabla \cdot \mathbf{F})]. \tag{2}$$

In particular, we find that the tangential field components are given by

$$E_y(x, z) = \frac{j}{2} \int_A E_y(x', 0)\, \frac{\partial}{\partial z}\, H_0^{(2)}(kR)\ dx',$$

$$H_x(x, z) = -\frac{1}{2\omega\mu_0} \left(k^2 + \frac{\partial^2}{\partial x^2}\right) \int_A E_y(x', 0)H_0^{(2)}(kR)\ dx'. \tag{3}$$

Notice that the integrals in equation (3) have to be evaluated carefully when z approaches 0. In particular, the differentiation and integration in the second equation may not be interchanged when $z \to 0$, because in doing so the integral becomes divergent.

The fields inside the waveguide are most conveniently expressed in terms of the waveguide modal functions. The presence of dielectric plugs near the aperture may be accounted for by using appropriate modal admittances which are derivable by applying the transmission line theory. Assuming that the incident wave originating in the region $z < -d$ has unit modal voltage, we may write the tangential electromagnetic fields at the aperture as

$$E_y(x, 0) = \sum_{n=1}^{\infty} \bar{V}_n \varphi_n(x),$$

$$H_x(x, 0) = -2\tilde{Y}_1 \varphi_1(x) + \sum_{n=1}^{\infty} \tilde{Y}_n \bar{V}_n \varphi_n(x),$$

(4)

where φ_n are the orthonormal modal functions, and

$$\tilde{Y}_n = Y_n^D \frac{Y_n + jY_n^D \tan \alpha_n^D d}{Y_n^D + jY_n \tan \alpha_n^D d},$$

$$\bar{Y}_1 = \frac{Y_1 Y_1^D}{Y_1^D \cos \alpha_1^D d + jY_1 \sin \alpha_1^D d},$$

(5)

with

$$Y_n = \frac{\alpha_n}{\omega \mu_0} \quad \text{and} \quad Y_n^D = \frac{\alpha_n^D}{\omega \mu_0}$$

(α_n^D and α_n being the nth propagation constants in the waveguide region with and without a dielectric, respectively). The $\bar{V}_n$ are the modal voltages at the aperture. When the modal voltages V_n in the empty waveguide region are desired, they may be obtained by using the following formula

$$V_n = \frac{Y_n^D}{Y_n^D \cos \alpha_n^D d + jY_n \sin \alpha_n^D d} \bar{V}_n$$

$$+ j \frac{2Y_1 \sin \alpha_1^D d}{Y_1^D \cos \alpha_1^D d + jY_1 \sin \alpha_1^D d} \delta_{1n} \quad (6)$$

where δ_{1n} is the Kronecker delta. The reflection coefficient R is obtainable from

$$1 + R = V_1. \tag{6a}$$

The orthonormality of the waveguide modal functions may be applied to the first equation of (4) to obtain

$$\bar{V}_n = \int_A E_y(x', 0)\varphi_n(x') \, dx'.$$

When the result is substituted into the second equation of (4), we find

$$H_x(x, 0) = -2\bar{Y}_1\varphi_1(x) + \sum_{n=1}^{\infty} \tilde{Y}_n\varphi_n(x) \int_A \varphi_n(x')E_y(x', 0) \, dx'. \tag{7}$$

Notice that the summation and integration in equation (7) are not interchangable. The reason is that when the summation is brought under the integral sign, the resulting kernel has a singularity of the form $1/(x - x')^2$, which is nonintegrable in the usual sense. In order to circumvent this difficulty and to put equation (7) into a form suitable for combination with equation (3) when the boundary condition is applied, we use the following relation

$$\tilde{Y}_n\varphi_n(x)\varphi_n(x') = \left(\frac{\partial^2}{\partial x^2} + k^2\right)\frac{\tilde{Y}_n}{\alpha_n^2}\varphi_n(x)\varphi_n(x'). \tag{8}$$

Equation (7) may then be written as

$$H_x(x, 0) = -2\bar{Y}_1\varphi_1(x) + \left(\frac{\partial^2}{\partial x^2} + k^2\right)$$
$$\cdot \int_A \left[\sum_{n=1}^{\infty} \frac{\tilde{Y}_n}{\alpha_n^2}\varphi_n(x)\varphi_n(x')\right]E_y(x', 0) \, dx'. \tag{8a}$$

An application of the continuity condition on H_x across the aperture leads to

$$2\bar{Y}_1\varphi_1(x) = \left(\frac{\partial^2}{\partial x^2} + k^2\right)\int_A \left[\sum_{n=1}^{\infty} \frac{\tilde{Y}_n}{\alpha_n^2}\varphi_n(x)\varphi_n(x')\right.$$
$$\left. + \frac{1}{2\omega\mu_0} H_0^{(2)}(k \mid x - x' \mid)\right]E_y(x', 0) \, dx' \quad \text{for} \quad x \, \varepsilon \, A. \tag{9}$$

This is the integral equation having as the unknown function the tangential electric field which is nonvanishing only over the aperture region.

Notice that the step introduced in equation (8) to facilitate the interchange of integration and summation is not essential in our later application of moment method for solution. The procedure, however, enables us to obtain a compact integro-differential equation from

which a pure integral equation may be derived, thus permitting a solution by different techniques.

2.3 Integral Equation for Tangential Magnetic Field

We next consider the integral equation using the tangential magnetic field at $z = 0$ as the unknown function. The derivation in this case follows the same procedure as discussed in Section 2.2. We first recognize that the fields in $z \geqq 0$ may be expressed in terms of the tangential magnetic field as follows

$$E_y(x, z) = -\frac{\omega \mu_0}{2} \int_{-\infty}^{\infty} H_0^{(2)}(kR) H_x(x', 0) \, dx',$$

$$H_x(x, z) = -\frac{j}{2} \int_{-\infty}^{\infty} \frac{\partial}{\partial z} H_0^{(2)}(kR) H_x(x', 0) \, dx', \tag{10}$$

$$H_z(x, z) = \frac{j}{2} \int_{-\infty}^{\infty} \frac{\partial}{\partial x} H_0^{(2)}(kR) H_x(x', 0) \, dx'.$$

The limits of integration extend from $-\infty$ to ∞ because $H_x(x', 0)$ has values over the entire $z = 0$ plane. The fields inside the waveguide are given by

$$H_x(x, 0) = \sum_{n=1}^{\infty} \bar{I}_n \varphi_n(x),$$

$$E_y(x, 0) = -2\tilde{Z}_1 \varphi_1(x) + \sum_{n=1}^{\infty} \tilde{Z}_n \bar{I}_n \varphi_n(x), \tag{11}$$

where

$$\tilde{Z}_n = 1/\tilde{Y}_n, \qquad \tilde{Z}_1 = \frac{Z_1 Z_1^D}{Z_1^D \cos \alpha_1^D d + j Z_1 \sin \alpha_1^D d}.$$

Again, the $\bar{I}$'s are the modal currents defined at the aperture, and the modal currents I_n for the empty waveguide region are related to $\bar{I}_n$ by

$$I_n = \frac{Z_n^D}{Z_n^D \cos \alpha_n^D d + j Z_n \sin \alpha_n^D d} \bar{I}_n$$

$$+ j \frac{2 Z_1 \sin \alpha_1^D d}{Z_1^D \cos \alpha_1^D d + j Z_1 \sin \alpha_1^D d} \delta_{1n}. \tag{12}$$

The reflection coefficient may be calculated using

$$1 - R = I_1. \tag{12a}$$

Equation (11) may be rewritten by making use of the orthonormality relation between the φ's. Thus,

$$E_y(x, 0) = -2\bar{Z}_1\varphi_1(x) + \int_A \left[\sum_{n=1}^{\infty} \tilde{Z}_n\varphi_n(x)\varphi_n(x') \right] H_x(x', 0) \, dx'. \quad (13)$$

In obtaining equation (13), the integration and summation have been interchanged. This is permissible because the kernel

$$\sum_{n=1}^{\infty} \tilde{Z}_n\varphi_n(x)\varphi_n(x')$$

behaves like $\ln |x - x'|$ so that the integral is absolutely convergent for physically acceptable solution H_x.

We are now ready to derive the integral equation by applying the boundary condition using equations (10) and (13). The limits of integration in equation (13) may be extended from A to $(-\infty, \infty)$ with the understanding that the φ's are defined to be identically zero outside the aperture. We thus obtain

$$2\bar{Z}_1\varphi_1(x) = \int_{-\infty}^{\infty} \left[\sum_{n=1}^{\infty} \tilde{Z}_n\varphi_n(x)\varphi_n(x') \right.$$

$$\left. + \frac{\omega\mu_0}{2} H_0^{(2)}(k \mid x - x' \mid) \right] H_x(x', 0) \, dx' \quad -\infty < x < \infty. \quad (14)$$

Notice that equation (9) and (14) may be cast into variational form for the input impedance and admittance, respectively.

III. SOLUTIONS OF THE INTEGRAL EQUATIONS

Equations (9) and (14) constitute a pair of alternative integral equations for the radiation from a parallel-plate waveguide into a half space. One of the equations has as the unknown function the tangential electric field, while the other has as the unknown function the tangential magnetic field. Since there is no known method for solving these equations analytically, we have to resort to approximate techniques. Because of the infinite limits associated with the equation for the magnetic field, which is usually rather difficult to handle numerically, the one for the electric field is much preferred.

Strictly speaking, equation (9) is an integro-differential equation. We may derive from it a pure integral equation in a similar vein as Hallen did for the dipole antenna. The usefulness of this approach is currently being investigated. We discuss solutions of equation (9) directly by the method of moments.[9,10] To do so, we first approximate the aperture electric field by the following representation

$$E_y(x', 0) \approx \sum_{n=1}^{N} b_n U_n(x'), \tag{15}$$

where $U_n(x')$ is a set of linearly independent functions which are chosen to satisfy the boundary conditions on E_y at both ends of the aperture, that is

$$U_n(0) = U_n(a) = 0. \tag{16}$$

Substituting equation (15) into equation (9) gives

$$2\bar{Y}_1\varphi_1(x) \approx \sum_{n=1}^{N} b_n\left(\frac{\partial^2}{\partial x^2} + k^2\right) \int_A \left[\sum_{n=1}^{\infty} \frac{\tilde{Y}_n}{\alpha_n^2} \varphi_n(x)\varphi_n(x') \right.$$

$$\left. + \frac{1}{2\omega\mu_0} H_0^{(2)}(k \mid x - x' \mid) \right] U_n(x') \, dx'. \tag{17}$$

We next require the difference between the left and right sides of equation (17) to be orthogonal to another set of functions

$$W_n(x), \qquad n = 1, 2, \cdots, N$$

with $W_n(0) = W_n(a) = 0$ (for reasons to become apparent presently). This last step then converts the integral equation into a set of algebraic equations

$$\sum_{p=1}^{N} A_{qp}b_p = f_q, \qquad q = 1, 2, \cdots, N, \tag{18}$$

where

$$A_{qp} = \sum_{n=1}^{\infty} \tilde{Y}_n(W_q, \varphi_n)(\varphi_n, U_p)$$

$$+ \frac{1}{2\omega\mu_0} \int_A dx \, W_q(x)\left(\frac{\partial^2}{\partial x^2} + k^2\right) \int_A dx' \, H_0^{(2)}(k \mid x - x' \mid)U_p(x'), \tag{19}$$

$$f_q = 2\bar{Y}_1 \int_A dx \, \varphi_1(x)W_q(x),$$

with

$$(W_q, \varphi_n) = \int_A dx \, W_q(x)\varphi_n(x).$$

For the evaluation of A_{qp}, it is desirable that $U_p(x)$ be chosen such that the integration of $U_p(x)$ and $H_0(k \mid x - x' \mid)$ can be carried out in closed form. Unfortunately, such functions which will also satisfy the boundary conditions (16) are not easy to find. This being the case, we shall manipulate the expression in equation (19) into forms which are

more convenient to implement for numerical integration. Thus, by interchanging one differentiation with the integral with respect to x', and then integrating by parts twice (once with respect to x' and once with respect to x), we find

$$
\int_A dx\ W_q(x) \frac{\partial^2}{\partial x^2} \int_A dx'\ H_0^{(2)}(k \mid x - x' \mid) U_p(x')
$$

$$
= - \int_A dx\ \frac{dW_q(x)}{dx} \int_A dx'\ H_0^{(2)}(k \mid x - x' \mid) \frac{dU_p(x')}{dx'} ,
$$

where we have used the relation

$$
\frac{\partial}{\partial x} H_0^{(2)}(k \mid x - x' \mid) = -\frac{\partial}{\partial x'} H_0^{(2)}(k \mid x - x' \mid)
$$

and the fact that the integrated terms vanish on account of the boundary conditions.

Using this result, we may rewrite equation (19) as

$$
A_{qp} = \sum_{n=1}^{\infty} \tilde{Y}_n(W_q , \varphi_n)(\varphi_n , U_p)
$$

$$
+ \frac{1}{2\omega\mu_0} \left[k^2 \int_A dx\ W_q(x) \int_A dx'\ H_0^{(2)}(k \mid x - x' \mid) U_p(x')
\right.
$$

$$
\left. - \int_A dx\ \frac{dW_q(x)}{dx} \int_A dx'\ H_0^{(2)}(k \mid x - x' \mid) \frac{dU_p(x')}{dx'} \right]. \tag{20}
$$

The double integrals in equation (20) may be converted into single integrals by a transformation of variables. If the waveguide modal functions are chosen as both the basis and testing functions and if the fact that only modes of even symmetry with respect to yz plane are excited is accounted, we obtain

$$
A_{qp} = \tilde{Y}_q\ \delta_{qp} + \frac{1}{2\omega\mu_0} \int_A ds\ H_0^{(2)}(ks)F_{qp}(s), \tag{21}
$$

where

$$
F_{qp}(s) = \begin{cases} \dfrac{2}{(q^2 - p^2)\pi} \left[k_p'^2 q \sin \dfrac{p\pi}{a} s - k_q'^2 p \sin \dfrac{q\pi}{a} s \right] & q \neq p \\[4ex] \dfrac{1}{a} \left\{ k_p'^2(a - s) \cos \dfrac{p\pi}{a} s + \left[k^2 + \left(\dfrac{p\pi}{a} \right)^2 \right] \dfrac{\sin \dfrac{p\pi}{a} s}{\dfrac{p\pi}{a}} \right\} & q = p \end{cases}
$$

with

$$k_p'^2 = k^2 - \left(\frac{p\pi}{a}\right)^2.$$

The last integrals in equation (21) may be evaluated numerically. We have found that a fast, accurate, and yet economical way is to apply the Simpson's rule with the values of the Hankel function obtained from the Tschebycheff representation.[11]

After the matrix elements are calculated, the set of equations (14) is ready for a solution. An advantage of choosing the waveguide modal functions as both the basis and testing functions in the application of the moments method is that the solutions are expressed directly in terms of the modal coefficients of the aperture field. The reflection coefficients are then easily calculated by using equations (6) and (6a).

The radiation patterns of the antenna may be obtained from equation (3). Introducing the asymptotic expression for large arguments for the Hankel function, we find that the electric field in the far zone is approximated by

$$E_y(r, \theta) \approx \left(\frac{k}{2\pi r}\right)^{1/2} e^{-i(kr-3\pi/4)} \cos \theta \int_A E_y(x', 0)e^{jk \sin \theta x'} dx'. \qquad (22)$$

It is easy to show that the magnetic field in the far zone is related to the electric field through the free space admittance. Thus,

$$H_\theta(r, \theta) = \eta_0 E_y(r, \theta), \quad \text{for} \quad kr \gg 1,$$

where η_0 is the characteristic admittance of free space. For comparison, it is often desirable to normalize the radiation patterns. A commonly used normalization is to make the amplitude unity in the direction of maximum radiation. We use a different normalization here, however. Our patterns are normalized such that the integral of the square of the amplitudes gives the radiated power when a unit power is supplied to the incident wave. This way of displaying the patterns is more advantageous because it shows the normalized radiation intensity in addition to the information contained in the usual pattern presentation; this provides a basis of comparison when the frequency is varied. Thus, using expression (15) with $\{U_p(x)\} = \{W_p(x)\} = \{\phi_p(x)\}$ and equation (22), we obtain for the normalized radiation pattern

$$T(\theta) = \frac{2k}{(2\pi\alpha_1)^{1/2}} \cos \theta \sum_{n=1,3,\cdots}^{N} b_n \frac{\cos \left(k \frac{a}{2} \sin \theta \right)}{\left(\frac{n\pi}{a} \right) \left[1 - \left(\frac{2a}{n\lambda} \sin \theta \right)^2 \right]}, \quad (23)$$

where α_1 is the propagation constant of the incident wave.

IV. RESULTS

We now present numerical results obtained by the method described in Section III. The computations are actually performed with exp $-j\omega t$ time convention. Table I shows the type of convergence one may expect for the reflection coefficient R versus N, the number of modes used to approximate the aperture electric field. The parameters used in this calculation are $\epsilon = 6$, $\lambda/a = 1.5$, and $d/a = 0.544$. This represents one of the worst situations encountered. Nevertheless, we find the convergence is quite rapid.

The variation of the reflection coefficients versus the thickness of the dielectric plug is considered first. Figure 2 shows such a calculation for $\lambda/a = 1.5$ and $\epsilon = 6$. With this value of a/λ, only one mode can propagate in an unloaded waveguide. The dielectric constant is chosen so that the third order mode is propagating inside the dielectric. (The second order mode will also be propagating; but this mode cannot be excited because of the symmetry in the geometry.)

The reflection coefficient shows a smooth standing wave like variation versus d/a over the entire range of d considered except in the vicinities of $d/a \approx 0.54$ and $d/a \approx 1.31$ (where sharp spikes appear). Figure 3 shows the details of the reflection coefficient near these spikes.

The maxima (or minima) of the standing wavelike pattern are equally displaced at a distance given by π/α_1^D, where α_n^D is the propagation constant of the nth mode of a dielectric loaded waveguide. The separa-

TABLE I—CONVERGENCE OF R VERSUS N

N	$|R|$	Phase of R (degrees)
1	0.8031	−162.8
3	0.9213	−169.8
5	0.9306	−169.2
7	0.9348	−168.9
9	0.9372	−168.6
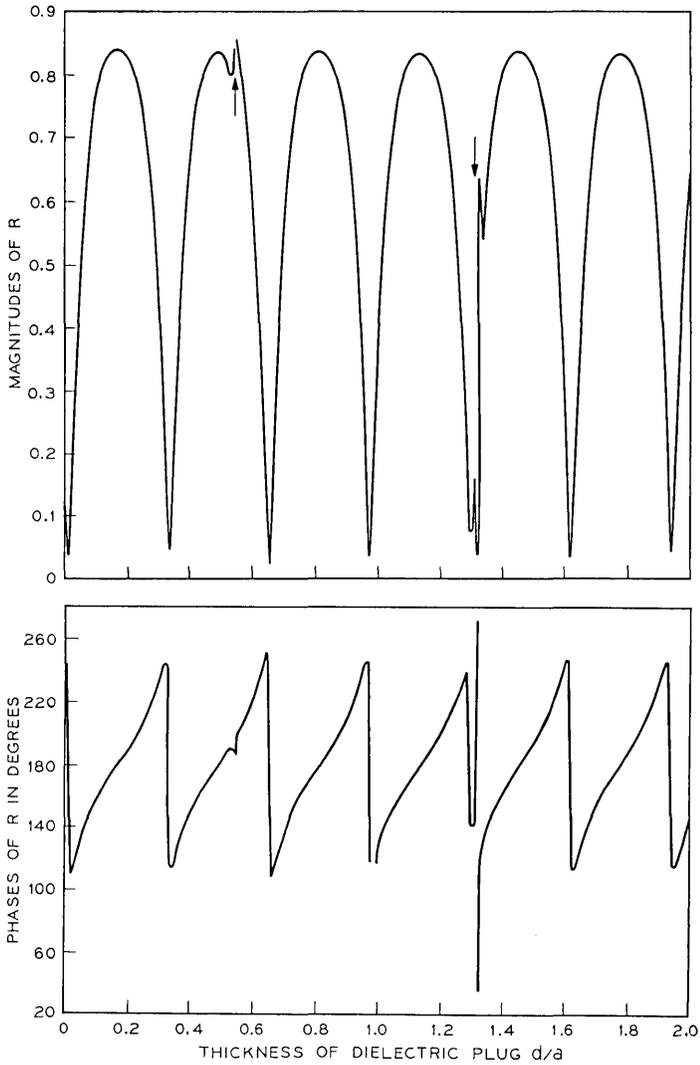
Fig. 2 — Reflection coefficient of a waveguide antenna with dielectric plug ($\epsilon = 6$ and $\lambda/a = 1.5$).

tion between the two spikes Δd is obtainable from the relation $\alpha_3^D(\Delta d) = \pi$. (Notice that the sharp spikes are frequently preceded by deep dips such that they may appear like close-by double spikes as displayed by the one at $d/a \approx 1.31$. See Fig. 3.) Figure 4 presents another calculation using a higher dielectric constant $\epsilon = 13$. Since the propagation
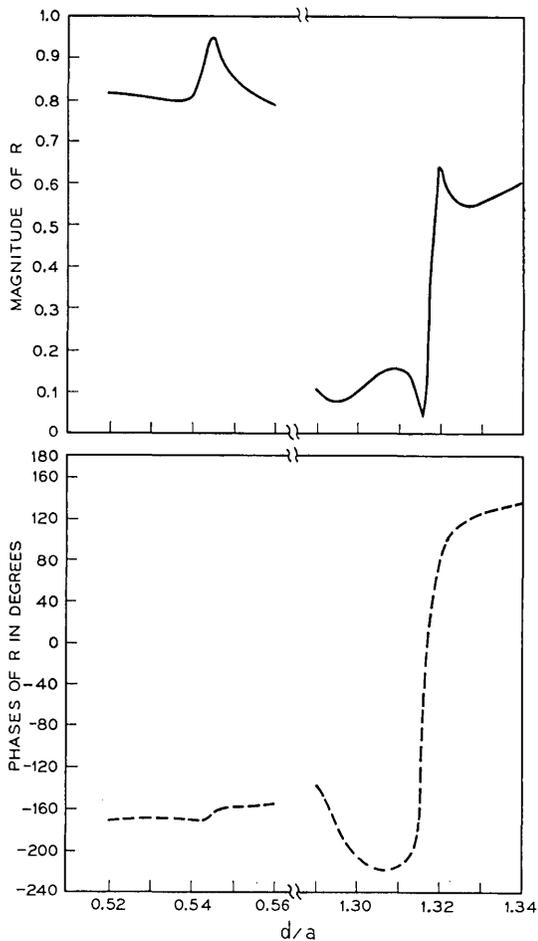
Fig. 3 — Details of R versus d/a for $\epsilon = 6$ and $\lambda/a = 1.5$.

constants α_1^D and α_3^D are larger when a higher dielectric constant is used, the maxima (or minima) and the spikes become more closely spaced. Otherwise, the relation stated above remains valid. This observation suggests that ordinarily the third order mode is only weakly excited so that the radiation impedance of the waveguide is determined primarily by the fundamental mode. Only when the dielectric plug has a certain thickness is the third order mode excited strongly enough to influence the reflection coefficient of the fundamental mode. Figure

5 shows the solutions for the third order modal coefficients versus d to demonstrate that indeed this is the case.

From the regularity of the intervals between the spikes at which the third order mode is excited sufficiently strongly to influence the radiation of the waveguide, it seems reasonable to assume that the dielectric plug forms a cavity for the third order mode. This cavity goes into resonance only at proper combinations of the wavelength and the thickness of the dielectric plug. To verify this conjecture we applied the transverse resonance technique at the waveguide aperture using the admittances pertinent to the third order mode. Let $\vec{Y}$ be the radiation admittance when a completely loaded waveguide is excited in the third order mode. The admittance looking toward the negative z direction, that is, into the waveguide is given by the appropriate modal admittance:

$$\bar{Y} = Y_3^D \frac{Y_3 + jY_3^D \tan \alpha_3^D d}{Y_3^D + jY_3 \tan \alpha_3^D d}.$$
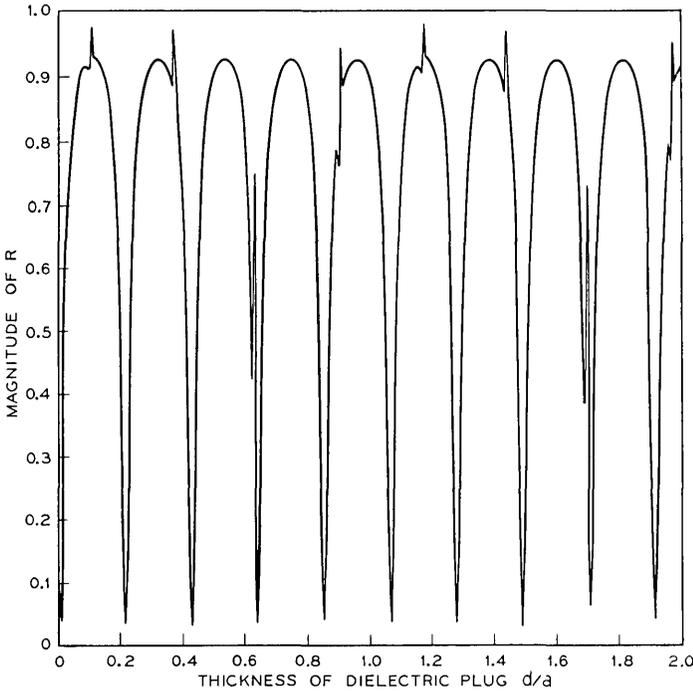
Fig. 4 — R versus d/a for $\epsilon = 13$, $\lambda/a = 1.5$.
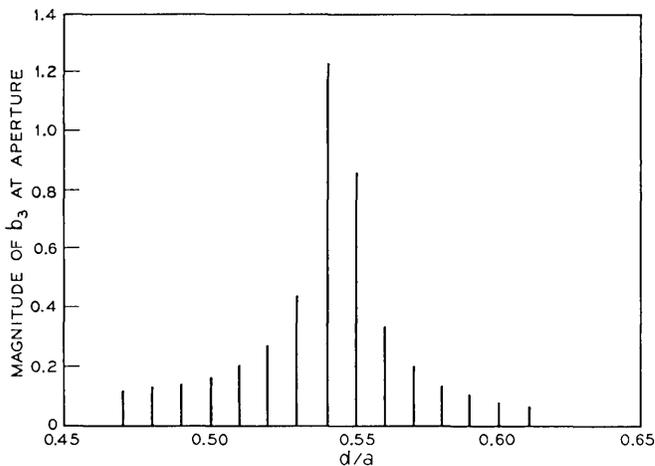
Fig. 5 — Magnitude of third model coefficients versus d/a for $\epsilon = 6$ and $\lambda/a = 1.5$.

The condition of resonance is given by*

$$\mathrm{Im}\,(\overleftarrow{Y} + \vec{Y}) = 0.$$

Figure 6 shows a calculation of the imaginary parts of $\overleftarrow{Y}$ and $\vec{Y}$ as functions of d. The graph clearly demonstrates that there are intersections occurring at the values for which resonance behavior is exhibited in the reflection coefficients.

We next consider the variation of the reflection coefficient when the frequency is varied. Figure 7 gives a calculation using $\epsilon = 6$ and $d/a = 0.55$. That there are two frequencies at which the reflection coefficient displays abrupt variations is quite evident. The details of one of the variations are illustrated in expanded scale in the inset. Examination of the admittances pertinent to the third order mode again shows that the transverse resonance condition is satisfied at both of these frequencies. Another salient feature shown in this calculation is that there are several frequencies at which the reflection coefficients are practically zero. Therefore, when the parameters are judiciously chosen, the use of a dielectric plug does not necessarily degrade the match characteristic of the antenna.

---

* Strictly speaking, because of the radiation from the waveguide aperture, the resonance condition should be $(\overleftarrow{Y} + \vec{Y}) = 0$. Since our interest is to obtain the condition for maximum excitation of the third order mode as d is varied, this should be a good approximation.
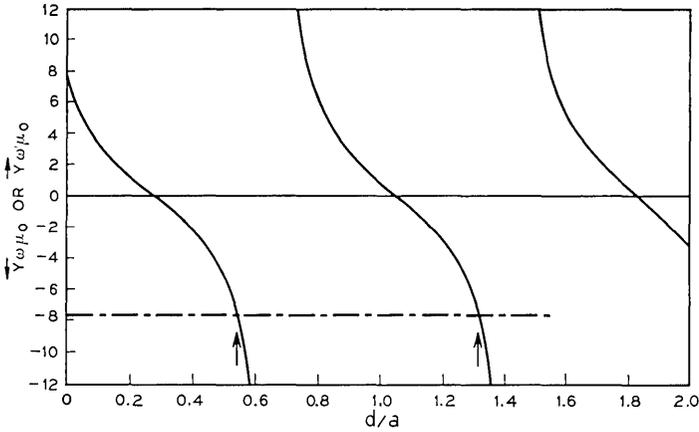
Fig. 6 — $\vec{Y}$ and $\overleftarrow{Y}$ of the third order mode for $\epsilon = 6$ and $\lambda/a = 1.5$ (———— $\overleftarrow{Y}$; —·—·— $-\mathrm{Im}\ \vec{Y}$).

The radiation patterns of the antenna have also been computed for the various values of parameters considered. The results in general display smooth variation versus the angle of observation θ. Only when the parameters are such that the resonating higher order mode is exceedingly strongly excited do dips appear in the radiation patterns. Figure 8 gives some typical results for smoothly varying patterns and Figure 9 illustrates the patterns with dips. Notice that the pattern dips are exhibited only over a very narrow range of the parameter
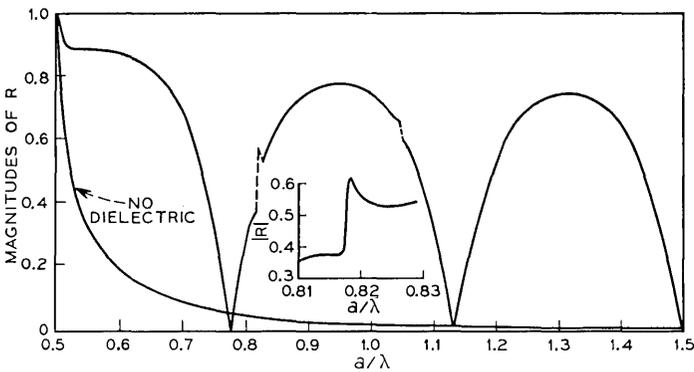
Fig. 7 — Variation of R with frequency for $\epsilon = 6$ and $d/a = 0.55$.
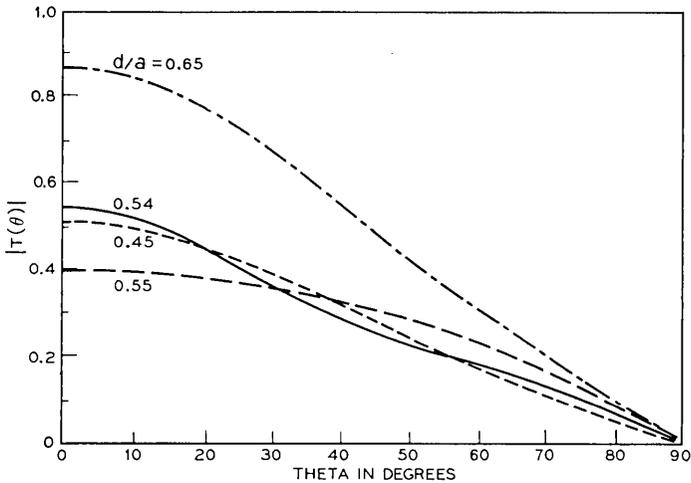
Fig. 8 — Normalized radiation patterns $T(\theta)$ of a waveguide with dielectric plug for $\epsilon = 6$ and $\lambda/a = 1.5$.

d/a. Moreover, the dips are rather broad and shallow because the aperture is small in wavelength, $\frac{1}{2} < a/\lambda < 1$.

Figure 9 also shows the patterns for the situation when the waveguide is *completely* loaded with a dielectric and is excited in the first or the third order mode. The aperture field in such situations consists primarily of the incident wave. We observe that a relatively small aperture with an aperture field distribution of the third order mode is capable of producing a dip in the radiation pattern. Now, when the third order mode is at resonance inside the dielectric plug so that it is strongly excited, the aperture field contains high content of both the incident dominant mode and the third order mode. The relative amplitudes and phases of these two modes determine the shape of the radiation pattern. The combination sometimes may be such as to generate a pattern which exhibits a considerably suppressed radiation in the broadside direction as shown in the curve for $d/a = 0.545$.

V. CONCLUSIONS AND DISCUSSIONS

The investigation of the effects of dielectric plugs on the radiation from a flush mounted waveguide has shown that dielectric plugs can function as a resonant cavity for the wave modes which are propagating inside the dielectric but evanescent in the unloaded waveguide region. Such wave modes have interesting effects on the radiation

impedances of the antenna. When one of these modes is at resonance, it is strongly excited by the incident wave; the presence of the resonance is manifested in the form of sharp spikes in the reflection coefficient.

Resonances have also been observed in the analysis of phased arrays using the present waveguide with dielectric plugs as the radiating elements. They appear in both infinite and finite arrays. The occurrence of these resonances may be identified by the conditions of total reflection of the incident power in infinite arrays[6] and rapid variation of the coupling coefficients in finite arrays.[7] Although there has been considerable discussion on array resonances in general, it appears that no consensus has been reached yet about the basic mechanism of this phenomenon. We hope that observation of resonances and our analysis of their causes may shed some light on this problem.

Another aspect which deserves some comment is the use of a single array element for the detection of potential difficulty due to resonances. This question is particularly important in array designs using antenna elements which are less susceptible to analysis. We realize that this is an ambitious question which cannot be answered completely without a more elaborate analysis. The calculation so far, however, has indicated that resonances observed in array configurations are often not exhibited by the radiation characteristics of a
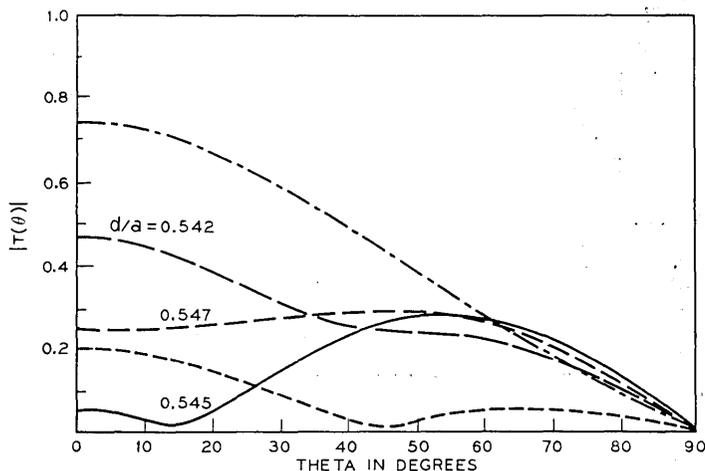
Fig. 9 — Pattern dips due to strong higher order mode excitation for $\epsilon = 6$ and $\lambda/a = 1.5$ (—·— 1st mode excitation; · · · · 3rd mode excitation).

single element. For example, in arrays of waveguides with dielectric plugs such as the one considered here,[6] resonances which are found to occur as a result of the interaction with the resonating second order mode are not displayed by a single element because this mode is usually not excited in the latter situation on account of geometric symmetry. When the dielectric constant is large enough to permit the third order mode to resonate, it is possible that the resonance conditions resulting from this mode may be uncovered. Even so, resonances which are caused by the second order mode are still undetectable. Moreover, there are other situations in which resonances do occur without the use of dielectrics such as planar arrays of rectangular and circular waveguides.[12,13] It therefore appears that it is not suitable to use a single element in the prediction of potential array resonances.

VI. ACKNOWLEDGMENT

REFERENCES

1. Noble, B., *Methods Based on the Wiener-Hopf Technique*, New York: Pergamon Press, 1958, Chapter III.
2. Marcuvitz, N., (Ed.) *Waveguide Handbook*, New York: McGraw-Hill, 1951, pp. 187–192.
3. Harrington, R. F., *Time Harmonics Electromagnetic Fields* New York: Mc-Graw-Hill, 1960, pp. 180–188.
4. Galindo, V., and Wu, C. P., "Numerical Solutions for an Infinite Phased Array of Rectangular Waveguides with Thick Walls," IEEE Trans. Antennas and Propagation, $AP-14$, No. 2 (March 1966), pp. 149–158.
5. Wu, C. P., "Integral Equation Solutions for the Radiation from a Waveguide Through a Dielectric Slab," presented at 1968 Fall URSI meeting in Boston, to be published in IEEE Trans. on Antennas and Propagation, $AP-17$, No. 6 (November 1969).
6. Wu, C. P., and Galindo, V., "Surface-Wave Effects on Phased Arrays of Rectangular Waveguides Loaded by Dielectric Plugs," IEEE Trans. Antennas and Propagation, $AP-16$, No. 3 (May 1968), pp. 358–360.
7. Wu, C. P., unpublished work.
8. Schelkunoff, S. A., *Electromagnetic Waves*, New York: Van Nostrand, 1943, pp. 158–159.
9. Harrington, R. F., *Field Computation by Moment Method*, New York: Macmillan, 1968, pp. 5–8.
10. Kantorovich, L. V., and Krylov, V. I., *Approximate Methods of Higher Analysis*, New York: Interscience, 1964.
11. Abramovitz, M., and Stegen, I. A., (Ed.) *Handbook of Mathematical Functions*, New York: Dover, 1965, pp. 369–370.
12. Farrel, G. F., Jr., and Kuhn, D. H., "Mutual Coupling Effects of Triangular-Grid Arrays by Modal Analysis," IEEE Trans. Antennas and Propagation, $AP-14$, No. 5 (September 1966), pp. 652–654.
13. Amitay, N., and Galindo, V., "The Analysis of Circular Waveguide Phased Arrays," B.S.T.J., *47*, No. 9 (November 1968), pp. 1903–1932.

Dielectric Loss in Integrated Microwave Circuits

By M. V. SCHNEIDER

Dielectric loss is important in integrated microwave and millimeter wave circuits which require small attenuation. Such circuits are usually built with microstrip or suspended microstrip transmission lines. This paper shows that the dielectric loss, the filling factor of the microstrip, and the stored field energy in the dielectric substrate can be computed from the partial derivative $\partial U/\partial \epsilon_r$ where U is the total electric field energy and ϵ_r the relative dielectric constant of the substrate. It also shows that the effective loss tangent is determined by the partial derivative $\partial \epsilon_{eff}/\partial \epsilon_r$ where ϵ_{eff} is the effective dielectric constant of the microstrip. Useful design formulas for computing the dielectric loss are given for the most important cases.

I. INTRODUCTION

The dielectric loss in microstrip or suspended microstrip transmission lines is an important parameter in the design of hybrid integrated circuits which require small attenuation. This loss can be calculated if one knows the loss tangent of the dielectric substrate and the electric field distribution inside the substrate. Electric field computations are usually complicated and not practical for design purposes. It is therefore important to find a simple and accurate method for calculating the dielectric loss from other well known properties of the microstrip transmission line.

The results of dielectric loss computations for microstrips, which have been made by other authors, are quoted in many recent papers on hybrid integrated circuit design.[1-6] It can be shown that these results are applicable only if the boundary between the dielectric substrate and air is parallel to an electric field line. This paper presents general design equations valid for all microstrip transmission lines provided that the propagating mode can be approximated by a TEM mode.

II. EFFECTIVE DIELECTRIC CONSTANT AND FILLING FACTOR OF MICROSTRIP LINES

The effective dielectric constant of a microstrip line partially filled with dielectric material is defined by

$$\epsilon_{eff} = \left(\frac{\lambda_o}{\lambda}\right)^2, \tag{1}$$

where λ_o is the vacuum wavelength and λ the wavelength of the propagating mode on the microstrip. If the propagating mode can be approximated by a TEM mode one can also define ϵ_{eff} by

$$\epsilon_{eff} = \frac{C}{C_o}, \tag{2}$$

where C is the capacitance per unit length with partial dielectric filling and C_o the capacitance per unit length without dielectric material.

The filling factor q of a microstrip is defined by

$$q = \frac{U_1}{U} \tag{3}$$

where U_1 is the electric field energy stored in the dielectric and U the total electric field energy of the microstrip. Notice that some authors do not use the same definition for q. Poole and Von Hippel use the ratio given by equation (3).[7,8] This definition is useful because it simplifies the loss calculation.

III. PARTIAL DERIVATIVES OF FIELD ENERGY AND EFFECTIVE DIELELCTRIC CONSTANT

If one computes the partial derivative of the total electric field energy U with respect to the relative dielectric constant ϵ_1 of the substrate, one obtains the basic result

$$\frac{\partial U}{\partial \epsilon_1} = \frac{U_1}{\epsilon_1}. \tag{4}$$

The Appendix gives the derivation of this equation. We assume that the conductor configuration remains the same and that the potential difference between the conductors is constant. From equations (2) and (4), and from $U = CV^2/2$ we obtain

$$\frac{\partial \epsilon_{eff}}{\partial \epsilon_1} = \frac{\epsilon_{eff}}{\epsilon_1} \frac{U_1}{U} \tag{5}$$

The filling factor q is now given by

$$q = \frac{\epsilon_1}{\epsilon_{eff}} \frac{\partial \epsilon_{eff}}{\partial \epsilon_1}, \tag{6}$$

and the effective loss tangent of the microstrip is

$$(\tan \delta)_{eff} = \frac{\epsilon_1}{\epsilon_{eff}} \frac{\partial \epsilon_{eff}}{\partial \epsilon_1} \tan \delta \tag{7}$$

with $\tan \delta$ being the loss tangent of the dielectric substrate. One can show that the effective loss tangent of microstrips with more than one single lossy substrate is given by

$$(\tan \delta)_{eff} = \frac{1}{\epsilon_{eff}} \sum_{n=1}^{N} \epsilon_n \frac{\partial \epsilon_{eff}}{\partial \epsilon_n} \tan \delta_n \tag{8}$$

where ϵ_n and $\tan \delta_n$ are the relative dielectric constants and loss tangents of each substrate respectively and N the total number of lossy dielectric materials in the microstrip.

IV. DIELECTRIC ATTENUATION AND UNLOADED Q

The unloaded dielectric quality factor Q_D of the microstrip is

$$Q_D = \frac{1}{(\tan \delta)_{eff}} = \frac{1}{q \tan \delta}, \tag{9}$$

and the dielectric attenuation in dB per unit length is

$$\alpha_D = \frac{20\pi}{\ln 10} \frac{q \tan \delta}{\lambda} = 27.3 \frac{(\tan \delta)_{eff}}{\lambda}, \tag{10}$$

with λ being the microstrip wavelength $\lambda = \lambda_o / (\epsilon_{eff})^{1/2}$.

The effective dielectric constant for the standard microstrip of Fig. 1a is known and can be approximated by[9]

$$\epsilon_{eff} = \frac{\epsilon_1 + 1}{2} + \frac{\epsilon_1 - 1}{2} \left(1 + 10 \frac{h}{w}\right)^{-\frac{1}{2}}. \tag{11}$$

By introducing $F(w, h) = (1 + 10 \ h/w)^{1/2}$ we obtain, from equation (6), the filling factor

$$q = \frac{1}{1 + \dfrac{F - 1}{\epsilon_1(F + 1)}}. \tag{12}$$

Figure 2 is a graph of the filling factor for the standard microstrip as a function of w/h with ϵ_1 as a parameter.
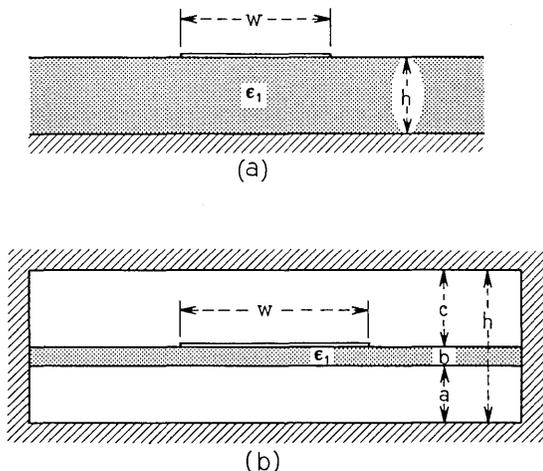
Fig. 1 — (a) Standard microstrip transmission line and (b) suspended microstrip transmission line.

Computation of q for the suspended microstrip shown in Fig. 1b is more difficult. An approximate value can be obtained if $w \gg h$, which means the fringe field contributions are small. The effective dielectric constant is

$$\epsilon_{\mathrm{eff}} = \frac{a + b}{a + b + c}\left(1 + \frac{c\epsilon_1}{a\epsilon_1 + b}\right) \qquad w \gg h, \qquad (13)$$

and the filling factor becomes

$$q = \frac{bc\epsilon_1}{(a\epsilon_1 + b)(a\epsilon_1 + b + c\epsilon_1)}. \qquad (14)$$

A different approach is necessary if the fringe field contribution cannot be neglected. Figure 3 shows a suspended microstrip which has been used in circuits built by Engelbrecht and Kurokawa, Saunders and Stark, and Tatsuguchi and Aslaksen.[10-12] The effective dielectric constant of the configuration with the dimensions given in Fig. 3 has been computed by Brenner.[13] It is possible to approximate the result by Brenner by the simple rational function

$$\epsilon_{\mathrm{eff}} = 1 + \frac{\epsilon_1 - 1}{0.38\epsilon_1 + 7.70}. \qquad (15)$$

From equation (6) we obtain

$$q = \frac{\epsilon_1}{6.38 + 1.63\epsilon_1 + 0.065\epsilon_1^2}. \tag{16}$$

Figure 3 is a graph of this filling factor as a function of the relative dielectric constant ϵ_1. The filling factor reaches a broad maximum for relative dielectric constants between 6 and 12. This maximum is obtained for structures with substantial fringe field contributions. If one neglects the fringe field the filling factor is substantially reduced and decreases if ϵ_r is increased.

V. DISCUSSION

There are several types of substrates which are useful for building integrated circuits. These substrates are

(i) borosilicate glasses and other commercial glasses with loss tangents of the order of 10^{-2} at microwave and millimeter wave frequencies,[14]

(ii) semiconductor substrates such as Si and GaAs with loss tangents determined by $\tan \delta = \sigma/\omega\epsilon_o\epsilon_1$ where σ is the substrate conductivity in mho per centimeter, ϵ_o the free space permittivity $\epsilon_o = 8.85 \cdot 10^{-14}$ F per cm, and ϵ_1 the relative dielectric constant of the semi-conductor,

(iii) ceramics such as alumina, beryllia, and rutile with loss tangents of about 10^{-4} at microwave and millimeter wave frequencies, and

(iv) fused silica with $\tan \delta = 10^{-4}$ in the same frequency range.
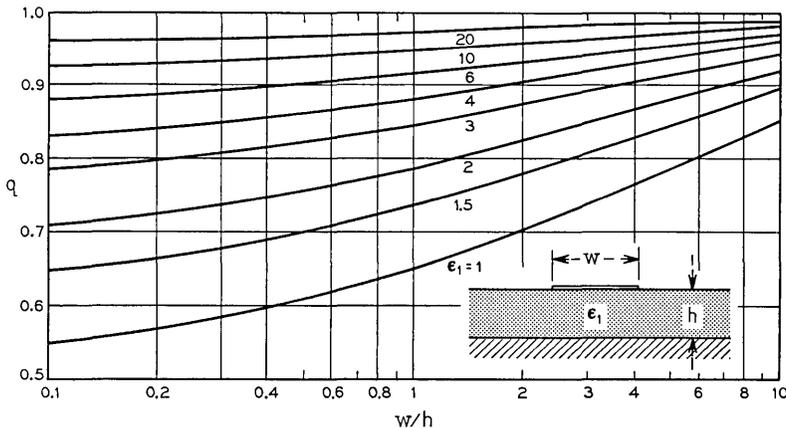
Fig. 2 — Filling factor q for standard microstrip transmission line as a function of the ratio w/h with relative dielectric constant ϵ_1 as parameter.
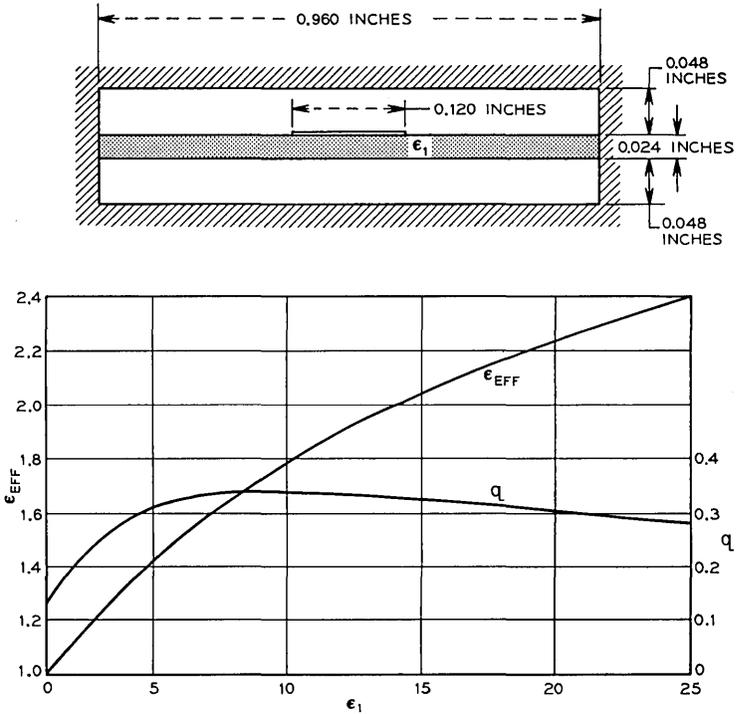
Fig. 3 — Effective dielectric constant ϵ_{eff} and filling factor q for suspended microstrip transmission line for $w = h = 0.120$ inch, $a = c = 0.048$ inch and $b = 0.024$ inch.

For glasses one can, therefore, expect an unloaded dielectric quality factor of $Q_D = 100/q$; with high quality ceramics and fused silica one obtains $Q_D = 10000/q$. However, loss tangents of many substrates above 30 GHz are presently not available.

The unloaded Q resulting from conductor loss alone is typically $Q_c = 100$ to 1000 for completely shielded microstrips at microwave and millimeter wave frequencies. The total unloaded Q is $Q_T = Q_D Q_c/(Q_D + Q_c)$. One concludes that the conductor loss is predominant for circuits built with high quality ceramics and quartz. For microstrips built on glass substrates and some semiconductor substrates, the filling factor is important for computing the total loss of the microstrip.

Partial Derivative of Field Energy U

The total electric field energy U stored in a microstrip is given by the volume integral

$$U = \int_V \frac{D^2}{2\epsilon} dV, \tag{17}$$

where D is the displacement, $D = \epsilon E$, and $\epsilon = \epsilon_o \cdot \epsilon_1(x, y, z)$ is an isotropic dielectric constant. We make a small perturbation subject to boundary conditions which follow equation (20).

$$\delta U = \int_V \frac{D \, \delta D}{\epsilon} dV - \int_V \frac{D^2 \, \delta\epsilon}{2\epsilon^2} dV. \tag{18}$$

By using $E = -\text{grad } \varphi$ and div $D = \rho$ one obtains, from div $(\varphi \, \delta D) = -E \, \delta D + \varphi$ div δD,

$$\delta U = \int_V \varphi \, \delta\rho \, dV - \int_V \text{div} \, (\varphi \, \delta D) \, dV - \frac{1}{2} \int_V E^2 \, \delta\epsilon \, dV, \tag{19}$$

and from the theorem by Gauss

$$\int_V \text{div} \, (\varphi \, \delta D) \, dV = \sum_{K=1}^N \varphi_K \int_{F_K} \delta D_n \, dF = \sum_{K=1}^N \varphi_K \, \delta Q_K, \tag{20}$$

where the surface integral is carried out over all conductor surfaces $K = 1, 2, \ldots, N$. We are interested in a perturbation subject to the following boundary conditions:

(i) The space charge is zero, $\delta\rho \equiv 0$.

(ii) The charge on each conductor remains constant, $\delta Q_K = 0$.

(iii) $\delta\epsilon$ is constant in the dielectric substrate, and $\delta\epsilon = 0$ outside the substrate.

If the dielectric constant of the substrate is ϵ from equation (18) we obtain

$$\delta U = -\delta\epsilon \, \frac{\int_{V_1} \frac{\epsilon}{2} E^2 \, dV}{\epsilon}. \tag{21}$$

The volume integral is the electric field energy U_1 stored in the dielectric substrate. For two conductors and $\Delta\varphi = \varphi_1 - \varphi_2 = $ constant one has $\delta U = +\delta\epsilon \cdot U_1/\epsilon$ and consequently

$$\frac{\partial U}{\partial \epsilon_1} = \frac{U_1}{\epsilon_1}. \tag{22}$$

REFERENCES

1. Welch, J. D., and Pratt, H. J., "Losses in Microstrip Transmission Systems for Integrated Microwave Circuits," Northeast Elec. Res. and Eng. Meeting Record, 8, Boston, Massachusetts, November 1966, pp. 100–101.
2. Pucel, R. A., Massé, D. J., and Hartwig, C. P., "Losses in Microstrip," IEEE Trans. Microwave Theory and Techniques, MTT-16, No. 6 (June 1968), pp. 342–350.
3. Presser, A., "RF Properties of Microstrip Line," Microwaves, 7, No. 3 (March 1968), pp. 53–55.
4. Hartwig, C. P., Lepie, M. P., Massé, D., Paladino, A. E., and Pucel, R. A., "Microstrip Technology," Proc. of the Nat. Elec. Conf., 24, (December 9–11, 1968), pp. 314–317.
5. Schilling, W., "The Real World of Micromin Substrates—Part 1," Microwaves, 7, No. 12 (December 1968), pp. 52–56.
6. Emery, E. F., and Noel, P. L., "Recent Experimental Work on Silicon Microstrip Microwave Transmission Lines," IEEE J. of Solid State Circuits, SC-3, No. 2 (June 1968), pp. 145–146.
7. Poole, C. P., Electron Spin Resonance, New York: Interscience Publishers, 1967, pp. 291–307.
8. Von Hippel, A. R., unpublished work.
9. Schneider, M. V., "Microstrip Lines for Microwave Integrated Circuits," B.S.T.J., 48, No. 5 (May–June 1969), pp. 1421–1444.
10. Engelbrecht, R. S., and Kurokawa, K., "A Wideband Low Noise L-Band Balanced Transistor Amplifier," Proc. IEEE, 53, No. 3 (March 1965), pp. 237–247.
11. Saunders, T. E., and Stark, P. D., "An Integrated 4-GHz Balanced Transistor Amplifier," IEEE J. Solid-State Circuits, SC-2, No. 1 (March 1967), pp. 4–10.
12. Tatsuguchi, I., and Aslaksen, E. W., "Intergrated 4-GHz Balanced Mixer Assembly," IEEE J. Solid-State Circuits, SC-3, No. 1 (March 1968), pp. 21–26.
13. Brenner, H. E., "Use a Computer to Design Suspended-Substrate Integrated Circuits," Microwaves, 7, No. 9 (September 1968), pp. 38–45.
14. Heinrich, W., "Die Komplexe Dielektrizitäts-Konstante einiger Gläser und Keramiken im Frequenzbereich zwischen 8.5 and 34.4 GHz," Zeitschrift für angewandte Physik, 22, No. 2 (February 1967), pp. 115–121.

Interchannel Interference Considerations in Angle-Modulated Systems

By V. K. PRABHU and L. H. ENLOE

This paper considers the deterioration in performance of angle-modulated systems resulting from interchannel interference. We show that with band-limited white gaussian noise modulation (simulating modulation by a frequency division multiplex signal), we can derive an explicit expression for the spectral density of the baseband interchannel interference when two or more PM waves interfere with each other.

We show that, if the interference is co-channel, maximum interference occurs at the lowest baseband frequency present in the system and we can derive upper and lower bounds to this minimum baseband signal-to-interference ratio. For high enough modulation index, we show that this minimum signal-to-interference ratio is proportional to the cube of the modulation index and that phase modulation can be used with advantage in interference limited systems. We do not consider the effects of linear filters on angle-modulated systems, but give some results about the effect of adjacent channel interference when the interference is in the passband of the receiver.

I. INTRODUCTION

The properties of frequency and phase modulation with respect to exchanging bandwidth for signal-to-noise ratio are well known,[1,2] but the type of noise considered is almost always limited to be random gaussian noise. In the design of any system, where the noise is likely to be interference limited, it is necessary to consider other kinds of disturbances such as co-channel and adjacent channel interference corrupting the desired received signal.

Consider the following situation. In the frequency bands above 10 GHz where the signal attenuation resulting from rain could be very severe, close spacings of the repeaters are almost always mandatory for reliable communication from point-to-point and for all periods of time.[3,4] If low noise receivers are used in the system, it is possible

that the total interference power received by the system may be very much larger than the noise power in the system. For all practical purposes, the performance of such a system is determined by the interchannel interference.[3,4] It is therefore desirable to evaluate the effect of co-channel and adjacent channel interference on the performance of any modulation system like FM or PM (or PCM) so that its advantages in combating interference can be determined, and any system parameters (such as rms phase deviation, channel separation, and so on) can be properly chosen to keep the baseband interference below a certain desired level. (It is possible to reduce adjacent channel interference by using suitable receiving filters, but co-channel interference occupies the same band as the signal.)

The problem of interference in angle-modulated systems has been considered by many authors.[5-12] In the analysis, most of these authors have given an approximate expression (the first term in the power series expansion) for the baseband interchannel interference, and have shown that it can be expressed as the convolution of the spectral densities of the angle-modulated waves. The accuracy in this approximation has not been determined previously. Also, in the calculation of interchannel interference in high index FM and PM systems, most of these authors use the quasistatic approximation, the accuracy of which is unknown.

We first consider a general method of evaluating the baseband interchannel interference when two angle-modulated waves interfere with each other. We assume that an ideal angle (frequency or phase) demodulator is used in the system. (An ideal angle demodulator does not respond to any variations in the amplitude of the wave. This can be achieved in practice by using an ideal limiter at the front end of the receiver. If $A(t)e^{i\Phi(t)}$ is the input to an ideal limiter, its output is given by $A_o e^{i\Phi(t)}$ where A_o is a constant.)

We obtain a general expression for the baseband interference when the modulating wave is gaussian. This expression can be utilized even when the baseband signal is passed through a linear network (such as a pre-emphasis—de-emphasis network).

We are specifically interested in calculating the baseband interchannel interference between two or more waves phase modulated (without pre-emphasis) by band-limited white gaussian random processes. It has been found in practice that such a random gaussian noise of appropriate bandwidth and power spectral density adequately simulates (for some purposes) a variety of signals such as a frequency division multiplex (FDM) signal, a composite speech

signal, and so on.[13] Since the determination of the power spectrum is fundamental to the evaluation of baseband interference, first we review briefly the methods of obtaining this spectrum for a wave phase modulated by band-limited white gaussian noise.

In the case of band-limited white gaussian noise modulation, if the bandwidths of the modulating waveforms for the desired and interfering carriers are the same, we show that the determination of baseband interference power is relatively simple, and requires only the computation of the spectral density of a phase-modulated carrier for a variety of values of rms phase deviation. For small values of interference and for band-limited white gaussian noise modulation, we also show that the first term in the series gives most of the contribution to the baseband interference, and that this first term can be used as a good approximation.

For a co-channel interferer, we show that maximum interference occurs at the lowest baseband frequency present in the system (we assume that this lowest frequency is $f = 0$)* and that we can derive upper and lower bounds to this minimum signal-to-interference ratio. For sufficiently high modulation index, we show that these bounds are proportional to the cube of the modulation index, and that phase modulation can be used to advantage in combating interference.[14]

We show that maximum interference with an adjacent channel interferer occurs at the highest baseband frequency present in the system if the carrier frequency separation f_d between the two channels is relatively large compared with the baseband bandwidth W. For a set of values of f_d/W and for different modulation indexes of the two channels, we compute this minimum signal-to-interference ratio and give the results in graphic form.

We then consider the case in which more than one interferer may corrupt the desired received carrier and show that we can derive an expression for the spectral density of the resulting baseband interference. This expression is in the form of an infinite series and for its evaluation, in the case of band-limited white gaussian noise modulation and equal modulation bandwidths, it is only necessary to be able to compute the spectral density of a sinusoidal carrier phase modulated by gaussian noise. In case all these interferers are co-channel and all of them have the same (high) modulation index Φ, we show that we can derive upper and lower bounds to the minimum baseband signal-to-interference ratio.

---

* We do not imply that maximum baseband interchannel interference always occurs at $f = 0$ for any general system angle modulated by gaussian noise.

II. INTERFERENCE BETWEEN TWO ANGLE-MODULATED WAVES

We first assume that there is only one interfering wave corrupting the desired received signal, and that both of them are angle modulated by two independent gaussian random processes. Let the desired angle-modulated wave be given by

$$s(t) = A \cos [\omega_o t + p(t) * \varphi(t)]$$
$$= \text{Re } A \exp \{j[\omega_o t + p(t) * \varphi(t)]\},$$

(1)

where A is the amplitude of the wave, $f_o = \omega_o/2\pi$ its carrier frequency, $p(t)$ the impulse response of the pre-emphasis network, and $\varphi(t)$ is a stationary gaussian random process with mean zero, and covariance function $R_\varphi(\tau)$. (We only assume that $p(t)$ is the impulse response of a linear network through which $\varphi(t)$ may be passed. Only for convenience, we refer to it as the impulse response of the pre-emphasis network.) The notation $A(x)*B(x)$ represents the convolution of function $A(x)$ with $B(x)$.

Let the interfering wave $i(t)$ be given by

$$i(t) = R_i A \cos [\omega_i t + p_i(t) * \varphi_i(t) + \mu_i]$$
$$= \text{Re } AR_i \exp \{j[\omega_i t + p_i(t) * \varphi_i(t) + \mu_i]\},$$

(2)

where AR_i is its amplitude (R_i is the relative amplitude of the interfering wave with respect to the desired wave), ω_i is its angular frequency, $p_i(t)$ is the impulse response of its pre-emphasis network, and $\varphi_i(t)$ is a stationary gaussian random process with mean zero and covariance function $R_{\varphi_i}(\tau)$.

Since $s(t)$ and $i(t)$ usually originate from two different sources, it seems reasonable to assume that μ_i is a uniformly distributed random variable with probability density $\pi_{\mu_i}(\mu)$ where

$$\pi_{\mu_i}(\mu) = \begin{cases} \dfrac{1}{2\pi}, & 0 \leq \mu < 2\pi \\ 0, & \text{otherwise.} \end{cases}$$

(3)

Further, we assume that $\varphi(t)$ and $\varphi_i(t)$ are independent of each other and independent of μ_i. (Reference 13 treats of the case in which μ_i is a deterministic constant, and $\varphi(t)$ and $\varphi_i(t)$ are not independent of each other.)

If we assume that $s(t)$ and $i(t)$ are both in the passband of the receiver used in the system, the total signal $r(t)$ incident at the re-

ceiver is given by[†]

$$r(t) = \text{Re } A(\exp\{j[\omega_o t + p(t) * \varphi(t)]\}$$

$$+ R_i \exp\{j[\omega_i t + p_i(t) * \varphi_i(t) + \mu_i]\})$$

$$= \text{Re } A(1 + R_i \exp\{j[(\omega_i - \omega_o)t + p_i(t) * \varphi_i(t) - p(t) * \varphi(t) + \mu_i]\})$$

$$\cdot \exp\{j[\omega_o t + p(t) * \varphi(t)]\}$$

$$= \text{Re } Aa(t)e^{j\lambda(t)} \exp\{j[\omega_o t + p(t) * \varphi(t)]\}$$

$$= \text{Re } Aa(t) \exp\{j[\omega_o t + p(t) * \varphi(t) + \lambda(t)]\}, \tag{4}$$

where

$$a(t)e^{j\lambda(t)} = 1 + R_i$$

$$\cdot \exp\{j[(\omega_i - \omega_o)t + p_i(t) * \varphi_i(t) - p(t) * \varphi(t) + \mu_i]\}. \tag{5}$$

Notice from equation (4) that the (excess) phase angle $\eta(t)$, as detected by an ideal angle demodulator, is given by

$$\eta(t) = \varphi(t) + \lambda(t). \tag{6}$$

(The gain—or proportionality factor—of the phase demodulator has been assumed to be unity.) Therefore, the spectral density of $\eta(t)$ can be written as

$$S_\eta(f) = \int_{-\infty}^{\infty} R_\eta(\tau)e^{-j2\pi f\tau} d\tau, \tag{7}$$

where $R\eta(\tau)$ is the covariance function of $\eta(t)$, and

$$R_\eta(\tau) = \langle \eta(t)\eta(t + \tau) \rangle. \tag{8}$$

(The notation $\langle x \rangle$ represents the ensemble average of random variable x.) If there is no interference, and if $q(t)$ is the impulse response of the de-emphasis network used in the system, the detected phase angle $\Omega(t)$ can be written as

$$[\Omega(t)]_{R_i=o} = q(t) * p(t) * \varphi(t). \tag{9}$$

If $R_i \neq 0$,

$$\Omega(t) = q(t) * p(t) * \varphi(t) + q(t) * \lambda(t). \tag{10}$$

Now if we assume that the de-emphasis network is the inverse of

the pre-emphasis network, we have

$$q(t) * p(t) = \delta(t), \tag{11}$$

and

$$\Omega(t) = \varphi(t) + q(t) * \lambda(t), \tag{12}$$

where $\delta(t)$ is the Dirac delta function.

From equation (5), we have

$$\lambda(t) = \text{Im} \ln (1 + R_i \exp \{j[\omega_d t + p_i(t) * \varphi_i(t)$$
$$- p(t) * \varphi(t) + \mu_i]\}), \tag{13}$$

where

$$\omega_d = \omega_i - \omega_o . \tag{14}$$

Notice that

$$\ln (1 + z) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{z^m}{m} , \qquad |z| < 1, \tag{15}$$

where z is any complex number.

Therefore, for $R_i < 1$, we have[†]

$$\lambda(t) = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}$$
$$\cdot R_i^m \left[\frac{\exp \{jm[\omega_d t + p_i(t) * \varphi_i(t) - p(t) * \varphi(t) + \mu_i]\}}{2j} \right.$$
$$\left. - \frac{\exp \{-jm[\omega_d t + p_i(t) * \varphi_i(t) - p(t) * \varphi(t) + \mu_i]\}}{2j} \right]$$
$$= \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} R_i^m \sin \{m[\omega_d t + p_i(t) * \varphi_i(t) - p(t) * \varphi(t) + \mu_i]\}. \tag{16}$$

Since $\varphi(t)$, $\varphi_i(t)$, and μ_i are statistically independent random variables and since $\langle \exp (jk\mu_i) \rangle = 0$ with $k \neq 0$, we can show from equations (6), (8), (13), and (16) that

$$R_\eta(\tau) = R_p(\tau) * R_\varphi(\tau) + \sum_{m=1}^{\infty} \frac{R_i^{2m}}{2m^2} \cos m\omega_d \tau$$
$$\cdot \exp (-m^2 \{[R_{\varphi p}(0) - R_{\varphi p}(\tau)] + [R_{\varphi_i p_i}(0) - R_{\varphi_i p_i}(\tau)]\}), \tag{17}$$

[†] For $R_i < 1$, notice that $a(t) > 0$.

where[‡]

$$R_p(\tau) = \int_{-\infty}^{\infty} p(t)p(t + \tau) \, dt, \tag{18}$$

$$R_{p_i}(\tau) = \int_{-\infty}^{\infty} p_i(t)p_i(t + \tau) \, dt, \tag{19}$$

$$R_{\varphi p}(\tau) = R_p(\tau) * R_\varphi(\tau), \tag{20}$$

and

$$R_{\varphi_i p_i}(\tau) = R_{p_i}(\tau) * R_{\varphi_i}(\tau). \tag{21}$$

Therefore, the spectral density of the output is given by

$$S_\Omega(f) = S_\varphi(f) + \frac{1}{|H_p(f)|^2} \sum_{m=1}^{\infty} \frac{R_i^{2m}}{4m^2} [T_m(f - mf_d) + T_m(f + mf_d)], \tag{22}$$

where $H_p(f)$ is the Fourier transform of $p(t)$, and

$$T_m(f) = \int_{-\infty}^{\infty} \exp\left(-m^2\{[R_{\varphi p}(0) - R_{\varphi p}(\tau)]\right.$$
$$\left. + [R_{\varphi_i p_i}(0) - R_{\varphi_i p_i}(\tau)]\}\right)e^{-j2\pi f \tau} \, d\tau. \tag{23}$$

From equation (23), we can show that

$$T_m(f) = U_m(f) * V_m(f) \tag{24}$$

where[¶]

$$U_m(f) = \int_{-\infty}^{\infty} \exp\left\{-m^2[R_{\varphi p}(0) - R_{\varphi p}(\tau)]\right\}e^{-j2\pi f \tau} \, d\tau, \tag{25}$$

and

$$V_m(f) = \int_{-\infty}^{\infty} \exp\left\{-m^2[R_{\varphi_i p_i}(0) - R_{\varphi_i p_i}(\tau)]\right\}e^{-j2\pi f \tau} \, d\tau. \tag{26}$$

Equation (22) gives a general expression for the baseband interchannel interference when two angle-modulated waves interfere with each other. To calculate this interchannel interference, equations (22) through (26) show that it is essential to determine the RF spectral density of a wave angle modulated by gaussian noise. Methods of

---

[‡] Since $\varphi(t)$ and $\varphi_i(t)$ are assumed to be gaussian, $p(t) * \varphi(t)$, and $p_i(t) * \varphi_i(t)$ are also gaussian.[2,15] Notice also that the Fourier transform of $R_p(\tau)$ is equal to $|H_p(f)|^2$, if $H_p(f)$ is the Fourier transform of $p(t)$.

[¶] Notice that $U_m(f)$ and $V_m(f)$ are the RF spectral densities of waves angle modulated by gaussian noise.

calculating this spectrum for low and medium index modulation are generally available, and the quasistatic approximation has been used for high index modulation.[2,13-16] Since the accuracy in the quasistatic approximation cannot often be determined, some rigorous methods of evaluating this spectrum for high index modulation have recently been developed.[2,16]

III. SPECTRAL DENSITY OF A PM WAVE

In this paper, we are specifically interested in determining the interchannel interference between two or more waves phase modulated by band-limited white gaussian random processes. Hence, we now review briefly the methods of obtaining the RF spectrum of such a wave. A sinusoidal wave of constant amplitude A phase modulated by a signal $n(t)$ can be written as

$$w(t) = A \cos [\omega_0 t + n(t) + \theta], \qquad (27)$$

$$= \mathrm{Re}\ A \exp \{j[\omega_0 t + n(t) + \theta]\}, \qquad (28)$$

where $f_o = \omega_o/2\pi$ is the carrier frequency of the wave, and θ is a random variable with probability density function

$$\pi_\theta(\theta) = \begin{cases} 1/2\pi, & 0 \leqq \theta < 2\pi \\ 0, & \text{otherwise.} \end{cases} \qquad (29)$$

If the modulating waveform is band-limited and white, its spectrum $S_n(f)$ is given by (see Fig. 1)

$$S_n(f) = \begin{cases} \Phi^2/2W, & |f| < W, \\ 0, & \text{otherwise.} \end{cases} \qquad (30)$$
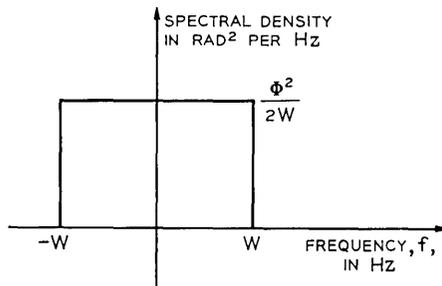
Fig. 1 — Spectral density of modulating wave.

Notice that Ref. 16 treats, in detail, the methods of obtaining the spectral characteristics of a sinusoidal carrier phase modulated by such a signal. From equation (30) we can show that (see Fig. 2)

$$R_n(\tau) = \Phi^2 \frac{\sin 2\pi W \tau}{2\pi W \tau}. \tag{31}$$

For $\Phi^2 \gg 1$ and for low frequencies, the quasistatic approximation yields[2,15]

$$S_V(f) \approx \exp\left(-\Phi^2\right) \delta(f) + \frac{1}{\Phi W}\left(\frac{3}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{3}{2}\frac{1}{\Phi^2}\left(\frac{f}{W}\right)^2\right]. \tag{32}$$

One can show that the approximation given by equation (32) is only good at low frequencies and that it is too small for large f.[16]

For large modulation indexes ($\Phi > 1.7432$ rad) and for all frequencies, we can show that[16]

$$S_V(f) = \exp\left(-\Phi^2\right)\left\{\delta(f) + \frac{\Phi^2}{2W}\left[u_{-1}(f + W) - u_{-1}(f - W)\right]\right\}$$

$$+ \frac{1}{2\pi W} \exp\left[-2\Phi^2\left(\cosh^2\frac{y_s}{2} - \frac{\sinh y_s}{y_s}\right)\right]\mu, \tag{33}$$

where

$$u_{-1}(x) = \begin{cases} 1, & x > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{34}$$

$$\left(\frac{2\pi}{\Phi^2 A_2}\right)^{\frac{1}{2}}(1 - C) < \mu < \left(\frac{2\pi}{\Phi^2 A_2}\right)^{\frac{1}{2}}(1 + D), \tag{35}$$

$$\frac{\cosh y_s}{y_s} - \frac{\sinh y_s}{y_s^2} = \frac{f}{\Phi^2 W}, \tag{36}$$

and

$$A_2 = \frac{\sinh y_s}{y_s} - \frac{2}{y_s}\frac{f}{\Phi^2 W}. \tag{37}$$

We can also show that C and D, appearing in equation (35), are less than 8 per cent for $\Phi > (10)^{\frac{1}{2}}$ rad. Further, for all f, one can show that[16]

$$C < 2\% \quad \text{for} \quad \Phi > 5\,\text{rad}, \tag{38}$$

and

$$D < 2\% \quad \text{for} \quad \Phi > 5\,\text{rad}. \tag{39}$$
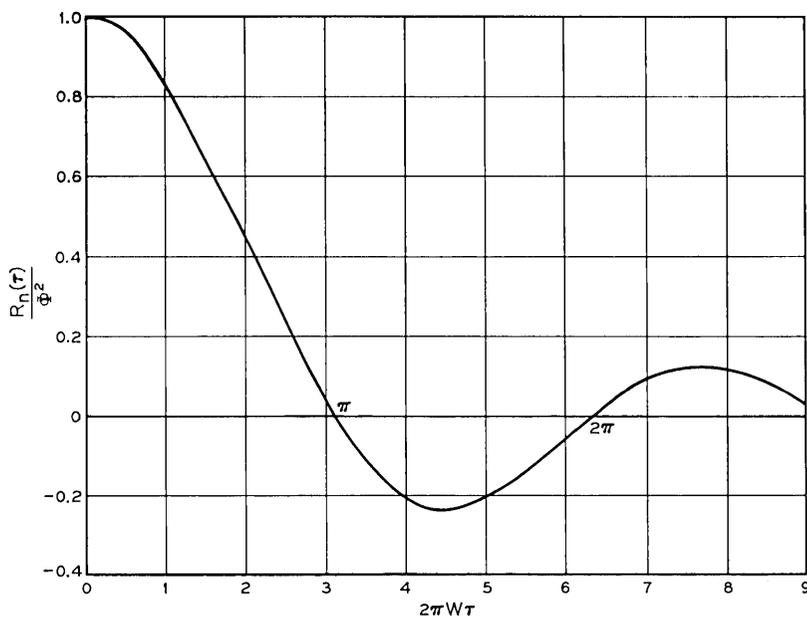
Fig. 2 — Covariance function $R_n(\tau)$. Since $R_n(\tau)$ is an even function of τ, we only show $R_n(\tau)$ for $\tau \geqq 0$.

Hence, we can say that

$$\mu \approx \left(\frac{2\pi}{\Phi^2 A_2}\right)^{\frac{1}{2}}, \tag{40}$$

and that the fractional error in this approximation is very much less than unity (less than 2 per cent, $\Phi > 5$ rad).

For $f = 0$, from equations (33) through (37) we can show that

$$0.92 \left(\frac{3}{2\pi}\right)^{\frac{1}{2}} \frac{1}{\Phi W} < S_V(f) - \exp\left(-\Phi^2\right) \delta(f) < 1.08 \left(\frac{3}{2\pi}\right)^{\frac{1}{2}} \frac{1}{\Phi W},$$

$$\Phi > (10)^{\frac{1}{2}} \text{ rad.} \tag{41}$$

For any f and Φ, the determination of the spectral density $S_V(f)$ from equations (33) through (40) is rather simple. For any given f, Φ^2, and W, we calculate y_s from equation (36), and A_2 from equation (37). The spectral density $S_V(f)$ is then calculated from equations (33) and (40).

IV. INTERFERENCE BETWEEN TWO PM WAVES

We now assume that $\varphi(t)$ and $\varphi_i(t)$ in Section II are band-limited white gaussian random processes with the same bandwidth W and rms phase deviations Φ and Φ_i. We also assume that $p(t) = p_i(t) = \delta(t)$, or that no pre-emphasis—de-emphasis networks are used in the system. Therefore, we have

$$R_\varphi(\tau) = \Phi^2 \frac{\sin 2\pi W \tau}{2\pi W \tau}, \tag{42}$$

and

$$R_{\varphi i}(\tau) = \Phi_i^2 \frac{\sin 2\pi W \tau}{2\pi W \tau}. \tag{43}$$

From equations (22), (23), (42), and (43) we can write†

$$S_\Omega(f) = S_\varphi(f) + \sum_{m=1}^{\infty} \frac{R_i^{2m}}{m^2} G_m(f), \tag{44}$$

where

$$G_m(f) = \tfrac{1}{4}[H_m(f - mf_d) + H_m(f + mf_d)], \tag{45}$$

and

$$H_m(f) = \int_{-\infty}^{\infty} \exp\left[-m^2(\Phi^2 + \Phi_i^2)\left(1 - \frac{\sin 2\pi W \tau}{2\pi W \tau}\right)\right]e^{-i2\pi f\tau} \, d\tau. \tag{46}$$

Notice that $G_m(f)$ is the spectral density of a sinusoidal carrier (at carrier frequency mf_d, and having unit amplitude) phase modulated by a band-limited white gaussian random process having mean square phase deviation $m^2(\Phi^2 + \Phi^2)$. Section III gives methods of obtaining this spectrum for all values of f; hence, $S_\Omega(f)$ can easily be calculated. In order to evaluate $S_\Omega(f)$ from equation (44), we must be able to determine the spectral density of a carrier phase modulated by gaussian noise for any arbitrary modulation index. In the case of band-limited white gaussian noise modulation the technique presented in Ref. 16 is very convenient to calculate this spectrum. The series method of determining this spectral density can become rather tedious when Φ or Φ_i is large.

When there is no interference, the signal as detected by an ideal phase demodulator is given by $\varphi(t)$, and its spectral density by $S_\varphi(f)$. Therefore, from equation (44), the spectral density $S_I(f)$ of the base-

---

† Notice that in this case $\Omega(t) = \eta(t)$, since $p(t) = p_i(t) = \delta(t)$.

band interchannel interference can be written as

$$S_I(f) = S_{\Omega}(f) - S_{\varphi}(f), \tag{47}$$

or

$$S_I(f) = \sum_{m=1}^{\infty} \frac{R_i^{2m}}{m^2} G_m(f). \tag{48}$$

Figure 3 is a graph of $S_I(f)$ for $f_d/W = 0, 1$, and 5; $\Phi = 3$ rad, and $\Phi_i = 2$ rad. Notice that, for $f_d/W = 1$, $S_I(f)$ is maximum at $f = 0$ or that maximum interchannel interference occurs at the lowest baseband frequency present in the system.

In practice the quantity of interest is usually the ratio of the average signal power to average interchannel interference power. In this case this signal-to-interference ratio $\sigma(f)$ can be written as

$$\sigma(f) = \frac{S_{\varphi}(f) \, \Delta f}{S_I(f) \, \Delta f} = \frac{S_{\varphi}(f)}{S_I(f)}, \tag{49}$$

where Δf is the spot frequency band of interest. Clearly, $\sigma(f)$ is a function of f and in designing an angle-modulated system one is usually
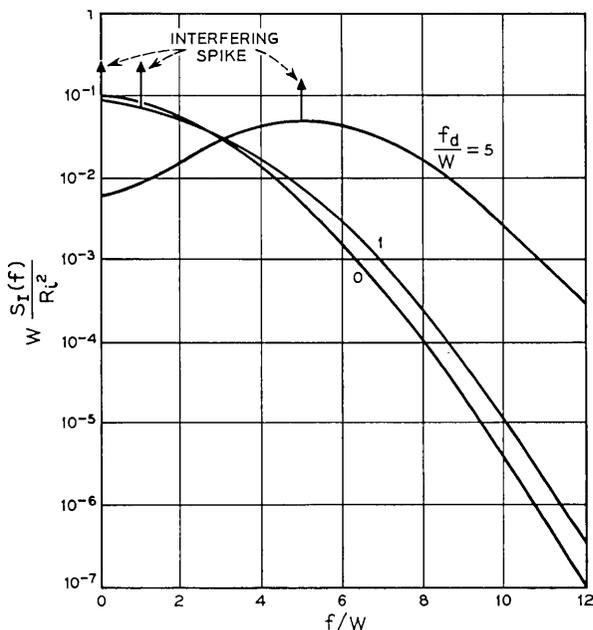
Fig. 3 — Spectral density $S_I(f)$ of baseband interference. $\Phi = 3$ rad; $\Phi_i = 2$ rad.

interested in the minimum value of $\sigma(f)$ for $0 < |f| \leq W$. We denote the minimum value of this signal-to-interference ratio by S/I. In practice a phase demodulator is followed by a linear low pass filter. We assume that this filter is ideal and that it removes all the frequency components outside the desired signal frequency band $0 < |f| \leq W$.

4.1 *Interference Between Two Co-Channel PM Waves*

In general, one can show (see Fig. 3) that $S_I(f)$ contains a (nonzero) Dirac delta function (corresponding to a line spectrum) at the frequency $\pm f_d$, and that the frequency division multiplex channel corresponding to this frequency may not be usable.[‡] In case the interference is co-channel, $f_d = 0$, and the line spectrum lies at the frequency $f = 0$. In systems usually encountered in practice, there is no frequency division multiplex channel at dc even though the lowest frequency present in the baseband signal may approach a frequency arbitrarily close to zero.[14]

Notice from equation (48) and Fig. 3 that, in the case of co-channel interference between two PM waves, maximum baseband interference occurs at the lowest frequency present in the system; we assume that this lowest baseband frequency lies arbitrarily close to zero. In this case the minimum signal-to-interference ratio therefore occurs at $f = 0$ and

$$S/I = \frac{\Phi^2}{2W} \frac{1}{S_I'(0)} , \tag{50}$$

where

$$S_I'(0) = \sum_{m=1}^{\infty} \frac{R_i^{2m}}{2m^2} \{H_m(f) - \exp[-m^2(\Phi^2 + \Phi_i^2)]\, \delta(f)\}_{f=0} . \tag{51}$$

Since the interference is co-channel we further assume that $\Phi = \Phi_i$ so that the rms phase deviations in the two PM waves are the same. We can now write

$$S_I'(0) = \sum_{m=1}^{\infty} \frac{R_i^{2m}}{2m^2} [H_m(f) - \exp(-2m^2\Phi^2)\, \delta(f)]_{f=0} . \tag{52}$$

Consider the case $\Phi > (5)^{\frac{1}{2}}$ radians. In this case one can show that[16]

$$[H_m(f) - \exp(-2m^2\Phi^2)\, \delta(f)]_{f=0} \approx \frac{1}{2m\Phi W} \left(\frac{3}{\pi}\right)^{\frac{1}{2}}, \tag{53}$$

and that the error in this approximation is less than 8 per cent. Hence,

---

[‡] We do not put any lower limit on the width of any frequency division multiplex channel present in the baseband signal.

we have

$$\frac{0.23}{\Phi W}\left(\frac{3}{\pi}\right)^{\frac{1}{2}}\sum_{m=1}^{\infty}\frac{R_i^{2m}}{m^3} < S_I'(0) < \frac{0.27}{\Phi W}\left(\frac{3}{\pi}\right)^{\frac{1}{2}}\sum_{m=1}^{\infty}\frac{R_i^{2m}}{m^3}, \quad \Phi > (5)^{\frac{1}{2}}\text{ rad. } (54)$$

It can be shown that

$$\sum_{m=1}^{\infty}\frac{R_i^{2m}}{m^3} = Q(R_i^2) = \int_0^{\infty}\frac{t^2\,dt}{e^t - R_i^2}. \tag{55}$$

Therefore, the signal-to-interference ratio at $f = 0$ is bounded by

$$\frac{1}{0.46}\left(\frac{\pi}{3}\right)^{\frac{1}{2}}\frac{\Phi^3}{Q(R_i^2)} > S/I > \frac{1}{0.54}\left(\frac{\pi}{3}\right)^{\frac{1}{2}}\frac{\Phi^3}{Q(R_i^2)}, \quad \Phi > (5)^{\frac{1}{2}}\text{ rad. } (56)$$

For any value of $R_i < 1$, equation (56) gives upper and lower bounds to S/I. We shall now investigate whether we can derive simpler upper and lower bounds to $Q(R_i^2)$.

From equation (55)

$$\sum_{m=1}^{\infty}\frac{R_i^{2m}}{m^3} = R_i^2 + \sum_{m=2}^{\infty}\frac{\exp\left[-m\ln\left(1/R_i^2\right)\right]}{m^3}. \tag{57}$$

Now one can show (see Fig. 4) that

$$0 < \sum_{m=2}^{\infty}\frac{\exp\left[-m\ln\left(1/R_i^2\right)\right]}{m^3} < \int_1^{\infty}\frac{\exp\left[-x\ln\left(1/R_i^2\right)\right]}{x^3}\,dx$$

$$= E_3[\ln\left(1/R_i^2\right)], \tag{58}$$

where[†]

$$E_3(z) = \int_1^{\infty}\frac{e^{-zt}}{t^3}\,dt, \quad z > 0. \tag{59}$$

We can show that for $R_i < 1$, $(\ln 1/R_i^2 > 0)$, (see Ref. 17)

$$0 < E_3(\ln 1/R_i^2) \leqq \frac{R_i^2}{2 + \ln\left(1/R_i^2\right)}, \tag{60}$$

or

$$Q(R_i^2) < R_i^2\left[1 + \frac{1}{2 + \ln\left(1/R_i^2\right)}\right]. \tag{61}$$

Since

[†] The function $E_3(z)$ is tabulated in Ref. 17 (see pp. 228–248). Notice also the inequality $E_n(z) \leqq e^{-z}/(z + n - 1)$ on p. 229.
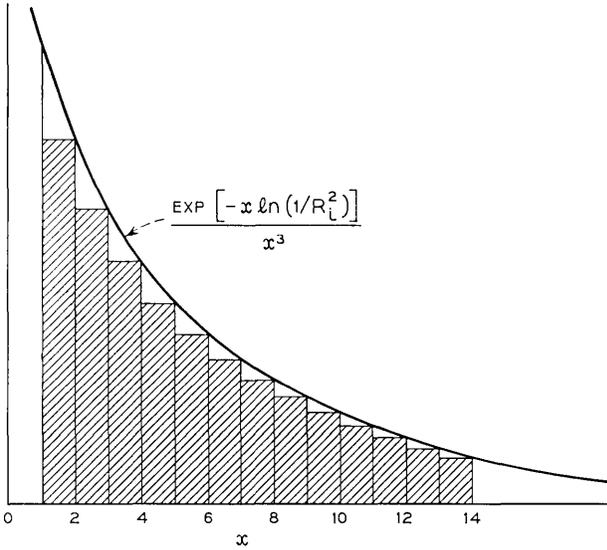
Fig. 4. — Function $\exp\left[-x \ln\left(1/R_i^2\right)\right]/x^3$ and $\Sigma^{\infty}{}_{m=2} R_i{}^{2m}/m^3$. The area in the shaded region is less than the area under the curve from $x = 1$.

$$\sum_{m=1}^{\infty} \frac{R_i^{2m}}{m^3} < \sum_{m=1}^{\infty} \frac{R_i^{2m}}{m} = -\ln\left(1 - R_i^2\right), \qquad Q(R_i^2) < -\ln\left(1 - R_i^2\right). \quad (62)$$

We are thankful to W. T. Barnett for having suggested another upper bound $R_i^2/(1 - R_i^2)$ to $Q(R_i^2)$.

One can show that the bound given in equation (62) is tighter than that given in equation (61) if

$$R_i < R_0 = 0.695573. \quad (63)$$

Let us write

$$U(R_i^2) = \begin{cases} 1 + \dfrac{1}{2 + \ln\left(1/R_i^2\right)}, & R_0 < R_i < 1, \\[2mm] -\dfrac{\ln\left(1 - R_i^2\right)}{R_i^2}, & 0 < R_i < R_0, \end{cases} \quad (64)$$

so that

$$R_i^2 < Q(R_i^2) < R_i^2 U(R_i^2), \qquad 0 < R_i < 1. \quad (65)$$

For carrier-to-interference ratio of 10 dB or for $R_i^2 = 0.1$

$$R_i^2 < \sum_{m=1}^{\infty} \frac{R_i^{2m}}{m^3} < 1.0536\, R_i^2. \quad (66)$$

From equations (56) and (65), we next write

$$\frac{1}{0.46}\left(\frac{\pi}{3}\right)^{\frac{1}{2}}\frac{1}{R_i^2}\,\Phi^3 > S/I > \frac{1}{0.54}\left(\frac{\pi}{3}\right)^{\frac{1}{2}}\frac{1}{R_i^2}\frac{\Phi^3}{U(R_i^2)},$$

$$\Phi > (5)^{\frac{1}{2}}\,\text{rad}, \quad R_i < 1. \qquad (67)$$

Since the physical characteristics of elements used in a PM receiver are far from being ideal, and since thermal noise (which is always present) further deteriorates the performance of any PM receiver, we often find that $R_i^2 < 0.1$ in systems currently in use. Equations (66) and (67) show that the error introduced in truncating the series at $m = 1$ is less than 5.36 percent if $R_i^2 < 0.1$. For any $R_i \ll 1$, we therefore need take only the $m = 1$ term in equation (54) to estimate the baseband interference. Equation (67) gives upper and lower bounds to S/I for any $R_i < 1$. Also, note from equation (67) that co-channel interference can be suppressed in PM systems by using a large modulation index Φ.[14]

4.2 *Interference between Two Adjacent-Channel PM Waves*

As mentioned in Section II we do not consider the effects of linear filters on angle-modulated systems. We assume that the desired and interfering wave are both in the passband of the PM receiver used in the system, and that no filters are used to reduce the adjacent channel interference.

In any multichannel angle-modulated system generally encountered in practice there is usually both adjacent channel and co-channel interference. Protection against adjacent channel interference is often obtained by proper choice of the channel separation frequency and the required (linear) filters generally used in such systems. The assumptions made in this section are, therefore, a little unrealistic; hence, the results given may serve only as a guide in the actual calculation of adjacent channel interference.

For $0 < f_d/W < 1$, one can show that $S_I(f)$ contains a (nonzero) Dirac delta function (corresponding to a line spectrum) at the frequency $\pm f_d$ and that the frequency division multiplex channel corresponding to f_d/W may not be usable.

For $f_d \neq 0$ we can show, from equations (44) through (46), that

$$S_I(f) = \sum_{m=1}^{\infty} \frac{R_i^{2m}}{m^2} G_m(f), \qquad (68)$$

where

$$G_m(f) = \tfrac{1}{4}[H_m(f - mf_d) + H_m(f + mf_d)], \qquad (69)$$

and

$$H_m(f) = \int_{-\infty}^{\infty} \exp\left[-m^2(\Phi^2 + \Phi_i^2)\left(1 - \frac{\sin 2\pi W\tau}{2\pi W\tau}\right)\right]e^{-j2\pi f\tau}\, d\tau. \quad (70)$$

For $0 \leqq |f| < W$, and $|f_d| \gg W$, one can show (by numerical methods) that $S_I(f)$ reaches its maximum at $f = W$, or that maximum baseband interchannel interference occurs at the highest frequency present in the baseband signal. For other values of channel separation frequency, this maximum is to be determined from equations (68) through (70).

For $(\Phi^2 + \Phi_i^2)^{\frac{1}{2}} > (30/\pi)^{\frac{1}{2}}$ rad, the saddle-point method of calculating $G_m(f)$ is very convenient;[16] and this method can be applied in a straight-forward manner to estimate $S_I(f)$. (Since one can show that the saddle-point approximation reduces to the quasistatic approximation for $f_d/W \ll (\Phi^2 + \Phi_i^2)^{\frac{1}{2}}$, the quasistatic approximation may be used for convenience if this condition is satisfied. However, the error introduced as a result of the use of quasistatic approximation cannot often be estimated.) For $R_i \ll 1$, we can also show that we need take only the $m = 1$ term in equation (68) to estimate S/I with a very small fractional error (less than 5.36 percent for $R_i < 0.1$).

For $f_d/W = 2, 4, 6, 8$, and 10 and for a set of values of Φ and Φ_i, we have calculated this minimum signal-to-interference ratio; Figs. 5 through 9 give these results. For any value of f_d/W and for any S/I, the required values of Φ and Φ_i may be obtained from these figures. Since the effects of linear filters on adjacent channel interference has not been taken into account in this paper, these values of Φ and Φ_i may serve only as a guide in the design of any angle-modulated system.

V. INTERFERENCE BETWEEN L+1 PM WAVES

We now assume that there are L interfering waves, and that all of them are phase modulated by mutually independent gaussian random processes.[†] Let the desired PM wave be given by

$$s(t) = \text{Re } A \exp\{j[\omega_o t + \varphi(t)]\}. \quad (71)$$

Let the kth interfering wave be represented as

$$i_k(t) = \text{Re } R_{ik}A \exp\{j[\omega_k t + \varphi_{ik}(t) + \mu_k]\}, \quad 1 \leqq k \leqq L. \quad (72)$$

[†] The analysis given in this section can suitably be modified for angle modulation by general gaussian random processes (see Section II).

Fig. 5 — Signal-to-interference ratio as a function of rms phase deviations and channel separation for $\Phi_i = 2$ rad.

Since the L interfering waves are assumed to originate from L different sources, we assume that the μ_k's are independent of each other, and that μ_k, $1 \leqq k \leqq L$ has a uniform probability density function $\pi_{\mu_k}(\mu)$ where

$$\pi_{\mu_k}(\mu) = \begin{cases} 1/2\pi, & 0 \leqq \mu < 2\pi, \quad 1 \leqq k \leqq L, \\ 0, & \text{otherwise.} \end{cases} \tag{73}$$

We further assume that $\varphi(t)$, the $\varphi_k(t)$'s, and the μ_k's (with $1 \leqq k \leqq L$) are mutually independent random variables.

If $s(t)$ and the $i_k(t)$'s are all in the passband of the PM receiver used in the system, the total signal incident at the receiver can be written as

$$r(t) = s(t) + \sum_{k=1}^{L} i_k(t)$$

$$= \text{Re } A\left(1 + \sum_{k=1}^{L} R_{ik} \exp \{j[\omega_{dk}t + \varphi_{ik}(t) - \varphi(t) + \mu_k]\}\right)$$

$$\cdot \exp \{j[\omega_0 t + \varphi(t)]\}, \tag{74}$$

where

$$\omega_{dk} = \omega_k - \omega_o = f_{dk}/2\pi. \tag{75}$$

From equation (74), we can show that the output $\theta(t)$ of an ideal phase demodulator can be represented as

$$\theta(t) = \varphi(t) + \text{Im ln}\left(1 + \sum_{k=1}^{L} R_{ik} \exp \{j[\omega_{dk}t + \varphi_{ik}(t) - \varphi(t) + \mu_k]\}\right). \tag{76}$$

Next we write

$$\ln\left(1 + \sum_{k=1}^{L} R_{ik} \exp \{j[\omega_{dk}t + \varphi_{ik}(t) - \varphi(t) + \mu_k]\}\right)$$

$$= \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \left(\sum_{k=1}^{L} R_{ik} \exp \{j[\omega_{dk}t + \varphi_{ik}(t) - \varphi(t) + \mu_k]\}\right)^m$$
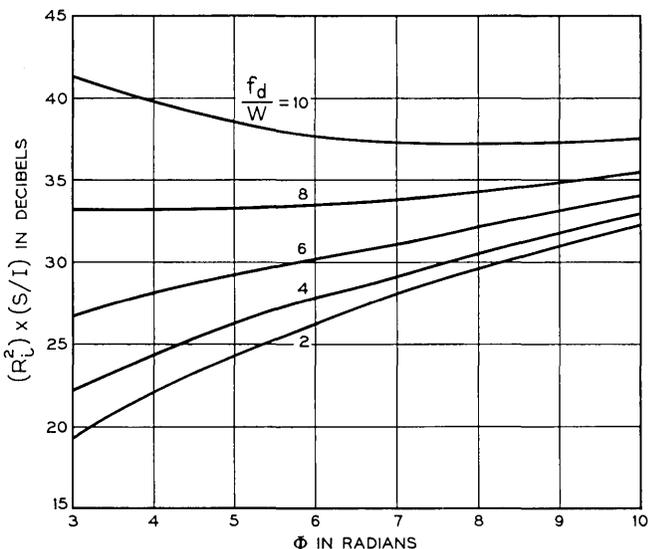
Fig. 6 — Signal-to-interference ratio as a function of rms phase deviations and channel separation for $\Phi_i = 4$ rad.
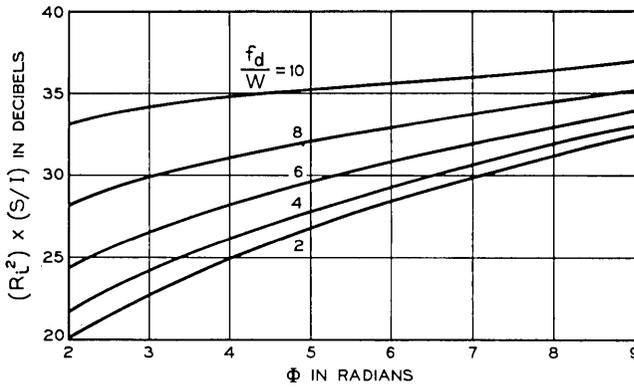
Fig. 7 — Signal-to-interference ratio as a function of rms phase deviations and channel separation for $\Phi_i = 6$ rad.

$$\sum_{k=1}^{L} R_{ik} < 1. \tag{77}$$

By the multinomial theorem, we have

$$\left(\sum_{k=1}^{L} R_{ik} \exp \{ j[\omega_{dk}t + \varphi_{ik}(t) - \varphi(t)] \} \right)^{m}$$

$$= \sum \frac{m!}{\prod\limits_{r=1}^{L} a_r!} \prod_{r=1}^{L} R_{ir}^{a_r} \exp \{ j a_r [\omega_{dr}t + \varphi_{ir}(t) - \varphi(t) + \mu_r] \}, \tag{78}$$

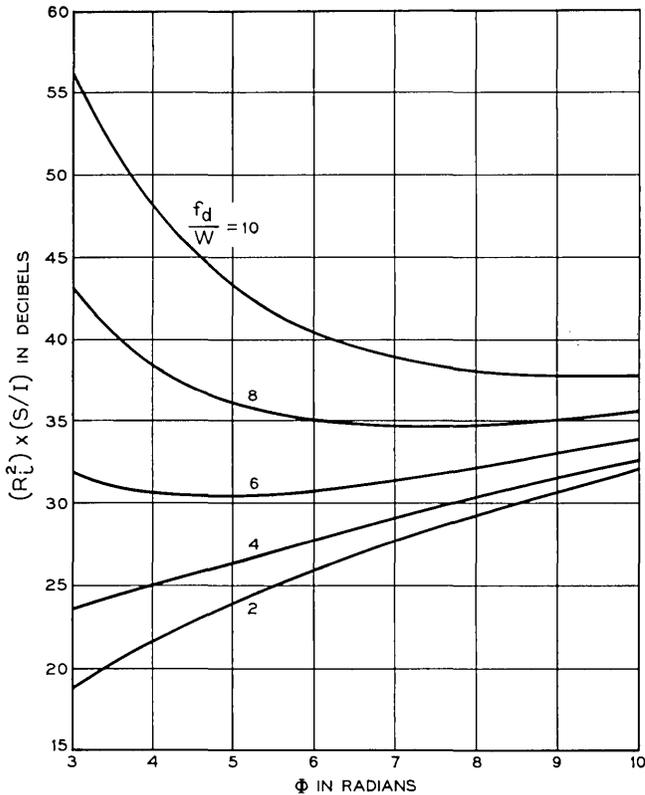where the a_r's are a set of nonnegative integers such that
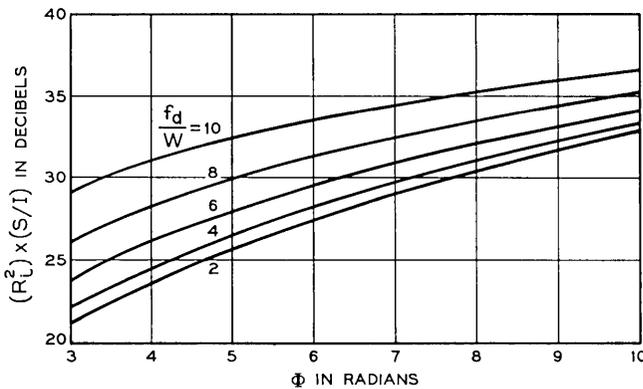
Fig. 8 — Signal-to-interference ratio as a function of rms phase deviations and channel separation for $\Phi_i = 8$ rad.
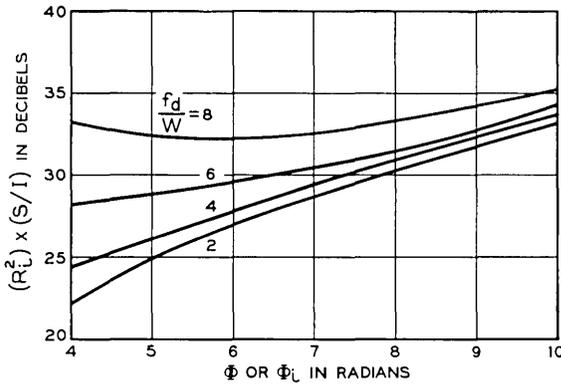
Fig. 9 — Signal-to-interference ratio as a function of rms phase deviations and channel separation for $\Phi = \Phi_i$.

$$\sum_{r=1}^{L} a_r = m. \tag{79}$$

From equations (73) and (76) through (79), one can show that the covariance function $R_\theta(\tau)$ of $\theta(t)$ can be written as

$$R_\theta(\tau) = \langle \theta(t)\theta(t + \tau) \rangle$$

$$= R_\varphi(\tau) + \sum_{m=1}^{\infty} \frac{1}{2m^2} \exp \{-m^2[R_\varphi(0) - R_\varphi(\tau)]\}$$

$$\cdot \left[\sum \left[\frac{m! \prod_{r=1}^{L} R_{ir}^{a_r}}{\prod_{r=1}^{L} a_r!} \right]^2 \exp \left(-\sum_{r=1}^{L} a_r^2[R_{\varphi_r}(0) - R_{\varphi_r}(\tau)] \right) \right.$$

$$\left. \cdot \cos \left(\tau \sum_{r=1}^{L} a_r\omega_{dr} \right) \right]. \tag{80}$$

If the random gaussian noise is band-limited and white, and if all the modulating waveforms have the same bandwidth W, we have

$$R_\varphi(\tau) = \Phi^2 \cdot \frac{\sin 2\pi W\tau}{2\pi W\tau}, \tag{81}$$

and

$$R_{\varphi_k}(\tau) = \Phi_k^2 \frac{\sin 2\pi W\tau}{2\pi W\tau}, \qquad 1 \leqq k \leqq L. \tag{82}$$

In this case, equation (80) can be written as

$$R_\theta(\tau) = R_\varphi(\tau) + \sum_{m=1}^{\infty} \frac{1}{2m^2} \exp\left[-m^2\Phi^2\left(1 - \frac{\sin 2\pi W\tau}{2\pi W\tau}\right)\right]$$

$$\cdot \left[\sum \frac{(m!)^2 \prod_{r=1}^{L} R_{ir}^{2a_r}}{\prod_{r=1}^{L} (a_r!)^2} \exp\left[-\left(1 - \frac{\sin 2\pi W\tau}{2\pi W\tau}\right)\sum_{r=1}^{L} a_r^2\Phi_r^2\right]\right.$$

$$\left. \cdot \cos\left(\tau \sum_{r=1}^{L} a_r\omega_{dr}\right)\right]. \tag{83}$$

Therefore, the spectral density of baseband interchannel interference is given by

$$S_I(f) = \sum_{m=1}^{\infty} \frac{1}{4m^2}\left[\sum_s \frac{(m!)^2 \prod_{r=1}^{L} R_{ir}^{2a_r}}{\prod_{r=1}^{L} (a_r!)^2}\right.$$

$$\left. \cdot \left[T_{ms}\left(f - \sum_{r=1}^{L} a_r f_{dr}\right) + T_{ms}\left(f + \sum_{r=1}^{L} a_r f_{dr}\right)\right]\right], \tag{84}$$

where

$$T_{ms}(f) = \int_{-\infty}^{\infty} \exp\left[-\left(1 - \frac{\sin 2\pi W\tau}{2\pi W\tau}\right)\left(m^2\Phi^2 + \sum_{r=1}^{L} a_r^2\Phi_r^2\right)\right]e^{-j2\pi f\tau}d\tau. \tag{85}$$

Next notice that the methods given in Section III can be used to calculate $T(f)$ for all values of Φ, and Φ_{ik}'s (with $1 \leq k \leq L$); hence, we can calculate $S_I(f)$ for all values of R_{ik}'s such that $\sum_{k=1}^{L} R_{ik} < 1$. The minimum signal-to-interference ratio S/I can then be obtained from equation (49).

Now assume that we have L co-channel interferers and that all have the same rms phase deviation Φ, or

$$\Phi_r = \Phi, \qquad 1 \leq r \leq L. \tag{86}$$

In this case equation (84) yields

$$S_I(f) = \sum_{m=1}^{\infty} \frac{1}{2m^2}\left[\sum_s \frac{(m!)^2 \prod_{r=1}^{L} R_{ir}^{2a_r}}{\prod_{r=1}^{L} (a_r!)^2} G_{ms}(f)\right], \tag{87}$$

where

$$G_{ms}(f) = \int_{-\infty}^{\infty} \exp\left[-\left(1 - \frac{\sin 2\pi W \tau}{2\pi W \tau}\right)\Phi^2\left(m^2 + \sum_{r=1}^{L} a_r^2\right)\right] e^{-j2\pi f\tau} \, d\tau.$$

(88)

From equations (87) and (88) and Refs. 2 and 16, one can show that the continuous part of $S_I(f)$ reaches its maximum at $f = 0$, and that[†]

$$0.92\left(\frac{3}{2\pi}\right)^{\frac{1}{2}} \frac{1}{\left(m^2 + \sum_{r=1}^{L} a_r^2\right)^{\frac{1}{2}}} \frac{1}{W\Phi} < G_{ms}(0)$$

$$< 1.08\left(\frac{3}{2\pi}\right)^{\frac{1}{2}} \frac{1}{\left(m^2 + \sum_{r=1}^{L} a_r^2\right)^{\frac{1}{2}}} \frac{1}{W\Phi}, \qquad \Phi > (5)^{\frac{1}{2}} \text{ rad.} \quad (89)$$

The expression $G_{ms}(0)$ in equation (89) does not include the delta function contained in $G_{ms}(f)$ at $f = 0$.

Since $\sum_{r=1}^{L} a_r = m$, one can prove that

$$\frac{m^2}{L} \leqq \sum_{r=1}^{L} a_r^2 \leqq m^2.$$

(90)

From equations (89) and (90) we have

$$0.46\left(\frac{3}{\pi}\right)^{\frac{1}{2}} \frac{1}{m\Phi W} < G_{ms}(0) < 0.54\left(\frac{3}{\pi}\right)^{\frac{1}{2}}\left(\frac{2L}{L+1}\right)^{\frac{1}{2}} \frac{1}{m\Phi W}.$$

(91)

Next

$$S_I(0) < \sum_{m=1}^{\infty} \frac{1}{2m^2} \left[\sum \frac{(m!)^2 \prod_{r=1}^{L} R_{ir}^{2a_r}}{\prod_{r=1}^{L} (a_r!)^2} \, 0.54\left(\frac{3}{\pi}\right)^{\frac{1}{2}}\left(\frac{2L}{L+1}\right)^{\frac{1}{2}} \frac{1}{m\Phi W}\right]. \quad (92)$$

If all x_i's are nonnegative, one can show that

$$\sum x_i^2 \leqq \left(\sum x_i\right)^2.$$

(93)

Using equation (93), equation (92) yields

$$S_I(0) < \sum_{m=1}^{\infty} 0.27\left(\frac{3}{\pi}\right)^{\frac{1}{2}}\left(\frac{2L}{L+1}\right)^{\frac{1}{2}} \frac{1}{\Phi W} \frac{1}{m^3} \left[\left[\sum \frac{m! \prod_{r=1}^{L} R_{ir}^{a_r}}{\prod_{r=1}^{L} a_r!}\right]^2\right]$$

[†] We consider only the continuous part of $S_I(f)$.

$$= 0.27 \left(\frac{3}{\pi}\right)^{\frac{1}{2}} \left(\frac{2L}{L+1}\right)^{\frac{1}{2}} \frac{1}{\Phi W} \sum_{m=1}^{\infty} \frac{\left(\sum_{r=1}^{L} R_{ir}\right)^{2m}}{m^3} \qquad (94)$$

or

$$S_I(0) < 0.27 \left(\frac{3}{\pi}\right)^{\frac{1}{2}} \left(\frac{2L}{L+1}\right)^{\frac{1}{2}} \frac{1}{\Phi W} Q(b^2), \qquad (95)$$

where

$$b^2 = \left(\sum_{r=1}^{L} R_{ir}\right)^2 < 1. \qquad (96)$$

We have shown in Section IV that

$$\sum_{m=1}^{\infty} \frac{b^{2m}}{m^3} < b^2 U(b^2), \qquad b^2 < 1. \qquad (97)$$

Therefore, the minimum baseband signal-to-interference ratio is bounded by

$$S/I > \frac{1}{0.54} \left(\frac{\pi}{3}\right)^{\frac{1}{2}} \left(\frac{L+1}{2L}\right)^{\frac{1}{2}} \frac{\Phi^3}{b^2 U(b^2)}, \qquad \Phi > (5)^{\frac{1}{2}} \text{ rad}, \quad b < 1. \quad (98)$$

From equation (87) we can also show that

$$S_I(0) > \frac{1}{2} \left(\sum_{r=1}^{L} R_{ir}^2\right) G_{1s}(0). \qquad (99)$$

Equations (89) and (99) yield

$$S_I(0) > 0.23 \left(\frac{3}{\pi}\right)^{\frac{1}{2}} \frac{1}{W\Phi} \left(\sum_{r=1}^{L} R_{ir}^2\right), \qquad (100)$$

or

$$S/I < \frac{1}{0.46} \left(\frac{\pi}{3}\right)^{\frac{1}{2}} \frac{\Phi^3}{\sum_{r=1}^{L} R_{ir}^2}. \qquad (101)$$

Hence we have

$$\frac{1}{0.46} \left(\frac{\pi}{3}\right)^{\frac{1}{2}} \frac{\Phi^3}{\sum_{r=1}^{L} R_{ir}^2} > S/I > \frac{1}{0.54} \left(\frac{\pi}{3}\right)^{\frac{1}{2}} \left(\frac{L+1}{2L}\right)^{\frac{1}{2}} \frac{\Phi^3}{b^2 U(b^2)},$$

$$\Phi > (5)^{\frac{1}{2}} \text{ rad}, \quad b = \sum_{k=1}^{L} R_{ik} < 1. \qquad (102)$$

For any set of values of R_{ik}'s, $1 \leq k \leq L$, and for any Φ, bounds to the signal-to-interference ratio S/I can be calculated from equation (102), and a proper Φ can then be chosen to keep the baseband interference below any desired level.

Notice that the upper bound is a function of the total interference power, and the lower bound a function of the sum of the amplitudes of all the interfering carriers. In such cases, the distribution of R_{ik}'s generally determines the closeness of the two bounds. However, it may be observed that both these bounds are proportional to the cube of the modulation index Φ (for a high index system).

VI. RESULTS AND CONCLUSIONS

In this paper we consider the effect of interchannel interference on angle-modulated systems. We also derive an expression for the baseband interchannel interference when two (or more) waves angle modulated by gaussian noise interfere with each other. This formula can be used even when the baseband signal is passed through a linear network such as a pre-emphasis—de-emphasis network. We show that the calculation of the RF spectral density is essential to the evaluation of the baseband interchannel interference.

We then consider band-limited white gaussian noise modulation and show that, in the case of co-channel interference, maximum baseband interference occurs at the lowest baseband frequency present in the system. For moderately high modulation index, we show that we can derive upper and lower bounds to this minimum signal-to-interference ratio and that these bounds are proportional to the cube of the modulation index. It therefore follows that co-channel interference in PM systems can be reduced by expanding bandwidth, and that phase modulation can be used with advantage in combating interference. We also show that the first term in the power series expansion for the baseband interchannel interference gives most of the contribution if the carrier-to-interference ratio is greater than about 10 dB (the error is less than 5.36 per cent for a carrier-to-interference ratio greater than 10 dB).

In this paper we also give some results about the effects of adjacent channel interference on angle-modulated systems. We assume that all the incident signals at the receiver are in the passband of the PM receiver used in the system. This assumption is justified in the case of co-channel interference, but is not realistic in the case of adjacent channel interference. However, we feel that the results given in this paper for the adjacent channel interference may serve as a guide in

determining the deterioration in performance produced by adjacent channel interference.

VII. ACKNOWLEDGMENT

Some of the results presented here were obtained earlier by Clyde L. Ruthroff. We are grateful to him for his consent to publish those results in this paper.

REFERENCES

1. Black, H. S., *Modulation Theory,* Princeton, New Jersey: D. Van Nostrand, 1953.
2. Rowe, H. E., *Signals and Noise in Communication Systems,* Princeton, New Jersey: D. Van Nostrand, 1965, pp. 98–203.
3. Tillotson, L. C., "Use of Frequencies above 10 GHz for Common Carrier Applications," B.S.T.J., *48,* No. 7 (July–August 1969), pp. 1563–1576.
4. Ruthroff, C. L., and Tillotson, L. C., "Interference in Dense Radio Networks," B.S.T.J., *48,* No. 7 (July–August), pp. 1727–1743.
5. Medhurst, R. G., Hicks, E. M., and Grossett, W., "Distortion in Frequency Division Multiplex FM Systems Due to an Interfering Carrier," Proc. IEE, *105B,* No. 5 (May 1958), pp. 282–292.
6. Medhurst, R. G., "FM Interfering Carrier Distortion: General Formula," Proc. IEE, *109B,* No. 3 (March 1962), pp. 149–150 and 519–523.
7. Medhurst, R. G., and Roberts, J. H., "Expected Interference Levels Due to Interactions between Line-of-Sight Radio-Relay Systems and Broadband Satellite Systems," Proc. IEE, *111,* No. 3 (March 1964), pp. 519–523.
8. Hamer, R., "Radio-Frequency Interference in Multi-Channel Telephone FM Radio Systems," Proc. IEE, *108B,* No. 1 (January 1961), pp. 75–89.
9. Curtis, H. E., and Rice, S. O., unpublished work.
10. Borodich, S. Y., "Calculating the Permissible Magnitude of Radio Interference in Multi-Channel Radio Relay Systems," Electrosviaz, *1,* No. 1 (January 1962), pp. 13–24.
11. Hayashi, S., "On the Interference Characteristics of the Phase Modulation Receiver for the Multiplex Transmission," IEE (Japan), *35,* No. 11 (November 1952), pp. 522–528.
12. Curtis, H. E., "Interference between Satellite Communication Systems and Common Carrier Surface Systems," B.S.T.J., *41,* No. 3 (May 1962), pp. 921–943.
13. Bennett, W. R., Curtis, H. E., and Rice, S. O., "Interchannel Interference in FM and PM Systems under Noise Loading Conditions," B.S.T.J., *34,* No. 3 (May 1955), pp. 601–636.
14. Ruthroff, C. L., unpublished work.
15. Middleton, D., *Introduction to Statistical Communication Theory,* New York: McGraw-Hill, 1960, pp. 599–678.
16. Prabhu, V. K., and Rowe, H. E., "Spectral Density Bounds of a PM Wave," B.S.T.J., *48,* No. 3 (March 1969), pp. 789–831.
17. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions,* Nat. Bureau of Standards, Washington, D. C., 1967, pp. 227–251.

Calculated Quantizing Noise of Single-Integration Delta-Modulation Coders

By J. E. IWERSEN

(Manuscript received March 28, 1969)

We calculate the granular quantizing noise for a delta modulator that has unequal positive and negative step sizes. The asymmetry leads to a highly colored noise spectrum. We perform this calculation by adding a ramp function of time to the input of a symmetrical coder. The resulting formulas can also be used for uniform DPCM and PCM coders. The idle-channel spectrum consists of discrete lines which scatter somewhat irregularly in amplitude and frequency; they can be regarded as the result of sampling (aliasing) a sawtooth wave. These lines are phase-modulated by a coder input. For a sinusoidal input, discrete side frequencies are produced which again have an irregular progression of amplitudes. Gaussian inputs lead to gaussian line shapes; the lines broaden as input power is increased. A totally white spectrum (as is often assumed in connection with delta-modulation-system considerations) cannot be attained, however, before the onset of slope overload. We give a numerical example that uses a coder suitable for telephone applications. One can see that step asymmetry can be very advantageous in attaining low noise.

I. INTRODUCTION

While Laane and Murphy[1] were investigating the encoding of speech using delta modulation $(\triangle M)^2$ it became apparent to us that existing theories of granular quantizing noise[3,4] were seriously deficient; they did not take into account, except in a very elementary way, the asymmetry of the positive and negative integrator step sizes. This work intends to correct this deficiency.

Figure 1 is a block diagram of a $\triangle M$ coder-decoder. An input signal is compared with a locally reconstructed version of itself and the differential, or error, is quantized into a one-bit code, transmitted, and integrated at a receiver to recover the original signal. Quantizing noise is produced by the coding process and is also recovered at the
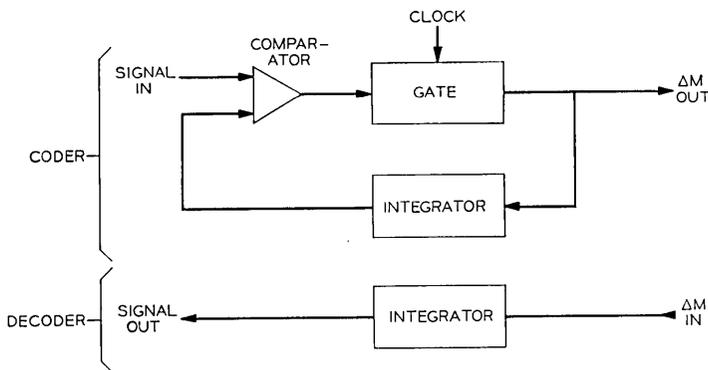
Fig. 1 — Delta-modulation coder plus decoder (codec).

receiver; this noise is the subject of this paper. We limit our consideration to single-integration systems.

In the past, considerations of ΔM noise have been broken into two distinct areas: calculation of quantizing noise accompanying a typical signal,[3] and calculation of idle-channel (zero-input-signal) noise.[4] As Fig. 2 shows, there is no idle-channel noise for a coder in which the plus and minus quanta (steps) fed to the integrator are exactly equal in magnitude. The integrator output spectrum contains only the out-of-signal-band Nyquist frequency, f_N (one half the sampling frequency, f_s) and its harmonics. In any real coder, however, it is impossible to balance the plus and minus steps perfectly, with the result that the output contains occasional double-plus (or double-minus) steps, as Fig. 3 shows. In general, this waveform has signal-band components. Wang calculated the noise for this case but his results, while adequate as far as they go, are incomplete and nonrigorous.[4] Van de Weg's calculation of noise in the presence of signal was for an equal-step (symmetrical) coder.[3] We do the calculation for an unequal-
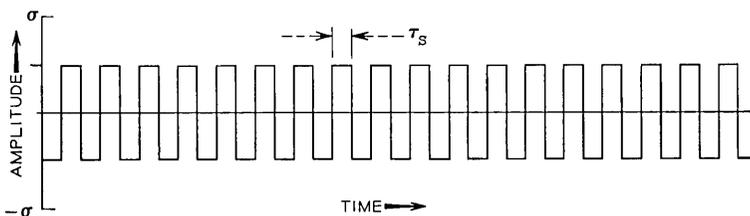
Fig. 2 — Integrator-output wave from a symmetrical (equal-step-size) coder.
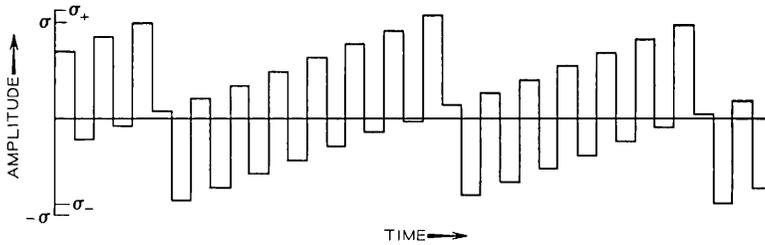
Fig. 3 — Integrator-output wave from an asymmetrical (unequal-step-size) coder, shown with $|\sigma_+| > |\sigma_-|$.

step (asymmetrical) coder and show that there are significant differences.

In much of the literature on delta modulators, where noise is treated casually, the assumption is made that the total average noise power is more or less uniformly distributed in the band from zero frequency to the Nyquist frequency. This assumption is a very good approximation for multibit PCM and DPCM but, as we show, can lead to colossal errors for $\triangle$M.

Results presented for gaussian input are in terms of time-averaged noise power. In this form they are directly useful for speech systems and typical data systems but are of more limited value for video systems, where details of the waveform are perceived.

In Sections II and III, we set up the method of attacking the problem. Then in Section IV we treat zero input (the idle channel), in Section V a sinusoidal input, and in Section VI a broadband gaussian input. The appendixes contain various mathematical developments necessary for logical completeness but not important to the reader interested in engineering understanding and application of the main results (with the possible exception of Appendix D).

II. QUANTIZING RULES

The function of the coder (Fig. 1) is, at each clock time or sampling instant, to add a positive step (σ_+) to the integrator output (q) if this output is less than the signal input (y) or to add a negative step (σ_-) if the output is greater than the input. If the integrator has instantaneous response and infinite time constant, the output is a sequence of rectangular pulses, as in Figs. 2 and 3.

If y_n is the value of the input at the nth sampling instant, and q_n is the output value just before this instant, the operation can be sum-

marized:

$y_n - q_n$	q_{n+1}
$+$	$q_n + \sigma_+$
$-$	$q_n + \sigma_-$

As we mentioned, it is not possible to make the magnitudes of σ_+ and σ_- exactly equal in a real coder. Let us therefore define

$$\sigma_+ \equiv \sigma + \epsilon, \qquad \sigma_- \equiv -\sigma + \epsilon, \tag{1}$$

where σ, the average step size, is a positive quantity. The coder operation can then be summarized in a single equation:

$$q_{n+1} = q_n + \sigma \operatorname{sgn} (y_n - q_n) + \epsilon. \tag{2}$$

We are actually interested in the error, or noise, $x \equiv q - y$, which accompanies the reconstructed signal. (Appendix A shows that x is usually uncorrelated with y and is therefore noise under any circumstances.) Substituting for q in equation (2) gives the noise as a function of the input:

$$x_{n+1} - x_n + \sigma \operatorname{sgn} x_n = -y_{n+1} + y_n + \epsilon \tag{3}$$

$$= -[y_{n+1} - (n + 1)\epsilon] + [y_n - n\epsilon] \tag{4}$$

$$\equiv -y'_{n+1} + y'_n . \tag{5}$$

Thus we are led to a crucial principle: *The noise output of an asymmetrical* $(\epsilon \neq 0)$ ΔM *coder can be calculated as the noise output of a symmetrical* $(\epsilon = 0)$ *coder, if the input is taken as the actual input plus an appropriate ramp or staircase function of time.*

If equation (5) is summed from $t = -\infty$ to just before the nth instant [assuming $x(-\infty) = y(-\infty) = 0$],

$$x_n + \sigma \sum_{i=-\infty}^{n-1} \operatorname{sgn} x_i = -y'_n \tag{6}$$

or

$$q'_n = -\sigma \sum_{i=-\infty}^{n-1} \operatorname{sgn} x_i . \tag{7}$$

The resulting summation in equation (7) must be an integer alternating between odd and even values as a function of n. We can, without loss of generality, take it even for even n. [In equation (11) we include an arbitrary initial value of amplitude for the ramp; this covers the

possibility that the odd-even assumption is consequential.] If we assume that the coder does not go into slope overload (that is the input slope stays between the limits σ_+/τ_s and σ_-/τ_s), then q_n'/σ is the nearest odd integer to $(y_n' + \epsilon)/\sigma$ for the odd sampling instants and the nearest even integer for the even instants. $y_n' + \epsilon$ appears, rather than y_n', because the error must range from $\sigma + \epsilon$ to $-\sigma + \epsilon$ rather than from $+\sigma$ to $-\sigma$. The effect of this added ϵ is simply that the coder transmits a dc level of ϵ in addition to other signals and noise. Since ΔM systems normally suppress dc, as is mentioned in Section III in connection with other reasons, this added ϵ is dropped in the succeeding mathematical development. If it is desired to include it, $x - \epsilon$ should be substituted for x in what follows.

We have seen that a ΔM coder has two quantizing functions which alternate in time. Figures 4 and 5 show these functions; both the input and output are normalized to σ.

$q_0'(y)$ and $q_e'(y)$ were called $E(y)$ and $O(y)$ by van de Weg who cast them into contour-integral form and used them directly.[3] We prefer to follow the suggestion of Rice and use the error functions, $x(y)$, also shown in Figs. 4 and 5.[5] These functions, periodic in y, are conveniently represented by their Fourier series:

$$x_e = \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp\left(\pi i l g'\right), \tag{8}$$

and

$$x_0 = \sum_{l \neq 0} (-1)^l \frac{\sigma}{\pi i l} \exp\left(\pi i l g'\right) \tag{9}$$

where $g' \equiv y'/\sigma$. For a ΔM coder then,

$$x_n = \sum_{l \neq 0} (-1)^{ln} \frac{\sigma}{\pi i l} \exp\left(\pi i l g_n'\right) \tag{10}$$

$$= \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp\left[\pi i l(n + g_n')\right] \tag{11}$$

$$= \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp\left\{\pi i l[\vartheta_0 + (1 - \vartheta)n + g_n]\right\}, \tag{12}$$

where we have introduced

$$g_n \equiv \frac{y_n}{\sigma}, \qquad \vartheta = \frac{\epsilon}{\sigma}, \qquad \vartheta_0 = \frac{\epsilon_0}{\sigma}. \tag{13}$$

The last part of equation (13) takes into account an arbitrary initial amplitude for the ramp.
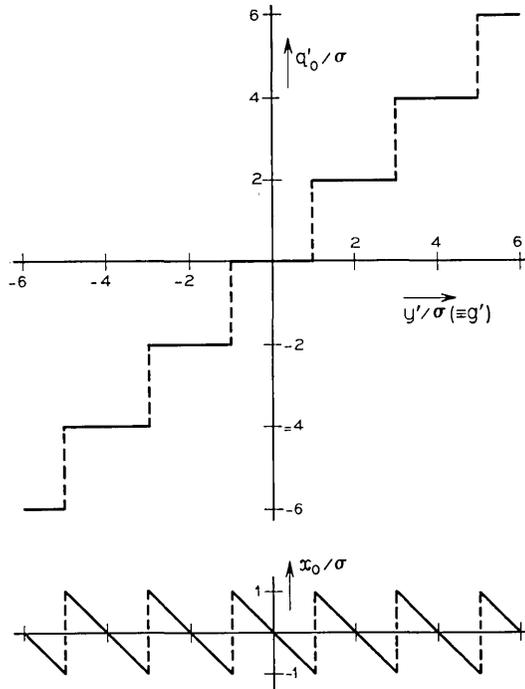
Fig. 4 — Quantizing and error functions for odd sampling intervals.

There is actually nothing in equation (12) which constrains the change in integrator output to be equal to one step per sampling interval. Indeed, the quantizing functions in Figs. 4 and 5 are perfectly valid for uniform DPCM systems where changes $\pm\sigma$, $\pm3\sigma$, $\cdots$, $\pm(2N-1)\sigma$ are allowed. Thus the formulas developed in this paper can be used for DPCM (and PCM—see Appendix B) with the provision that they are useful for input signals with up to $2N-1$ times the maximum slope of the ΔM system. This provision is not trivial, however; when signals range over many steps per sampling interval the errors tend to be uncorrelated, the noise spectrum tends to be white, and the structure (important for ΔM) calculated here is negligible.

III. NOISE FORMULA

Results in this paper are given in terms of frequency spectra of noise (two-sided unless otherwise identified). It is well known that

the spectrum of a pulse sequence can be broken into a factor which contains the information about the pulse shape and a factor which contains information about the area of each pulse and the periodicity. The "shape" factor (called also the "structure" or "aperture" factor) depends on the details of the coder circuit response. This factor is frequently negligible, because the low pass filter it represents normally does not contribute any significant distortion in the signal band. Thus we need only consider a δ-function representation of the sampled-signal, integrator-output, and noise pulse trains, as did van de Weg.[3]

The noise wave with the proper area for each pulse is

$$\nu(t) = \tau_s \sum_{n=-\infty}^{\infty} x_n \, \delta(t - n\tau_s), \tag{14}$$
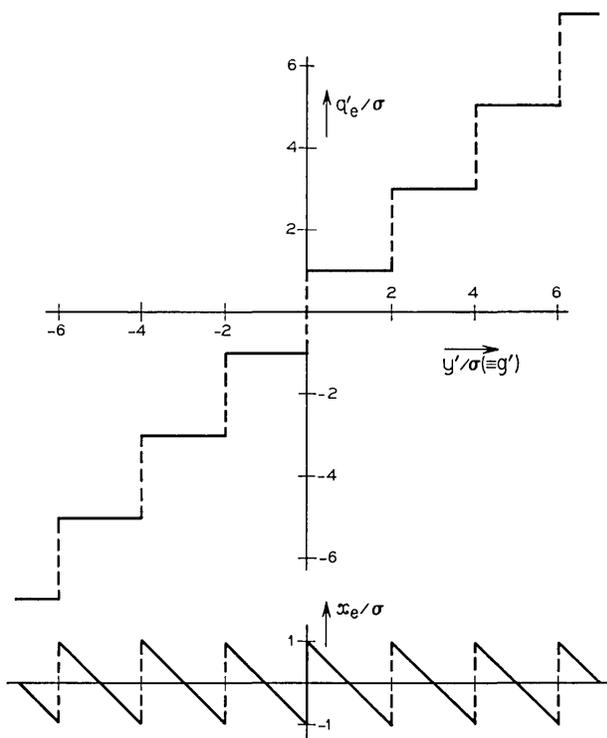
Fig. 5 — Quantizing and error functions for even sampling intervals.

where τ_s is the sampling interval $(\equiv 1/f_s)$. If $x(t)$ is defined as a continuous wave with samples $x_n = x(n\tau_s)$, equation (14) can be written

$$\nu(t) = \tau_s x(t) \sum_{n=-\infty}^{\infty} \delta(t - n\tau_s) \tag{15}$$

$$= x(t) \sum_{k=-\infty}^{\infty} \exp(2\pi i k f_s t). \tag{16}$$

A convenient form for $x(t)$ is, using equation (12),

$$x(t) = \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp\{\pi i l[\vartheta_0 + (1 - \vartheta)f_s t + g(t)]\}, \tag{17}$$

where

$$g(t) \equiv \frac{y(t)}{\sigma}. \tag{18}$$

Combining equations (16) and (17) gives

$$\nu(t) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \frac{\sigma}{\pi i l} \exp\left[\pi i l \vartheta_0 + 2\pi i\left(\frac{l(1-\vartheta)}{2} + k\right)f_s t + \pi i l g(t)\right]. \tag{19}$$

Thus the noise wave (before filtering by the shape factor) consists of a collection of lines of frequency

$$\left(\frac{l(1-\vartheta)}{2} + k\right)f_s ,$$

each phase-modulated by the input signal through a time-dependent angle, $\pi l g(t)$. These lines are examined in Section IV.

It is well known that the power spectrum of equation (14), and therefore equation (19), is periodic in frequency, f, with period f_s. We can thus concentrate on the band from $-f_N$ to f_N. Because of the aliasing or folding problem, all useful signals lie in this band (or any band of width f_s). The total power in $(-f_N, f_N)$ is $\sigma^2/3$, which is also equal to the mean square error, $\langle x^2 \rangle$. Appendix D treats these matters explicitly.

Equation (19) gives the noise generated at the coder, while one is ordinarily interested in the noise at a (distant) decoder. Unless the decoder has exactly the same step sizes as the coder, the noises are different. If the σ's are different, there is some linear gain or loss in the system; signal and noise are affected equally and their ratio, the really significant figure of merit, is not affected. If the ϑ's are different, the noises will differ only by a drift or ramp function of time. To get

rid of this ramp (and also because it is necessary to damp out the effect of errors in transmission) real systems have low-frequency (below-signal-band) cutoffs, called leaks, built into the integrators. The high-frequency cutoffs of the coder and decoder integrators may also be different. In the event that these cutoffs affect the signal band they can be taken into account as separate factors in determining the spectrum. Thus equation (19) can be used to calculate noise at the decoder output.

IV. IDLE-CHANNEL NOISE

The term "idle-channel noise" is used here as if it were synonymous with "zero-input noise." We recognize that this terminology is somewhat loose, in that an idle channel is actually characterized by a thermal or other noise input. Nevertheless, this usage seems established in the literature and the distinction is not significant for most cases of practical interest.

Figure 3 shows the integrator output of an asymmetrical coder. An approximately sawtooth-shaped wave with peak-to-peak amplitude $\approx \sigma$ is clearly visible (Wang's "first envelope function").[4] Other not-so-evident sawteeth are also usually present.

Putting $y = 0$ into equation (4), we see that the idle-channel noise output of an asymmetrical coder can be calculated as the noise output of a symmetrical coder with a ramp input. Figure 6 illustrates this. The error wave in Fig. 6 is the same as the wave in Fig. 3 except for an inconsequential difference in the pulse shapes.

The idle-channel output is calculated by setting $g = 0$ in equation (19):

$$\nu_0(t) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \frac{\sigma}{\pi i l} \exp\left[\pi i l \vartheta_0 + 2\pi i\left(\frac{l}{2}(1 - \vartheta) + k\right)f_s t\right] \qquad (20)$$

which describes a collection of discrete lines. Figure 7 shows a number of these lines; the symmetry of the spectrum about all integral multiples of f_N is apparent.

For any given value of l, there is only one value of k which leads to a line in the Nyquist interval $(-f_N, f_N)$. This line, of frequency f_l, can be defined: Let

$$Q(\alpha) = \alpha - N(\alpha) \qquad (21)$$

where

$$N(\alpha) = \text{integer nearest } \alpha.$$
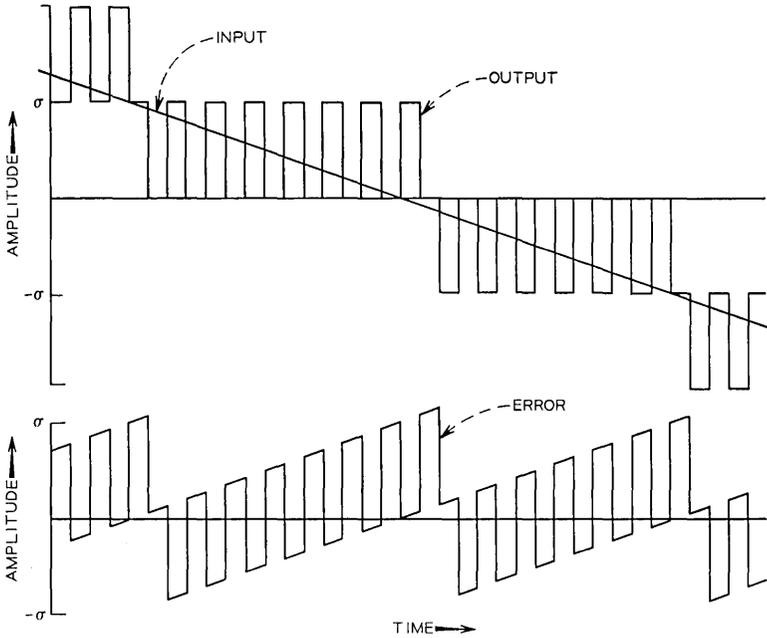
Fig. 6 — Input, output, and error waves for a (negative) ramp input to a symmetrical coder.

Then

$$f_l = Q\left(\frac{l(1 - \vartheta)}{2}\right)f_s .$$ (22)

Ignoring lines outside the Nyquist interval, and combining terms of $+l$ and $-l$,

$$\nu_{0N}(t) = \sum_{l=1}^{\infty} \frac{2\sigma}{\pi l} \sin (\pi l \vartheta_0 + 2\pi f_l t).$$ (23)

If we now think of the spectrum as one-sided, we have a collection of lines of frequency

$$f = |\, f_l \,|$$ (24)

and power

$$P_l = \frac{2\sigma^2}{\pi^2 l^2}.$$ (25)

These lines will subsequently be referred to as "main lines," "original lines," "carriers," or "l-lines" (2-line, 5-line, and so on).

Figure 8 is an example of the spectrum, bringing out some of the important qualitative features. One can see that the terms for which $l = 2, 4$, and so on, are the components of the sawtooth, of peak-to-peak amplitude σ and fundamental frequency ϑf_s, evident in Fig. 3. If we choose those values of l which equal aN, where a is a positive integer and N is the odd integer nearest $1/\vartheta$, we have the components of another sawtooth of peak-to-peak amplitude $2\sigma/N \approx 2\epsilon$ and fundamental frequency $|1 - N\vartheta| f_s/2$ (Wang's "second envelope function"). In Fig. 8, $N = 19$.

Notice that either the $(N - 2)$-line or the $(N + 2)$-line has a frequency equal to that of the 2-line minus that of the N-line and a power about equal to that of the N-line (the 21-line in Fig. 8). This line may also be thought of as the fundamental of a sawtooth, as may all lines at the lower end of the spectrum.

There is another interesting way of looking at the idle-channel noise spectrum. Recalling equations (15) and (17), it is apparent that $\nu_0(t)$ can be thought of as the result of sampling, at a rate f_s, the wave

$$x_0(t) = \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp \{\pi i l[\vartheta_0 + (1 - \vartheta)f_s t]\} \tag{26}$$

which describes a sawtooth of peak-to-peak amplitude 2σ and funda-
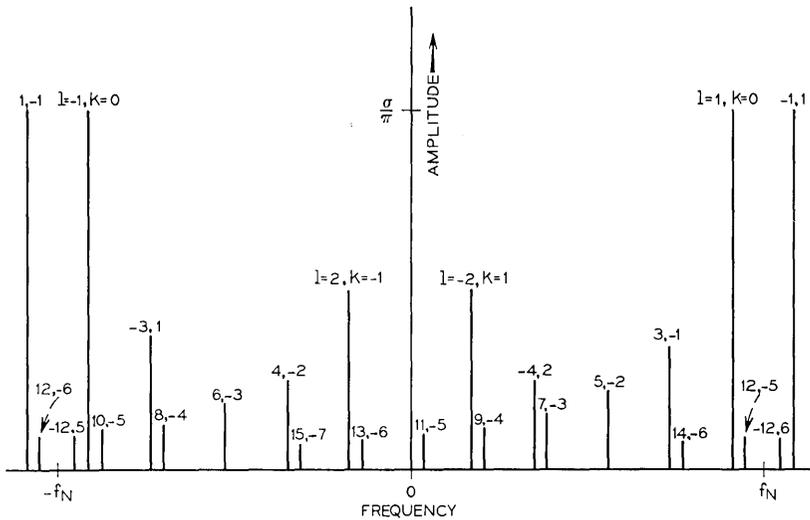
Fig. 7 — Example of an idle-channel noise spectrum. All lines for $|l| = 1, 2, 3, 4$, and 12 are given; notice their symmetries. Selected other lines are included to show the progressions involved.
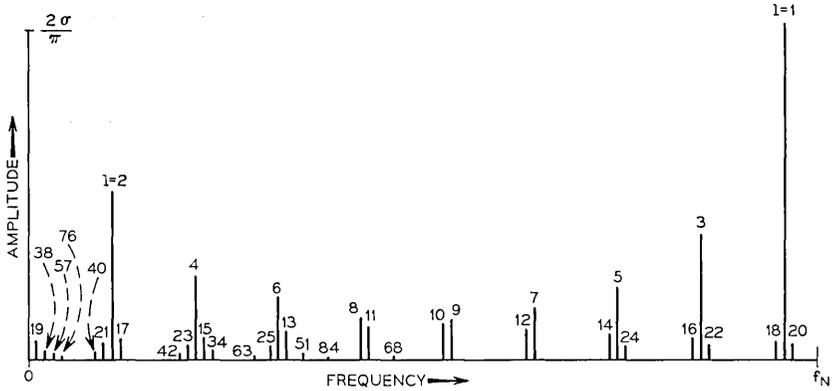
Fig. 8 — Example of a one-sided idle-channel noise spectrum, for $\vartheta = 5/96$. The most powerful 25 lines are shown along with selected others, in particular the harmonics of the 17-line, 19-line, and 21-line. These and the 2-line plus its harmonics make up sawtooth waves.

mental frequency $(1 - \vartheta)f_N$. In Fig. 9 this sawtooth is superimposed on the wave of Fig. 3.

All the l-lines are distinct if ϑ is irrational, which is expected to be the normal case. Appendix B treats rational ϑ.

V. SINUSOIDAL INPUT

Let us calculate the noise output of a $\triangle M$ coder for a pure sinusoidal input, setting

$$g(t) = A \sin (2\pi f_0 t + \varphi), \tag{27}$$

where φ is an arbitrary constant phase angle. We put equation (27)
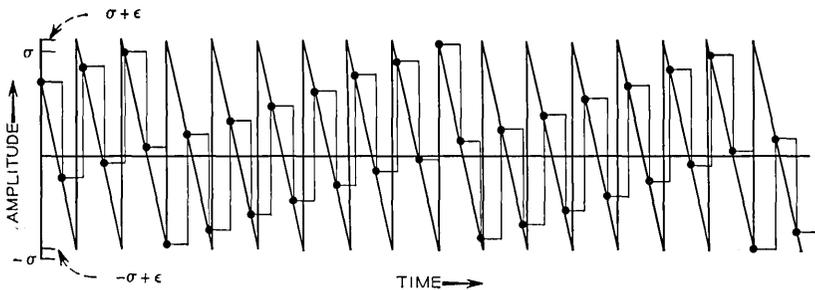
Fig. 9 — Error wave of Fig. 3 with superimposed sawtooth. The heavy dots are the sampling points. Section II explains the vertical offset (ϵ).

into equation (19) and make use of the Jacobi-Anger formula:[6]

$$\exp\left[\pi ilA\sin\left(2\pi f_0 t + \varphi\right)\right] = \sum_{m=-\infty}^{\infty} J_m(\pi lA)\exp\left(2\pi i m f_0 t + im\varphi\right), \quad (28)$$

where the J_m are the Bessel functions of integral order of the first kind. The result is

$$\nu_{\sin}(t) = \sum_{l\neq 0}\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\frac{\sigma}{\pi il}\,J_m(\pi lA)$$

$$\cdot\exp\left(\pi il\vartheta_0 + im\varphi\right)\exp\left\{2\pi i\left[\left(\frac{l(1-\vartheta)}{2}+k\right)f_s + mf_0\right]t\right\}. \quad (29)$$

If we define

$$f_l' \equiv \left(\frac{l(1-\vartheta)}{2}+k(l,\,m)\right)f_s\,, \quad (30)$$

where $k(l,\,m)$ is chosen so that $f_l' + mf_0$ is in the Nyquist interval, we can write (for this interval)

$$\nu_{N\sin}(t) = \sum_{l=1}^{\infty}\sum_{m=-\infty}^{\infty}\frac{2\sigma}{\pi l}\,J_m(\pi lA)\sin\left[2\pi(f_l' + mf_0)t + \pi l\vartheta_0 + m\varphi\right]. \quad (31)$$

Equation (31) describes a collection of lines consisting of the original lines of the idle-channel noise spectrum (or their replicas, $f_l + kf_s$), each with a set of uniformly spaced ($\pm mf_0$) satellites. The total power in an l-group (all the lines governed by the index l) is constant. $J_0^2(\pi lA)$ of the power remains in the main line; the mth satellite gets $J_m^2(\pi lA)$ of the total power. From the symmetry of the spectrum one can see that for every l-line satellite which falls outside the Nyquist interval there is a corresponding satellite in the Nyquist interval arising from a carrier outside the interval.

The nature of the Bessel function is such that main lines go through a series of peaks and nulls as a function of A for a given l, and l for a given A. The satellites, in addition, fluctuate in amplitude as a function of the index m. As a result, for the typical case of a signal band which is a small fraction of the Nyquist interval and for an input frequency of the same order as the signal bandwidth, the signal-band noise power is determined by relatively few lines and can be expected to fluctuate irregularly as a function of input amplitude and frequency.

$J_m(z)$ goes fairly rapidly to zero as a function of m for $m > z$. Thus the full width of an l-group is

$$\Delta f \approx 2\pi lAf_0\,. \quad (32)$$
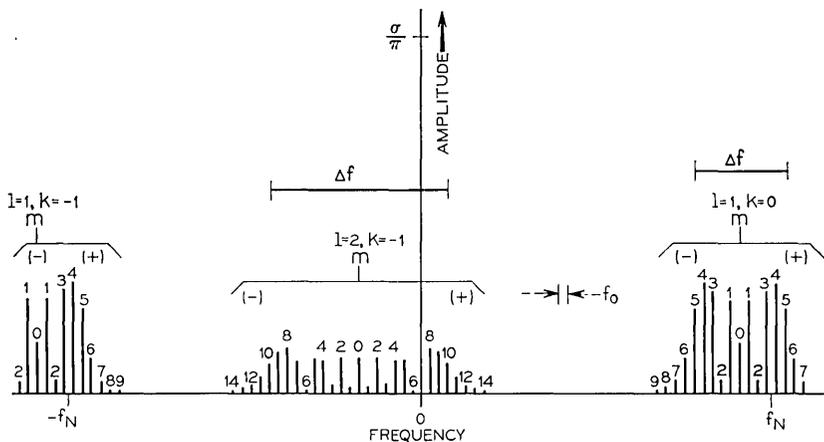
Fig. 10 — The $l = 1$ and $l = 2$ lines of the spectrum of Fig. 7, modulated by a sinusoidal input with $f_0 \approx f_N/40$ and $\pi A = 5$. The other lines are omitted for the sake of clarity. (The lack of symmetry in this figure, and in Fig. 11, is due to the omission of the image groups: $l = -1, -2$.) Equation (32) gives $\triangle f$.

Figure 10 gives an example of the spectrum which attempts to bring out the points made above. If f_0 and f_s are not rationally related, the lines are all distinct. Appendix C treats the case of rational f_0/f_s.

VI. BROADBAND INPUT

As first discussed by Bennett, the average noise performance of a coder in the presence of a broadband input signal is best calculated by using an input signal of random phase.[7] This test signal should have the same power spectrum as the input signal under consideration. Appendix D gives the mathematical manipulations.

Briefly the procedure is to calculate the noise power spectrum, $W(f)$, by finding the Fourier transform of the autocorrelation function, $R(\tau)$, of the noise wave. The averaging procedure in the definition of $R(\tau)$ is carried out assuming the input, $g(t)$, is a gaussian variate. This gives $R(\tau)$ in terms of the autocorrelation coefficients, a_k, and the mean power of the sequence of input samples; we determine the a_k from the Fourier transform of the input power spectrum, $U(f)$. We show that, under an assumption that usually holds in practice, the dependence on $U(f)$ reduces to a dependence on the rms time derivative of the input. A parameter S, which is this time derivative normalized to the average maximum slope of the coder, σf_s, characterizes the input in the following (equivalent) formulas:

$$W(f) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \frac{\tau_s \sigma^2}{\pi^2 l^2} \exp\left[-\frac{(\pi l k S)^2}{2} + 2\pi i k\left(f\tau_s + \frac{l(1-\vartheta)}{2}\right)\right], \quad (33)$$

$$W(f) = \sum_{l \neq 0} \sum_{p=-\infty}^{\infty} \frac{2^{1/2}\tau_s \sigma^2}{\pi^{5/2} |l|^3 S} \exp\left[-\frac{2}{(lS)^2}\left(f\tau_s + \frac{l(1-\vartheta)}{2} + p\right)^2\right].$$

$$(34)$$

Terms of different l do not interact in equations (33) and (34); therefore, one may use either formula to calculate the power density for a given l. Equation (33) converges faster for high values of l; equation (34) for low values. The crossover occurs at $l \approx 1/(\pi)^{\frac{1}{2}}S$.

One can see [most easily from (34)] that the spectrum consists of the lines given in Section IV for the idle-channel noise spectrum, each now broadened to a gaussian. Notice that some of the power in the wings of each gaussian falls outside $(-f_N, f_N)$. Conversely, lines centered outside this band have in-band wings. The total in-band power of each l-group is constant.

One can easily see that the full width of an l-group is

$$\Delta f = lSf_s. \quad (35)$$

Thus for $lS \ll 1$ one has a relatively sharp line, while all groups for which $lS \gtrsim 1$ sum to a white background. Figure 11 shows the important qualitative features of the spectrum.

As S approaches one, equations (33) and (34) lose their usefulness because of the onset of so-called slope-overload noise. (Strictly speaking equations (8) and (9) do not apply under overload conditions; but because the errors resulting from overload and quantization are prob-
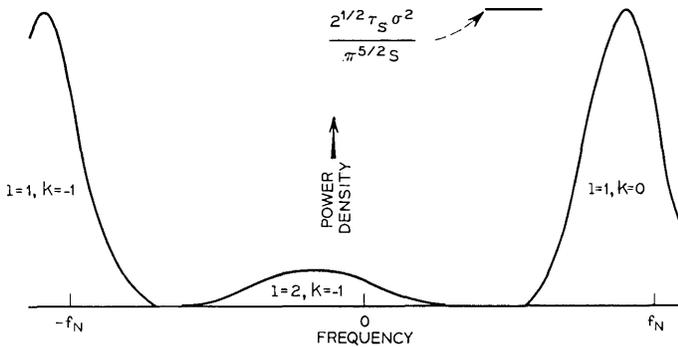
Fig. 11 — The $l = 1$ and $l = 2$ lines of the spectrum of Fig. 7, modulated by a gaussian input with $S \approx \frac{1}{8}$. The other lines are omitted for the sake of clarity.

ably largely uncorrelated, one should be able to calculate them separately with reasonable precision.) To give some idea of the effect, we quote Protonotarios' signal-to-overload-noise ratios for various input spectra: 3 to 17 dB for $S = 1$ and 16 to 31 dB for $S = \frac{1}{2}$.[8] These values are lower limits, and probably poor approximations for high-quality voice systems, because the total noise was used. Nevertheless, it seems safe to say that the occurrence of slope overload will prohibit inputs strong enough to whiten the 1-group and, usually, the 2-group as well.

Figure 12 illustrates calculated noise spectra of a coder suitable for telephone applications.[1] Notice the enormous differences in power in the voice band for different ϑ's. For $\vartheta = 0.02$ the 2-line, 4-line, and 6-line centered at 30.88, 61.76, and 92.64 kHz, respectively, can be seen. The broad line centered at 38.6 kHz for $\vartheta = 0.05$ and $S = 2^{-10}$ is the sum of the 19- and 21-lines. Although the spectra are white for $S = 2^{-2}$ in the frequency range shown, they are not independent of S. The 2-line is still spreading out and the 1-line is just starting to spread in.

Figure 13 presents the results of Fig. 12 in the form of noise power in dBrnC versus speech input power in dBm. The unit "dBrnC" means dB above one picowatt of integrated noise passing through a filter with C-message weighting.[9] Briefly, this filter, which weights noise according to its subjective effect in a telephone circuit, has a pass band with a transmission averaging about -0.5 dB from ≈ 800 to ≈ 2500 Hz; the noise bandwidth is ≈ 2070 Hz.

The parameter S is turned into speech power as follows: de Jager (see p. 447 of Ref. 2) showed that the ratio of the rms slope of the average speech spectrum[10] to the rms amplitude is given by

$$r \equiv \left(\frac{\langle \dot{y}^2 \rangle}{\langle y^2 \rangle} \right)^{\frac{1}{2}} \approx 2\pi \cdot 800 \text{ Hz} \approx 5000 \text{ rad per s.} \tag{36}$$

Thus, speech power is given by

$$P_{sp} = \left(\frac{S \sigma f_s}{r} \right)^2. \tag{37}$$

In Fig. 13, this quantity is plotted in units of dB above 1 mW. The structure in Fig. 13 is best interpreted by referring to Fig. 12.

VII. SUMMARY AND REMARKS

We have developed a $\triangle$M quantizing-noise formalism for the case of unequal positive and negative integrator step sizes and have given the noise spectrum for zero, sinusoidal, and gaussian coder inputs.
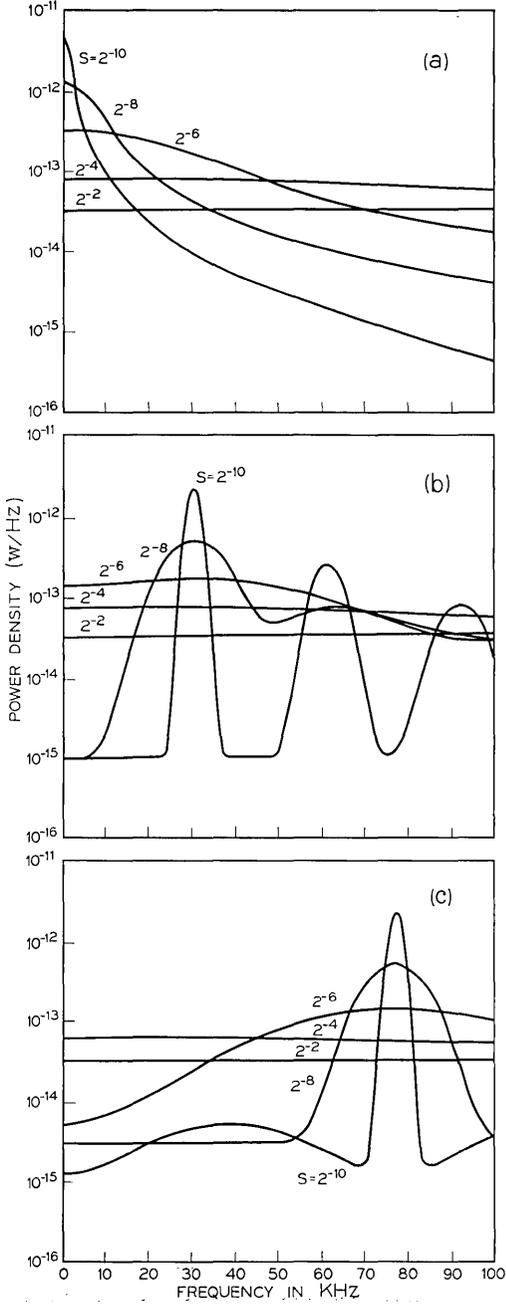
Fig. 12 — Partial one-sided noise spectra for gaussian inputs with $\sigma = 0.4$ mV/$\Omega^{1/2}$ (12-mV steps in 900Ω), $f_s = 1.544$ MHz, and various values of ϑ and S. (a) $\vartheta = 0$, (b) $\vartheta = 0.02$, and (c) $\vartheta = 0.05$.
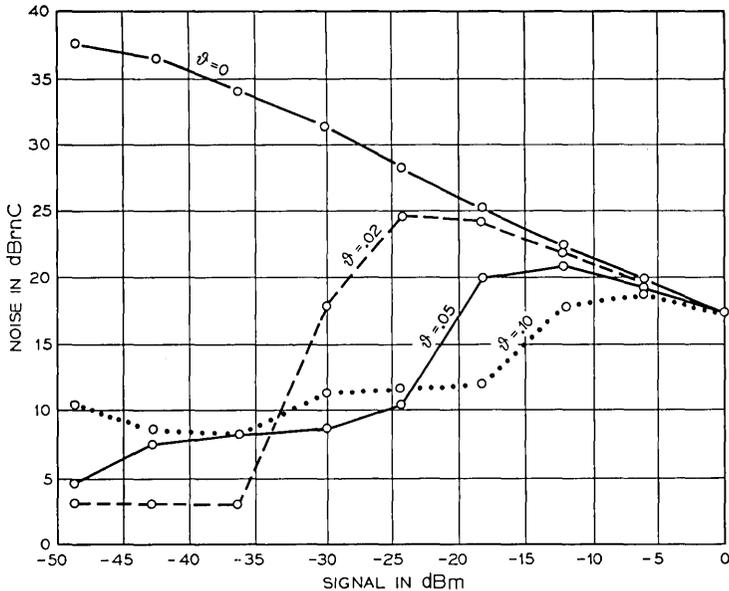
Fig. 13 — Noise versus average speech power for the coder of Fig. 12. See text for explanation of units. The circles are the calculated points; the lines are only to connect points of the same parameter.

In ΔM systems, as contrasted with multibit PCM and DPCM, the signal typically does not change more than a small fraction of a step size in one sampling interval. As a result, the sample-to-sample errors are strongly correlated and the noise spectrum is highly colored. The main contributions of this paper are to point out that the spectral distribution of power is strongly dependent on the step unbalance and to provide a means of calculating the spectrum precisely.

A typical ΔM system has a signal bandwidth very much smaller than the Nyquist bandwidth. The consequences of this situation for the idle channel (zero input) are best seen by referring to Figs. 7 and 8. There are extreme system-noise variations depending on whether or not the system parameters are such as to bring into the signal band one of the stronger spectral lines. The $\mid l \mid = 2$-line, which has ≈ 15 percent of the total Nyquist-interval power, is especially important in this regard.

Coder inputs phase-modulate the idle-channel lines; the frequency breadth of the sideband structure is proportional to the rms slope (roughly, root power times frequency) of the input. Thus, as power is

increased, there may be an abrupt increase in noise as the sidebands of a strong line come into the signal band. Figure 13 illustrates such a situation.

At very high input powers most of the idle-channel lines are broadened to the point where they make an easily calculable white contribution to the spectrum. Unfortunately, the most powerful lines ($|l| = 1$ with 61 percent of the Nyquist-interval power, $|l| = 2$, and so on) can be broadened to whiteness only by inputs powerful enough to force the coder into slope overload. It is possible, however, to minimize noise in a given system by dithering, that is, the deliberate injection of certain appropriate signals into the coder (including the judicious choice of step unbalance). Dithering requires extensive treatment and will be the subject of a future paper.

VIII. ACKNOWLEDGMENT

APPENDIX A

Correlation of Error and Input

The error wave, $v(t)$, consists in general of a part fully correlated with the input, $y(t)$, and an uncorrelated part. The uncorrelated part is noise in almost any conceivable system; whether the fully correlated part is considered noise or not depends on the use to which the system is put. Let us investigate the correlation by forming the cross-correlation function of v and g (assuming zero mean for each):

$$R_{vo}(\tau) \equiv \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} v(t)g(t + \tau) \, dt. \tag{38}$$

Inserting equation (19) gives

$$R_{vo}(\tau) = \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp(\pi i l \vartheta_0) \lim_{T \to \infty} \frac{1}{2T}$$

$$\cdot \int_{-T}^{T} g(t + \tau) e^{\pi i l g(t)} \sum_{k=-\infty}^{\infty} \exp\left[2\pi i\left(\frac{l(1 - \vartheta)}{2} + k\right)f_s t\right] dt. \tag{39}$$

The integral in equation (39) is zero unless $g(t)$ contains components locked to the idle-channel-noise frequencies (Section IV) or their subharmonics. Thus for typical $\triangle$M systems

$$R_{\nu_g}(\tau) = 0 \quad \text{for all} \quad \tau, \tag{40}$$

and $\nu(t)$ is a noise wave under any circumstances.

This conclusion is not applicable to the case of rational ϑ, treated in Appendix B, where $l(1 - \vartheta)/2 + k = 0$ for some values of l and k; that is, some of the idle-channel-noise frequencies are zero. The most extreme case is that of PCM ($\vartheta = 1$) where every value of l contributes a dc term to the summation in equation (39) and the summation is therefore replaced by one. Let us calculate the two parts of $\nu(t)$ for this case.

It is easy to show that $R_{\nu_g}(\tau)$ is a maximum for $\tau = 0$ and that we need consider only instantaneous correlations. This result is physically reasonable when one considers that no delay from input to output was introduced in the formulation of $\nu(t)$. If we let $\langle \ \ \rangle_{\mathrm{av}}$ stand for the integrating-limiting (averaging) process defined in equation (38) we can write

$$\langle \nu g \rangle_{\mathrm{av}} \equiv R_{\nu_g}(0). \tag{41}$$

The correlated part of $\nu(t)$ is $\alpha y(t)$, where α is a constant for a given $y(t)$. The uncorrelated part is then $\nu(t) - \alpha y(t)$, and the condition for determining α is

$$\langle (\nu - \alpha y)y \rangle_{\mathrm{av}} = 0, \tag{42}$$

or

$$\alpha_1 = \frac{\langle \nu y \rangle_{\mathrm{av}}}{\langle y^2 \rangle_{\mathrm{av}}} = \frac{\langle \nu g \rangle_{\mathrm{av}}}{\sigma \langle g^2 \rangle_{\mathrm{av}}}. \tag{43}$$

Substitution of a specific input waveform into

$$\langle \nu g \rangle_{\mathrm{av}} = \sum_{l \neq 0} \frac{\sigma}{\pi i l} \exp{(\pi i l \vartheta_0)}\langle g(t)e^{\pi i l g(t)} \rangle_{\mathrm{av}} \tag{44}$$

will show that αy is generally negligible compared with ν unless $\langle g^2 \rangle_{\mathrm{av}} < 1$; that is, the signal-to-noise ratio is low. This conclusion is quite plausible because of the $\langle g^2 \rangle^{-1}$ dependence of α and because the oscillatory character of the second factor in the averaging bracket in equation (44) makes the bracket tend toward zero as g increases.

For rational values of $\vartheta \neq 1$, the summation in equation (44) is only over multiples of an integer L (as shown in Appendix B) and αy is negligible for even smaller values of $\langle g^2 \rangle$. $L = 2$ for the van de Weg case, $\vartheta = 0$.

APPENDIX B

Rational Step Unbalance

If the fractional step unbalance, ϑ, is a rational number, the l-lines of Section IV are not all distinct. Indeed, if L is the least positive integer for which $(L/2)(1 - \vartheta)$ is an integer, it is easy to see from equation (22) that

$$f_{l+L} = f_l \quad \text{and} \quad f_{L-l} = -f_l . \tag{45}$$

Let us sum up terms of frequency $\pm f_l$ in equation (23), ignoring for the moment the cases $l = L$ and $l = L/2$ (if it exists). Then

$$v_{ONl}(t) = \sum_{l'=0}^{\infty} \frac{2\sigma}{\pi(l + l'L)} \sin \left[\pi(l + l'L)\vartheta_0 + 2\pi f_l t\right]$$

$$+ \sum_{l'=0}^{\infty} \frac{2\sigma}{\pi(L - l + l'L)} \sin \left[\pi(L - l + l'L) - 2\pi f_l t\right]. \tag{46}$$

A little manipulation of the indices in the second summation gives

$$v_{ONl}(t) = \sum_{l'=-\infty}^{\infty} \frac{2\sigma}{\pi(l + l'L)} \sin \left[\pi(l + l'L)\vartheta_0 + 2\pi f_l t\right] \tag{47}$$

$$= \sum_{l'=-\infty}^{\infty} \frac{\sigma}{\pi i(l + l'L)} \exp \left\{i[\pi(l + l'L)\vartheta_0 + 2\pi f_l t]\right\} + \text{c.c.}^*. \tag{48}$$

From Jolley's series Nos. 534 and 535 it is easily established that[11]

$$\sum_{n=-\infty}^{\infty} \frac{e^{in\psi}}{a + n} = \pi \csc a\pi e^{ia(\pi - \psi)} \quad \text{for} \quad 0 < \psi < 2\pi. \tag{49}$$

Using this to do the sum in equation (48),

$$v_{ONl}(t) = \frac{\sigma \csc \left(\dfrac{l}{L}\pi\right)}{iL} \exp \left[i\left(\frac{l}{L}\pi - \pi l\vartheta_1 + \pi l\vartheta_0 + 2\pi f_l t\right)\right] + \text{c.c.} \tag{50}$$

$$= \frac{2\sigma \csc \left(\dfrac{l}{L}\pi\right)}{L} \sin \left(\pi \frac{l}{L} - \pi l\vartheta_1 + \pi l\vartheta_0 + 2\pi f_l t\right), \tag{51}$$

where

$$L\vartheta_1 = \text{least positive quantity} \equiv L\vartheta_0 \text{ mod 2.} \tag{52}$$

---

* By c.c. we mean complex conjugate.

The power at $f = |f_l|$ is thus

$$P_l = \frac{2\sigma^2 \csc^2 \left(\frac{l}{L}\pi\right)}{L^2}. \tag{53}$$

Let us compare this with the sum of the powers of the same lines as given by equation (25):

$$P'_l = \sum_{l'=0}^{\infty} \left(\frac{2\sigma^2}{\pi^2(l + l'L)^2} + \frac{2\sigma^2}{\pi^2(L - l + l'L)^2}\right). \tag{54}$$

Again, index manipulation yields

$$P'_l = \sum_{l'=-\infty}^{\infty} \frac{2\sigma^2}{\pi^2(l + l'L)^2}. \tag{55}$$

This series is easily summed by means of cotangent residues (see Section 7.4-4 of Ref. 6) to give

$$P'_l = \frac{2\sigma^2 \csc^2 \left(\frac{l}{L}\pi\right)}{L^2}, \tag{56}$$

which is identical to equation (53). Thus, all lines with the same frequency, f, where $0 < f < f_N$, are phased such that their powers add. As a result one need not take line degeneracies into special account when considering the noise power spectrum. We assume, without proof, that this statement is true of sidebands as well as main lines; it is elementary that the power in a sine wave is not changed when its phase is modulated.

If L is even there is a line at $f_{L/2} = f_N$, on the border between the Nyquist interval and higher frequencies. If one starts from equation (20) rather than equation (23), so that the higher frequencies are taken into account, equation (51) with $l = L/2$ will result.

For $l = l' L$ we get $f_l = 0$, that is, a dc component. Summing up these terms of equation (23),

$$\nu_{ONL}(t) = \sum_{l'=1}^{\infty} \frac{2\sigma}{\pi l'L} \sin \pi l'L\vartheta_0. \tag{57}$$

Starting from equation (49), subtracting the $n = 0$ term $(1/a)$ from both sides, letting $a \to 0$, and combining terms of $+n$ and $-n$, gives

$$\sum_{n=1}^{\infty} \frac{\sin n\psi}{n} = \frac{\pi - \psi}{2} \quad \text{for} \quad 0 < \psi < 2\pi. \tag{58}$$

Using equation (58) in equation (57) gives

$$\nu_{ONL} = \frac{\sigma}{L}(1 - L\vartheta_1), \tag{59}$$

where ϑ_1 is defined in equation (52). Thus, the dc component fluctuates as a function of ϑ_1 or ϑ_0. If it is averaged uniformly over this parameter, the mean value is zero and the mean square is $\sigma^2/3L^2$. The latter is the sum of the powers of the $l'L$-lines. The dc power varies from zero to σ^2/L^2. Thus, for rational ϑ, the total noise power in the Nyquist interval (counting one-half the power at f_N) is not always $\sigma^2/3$ but can vary from this total by $-\sigma^2/3L^2$, $+2\sigma^2/3L^2$. For small ϑ and correspondingly large L this variation is not very significant. In any case, as Section III explains, the usual ΔM system suppresses dc at the decoder.

If the $l'L$-lines are modulated, we have

$$\nu_{NL}(t) = \sum_{l'=1}^{\infty} \frac{2\sigma}{\pi l'L} \sin \pi l'L[\vartheta_0 + g(t)] \tag{60}$$

$$= \frac{\sigma}{L}[1 - L\vartheta_1(t)], \tag{61}$$

where $\vartheta_1(t)$ is defined by using $\vartheta_0 + g(t)$ in equation (52) in place of ϑ_0. If the excursions of $g(t)$ are significantly greater than $1/L$ ($\langle g^2 \rangle \gg 1/L^2$, which covers nearly all cases of practical interest) equation (61) can be time-averaged uniformly over $\vartheta_1(t)$. As stated above, the mean square will be $\sigma^2/3L^2$. That is, for a practical input signal, the power in the $l'L$-lines is dispersed into sidebands, and the total power in these sidebands is equal to that which would be calculated using the formulas for irrational ϑ. This argument is used to justify the assumption that, except for dc power, the noise spectrum for rational ϑ can be calculated as indicated in the text for irrational ϑ.

There are two cases of rational ϑ which are of special interest. One is $\vartheta = 0$, calculated for gaussian input by van de Weg[3]. In this case all the even-l lines are centered at zero frequency and all the odd-l lines at f_N. As one can see from Section VI and Figs. 12 and 13, this is a good approximation only for a baseband with a width much greater than ϑf_s.

The other case of interest is $\vartheta = 1$. This is equivalent to using the even-instant law of equation (8) for all sampling instants, which is equivalent to ordinary uniform PCM (with a step size of 2σ). Let us consider a typical PCM speech system, for which (see Section VI)

$$\langle \dot{g}^2 \rangle = \langle g^2 \rangle \cdot (2\pi \cdot 800 \text{ Hz})^2 \tag{62}$$

and

$$f_s = 8 \text{ kHz}. \tag{63}$$

Then (Appendix D)

$$S \approx 0.63 \langle g^2 \rangle^{\frac{1}{3}}. \tag{64}$$

Useful input signals are many steps high in amplitude ($\langle g^2 \rangle \gg 1$); thus $S > 1$ and the noise spectrum is substantially white. (See Section VI. This range of S is permissible for PCM, since slope overload does not occur.)

If $\vartheta = 1$ is inserted into equation (31) the result can be shown to be equivalent to that of Schouten and van't Groenewout for a sinusoidal input into a PCM coder if one allows for:[12] (*i*) their nonunity shape factor, (*ii*) their particular choice of phase (φ), (*iii*) replacing the last sine factor in their expression (17) by a cosine, and (*iv*) multiplying their expressions (15), (16), and (17) by 2.

APPENDIX C

Rational Input-to-Sampling Frequency Ratio

If the ratio of f_0 to f_s is a rational fraction, there exists a least positive integer M for which Mf_0/f_s is integral. In this case it is easily seen, from equation (29), that terms for which the values of m differ by a multiple of M have the same Nyquist-interval frequency. Summing up these equal-frequency terms, we replace equation (29) with

$$v_{\sin}(t) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \sum_{m=1}^{M} \frac{\sigma}{\pi i l} \sum_{m'=-\infty}^{\infty} J_{m+m'M}(\pi l A) \exp\left[i(m + m'M)\varphi\right]$$
$$\cdot \exp\left\{\pi i l \vartheta_0 + 2\pi i \left[\left(\frac{l(1 - \vartheta)}{2} + k\right)f_s + mf_0\right]t\right\}. \tag{65}$$

One can therefore see that a given l-group consists of a total of M lines, the original and $M-1$ satellites spaced uniformly throughout $(-f_N, f_N)$.

The sum over m' in equation (65), which is the relative amplitude coefficient of a satellite [$B_m(z, \varphi, M)$, where $z = \pi l A$] can be turned into a finite sum:

It is easy to verify that

$$\frac{1}{M} \sum_{n=1}^{M} \exp\left(\frac{2\pi i n(m - p)}{M}\right) = 1 \quad \text{if} \quad p = m + m'M$$
$$= 0 \quad \text{otherwise}, \tag{66}$$

where m, m', M, and p are integers.

Thus

$$B_m(z, \varphi, M) \equiv \sum_{m'=-\infty}^{\infty} J_{m+m'M}(z) \exp \left[i(m + m'M)\varphi \right] \tag{67}$$

$$= \sum_{p=-\infty}^{\infty} J_p(z)e^{ip\varphi} \left[\frac{1}{M} \sum_{n=1}^{M} \exp \left(\frac{2\pi i n(m - p)}{M} \right) \right] \tag{68}$$

$$= \frac{1}{M} \sum_{n=1}^{M} \sum_{p=-\infty}^{\infty} J_p(z) \exp \left[ip\left(\varphi - \frac{2\pi n}{M} \right) \right] \exp \left(\frac{2\pi i n m}{M} \right). \tag{69}$$

The sum over p is given by equation (28). Thus

$$B_m(z, \varphi, M) = \frac{1}{M} \sum_{n=1}^{M} \exp \left[iz \sin \left(\varphi - \frac{2\pi n}{M} \right) + \frac{2\pi i n m}{M} \right]. \tag{70}$$

We note the dependence on the phase (φ) of the input signal, which indicates that this result could not have been obtained by adding the powers of the equal-frequency terms. Indeed, it can be established easily from equation (67) that

$$\frac{1}{2\pi} \int_0^{2\pi} |B_m(z, \varphi, M)|^2 \, d\varphi = \sum_{m'=-\infty}^{\infty} J_{m+m'M}^2(z); \tag{71}$$

that is, the sum of the term powers is given by the true satellite power averaged uniformly over φ.

In order to emphasize this dependence on phase, let us examine the highly artificial but simplest nontrivial case, $f_0 = f_N$. In this case $M = 2$ and each main line has one satellite spaced f_N away. Then

$$B_0(z, \varphi, 2) = B_2(z, \varphi, 2) = \cos (z \sin \varphi), \tag{72}$$

and

$$B_1(z, \varphi, 2) = i \sin (z \sin \varphi). \tag{73}$$

For $\varphi = 0$, the satellite power is always zero and we get the undisturbed idle-channel-noise spectrum. That this is to be expected can be seen from equation (27) where $\varphi = 0$ is the condition under which the sampling instants fall precisely on the zeroes of the input wave. For $\varphi = \pi/2$, where the sampling instants fall on the crests, the power for any given l oscillates between main and satellite as a function of amplitude. This result may be compared with the incorrect one obtained by summing powers:

$$\frac{1}{2\pi} \int_0^{2\pi} |B_0(z, \varphi, 2)|^2 \, d\varphi = \frac{1 + J_0(2z)}{2}, \tag{74}$$

and

$$\frac{1}{2\pi} \int_0^\pi |B_1(z, \varphi, 2)|^2 \, d\varphi = \frac{1 - J_0(2z)}{2}.$$ (75)

APPENDIX D

Calculations for Broadband Input

The autocorrelation function of $\nu(t)$ [defined in equation (14)] is given by

$$R(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \left(\sum_{n=-\infty}^{\infty} x_n\tau_s \, \delta(t - n\tau_s) \right)$$

$$\cdot \left(\sum_{m=-\infty}^{\infty} x_m\tau_s \, \delta(t - m\tau_s + \tau) \right) dt$$ (76)

$$= \lim_{T \to \infty} \frac{\tau_s^2}{2T} \sum_{n \geq -T/\tau_s}^{\leq T/\tau_s} \sum_{m \geq (-T+\tau)/\tau_s}^{\leq (T+\tau)/\tau_s} x_n x_m \, \delta[\tau + (n - m)\tau_s].$$ (77)

We can replace T/τ_s with a positive integer N without loss of generality. Let us concentrate on values of τ lying between $(k - 1/2)\tau_s$ and $(k + 1/2)\tau_s$ where k is an integer. Only terms for which $n-m = -k$ fall within this interval. Thus

$$R(\tau) = \tau_s \lim_{N \to \infty} \frac{1}{2N} \sum_{n=-N}^{N} x_n x_{n+k} \, \delta(\tau - k\tau_s)$$

$$\text{for} \quad (k - \tfrac{1}{2})\tau_s < \tau \leq (k + \tfrac{1}{2})\tau_s \, .$$ (78)

Defining

$$\langle x_n x_{n+k} \rangle \equiv \lim_{N \to \infty} \frac{1}{2N} \sum_{n=-N}^{N} x_n x_{n+k} \, ,$$ (79)

and joining together the segments of the function given by equation (78), we have

$$R(\tau) = \tau_s \sum_{k=-\infty}^{\infty} \langle x_n x_{n+k} \rangle \, \delta(\tau - k\tau_s).$$ (80)

The Fourier transform of $R(\tau)$ is the noise-power spectral density, first given by Bennett (see pp. 460–464 of Ref. 7):

$$W(f) = \tau_s \sum_{k=-\infty}^{\infty} \langle x_n x_{n+k} \rangle \exp(2\pi i k f \tau_s).$$ (81)

It is easy to see that

$$\int_{-f_N}^{f_N} W(f) \, df = \langle x_n x_n \rangle \equiv \langle x^2 \rangle;$$ (82)

that is, the total noise power in the Nyquist interval is given by the mean square of the sequence $\{x_n\}$.

Next, let us connect $\langle x_n x_{n+k} \rangle$ with the input signal. Using equation (12)

$$\langle x_n x_{n+k} \rangle = \left\langle \sum_{l \neq 0} \sum_{\lambda \neq 0} - \frac{\sigma^2}{\pi^2 l \lambda} \right.$$

$$\left. \cdot \exp \{\pi i[(l + \lambda)\vartheta_0 + (l + \lambda)n(1 - \vartheta) + lk(1 - \vartheta) + lg_{n+k} + \lambda g_n]\} \right\rangle.$$

(83)

We carry the averaging bracket inside the summations and examine the various factors of the exponential. Since n and g_n are uncorrelated, the factors containing them can be averaged separately. Let us examine the factor

$$\langle e^{\pi i (l + \lambda) n (1 - \vartheta)} \rangle.$$

For irrational ϑ this expression is zero unless $l + \lambda = 0$, in which case its value is one. Thus equation (83) reduces to

$$\langle x_n x_{n+k} \rangle = \sum_{l \neq 0} \frac{\sigma^2}{\pi^2 l^2} e^{\pi i l k (1 - \vartheta)} \langle \exp [\pi i l (g_{n+k} - g_n)] \rangle.$$ (84)

Notice that (see p. 66 of Ref. 12)

$$\langle x^2 \rangle = \sum_{l \neq 0} \frac{\sigma^2}{\pi^2 l^2} = \frac{\sigma^2}{3}.$$ (85)

Let us now find the value of the averaging bracket in equation (84) for a gaussian input:

$$\langle \exp [\pi i l (g_{n+k} - g_n)] \rangle$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp [\pi i l (g_{n+k} - g_n)] P(g_{n+k}, g_n) \, dg_{n+k} \, dg_n,$$ (86)

where $P(g_{n+k}, g_n)$ is the joint probability density of two gaussian variates, which is (see Section 18.8–6 of Ref. 6):

$$P(g_{n+k}, g_n) = \frac{1}{2\pi \langle g^2 \rangle (1 - a_k^2)^{\frac{1}{2}}} \cdot \exp \left(-\frac{g_n^2 - 2a_k g_n g_{n+k} + g_{n+k}^2}{2\langle g^2 \rangle (1 - a_k^2)} \right),$$ (87)

where

$$a_k = \frac{\langle g_n g_{n+k} \rangle}{\langle g^2 \rangle}, \tag{88}$$

$$\langle g_n g_{n+k} \rangle \equiv \lim_{N \to 0} \frac{1}{2N} \sum_{m=-N}^{N} g_n g_{n+k}, \tag{89}$$

and

$$\langle g^2 \rangle \equiv \langle g_n g_n \rangle. \tag{90}$$

Combining equations (86) and (87) gives

$$\langle \exp [\pi i l (g_{n+k} - g_n)] \rangle = \exp [-\pi^2 l^2 \langle g^2 \rangle (1 - a_k)], \tag{91}$$

a result first obtained by Rice.[13] Substituting equation (91) into equation (84), and the result into equation (81), gives

$$W(f) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \frac{\tau_s \sigma^2}{\pi^2 l^2}$$

$$\cdot \exp \left[-\pi^2 l^2 \langle g^2 \rangle (1 - a_k) + 2\pi i k \left(f\tau_s + \frac{l(1 - \vartheta)}{2} \right) \right]. \tag{92}$$

This result, with $\vartheta = 0$, was given by van de Weg.[3] He also used as an input a flat signal, band-limited to $(-f_m, f_m)$, for which

$$a_k = \frac{\sin (2\pi k f_m \tau_s)}{2\pi k f_m \tau_s}, \tag{93}$$

and inserted

$$1 - a_k \approx \frac{(2\pi k f_m \tau_s)^2}{6}. \tag{94}$$

This approximation is made possible by observing that the real exponential factor in equation (92) is appreciable only for small values of the exponent. Thus, if one ignores the region of low signal-to-noise ratios (that is, small $\langle g^2 \rangle$), $1 - a_k$ need only be accurately approximated for small values. $\langle g^2 \rangle > 0.1$ is high enough for the approximation to be good for most purposes, and appears to cover nearly all cases of practical interest.

It is not necessary for the input spectrum to be flat. We take a spectrum, $U(f)$, even in f and confined to the Nyquist interval. Appendix E shows that $\langle g_n g_{n+k} \rangle$ is given by $R_g(k\tau_s)$, the autocorrelation function of $g(t)$. This autocorrelation function is the Fourier trans-

form of the spectral density.

$$\langle g_n g_{n+k} \rangle = \int_{-\infty}^{\infty} U(f) \exp\left(2\pi i k f \tau_s\right) df \tag{95}$$

and [using the evenness of $U(f)$]

$$\langle g^2 \rangle (1 - a_k) = \int_{-\infty}^{\infty} U(f)(1 - \cos 2\pi k f \tau_s) \, df. \tag{96}$$

Using the reasoning given in the previous paragraph, and specifying that $U(f)$ be a smooth function of frequency (free of strong narrow-band components which could make $1 - a_k \approx 0$ for isolated high values of k), gives

$$\langle g^2 \rangle (1 - a_k) \approx \frac{k^2}{2f_s^2} \int_{-\infty}^{\infty} (2\pi f)^2 U(f) \, df. \tag{97}$$

It is well known that the integral in equation (97) gives the mean square of the time derivative of $g(\equiv \dot{g})$. Thus

$$\langle g^2 \rangle (1 - a_k) \approx \frac{k^2 \langle \dot{g}^2 \rangle}{2f_s^2} = \frac{k^2 \langle \dot{y}^2 \rangle}{2\sigma^2 f_s^2} \equiv \frac{k^2 S^2}{2} , \tag{98}$$

where S is the rms time slope of the input normalized to the maximum average slope of the coder (σf_s). Inserting equation (98) into equation (92) gives

$$W(f) = \sum_{l \neq 0} \sum_{k=-\infty}^{\infty} \frac{\tau_s \sigma^2}{\pi^2 l^2} \exp\left[-\frac{(\pi l k S)^2}{2} + 2\pi i k \left(f \tau_s + \frac{l(1 - \vartheta)}{2} \right) \right] \tag{99}$$

which is equation (33).

Making use of the Fourier-series expansion of a picket fence of gaussians,

$$\sum_{p=-\infty}^{\infty} \exp\left[-\alpha^2 (x - p x_0)^2 \right] = \sum_{k=-\infty}^{\infty} \frac{\pi^{\frac{1}{2}}}{|\alpha| x_0} \exp\left[-\left(\frac{\pi k}{\alpha x_0} \right)^2 + \frac{2\pi i k x}{x_0} \right], \tag{100}$$

we can rewrite equation (99) as

$$W(f) = \sum_{l \neq 0} \sum_{p=-\infty}^{\infty} \frac{2^{1/2} \tau_s \sigma^2}{\pi^{5/2} |l|^3 S} \exp\left[-\frac{2}{(lS)^2} \left(f \tau_s + \frac{l(1 - \vartheta)}{2} + p \right)^2 \right] \tag{101}$$

which is equation (34).

Computationally more convenient versions of equations (99) and (101) are, in one-sided form,

$$P(f) = \frac{8\tau_s\sigma^2}{\pi^2}$$

$$\cdot \left\{ \sum_{l=1}^{\infty} \frac{1}{l^2} \left[\frac{1}{2} + \sum_{k=1}^{\infty} \exp\left(-\frac{(\pi lkS)^2}{2} \right) \cos \pi lk(1 - \vartheta) \cos 2\pi kf\tau_s \right] \right\},$$

(102)

and

$$P(f) = \frac{2^{3/2}\tau_s\sigma^2}{\pi^{5/2}S} \sum_{l=1}^{\infty} \left(\frac{1}{|l|^3} \sum_{p=-\infty}^{\infty} \left\{ \exp\left[-\frac{2}{(lS)^2} \left(f\tau_s + p + \frac{l(1-\vartheta)}{2} \right)^2 \right] \right. \right.$$

$$\left. \left. + \exp\left[-\frac{2}{(lS)^2} \left(f\tau_s + p - \frac{l(1-\vartheta)}{2} \right)^2 \right] \right\} \right).$$

(103)

In equation (102), $1/2 \sum(1/l^2)$ is left in that form in order that the power density can be calculated separately for each l.

APPENDIX E

Autocorrelation—Function and Its Samples

In the absence of aliasing, the autocorrelation coefficients of a sequence of samples of a function are equal to the appropriate samples of the autocorrelation of the function:

$$\langle g_n g_{n+k} \rangle = \lim_{N\to\infty} \frac{1}{2N} \sum_{n=-N}^{N} g_n g_{n+k}$$

(104)

$$= \lim_{N\to\infty} \frac{1}{2N} \int_{-N\tau_s}^{N\tau_s} \sum_{n=-\infty}^{\infty} g(n\tau_s)g(n\tau_s + k\tau_s)\, \delta(t - n\tau_s)\, dt$$

(105)

$$= \lim_{N\to\infty} \frac{1}{2N} \int_{-N\tau_s}^{N\tau_s} g(t)g(t + k\tau_s) \sum_{n=-\infty}^{\infty} \delta(t - n\tau_s)\, dt$$

(106)

$$= \lim_{N\to\infty} \frac{1}{2N\tau_s} \int_{-N\tau_s}^{N\tau_s} g(t)g(t + k\tau_s) \sum_{p=-\infty}^{\infty} \exp(2\pi ipf_s t)\, dt$$

(107)

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} g(t)g(t + k\tau_s)\, dt$$

(108)

$$\equiv R_g(k\tau_s)$$

(109)

where the transition from equation (107) to equation (108) is made by assuming that $g(t)$ is confined to the Nyquist interval. For any given

value of $p \neq 0$, equation (107) can be regarded as the correlation function of $g(t)$ and $g(t)$ exp $(2\pi i p f_s t)$. The latter represents a carrier wave at $f = p f_s$ amplitude-modulated by $g(t)$. None of the sidebands resulting from this modulation overlap $g(t)$ in frequency if $g(t)$ is confined to the Nyquist interval. It is well known that two signals are uncorrelated if their frequency bands do not overlap (but not, in general, otherwise).[14]

Dividing equation (109) by $\langle g^2 \rangle = R_g(0)$ gives the lemma in normalized form. The procedure given here is a slight generalization of one given by Bennett for $k = 0$ (see pp. 468–469 of Ref. 5).

REFERENCES

1. Laane, R. R., and Murphy, B. T., unpublished work. There have been other recent expressions of interest in $\triangle$M-coded speech, namely, Tomozawa, A., and Kaneko, H., "Companded Delta Modulation for Telephone Transmission," IEEE Trans. Commun. Techniques, *16*, No. 2 (February 1968), pp. 149–157, and Brolin, S. J., and Brown, J. M., "Companded Delta Modulation for Telephony," IEEE Trans. Commun. Techniques, *16*, No. 2 (February 1968), pp. 157–162.
2. de Jager, F., "Deltamodulation, A Method of PCM Transmission Using the 1-Unit Code," Philips Res. Rep. 7, 1952, pp. 442–466.
3. van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep. *8*, 1953, pp. 367–385.
4. Wang, P. P., "Idle Channel Noise of Delta Modulation," IEEE Trans. Commun. Techniques, *16*, No. 10 (October 1968), pp. 737–742.
5. Rice, S. O., quoted in Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J., *27*, No. 3 (July 1948), p. 466.
6. Korn, G. A., and Korn, T. M., *Mathematical Handbook for Scientists and Engineers*, Section 21.8-4, New York: McGraw-Hill, 1961.
7. Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J., *27*, No. 3 (July 1948), p. 450.
8. Protonotarios, E. N., "Slope Overload Noise in Differential Pulse Code Modulation Systems," B.S.T.J., *46*, No. 9 (November 1967), pp. 2119–2161.
9. Cochran, W. T., and Lewinski, D. A., "A New Measuring Set for Message Circuit Noise," B.S.T.J., *39*, No. 4 (July 1960), p. 916.
10. French, N. R., and Steinberg, J. C., "Factors Governing the Intelligibility of Speech Sounds," J. Acoust. Soc. Amer., *19*, No. 1 (January 1947), p. 94.
11. Jolley, L. B. W., *Summation of Series*, New York: Dover Publications, 1961, p. 100.
12. Schouten, J. P., and van't Groenewout, H. W. F., "Analysis of Distortion in Pulse-Code Modulation Systems," Appl. Sci. Res., *B2*, (1951-52), pp. 277–290.
13. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *24*, No. 1 (January 1945), p. 51.
14. Bennett, W. R., *Electrical Noise*, McGraw-Hill Book Company, New York, New York, 1960, pp. 209–210.

Computer Study of Quantizer Output Spectra

By G. H. ROBERTSON

(Manuscript received January 25, 1969)

This article describes a method for accurately calculating the output spectrum of a quantizer. The method was developed for known expressions defining the output spectrum of an arbitrary quantizer with gaussian input of arbitrary bandshape. Results obtained for a variety of conditions, however, suggest that the calculations are valid even though the input has only a minor gaussian component. When sampling is also used, at the Nyquist rate or a little higher, the quantizing noise folded into the input band is almost flat even when the input bandshape is sharply peaked. When interference at the input is increased, the quantizer (preceded by AGC) appears to operate like an increasingly noisy linear transducer up to a breaking point beyond which its performance (for small signals) degrades rapidly and becomes difficult to analyze.

I. INTRODUCTION

Several authors have described formulas for calculating the output noise spectrum from a quantizer when the input is a gaussian waveform. References 1 through 4 are characteristic and contain representative bibliographies. Evaluation of the resulting expressions is difficult because they contain multiple infinite sums of terms containing Hermite polynomials whose order increases without limit. Consequently simplifying assumptions are made about the input spectrum and quantizer characteristics, or only a few terms are evaluated and the rest assumed negligible, to get results.

This article describes a more fruitful approach in which the Hermite polynomials are evaluated in conjunction with other parts of the expression such that the combination tends to zero as the order increases to infinity. The convergence is slow and many terms are needed to get sufficient accuracy in the noise spectrum. It is possible to get results even when the quantizer is not linear or symmetrical, and for arbitrary input spectrum shapes.

For quantizing steps no greater than σ (the rms gaussian component) an interesting and useful result is that even when the input spectrum is sharply peaked, if the quantized waveform is also sampled uniformly at up to a few times the Nyquist rate for the input band, the resulting quantizing noise appearing within the input band is nearly flat. Many systems can therefore be evaluated quite accurately with much simpler calculations than those needed to define the quantizing noise spectrum.

Study of quantizers having uniform steps less than σ in amplitude show the output spectrum to be practically independent of the location of the gaussian mean if it is at least σ from the overload limit. Consequently, added signals (whose waveform defines the gaussian mean) have a negligible effect on the quantizer output noise as long as they do not approach within σ of the limit. A quantizer with many steps activated thus produces a noise spectrum virtually independent of relatively large signals added to the gaussian component.

II. DEFINITION OF QUANTIZER

Figure 1 shows the transfer characteristic of the quantizer where the "staircase" relates the output voltage (ordinate) scale to the input voltage (abscissa) scale. Assuming that the input waveform is gaussian about some arbitrary mean value, the probability that it is Z or more above the mean value is

$$Q_Z = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_Z^\infty \exp\left(-t^2/2\sigma^2\right) dt. \tag{1}$$

Figure 1 shows that when the input waveform reaches a "riser" of the staircase, the output waveform changes abruptly from the value on one tread to the value of the one on the other side of the riser. For convenience, number the treads and risers starting with 1 at the left. There is one more tread than the number of risers, so if the last riser is k, the last tread is $k + 1$. Let Q_r be the probability that the input waveform is greater than riser r, and the output voltage of step r be W_r. The mean value of the output is

$$S = W_1(1 - Q_1) + W_2(Q_1 - Q_2) + \cdots + W_{k+1}Q_k. \tag{2}$$

The mean squared value is

$$V^2 = W_1^2(1 - Q_1) + W_2^2(Q_1 - Q_2) + \cdots + W_{k+1}^2 Q_k. \tag{3}$$

The variance of the output is

$$V^2 - S^2 = P. \tag{4}$$

Assuming unity impedance, P is the output power after subtracting the component caused by the displacement of the mean value of the input waveform from zero.

III. QUANTIZING NOISE SPECTRUM

Velichkin showed that the correlation function of the quantizer output can be written[4]

$$R_y(\tau) = \sum_{n=1}^{\infty} \left[\sum_{k=1}^{\nu-1} \Delta_k \exp\left(-a_k^2/2\sigma^2\right) H_{n-1}\left(\frac{a_k}{\sigma}\right) \right]^2 \frac{R_x^n(\tau)}{2\pi\sigma^{2n}n!}. \tag{5}$$

$R_x(\tau)$ is the input correlation function, σ^2 is the input variance, there are ν treads, Δ_k is the output voltage difference between treads $k + 1$ and k, a_k is the input voltage at riser k, and $H_r(z)$ is the Hermite polynomial

$$H_r(z) = (-1)^r \exp\left(z^2/2\right) \frac{d^r}{dz^r} \left[\exp\left(-z^2/2\right)\right]. \tag{6}$$

Also, where $[r/2]$ is the greatest integer $\leq r/2$,[5]
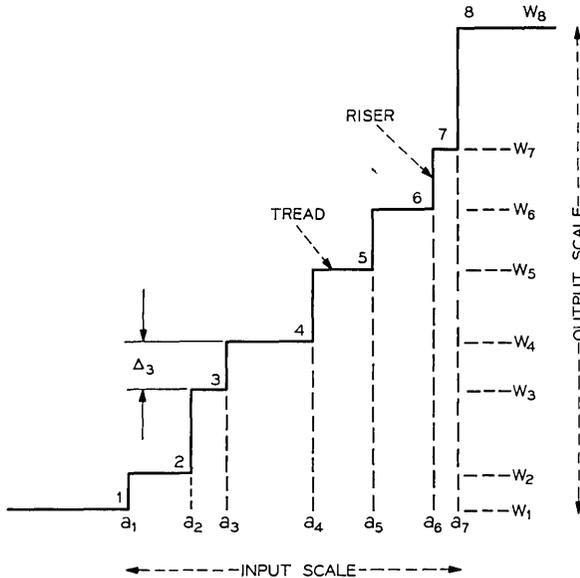
Fig. 1 — Quantizer transfer characteristics.

$$H_r(z) = \sum_{j=0}^{[r/2]} (-1/2)^j z^{r-2j} \frac{r!}{j! \, (r - 2j)!}. \tag{7}$$

By the Wiener-Khinchine theorem the power spectrum of the quantizer output is

$$\Omega(f) = 4 \int_0^\infty R_y(\tau) \cos (2\pi f \tau) \, d\tau$$

$$= \frac{2}{\pi} \sum_{n=1}^\infty \left\{ \sum_{k=1}^{\nu-1} \Delta_k \exp (-a_k^2/2\sigma^2) \frac{H_{n-1}\left(\dfrac{a_k}{\sigma}\right)}{[(n-1)!]^{\frac{1}{2}}} \right\}^2 \frac{1}{n\sigma^{2n}}$$

$$\cdot \int_0^\infty R_x^n(\tau) \cos (2\pi f \tau) \, d\tau. \tag{8}$$

Equation (8) can be written

$$\Omega(f) = \sum_{n=1}^\infty \frac{F_n}{\sigma^{2n}} 4 \int_0^\infty R_x^n(\tau) \cos (2\pi f \tau) \, d\tau, \tag{9}$$

in which the quantizing factor terms F_n depend only on the properties of the quantizer and n. When $n = 1$ the component $\Omega_o(f)$ is the input spectrum multiplied by F_1/σ^2. All other n give components whose bandwidth exceeds that of the input [because the integral in equation (9) then represents multiple convolutions of the input band], and their sum $\Omega_e(f)$ may be called the quantizer error spectrum. The quantizer output spectrum is

$$\Omega(f) = \Omega_o(f) + \Omega_e(f). \tag{10}$$

So far only amplitude quantizing has been considered. Sampling, at a rate f_s, is generally also used,[*] and the output spectrum becomes proportional to[1]

$$\Omega_s(f) = \Omega(f) + \sum_{n=1}^\infty \Omega(nf_s \pm f). \tag{11}$$

If f_s is at least twice the highest frequency of the input band, only Ω_e can add more noise by fold-over into the range of the input band.

These results are all known but now follow what are thought to be new contributions enabling equation (9) to be evaluated for an arbitrary choice of input spectrum shape. Equation (9) can be written

---

[*] The result is independent of which is done first.

$$\Omega(f) = \sum_{n=1}^{\infty} \frac{F_n}{\sigma^{2n}} 2 \int_{-\infty}^{\infty} R_x^n(\tau) e^{-i2\pi f\tau} \, d\tau$$

$$= \sum_{n=1}^{\infty} \frac{F_n}{\sigma^{2n}} 2C_{n-1}[s(f)/2], \tag{12}$$

where $C_{n-1}[s(f)/2]$ is the $(n-1)$th convolution of the cisoid power spectrum $s(f)/2$. The sinusoid power spectrum is $s(f)$, corresponding to the autocorrelation function $R_x(\tau)$.

The significance of equation (12) is that whereas the direct calculation of $R_x^n(\tau)$ may be impossible for an arbitrary spectrum shape, $C_{n-1}[s(f)/2]$ can always be calculated if $s(f)$ is defined. Appendix A describes the methods used to calculate $C_{n-1}[s(f)/2]$ in the computer program written to evaluate equation (9).

If $G(t)$ represents the input waveform, the autocorrelation function at zero lag is

$$R_x(0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} G(t)^2 \, dt$$

$$= \sigma^2 + S^2, \tag{13}$$

where S is the mean value of the input waveform and σ^2 is the variance as used in equation (1). Figure 1 shows that the value of the quantizer output for a given input waveform is independent of the scale on the input axis. For convenience, relabel this scale so that the input mean is zero. Consequently,

$$R_x(0) = \sigma^2. \tag{14}$$

Normalizing the input power that now contains no dc, so that $\sigma^2 = 1$, gives

$$R_x^n(0) = 1 \tag{15}$$

for all n. By the Wiener-Khinchine theorem the total output power P_T is

$$P_T = \int_0^{\infty} \Omega(f) \, df$$

$$= R_y(0)$$

$$= \frac{1}{2\pi} \sum_{n=1}^{\infty} \frac{1}{n} \left[\sum_{k=1}^{\nu-1} \Delta_k \exp\left(-a_k^2/2\right) \frac{H_{n-1}(a_k)}{[(n-1)!]^{\frac{1}{2}}} \right]^2 \tag{16}$$

using equation (5). P_T is the same as V^2 given by equation (3) so the

accuracy of computing the quantizing factor terms F_n can be checked by computing the total power by both these methods. The method of equation (3) gives high accuracy very easily, but the F_n terms are needed to compute the quantizer error spectrum $\Omega_e(f)$ of equation (10). Appendix B describes the methods used to compute F_n for values of n up to 10,000, the limit used in the program.

Recall [after equation (9)] that when $n = 1$ the resulting component of $\Omega(f)$ is the input spectrum multiplied by F_1/σ^2, where the gain factor is

$$F_1 = \frac{1}{2\pi} \left[\sum_{k=1}^{\nu-1} \Delta_k \exp\left(-a_k^2/2\sigma^2\right) \right]^2 . \tag{17}$$

When $\sigma^2 = 1$, the total quantizer error power is

$$P_E = P - F_1 , \tag{18}$$

where P is given in equation (4). Both P and F_1 can be computed easily and accurately, so P_E can be determined accurately with little computational effort. Note that this shows P_E to be independent of the input spectrum shape.

A computer program, using the techniques described in Appendixes A and B to compute $\Omega(f)$, simulated the effect of sampling (without holding) by pivoting $\Omega(f)$ about the sampling frequency and its harmonics, and computing the contributions thus folded into the original band. The total P_E is folded into a bandwidth equal to half the sampling frequency; and when the latter was less than a few times the Nyquist rate for the input band, the level of the error component resulting from P_E was nearly flat over the input band even when the input spectrum was sharply peaked.

This result is very useful because the performance of quantizers can now be evaluated quite accurately using only the simple calculations indicated by equations (4) and (17). The error spectrum after sampling was flatter when more levels were used in the quantizer.

IV. SIGNALS ADDED TO INPUT

Signals added to the gaussian noise at the input cause the mean value of the latter to vary according to the signal waveform. Computation shows that under static conditions the gain factor F_1 and the total error power P_E remain nearly constant when the step size is about σ and the mean is no closer than σ to the overload limit. Under these conditions the position of the mean has negligible effect on the shape of the quantizing noise spectrum. Assuming a signal wave-

form uncorrelated with the gaussian noise, and of a magnitude such that the mean rarely approaches within σ of the overload limit, it is thus quite accurate to assume that the quantizing error is independent of the signal when the step-to-σ ratio is constant and less than unity. A sampling rate, up to a few times the high end of the input band, further improves the accuracy of this assumption as the quantizing noise then becomes almost flat across the input band even when the input spectrum is sharply peaked.

Assume now that an AGC unit is used to maintain constant power into the quantizer so that the waveform representing the sum of the gaussian component and large signal (interference) very rarely exceeds the overload limits. As the level of the interference increases, the ratio of quantizer step to gaussian rms (rms_g) also increases. Assuming no correlation between the interference and gaussian components, the degradation from quantizing noise can be estimated from the way the parameters F_1 and P_E vary with the position of the mean. The greatest variation in these parameters occurs between the values when the mean is at a riser (see Fig. 1) and when it is midway between risers.

Figures 2 and 3 show the results obtained for a 16-level quantizer with a flat input spectrum and with a sharply peaked input spectrum, respectively, in calculations carried out for these limiting cases. Up to a breaking point (where the two curves diverge) the quantizer appears to act like a linear but noisy transducer for input signals. Note that the breaking point seems to be independent of the spectrum shape. When the interference level is high enough to cause operation beyond the breaking point, the spectrum becomes difficult to analyze and depends on the interference waveform. At all points on the abscissas of Figs. 2 and 3 below the breaking point, F_1 and P_E were found virtually constant for all positions likely to be occupied by the input mean (determined by the AGC unit). Since the quantizing noise level was flat it was therefore proportional to P_E. The input copy was proportional to F_1; the curves in Figs. 2 and 3 show the ratio of the level of input copy plus quantizing noise to the level of the input copy alone. The degradation these curves indicate, as the interference increases, results from the decreasing ratio of σ to quantizing step size caused by the AGC unit preceding the quantizer.

V. COMPARISON WITH MEASUREMENTS

A sharply peaked spectrum was produced in the laboratory by filtering the output of a noise generator, and the resulting waveform
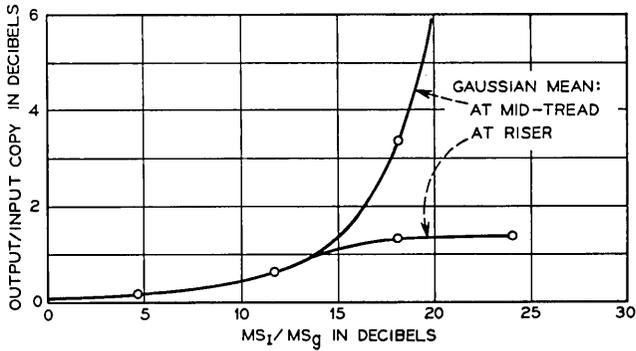
Fig. 2 — Quantizer performance (static) versus relative interference (flat input) for 16-level quantizer with flat input spectrum (gaussian); overload at $3 \times (MS_g + MS_I)^{1/2}$ set by AGC; MS_g = gaussian component $(rms)^2$; MS_I = interference $(rms)^2$; output sampled at $3 \times$ high end.

was radically clipped before being submitted to a spectrum analyzer. A 1910-A recording wave analyzer (made by General Radio Company) was used, and several successive traces were superimposed by the recorder as the narrowband (10 Hz) filter was slowly swept across the spectrum. The spectrum before and after clipping were determined in this way; the final results were obtained by drawing a smooth curve through the mean of the superimposed traces. Figure 4, where the solid
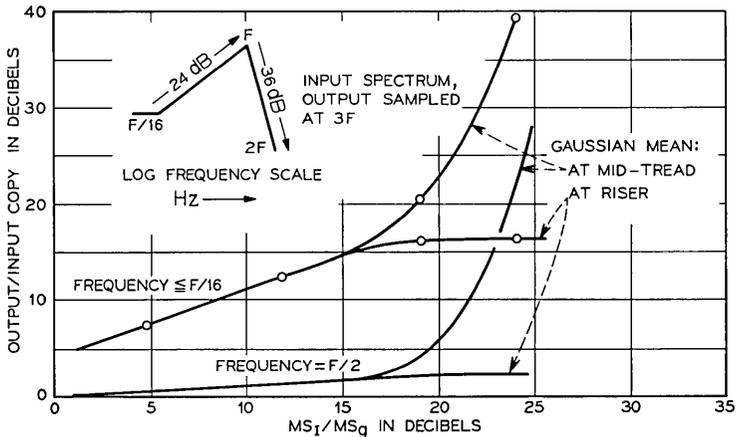
Fig. 3 — Quantizer performance (static) versus relative interference (peaked input) for 16-level quantizer with peaked input spectrum (gaussian); overload at $3 \times (MS_I + MS_g)^{1/2}$ set by AGC.

curve is the computed clipper output spectrum when a copy of the input is given by the dashed curve, shows the results. Values of the measured output spectrum appear as circles and agree well with the computed curve.

Another check between computed and measured results can be obtained for a uniform step 16-level quantizer. A band of noise, nearly flat from zero to about 330 Hz and falling rapidly at higher frequencies, is added to a sinewave at 160 Hz and passes through an AGC unit before quantization. The quantizer overload limit is set near four times the rms value of the AGC output, and the results are recorded on a magnetic tape for various ratios of the sinewave-to-noise power. In this capacity the sine wave acts as an interfering signal. A computer program processes the tape using a version of the east Fourier transform algorithm to produce estimates of the spectrum level at the quantizer output up to half the sampling rate of 1024 Hz.[6] Since the input spectrum level at 500 Hz is much lower than in the flat part below 300 Hz, the increase in noise level estimated at 500 Hz is taken as a measure of the quantizing noise introduced as the interfering signal increases. Assuming this noise to be flat from 512 Hz to zero it is possible to estimate the degradation in signal-to-noise power suffered by a small signal in the flat part of the input band.

Figure 5 shows the results, as circles superimposed on the solid curves, which are computed for a 16-level quantizer sampled at three times the high end of an input band of noise flat to zero frequency. The quantizer is preceded by an AGC unit and its overload is four times the rms input.
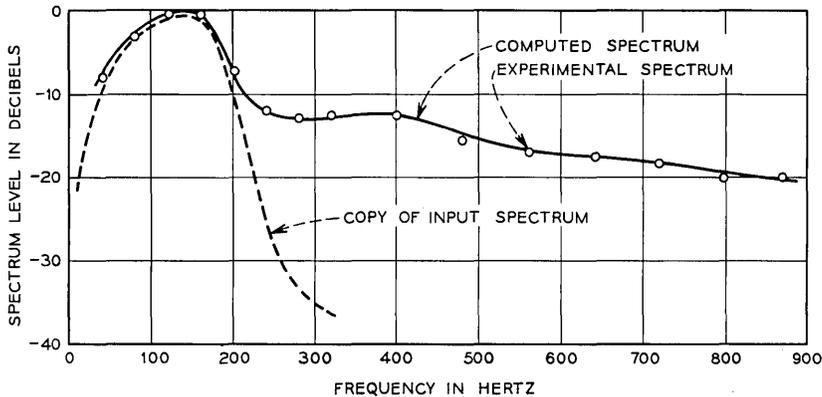
Fig. 4 — Clipper output spectrum.
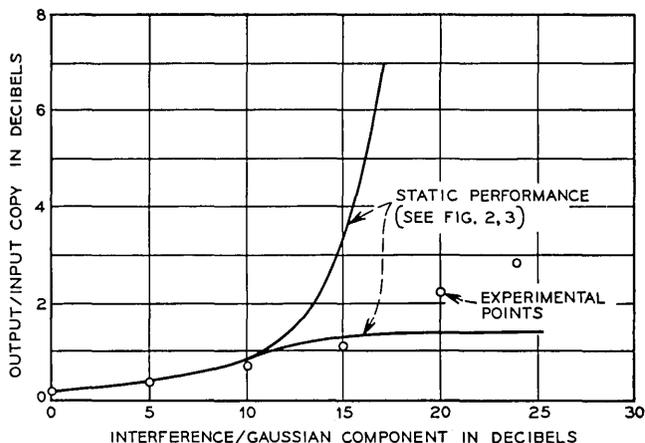
Fig. 5 — Quantizer degradation for flat input spectrum for 16-level quantizer; overload = 4 × input (rms) set by AGC; flat input gaussian component (0 to 330 Hz); sinewave interference (160 Hz); output sampled at 1024 Hz.

VI. CONCLUSION

This article describes a new method of calculating the quantizing noise spectrum when gaussian noise with arbitrary spectrum shape is applied to an arbitrary quantizer. The novelty is not in the form of the expressions that describe the noise spectrum but in the techniques used to compute the results. Applying the method to a sharply peaked spectrum shows that if the output is sampled at the Nyquist rate, or a little higher, the quantizing noise is folded back to cover the input band with almost uniform intensity. A clipper (2-level quantizer) and a 16-level quantizer, preceded by AGC to keep the overload at three times the rms input, operate like noisy but linear transducers for added signals of power less than one tenth and less than twenty times, respectively, that of the broadband background. These useful results indicate that the performance of quantizers under such conditions can be evaluated without the lengthy computations required to delineate the quantizing noise spectrum.

APPENDIX A

Calculating the Input Spectrum Convolutions

The input cisoid spectrum is defined and convolved with itself to calculate $C_{n-1}[s(f)/2]$ when n is small. Because the input spectrum is of finite width, the convolutions tend to take the form of a gaussian

distribution as n increases. Since direct computation of the convolutions becomes very lengthy when n is large, it is profitable to compute a Gram-Charlier approximation instead (see pp. 257–260 of Ref. 5.) This can be done if the moments for the desired convolution can be obtained. The input cisoid spectrum is symmetrical about zero, and is defined up to its limiting bandwidth, so all the moments desired can be computed for it. If the input spectrum shape is normalized so that it covers unit area and it is considered to define a probability distribution from which random samples are drawn, the nth convolution is the same as the probability distribution of $(n + 1)$ independent samples of the original distribution.[7] The moments of the nth convolution can thus be obtained from the moments of the input spectrum shape as follows. Since we desire ultimately standardized central moments, note that the standardized central moments for the sum and for the average of N independent samples are the same. Using the appropriate multinomial expansion the general term for the νth such moment is

$$M_\nu = \left(\frac{1}{N^\nu}\right)^{\frac{1}{2}} \sum \left(\frac{\mu_p}{p!}\right)^i \left(\frac{\mu_q}{q!}\right)^j \left(\frac{\mu_r}{r!}\right)^k \frac{\nu! \, N(N-1) \cdots (N-J)}{i! \, j! \, k!} , \qquad (19)$$

where

$$\nu = ip + jq + kr, \qquad (20)$$

the right side being a partition of ν, and*

$$J = i + j + k - 1.$$

The sum is taken over all the partitions of ν except those containing unity (because the first central moment is zero). The term μ_p is the pth standardized central moment of the original distribution. A program was developed to compute such moments; but since the computation rapidly becomes very lengthy when ν increases, the number of moments used to get the Gram-Charlier approximation was reduced as the convolution order increased. This can be done without undue sacrifice in accuracy since the distribution tends to become gaussian with increasing convolution order.

APPENDIX B

Calculation of Quantizer Factor Terms F_n

Equation (8) shows that F_n requires computation of terms like

---

* A partition of ν is a set of positive integers whose sum is ν. The terms i, j, k, p, q, and r are integers.

$$F_{kn} = \exp\,(-x_k^2/2)H_{n-1}(x_k)/[(n-1)\,!]^{\frac{1}{2}}, \qquad (21)$$

where $H_r(x)$ is a Hermite polynomial for which the recurrence relation exists[8]

$$H_{r+1}(x) = xH_r(x) - rH_{r-1}(x). \qquad (22)$$

Therefore

$$F_{k(n+1)} = x_k F_{kn}/n^{\frac{1}{2}} - F_{k(n-1)}[(n-1)/n]^{\frac{1}{2}}. \qquad (23)$$

Since $H_o(x) = 1$ and $H_1(x) = x$, from equation (21)

$$F_{k1} = \exp\,(-x_k^2/2) \qquad (24)$$

and

$$F_{k2} = x_k F_{k1}\,. \qquad (25)$$

Therefore, by using equations (23), (24), and (25), a straightforward method exists for finding any F_{kn}. When values are to be calculated using the same x_k and many successive values of n, the programming can be simplified by saving the computed values for n and $(n-1)$ to be used in equation (23) when the value for $(n+1)$ is desired. Taking advantage of this way of arranging the computations values were computed for n up to 10,000, enabling determination of the quantizing noise level at greater than 100 times the input bandwidth for a 16-level quantizer. Since recurrence relations like that in equation (23) sometimes result in rapid loss of accuracy, a few values of F_{kn} were computed by an independent method, for high values of n.

Hermite polynomials can be evaluated in terms of confluent hypergeometric functions;[8] a suitable asymptotic formula for these

TABLE I—VALUES OF F_{kn}

x_k	n	Recurrence Relation	Asymptotic Formula
1.0	9999	0.3540125940E-01	0.35401259262E-01
1.0	10,000	0.60060871554397E-01	0.60060871554399E-01
1.0	10,001	−0.34798910623E-01	−0.34798910644E-01
2.0	9999	0.28830572153E-01	0.28830572171E-01
2.0	10,000	0.16057291188981E-01	0.16057291188984E-01
2.0	10,001	−0.28508000965E-01	−0.28508000983E-01
10.0	9,999	−0.62935376617E-12	−0.629353766E-12
10.0	10,000	0.1037121050651E-12	0.1037121050655E-12
10.0	10,001	0.73302922069E-12	0.73302922115E-12

functions was obtained in Ref. 9. Although the asymptotic formula would give adequate accuracy when n is large, the recurrence relation permits much faster evaluations when values are needed over a large range of n. Table I compares a few values of F_{kn} calculated by the recurrence relation and the asymptotic formula. Very good agreement is obtained justifying the use of the recurrence relation.

REFERENCES

1. Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J., *27*, No. 4 (July 1948), pp. 446–472.
2. Bruce J. D., "Correlation Functions of Quantized Signals," *Quarterly Progress Report No. 76,* Massachusetts Institute of Technology Research Laboratory of Electronics, January, 1965, pp. 192–198.
3. Hurd, W. J., "Correlation Function of Quantized Sine Wave Plus Gaussian Noise," IEEE Trans. Inform. Theory, *IT-13,* No. 1 (January 1967), pp. 65–68.
4. Velichkin, A. I., "Correlation Function and Spectral Density of a Quantized Process," *Telecommunications and Radio Engineering,* Part II: Radio Engineering, (July 1962), pp. 70–77.
5. Fry, T. C., *Probability and Its Engineering Uses,* 2nd ed., New York: Van Nostrand, 1965, p. 268.
6. Cooley, J. W., and Tukey, J. W., "An Algorithm for the Machine Calculation of Complex Fourier Series," Math. of Computation, *19,* No. 90 (April 1965), pp. 297–301.
7. Woodward, P. M., *Probability and Information Theory with Applications to Radar,* New York: Pergamon Press, 1960, Section 1.4.
8. Magnus, W., and Oberhettinger, F.. *Formulas and Theorems for the Functions of Mathematical Physics,* New York: Chelsea Publishing, 1954, Chapter V, Section 2.
9. Slater, L. J., *Confluent Hypergeometric Functions,* London: Cambridge University Press, 1960, Section 4.5.2.

Adding Two Information Symbols to Certain Nonbinary BCH Codes and Some Applications

By JACK KEIL WOLF

This paper is a compendium of results based on a simple observation: two information symbols can be appended to certain nonbinary BCH codes without affecting the guaranteed minimum distance of these codes. We give two formulations which achieve this result; the second yields information regarding the weights of coset *leaders for the original BCH codes.*

Single-error-correcting Reed-Solomon codes with the added information symbols yield perfect codes for the Hamming metric. We use these lengthened Reed-Solomon codes as building blocks for perfect single-error-correcting codes in another metric.

I. INTRODUCTION

This paper is a compendium of results based upon a simple observation: two information symbols can be appended to the code words of certain BCH codes without weakening the error correction capability of these codes.

We define a class of BCH codes called "maximally redundant codes" in Section II; for codes in this class a simple method is given for appending two columns to the check matrix which does not increase the number of check symbols for the code nor decrease the error correction capability of these codes. Section III gives the parameters for lengthened Reed-Solomon codes and shows that such codes are perfect for single error correction. Section IV discusses a general decoding algorithm for the lengthened codes and shows that these codes are invariant under certain permutation operations.

Section V discusses a method for constructing the lengthened codes from cosets of the original code. We use this approach in Section VI to determine the lower bounds on the number of high weight cosets

for the original BCH codes. Section VII defines a new metric and gives a procedure for constructing some perfect codes in this metric. These codes are based upon the lengthened Reed-Solomon codes. The appendix shows that a necessary and sufficient condition for the nonzero elements of $GF(p)$ to be partitioned into mutually exclusive and exhaustive four element subsets of the form

$$\{x, \beta x, -x, -\beta x\}, \qquad \beta, x \ \varepsilon \ GF(p)$$

is that there exists an integer t such that

$$\beta^{2t} \equiv -1 (\text{mod } p).$$

II. BCH CODES

BCH codes are random-error-correcting codes for symbols from $GF(q)$ where q is a prime (in which case q is replaced by p) or a power of a prime.[1-3] Let α be an element of $GF(q^m)$ and let the order of α be n. That is, $\alpha^n = 1$ and $\alpha^i \neq 1$ for $i < n$. The check matrix of a BCH code with designed distance d can then be given as

$$\mathbf{H} = \begin{bmatrix} 1 & \alpha^{m_0} & (\alpha^{m_0})^2 & \cdots & (\alpha^{m_0})^{n-1} \\ 1 & \alpha^{m_0+1} & (\alpha^{m_0+1})^2 & \cdots & (\alpha^{m_0+1})^{n-1} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & \alpha^{m_0+d-2} & (\alpha^{m_0+d-2})^2 & \cdots & (\alpha^{m_0+d-2})^{n-1} \end{bmatrix}.$$

The code words are all n-vectors, $\mathbf{C}$, with entries from $GF(q)$ which satisfy the equation

$$\mathbf{HC} = \mathbf{O}.$$

(Unless stated to the contrary, all vectors are column vectors.)

The proof that such codes have minimum distance at least d follows from demonstrating that all sets of $d-1$ or fewer columns of $\mathbf{H}$ are linearly independent over $GF(q)$. Actually, the proof shows more than this: it shows that all sets of $d-1$ or fewer columns of $\mathbf{H}$ are linearly independent over any extension field of $GF(q)$. To establish this linear independence let us consider the columns $j_1, j_2, \cdots, j_{d-1}$ and the determinant of the corresponding $(d-1)$ by $(d-1)$ array of symbols from $GF(q^m)$. Then,

$$\det \begin{vmatrix} (\alpha^{m_0})^{j_1} & (\alpha^{m_0})^{j_2} & \cdots & (\alpha^{m_0})^{j_{d-1}} \\ (\alpha^{m_0+1})^{j_1} & (\alpha^{m_0+1})^{j_2} & \cdots & (\alpha^{m_0+1})^{j_{d-1}} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (\alpha^{m_0+d-2})^{j_1} & (\alpha^{m_0+d-2})^{j_2} & \cdots & (\alpha^{m_0+d-2})^{j_{d-1}} \end{vmatrix}$$

$$= \alpha^{m_0(j_1+j_2+\cdots+j_{d-1})} \det \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \alpha^{j_1} & \alpha^{j_2} & \cdots & \alpha^{j_{d-1}} \\ \vdots & \vdots & \vdots & \vdots \\ (\alpha^{j_1})^{d-2} & (\alpha^{j_2})^{d-2} & \cdots & (\alpha^{j_{d-1}})^{d-2} \end{vmatrix}.$$

The latter determinant is a Vander Monde determinant and is known to be nonzero if $\alpha^{j_i} \neq \alpha^{j_k}$ for $i \neq k$. Since the elements of the matrices in question are elements from $GF(q^m)$, the nonvanishing of the determinant ensures that any set of $d - 1$ columns of the check matrix are linearly independent over $GF(q^m)$. The special case of $m = 1$ defines a subset of BCH codes called Reed-Solomon codes.[4]

The number of check symbols in the code is upper bounded by $m(d - 1)$ since these are the number of rows in the check matrix after each symbol from $GF(q^m)$ is replaced by an m-vector with elements from $GF(q)$. The reason that $m(d - 1)$ is merely an upper bound is that the number of check symbols is equal to the number of linearly independent rows in the check matrix [when expressed in terms of elements from $GF(q)$]; in general this number can be less than $m(d - 1)$. In this paper, codes for which the number of check symbols is equal to $m(d - 1)$ are called "maximally redundant" BCH codes. Binary codes (codes for which $q = 2$) are examples of nonmaximally redundant codes while Reed-Solomon codes (codes for which $m = 1$) are examples of maximally redundant codes.

Let us now consider appending two columns to the check matrix, **H**, to form the new check matrix, **H′**,

$$\mathbf{H'} = \begin{bmatrix} 1 & 0 & \\ 0 & 0 & \\ \vdots & \vdots & \mathbf{H} \\ 0 & 0 & \\ 0 & 1 & \end{bmatrix}.$$

It is now easy to see that any $(d - 1)$ columns of **H′** are linearly independent over $GF(q^m)$. [Determinants formed from $(d - 1)$ columns, excluding the first two columns, are $(d - 1)$ by $(d - 1)$ Vander Monde. Determinants formed from $(d - 1)$ columns, including one of the first two columns, are $(d - 2)$ by $(d - 2)$ Vander Monde after expansion about the column in question. Determinants formed from $(d - 1)$

columns, including both the first and second column of $\mathbf{H}'$, are $(d - 3)$ by $(d - 3)$ Vander Monde after expansion.]

The number of symbols per block in the lengthened code is thus two more than the corresponding number for the BCH code. The number of check symbols may or may not be increased in accordance with whether or not the number of linearly independent rows remains the same after the addition of these two columns. One class of BCH codes for which the number of check symbols does not increase is the maximally redundant codes. This class includes all Reed-Solomon codes as well as other codes.

It is possible that in some cases more than two columns can be appended to the parity check matrix while preserving the designed distance of the code. No general results have been found, however, for such cases.* For example, if a column is appended which contains a single 1 in the $(l + 1)$th position of the column vector, the resultant determinant after expansion and factoring is of the form

$$
D_l = \begin{vmatrix}
1 & 1 & \cdots & 1 \\
\alpha^{i_1} & \alpha^{i_2} & \cdots & \alpha^{i_{d-2}} \\
\cdot & \cdot & \cdots & \cdot \\
(\alpha^{i_1})^{l-1} & (\alpha^{i_2})^{l-1} & \cdots & (\alpha^{i_{d-2}})^{l-1} \\
(\alpha^{i_1})^{l+1} & (\alpha^{i_2})^{l+1} & \cdots & (\alpha^{i_{d-2}})^{l+1} \\
\cdot & \cdot & \cdots & \cdot \\
(\alpha^{i_1})^{d-2} & (\alpha^{i_2})^{d-2} & \cdots & (\alpha^{i_{d-2}})^{d-2}
\end{vmatrix}.
$$

Such a determinant can be evaluated as

$$
D_l = \prod_{i>k}^{d-2} (\alpha^{i_i} - \alpha^{i_k}) \ [\text{sum of all products of } (d - 2 - l) \text{ distinct } \alpha^{i_i}].
$$

The latter sum of products can be zero even if all the α^{i_i} are distinct.

III. LENGTHENED REED-SOLOMON CODES

The Reed-Solomon codes codes with symbols from $GF(q)$ are BCH codes formed by choosing the parameter $m = 1$. These codes have parameters

$$\text{block length} \qquad n = q - 1,$$

$$\text{check symbols per block } r = d - 1,$$

---

* An exception is $d = 4$ and q even where three columns can be appended to the parity check matrix. The appended columns are then the 3×3 identity matrix.

and correct any pattern of $[(d - 1)/2]$ or fewer errors in a block of length n. Any t error-correcting linear code can have no fewer than $2t$ check symbols; this bound is achieved by the Reed-Solomon codes if d is an odd integer. This is not to say that the codes cannot be improved upon: in particular, the lengthened codes formed as described in Section II represent a minor improvement.

The lengthened code has parameters:

$$\text{block length} \qquad n' = q + 1,$$

$$\text{check symbols per block } r' = (d - 1),$$

and corrects any pattern of $[(d - 1)/2]$ or fewer errors in a block of length n' symbols. The lengthened codes are maximum distance separable (MDS) in that they have the maximum possible minimum distance for a given block length n', and code size $q^{(n'-r')}$. These codes complement the set of maximum distance separable codes given by Singleton.[5] The weight distributions of the code words of maximum distance separable codes are given by Berlekamp.[6] The case of single error-correcting lengthened Reed-Solomon codes (that is, $d = 3$) are of particular interest in that they are perfect codes. That is, bounded distance decoding results in the use of every syndrome. Specifically, there are q^2 distinct syndromes. There are $(q - 1)$ different errors which can occur in any of the $(q + 1)$ different positions resulting in $q^2 - 1$ different error patterns. The all zero error pattern (no errors) in addition to the $(q - 1)(q + 1) = q^2 - 1$ single error patterns use all q^2 syndromes.

IV. DECODING AND SYMMETRY OF LENGTHENED MAXIMALLY REDUNDANT BCH CODES[*]

The columns of the parity check matrix are conveniently labeled:

$$\overleftarrow{\hspace{2cm} n' \hspace{2cm}}\rightarrow$$

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \alpha & (\alpha)^2 & \cdots & (\alpha)^{n'-3} \\ 0 & 0 & 1 & \alpha^2 & (\alpha^2)^2 & \cdots & (\alpha^2)^{n'-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 1 & \alpha^{d-2} & (\alpha^{d-2})^2 & \cdots & (\alpha^{d-2})^{n'-3} \end{bmatrix}$$

$$\text{label} \quad 0 \quad \infty \quad 1 \quad \alpha \quad \alpha^2 \qquad \alpha^{n'-3}$$

---

[*] This section is based on suggestions from E. R. Berlekamp of Bell Telephone Laboratories.

where α is a primitive element of $GF(q^m)$ and m_o has been taken equal to zero. If the second column were omitted from this matrix, the resultant code would be an extension of a BCH code of designed distance $d - 1$. That is, the resultant code is obtained by appending an overall parity check digit to a BCH code of designed distance $d - 1$. The code with the second digit omitted (block length $n' - 1$) is called a "singly-lengthened" BCH code. The code of block length n' (which includes all digits) is called a "doubly-lengthened" BCH code.

For d odd, one decoding algorithm for the correction of $[(d - 1)/2]$ or fewer errors for the doubly lengthened BCH codes is:

(*i*) Ignore the last syndrome digit (the only equation involving the symbol in position labeled ∞) and decode as in Section 10.3 of Ref. 6 for extended BCH codes. Let D be the number of errors indicated by the decoding algorithm. If $D < (d - 1)/2$, decode all positions except the position labeled ∞ and then use the last parity check equation to decode the position labeled ∞.

(*ii*) If $D = (d - 1)/2$, assume that the digit in position ∞ is correct, modify the syndrome accordingly, and decode as in Ref. 6 using all digits in the modified syndrome.

The lengthened primitive BCH codes have interesting symmetry properties. Since the singly-lengthened primitive BCH code is an extension of a primitive BCH code with designed distance one less, it is invariant under the affine permutation group on $GF(q)$, as Theorem 10.37 of Ref. 6 shows.

One might hope that the doubly-lengthened BCH code would be invariant under the triply-transitive linear fractional group on $GF(q) \cup \infty$ (page 358 of Ref. 6). This is not really the case since the code is not invariant under the simple permutation $x \rightarrow 1/x$. The doubly-lengthened BCH code is invariant, however, under the multiply and permute operation of order two specified:

(*i*)   Exchange digits at 0 and ∞.

(*ii*) Multiply digit at α^i by $\alpha^{-i(d-2)}$ and then move it to position α^{-i}.

This operation transforms the **H′** matrix into the same matrix with the rows listed in reverse order. Since this operation preserves Hamming weights, it ensures considerable symmetry.

V. ALTERNATIVE FORMULATION OF LENGTHENED MAXIMALLY REDUNDANT BCH CODES

We will now describe an alternative formulation of lengthened maxi-

mally redundant BCH codes which is more complicated than that described in Section III. However, its real utility is that it gives insight to the problem of determining the weight distribution of coset leaders for the (unlengthened) BCH codes (a subject discussed in Section V).

Consider an (unlengthened) maximally redundant BCH code [with symbols from $GF(q)$] with check matrix

$$
\mathbf{H} = \begin{bmatrix}
1 & \alpha^{m_0} & (\alpha^{m_0})^2 & \cdots & (\alpha^{m_0})^{n-1} \\
1 & \alpha^{m_0+1} & (\alpha^{m_0+1})^2 & \cdots & (\alpha^{m_0+1})^{n-1} \\
\cdot & \cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdot & \cdots & \cdot \\
1 & \alpha^{m_0+d-2} & (\alpha^{m_0+d-2})^2 & \cdots & (\alpha^{m_0+d-2})^{n-1}
\end{bmatrix},
$$

where α is an element of $GF(q^m)$. Consider an n-vector $\mathbf{X}$ [with entries from $GF(q)$] such that

$$
\mathbf{HX} = \begin{bmatrix}
\sigma_1 \\
0 \\
0 \\
\vdots \\
0 \\
0 \\
\sigma_2
\end{bmatrix},
$$

where σ_1 and σ_2 are elements from $GF(q^m)$. We now prove the following inequalities regarding the weight of $\mathbf{X}$, denoted $W(\mathbf{X})$.

Inequality 1: If $\sigma_1 = \sigma_2 = 0$, $W(\mathbf{X}) \geqq d$ for $\mathbf{X} \neq 0$.

Proof: The vectors $\mathbf{X}$ which satisfy $\mathbf{HX} = \mathbf{O}$ are the code words of the code with check matrix $\mathbf{H}$ and have minimum distance at least d. Thus the weight of any nonzero code word is greater than or equal to d.

Inequality 2: If $\sigma_1 = 0$ and $\sigma_2 \neq 0$ or if $\sigma_1 \neq 0$ and $\sigma_2 = 0$, then $W(\mathbf{X}) \geqq d - 1$.

Proof: We first note that $\mathbf{X} \neq 0$ since either σ_1 or σ_2 is nonzero. Next consider the case where $\sigma_1 \neq 0$ and $\sigma_2 = 0$ and form a new check matrix $\mathbf{H}_{(1)}$ obtained by deleting the first row of $\mathbf{H}$. Now $\mathbf{H}_{(1)} \mathbf{X} = 0$ so that $\mathbf{X}$ is a code word corresponding to the check matrix $\mathbf{H}_{(1)}$. But any $(d - 2)$ columns of $\mathbf{H}_{(1)}$ form a $(d - 2)$ by $(d - 2)$ Vander Monde determinant

so that the weight of $\mathbf{X}$ is at least $(d - 1)$. The proof for the case where $\sigma_1 = 0$ and $\sigma_2 \neq 0$ follows similarly by noticing that $\mathbf{X}$ is a code word in a code corresponding to a check matrix formed by deleting the last row of $\mathbf{H}$.

Inequality 3:   If $\sigma_1 \neq 0$ and $\sigma_2 \neq 0$, then $W(\mathbf{X}) \geqq d - 2$.

Proof: Again $\mathbf{X} \neq 0$ since both σ_1 and σ_2 are nonzero. Now consider a check matrix formed by deleting the first and last rows of $\mathbf{H}$. Since $\mathbf{X}$ is in the null space of this new check matrix, every such nonzero vector must have weight at least $(d - 2)$.

The lengthened code is now formed of $(n + 2)$-tuples of the form $\begin{bmatrix} -\sigma_1 \\ -\sigma_2 \\ \mathbf{X} \end{bmatrix}$.

From before we see that all such nonzero vectors must have weight at least d. It is easy to verify that the set of code words from a linear code and indeed that such a linear code is the null space of the check matrix

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & \\ 0 & 0 & \\ 0 & 0 & \mathbf{H} \\ \vdots & \vdots & \\ 0 & 1 & \end{bmatrix}.$$

VI. WEIGHTS OF COSETS OF MAXIMALLY REDUNDANT BCH CODES

In this section we digress from the main theme of this paper to present some results on another problem: determining the weights of cosets (that is, coset leaders) for maximally redundant BCH codes. It should be emphasized that this problem differs from the widely researched problem of determining the weights of the code words themselves.

The complete weight enumeration of the cosets is known only for a very few classes of codes.[6] This knowledge is crucial to determining the performance of codes using a complete decoding algorithm (that is, maximum likelihood decoding).

In this section we are not able to determine the complete weight enumeration for the codes under consideration. Rather we can only give lower bounds to the number of coset leaders whose weight exceeds

certain values. However, we believe that this knowledge is both new and useful.

Specifically we are concerned with the weights of coset leaders of maximally redundant primitive BCH codes. Our main result is:

$$[\text{Number of coset leaders of weight} \geq d - j]$$

$$\geq (q^m)^{i-1}[(j + 1)q^m - j] - 1 \quad \text{for} \quad \begin{cases} j \geq 1 \\ 2j < d. \end{cases}$$

This result shows that for a maximally redundant BCH code of minimum (designed) distance d, in addition to having as coset leaders all vectors of weight less than or equal to $[(d - 1)/2]$, coset leaders exist for all weights up to and including $(d - 1)$. The actual minimum distance of the code, d_{ACT}, may exceed the designed distance d. If $[(d_{ACT} - 1)/2] < d - 1$, the codes cannot be perfect codes and if $[(d_{ACT} - 1)/2] < d - 2$, the codes cannot be quasiperfect. For Reed-Solomon codes $d_{ACT} = d$ and the codes are not perfect for any d and not quasiperfect for $d > 3$.

Proof: Consider a coset leader $\mathbf{X}'$ corresponding to the syndrome, $\mathbf{S}$, where

$$\mathbf{HX}' = \mathbf{S} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \cdot \\ \cdot \\ \sigma_i \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ \sigma_{i+1} \\ \sigma_{i+2} \\ \cdot \\ \cdot \\ \sigma_j \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ d - 1 \\ \\ \\ \\ \\ \end{matrix} \quad \text{where} \quad \sigma_i \,\varepsilon\, GF(q^m) \qquad i = 0, 1, \cdots, j.$$

Consider a new check matrix obtained by deleting the first i rows and the last $(j - i)$ rows of $\mathbf{H}$. $\mathbf{X}'$ must be a vector in the null space of this new check matrix and will be nonzero unless $\sigma_1 = \sigma_2 = \cdots = \sigma_j = 0$.

Furthermore every such nonzero vector must have weight at least $d - j$ since any set of $d - 1 - j$ columns of this new check matrix forms a Vander Monde determinant. A counting problem remains: counting the number of distinct nonzero syndromes having a run of $d - 1 - j$ consecutive zeros. For $i = j$, there are $(q^m)^i - 1$ such patterns corresponding to the q^m different values for each σ_i (excluding $\sigma_1 = \sigma_2 = \cdots = \sigma_i = 0$). For each $i < j$, there are $(q^m - 1)(q^m)^{i-1}$ such patterns corresponding to the $(q^m - 1)$ distinct nonzero values for σ_{i+1} and the q^m distinct values for all other σ_k, $k \neq i + 1$. Counting in this fashion, if $2j < d$ we include each such pattern once and only once resulting in a total of

$$(q^m)^i - 1 + j(q^m - 1)(q^m)^{i-1} = (q^m)^{i-1}[(j + 1)q^m - j] - 1$$

such patterns.

The above proof not only yields a bound to the number of high weight coset leaders but also gives an easy way of recognizing their occurrence from their respective syndromes. Thus if one were to use bounded distance decoding (decoding only coset leaders of weight $\leq [(d - 1)/2]$), many nondecodable cosets would be easily recognizable by the form of the syndrome.

A tighter bound can sometimes be obtained by noticing that the parity check matrix

$$\mathbf{H} = \begin{bmatrix} 1 & \alpha^{m_0} & (\alpha^{m_0})^2 & \cdots & (\alpha^{m_0})^{n-1} \\ 1 & \alpha^{m_0+a} & (\alpha^{m_0+a})^2 & \cdots & (\alpha^{m_0+a})^{n-1} \\ 1 & \alpha^{m_0+2a} & (\alpha^{m_0+2a})^2 & \cdots & (\alpha^{m_0+2a})^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \alpha^{m_0+(d-2)a} & (\alpha^{m_0+(d-2)a})^2 & \cdots & (\alpha^{m_0+(d-2)a})^{n-1} \end{bmatrix}$$

yields a code with a minimum distance of at least d if a and n are relatively prime. Thus the zeros in the syndrome that signify a high weight coset need not occur as a single burst but rather can occur with a fixed periodicity.

VII. SOME PERFECT SINGLE-ERROR-CORRECTING CODES FOR ANOTHER METRIC

In this section we use the lengthened Reed-Solomon codes to construct codes for a new metric. In particular, we consider the case where $q = p$, a prime, and we are interested in codes that correct errors of the form $\pm 1, \pm 2, \cdots, \pm T$ in a "single position" of a code word. In particular,

codes are given for $T = 1$ and $T = 2$. For $T = 1$, these codes are single-error-correcting Lee metric codes.[7]

The lengthened Reed-Solomon code used in the construction of these codes has a check matrix

$$\xleftarrow{\hspace{1cm}} n' = p + 1 \xrightarrow{\hspace{1cm}}$$

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \alpha & \alpha^2 & \cdots & \alpha^{p-2} \end{bmatrix}$$

where α is a primitive element from $GF(p)$. The null space of this matrix is a perfect single-error-correcting code for the Hamming metric. That is, it corrects any error $[\pm 1, \pm 2, \cdots, \pm(p - 1/2)]$ which occurs in any one position in a code word.

Consider the case where it is required only to correct an error of the form ± 1. Also consider the new check matrix

$$\xleftarrow{\hspace{0.5cm}} n'' = \left(\frac{p - 1}{2}\right) n' = \frac{p^2 - 1}{2} \xrightarrow{\hspace{0.5cm}}$$

$$\mathbf{H}'' = \begin{bmatrix} \mathbf{H}' & 2\mathbf{H}' & 3\mathbf{H}' & \cdots & \left(\frac{p - 1}{2}\right)\mathbf{H}' \end{bmatrix}.$$

To show that the null space of $\mathbf{H}''$ will correct any single error of the form ± 1, we need only show that all columns of $\mathbf{H}''$ are distinct from each other after multiplication by ± 1. This follows immediately from noticing that all pairs of columns of $\mathbf{H}'$ are linearly independent over $GF(p)$.

We prove the code is perfect by noting that $2n'' + 1 = p^2$ syndromes are needed to correct a ± 1 error in each of the n'' positions (plus the all zero error pattern). But since $\mathbf{H}''$ has two rows, there are exactly p^2 syndromes; every syndrome is used to correct the required error patterns.

The above code has the same block length, number of check symbols, and error corrections capability as Berlekamp's perfect megacyclic single-error-correcting Lee metric codes.[6]

The form chosen for $\mathbf{H}'$ with the first row consisting of all ones and a single zero makes the decoding algorithm easy. Let

$$\mathbf{S} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} \quad \text{where} \quad -\frac{(p - 1)}{2} \leqq \sigma_i \leqq \frac{p - 1}{2} \; i = 1, 2.$$

The algorithm is as follows.

(i) If $\sigma_1 = 0$, the error is in position $2 + (|\sigma_2| - 1)n'$ and has value sgn (σ_2).

(ii) If $\sigma_2 = 0$, the error is in position $1 + (|\sigma_1| - 1)n'$ and has value sgn (σ_1).

(*iii*) If $\sigma_1 \neq 0$ and $\sigma_2 \neq 0$, let x be the solution to the congruence $\sigma_1 \alpha^x \equiv \sigma_2 \bmod (p)$. The error is then in position $(x + 3) + (|\sigma_1| - 1)n'$ and has value sgn (σ_1).

As an example, consider the code for $p = 5$ with $\alpha = 2$, a primitive root. Then

$$\mathbf{H'} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 4 & 3 \end{bmatrix}$$

and

$$\mathbf{H''} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 & 2 & 0 & 2 & 2 & 2 & 2 \\ 0 & 1 & 1 & 2 & 4 & 3 & 0 & 2 & 2 & 4 & 3 & 1 \end{bmatrix}.$$

Consider an error pattern resulting in the syndrome $\left[{-2 \atop 2}\right]$. Solving for x [in accordance with (*iii*) above] we have

$$(-2)2^x \equiv 2 \bmod (5)$$

$$2^x \equiv -1 \bmod 5,$$

which has the solution $x = 2$. Thus we have an error in position

$$(x + 3) + (|\sigma_1| - 1)n' = (2 + 3) + (|-2| - 1)\, 6 = 11$$

having the value -1.

A more interesting case arises when one desires to correct a single error of magnitude $+1$, -1, $+2$, or -2. We give a construction procedure which results in perfect codes for the case where the prime p is such that there exists a least positive integer t which satisfies the congruence

$$2^{2t} \equiv -1 \bmod p.$$

Form the multiplicative subgroup

$$1\ 2\ 4\ 8 \cdots 2^{2t} \equiv -1 \qquad 2^{2t+1} \equiv -2 \cdots 2^{4t-1}.$$

Let $a_0 = 1$, and consider the coset table:

a_0	$2a_0$	$4a_0$	$8a_0$	$\cdots$	$2^{2t}a_0 \equiv -a_0$	$a_0 2^{2t+1} \equiv -2a_0$	$\cdots$	$(2^{4t-1})a_0$
a_1	$2a_1$	$4a_1$	$8a_1$	$\cdots$	$-a_1$	$-2a_1$	$\cdots$	$(2^{4t-1})a_1$
a_2	$2a_2$	$4a_2$	$8a_2$	$\cdots$	$-a_2$	$-2a_2$	$\cdots$	$(2^{4t-1})a_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
a_{l-1}	$2a_{l-1}$	$4a_{l-1}$	$8a_{l-1}$	$\cdots$	$-a_{l-1}$	$-2a_{l-1}$	$\cdots$	$(2^{4t-1})a_{l-1}$

where $4tl = p - 1$.

Now again begin with the check matrix for the lengthened Reed-Solomon single error-correcting code

$$\overleftarrow{\hspace{2em}} n' = p + 1 \overrightarrow{\hspace{2em}}$$

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \alpha & \alpha^2 & \cdots & \alpha^{p-2} \end{bmatrix},$$

and form the new check matrix

$$\mathbf{H}''' = [a_0\mathbf{H}' \quad 2^2 a_0\mathbf{H}' \quad 2^4 a_0\mathbf{H}' \quad \cdots \quad 2^{2(t-1)} a_0\mathbf{H}' \quad a_1\mathbf{H}' \quad 2^2 a_1\mathbf{H}' \quad \cdots$$

$$\cdot 2^{2(t-1)} a_1\mathbf{H}' \quad \cdots \quad a_{l-1}\mathbf{H}' \quad 2^2 a_{l-1}\mathbf{H}' \quad \cdots \quad 2^{2(t-1)} a_{l-1}\mathbf{H}'].$$

The block length of this code, n''', is

$$n''' = ltn' = \frac{(p-1)}{4} n' = \frac{p^2 - 1}{4}.$$

In order for the code to correct all single errors of the form ± 1, ± 2, we would require $4n''' + 1 = p^2$ syndromes. Since the code has two check symbols, it has exactly p^2 syndromes available for error correction. Thus the code will be a perfect code if we can prove that its error correction capability is as asserted.

Proof: We must prove that any column of $\mathbf{H}'''$, when multiplied by $+1$, -1, $+2$, or -2, is distinct from any other column of $\mathbf{H}'''$ when multiplied by $+1$, -1, $+2$, or -2. If the two columns in question come from two distinct columns of $\mathbf{H}'$, then this is certainly the case since the columns of $\mathbf{H}'$ are linearly independent over $GF(p)$. Let the pair of columns in question be derived from the same column of $\mathbf{H}'$, say $\mathbf{h}$. One such column is of the form $2^{2(l_1)} a_{j_1}\mathbf{h}$ and the other is of the form $2^{2(l_2)} a_{j_2}\mathbf{h}$ where $(0 < l_1, l_2 \leq t - 1)$ and $(0 \leq j_1, j_2 \leq l - 1)$. Now let z be any member of one of the cosets. Then $-z$, $+2z$, and $-2z$ are also members of that coset; so we need only consider the case $j_1 = j_2$. But

$$(1)(2^{2l_1}) = 2^{2l_1} \qquad\qquad (1)(2^{2l_2}) = 2^{2l_2}$$

$$(2)(2^{2l_1}) = 2^{2l_1+1} \qquad\qquad (2)(2^{2l_2}) = 2^{2l_2+1}$$

$$(-1)(2^{2l_1}) = 2^{2(l_1+t)} \qquad (-1)(2^{2l_2}) = 2^{2(l_2+t)}$$

$$(-2)(2^{2l_1}) = 2^{2(l_1+t)+1} \qquad (-2)(2^{2l_2}) = 2^{2(l_2+t)+1},$$

and no term in the left four equations can equal a term in the right four equations for $l_1 \neq l_2$, $0 \leq l_1, l_2 \leq t - 1$. Thus the assertion is proved.

As an example, let $p = 13$ where 2 is a primitive element of order

$4t = 12$. Thus $t = 3$ and $l = 1$. The matrix $\mathbf{H}'$ can be taken as

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 4 & 8 & 3 & 6 & 12 & 11 & 9 & 5 & 10 & 7 \end{bmatrix}$$

and $\mathbf{H}'''$ is

$$\mathbf{H}''' = [\mathbf{H}' \quad 4\mathbf{H}' \quad 3\mathbf{H}'].$$

As a second example let $p = 17$. The coset table is

$$\begin{array}{ccccccccc} 1 & 2 & 4 & 8 & 16 \equiv -1 & 15 & 13 & 9 \\ 3 & 6 & 12 & 7 & 14 & 11 & 5 & 10 \end{array}.$$

The check matrix $\mathbf{H}'''$ is

$$\longleftarrow\!\!\!\!\!\!\!\!\!- 72 -\!\!\!\!\!\!\!\!\!\longrightarrow \quad \uparrow$$
$$\mathbf{H}''' = [\mathbf{H}' \quad 4\mathbf{H}' \quad 3\mathbf{H}' \quad 12\mathbf{H}'] \quad 2$$
$$\downarrow$$

where $\mathbf{H}'$ is a two row by 18 column check matric formed in the manner described. Berlekamp has given a code for $p = 17$ with block length 72 with Lee distance 5 that requires four check symbols.[6] The above code requires only two check symbols but corrects only a small subset of the class of errors correctable by Berlekamp's code. Wyner has found several classes of codes which correct two errors per block, each error of the form ± 1.[8] One such class has a block length of p and requires three check symbols.

In the proof we have given a decomposition of the integers $1, 2, \cdots,$ $p - 2, p - 1 = 4m$ into disjoint sets $S_1 S_2 \cdots S_m$ each containing four elements, such that the elements of each set are of the form $x, 2x, -x,$ and $-2x \pmod{p}$. A sufficient condition for this decomposition was that there exists a least positive integer t such that $2^{2t} \equiv -1 \pmod{p}$. The appendix shows that this condition is necessary for this decomposition.

In particular we consider the following question in the appendix: For which primes p and elements β from $GF(p)$ is it possible to partition the nonzero field elements $(1, 2, \cdots, p - 1)$ into four element subsets, S_i, such that $S_i = \{x_i, \beta x_i, -x_i - \beta x_i\} \pmod{p}$, where each nonzero field element occurs in one and only one subset? We show that the answer is: Such a partition can be achieved if and only if there exists a least positive integer t such that $\beta^{2t} \equiv -1 \pmod{p}$. Stein has considered

a more general version of this problem.[9] The results in the appendix were proved independently of Stein.

VIII. ACKNOWLEDGMENTS

APPENDIX

On a Partitioning of the Nonzero Elements of GF (p)

By R. L. Graham and J. K. Wolf

A.1 *Introduction*

The problem we consider is: For which primes, p, and elements, β, from $GF(p)$ is it possible to partition the nonzero field elements of $GF(p)$ into mutually exclusive and exhaustive four element subsets, S_i, such that

$$S_i = \{x_i, \beta x_i, -x_i, -\beta x_i\}, \pmod{p}?$$

A necessary condition for the existence of such a partition is that

$$\beta \neq \begin{cases} \pm 1 \\ 0 \end{cases} \bmod p;$$

otherwise the subsets would not contain four distinct elements.

We will show that a necessary and sufficient condition for this partition is that there exists a t such that $\beta^{2t} \equiv -1 \pmod{p}$. Further we will show that for a prime of the form $p = 8k + 5$ such a partition always exists for $\beta = 2$.

A.2 *Proof of Assertion*

First notice that a necessary condition for this partition to exist is that $p = 4m + 1$ for some $m \geq 1$, since $p - 1$ must be divisible by four. A second necessary condition is that

$$\beta \neq \begin{cases} 1 \\ 0 \\ -1 \end{cases} \pmod{p}.$$

Let r be a primitive root of p so that $r^{2m} \equiv -1 \pmod{p}$. Define α as the smallest positive integer such that $r^\alpha \equiv \beta \pmod{p}$. By assumption

$\beta \neq -1 \pmod{p}$, so that $\alpha \neq 2m$. The subset S_i must then consist of the four elements

$$S_i = \{r^{y_i}, r^{y_i+\alpha}, r^{y_i+2m}, r^{y_i+2m+\alpha}\}$$

so that an equivalent problem is to decompose the additive group $Z_{p-1} = \{0, 1, 2, \cdots, p - 2 = 4m - 1\}$ into mutually exclusive and exhaustive subsets of the form $S_i' = \{y_i, y_i + \alpha, y_i + 2m, y_i + 2m + \alpha\}$ mod $(4m)$].

This problem can be viewed geometrically as that of covering the vertices of a regular $4m$-gon placed on a circle by translates of the pattern $\{0, \alpha, 2m, \alpha + 2m\}$. This pattern is symmetric modulo $2m$, so the problem reduces to covering the vertices of a regular $2m$-gon placed on a circle by translates of the pattern $\{0, \alpha\}$. This pattern $\{0, \alpha\}$ can be viewed as a chord spanning α vertices. For example, for $m = 6$ and $\alpha = 5$, this covering is shown in Fig. 1 while for $m = 6$ and $\alpha = 4$, no such covering is possible.
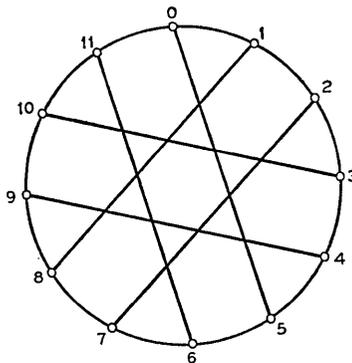
Fig. 1 — A covering for $m = 6$ and $\alpha = 5$.

In terms of sets, the problem now is to decompose the additive group Z_{2m} into m mutually exclusive and exhaustive two-element subsets, S_i'', of the form $S_i'' = \{y_i, y_i + \alpha\} \pmod{2m}$.

In the following, we denote by $[2m, \alpha]$ a covering of the $2m$-gon by chords spanning α vertices. Letting $2m = 2^\tau(2v + 1)$ we now prove the following theorem.

Theorem 1: *A $[2m, \alpha]$ covering exists if and only if $2^\tau \nmid \alpha$. We prove this theorem by first proving the following lemmas.*

Lemma 1: *A $[2m, \alpha]$ covering exists if *$(2m, \alpha) = 1$.*

---

* (x, y) = greatest common divisor of x and y.

Proof: Let $S_i'' = \{2(i - 1)\alpha, (2i - 1)\alpha\}$ (mod $2m$), $i = 1, 2, \cdots, m$. Then the subset S_i'' is of the proper form and it remains to show that each element of Z_{2m} appears in one and only one subset. Assume that an element of Z_{2m} appears in more than one subset. Then for $0 \leqq i \leqq j \leqq 2m - 1$,

$$i\alpha \neq j\alpha \ (\text{mod } 2m)$$

or

$$(j - i)\alpha \neq 0 \text{ mod } 2m.$$

But by assumption $(2m, \alpha) = 1$ so $2m$ and α have no common factors. Thus $2m \mid (j - i)$ which is impossible since $(j - i) < 2m$. We have then shown that no element of Z_{2m} appears in more than one subset. But there are $2m$ elements in the m subsets so that each element of Z_{2m} must appear once and only once in those subsets.

The decomposition used in the proof of Lemma 1 can also be viewed as taking alternate edges of the regular star of step size α. For example, the covering in Fig. 1 can be viewed as taking alternate edges of a star of step size α. In Fig. 2, this star is shown for $m = 6$ and $\alpha = 5$ with the alternate edges as solid lines.

Lemma 2: A $[x, \alpha]$ *covering exists if and only if a* $[kx, k\alpha]$ *covering exists, where* $k \geqq 1$.

Proof: If a $[x, \alpha]$ covering exists, a $[kx, k\alpha]$ covering can be obtained by simply interleaving the $[x, \alpha]$ covering k times. If a $[kx, k\alpha]$ covering exists, the chords must span exactly $k\alpha$ vertices. Thus deleting all
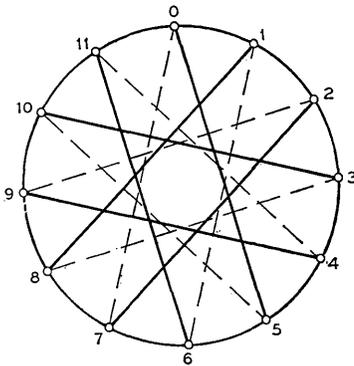
Fig. 2 — A star of step size 5 for $m = 6$.

vertices except those congruent to zero modulo k, we have a $[x, \alpha]$ covering.

Lemma 3:   *A $[x, \alpha]$ covering does not exist for x odd.*

Proof: The covering problem is that of partitioning the integers $\{0, 1, \ldots, x - 1\}$ into two element subsets. For this to be possible two must divide x.

Lemma 4:   *Let $2m = dM$ and $\alpha = dA$, where $(A, M) = 1$. Then a $[2m, \alpha]$ covering exists if and only if M is even.*

Proof: From Lemma 2, a $[2m, \alpha] = [dM, dA]$ covering exists if and only if a $[M, A]$ covering exists. But from Lemma 3, a $[M, A]$ covering will not exist if M is odd. If M is even, since $(A, M) = 1$, Lemma 1 insures the existence of a $[M, A]$ covering.

Proof of Theorem 1:   Let $2m = 2^{\gamma}(2v + 1) = dM$ and $\alpha = dA$ where $(A, M) = 1$. If $2^{\gamma} \mid \alpha$, then $2^{\gamma} \mid d$ and M will be odd. By Lemma 4, a $[2m, \alpha]$ covering will not exist if M is odd. Conversely, assume that $2^{\gamma} \nmid \alpha$. Then $2 \mid M$ and M is even and by Lemma 4, a $[2m, \alpha]$ covering exists.   Q.E.D.

Using Theorem 1 we now prove the main result, which is given as Theorem 2.

Theorem 2:   *The nonzero elements of $GF(p)$ can be partitioned into mutually exclusive and exhaustive 4 element subsets, S_i, such that $S_i = \{x_i, \beta x_i, -x_i, -\beta x_i\}$ (mod p) if and only if there exists a positive interger t such that $\beta^{2t} \equiv -1$ (mod p).*

Proof of Theorem 2:   From Theorem 1, such a partition is possible if and only if $2^{\gamma} \nmid \alpha$. Let us first assume the existence of a positive integer t such that $\beta^{2t} \equiv -1$ (mod p). But $r^{\alpha} \equiv \beta$ (mod p) so that $r^{\alpha 2t} \equiv -1$ (mod p). Since $r^{\delta} \equiv -1$ (mod p) implies

$$\delta = \frac{p - 1}{2} + l(p - 1) = (2l + 1)\left(\frac{p - 1}{2}\right) = (2l + 1)(2m),$$

for some l, then $\alpha 2t = (2l + 1)(2m) = (2l + 1)(2v + 1)2^{\gamma}$. Thus $\alpha t = 2^{\gamma - 1}(2l + 1)(2v + 1)$ and $2^{\gamma} \nmid \alpha$.

Next assume that $2^{\gamma} \nmid \alpha$. There exists a y such that $r^{\alpha y} \equiv \beta^{y} \equiv -1$ (mod p) if and only if

$$\alpha y = \left(\frac{p - 1}{2}\right)(2q + 1) = 2m(2q + 1) = 2^{\gamma}(2v + 1)(2q + 1)$$

for some q. But $2^\gamma \nmid \alpha$, so y must have an even factor, that is $2 \mid y$. Thus y can be written as $y = 2t$ and $\beta^{2t} \equiv -1 \pmod{p}$. Further notice that the condition

$$\beta \neq \begin{cases} +1 \\ \quad 0 \quad \pmod{p} \\ -1 \end{cases}$$

is subsumed by the condition $\beta^{2t} \equiv -1 \pmod{p}$.   Q.E.D.

A.3 *A Special Set of Primes with the Desired Partition*

Each of the two theorems in Section A.2 give a necessary and sufficient condition for the desired partition. Either condition, however, requires some calculation to discover whether p admits such a partition. The following discussion yields an easily recognizable class of primes, p, for which the partition will always be possible if $\beta = 2$.

The Legendre symbol (a/p) is defined as

$$(a/p) = \begin{cases} \quad 1 \quad \text{if} \quad x^2 = a \text{ has a solution in } GF(p) \text{ (that is, } a \text{ is a} \\ \qquad\qquad \text{quadratic residue mod } p) \\ -1 \quad \text{if} \quad x^2 = a \text{ does not have a solution in } GF(p) \text{ (that is,} \\ \qquad\qquad a \text{ is a quadratic nonresidue mod } p) \\ \quad 0 \quad \text{if} \quad a = 0. \end{cases}$$

Lemma 5: *A sufficient condition for the partition to exist is (β/p) $= -1$.*

Proof: By Euler's criterion

$$a^{(p-1)/2} \equiv (a/p) \bmod p.$$

Since $p - 1 = 4m$, if $(\beta/p) = -1$ then $\beta^{2m} \equiv -1 \bmod p$, and the partition is possible for that β and p.

One can show (p. 172 of Ref. 6) that $(2/p) = -1$ if $p = 8k + 5$ for some k. Thus if $\beta = 2$ and the prime p is of the form $p = 8k + 5$ such a partition can be achieved.

REFERENCES

1. Bose, R. C., and Ray-Chaudhuri, D. K., "On a Class of Error Correcting Binary Group Codes," Inform. and Control, *3*, No. 1 (March 1960), pp. 68–79.
2. Bose, R. C., and Ray-Chaudhuri, D. K., "Further Results on Error Correcting Binary Group Codes," Inform. and Control, *3*, No. 3 (September 1960), pp. 279–390.
3. Hocquenghem, A., "Codes correcteurs d'erreurs," Chiffres, *2*, (September 1959), pp. 147–156.

4. Reed, I. S., and Solomon, G., "Polynomial Codes over Certain Finite Fields," J.S.I.A.M., *8*, No. 2 (June 1960), pp. 300–304.
5. Singleton, R., "Maximum Distance Q-Nary Codes," IEEE Trans. Inform. Theory, *IT-10*, No. 2 (April 1964), pp. 116–118.
6. Berlekamp, E. R., *Algebraic Coding Theory*, New York: McGraw-Hill, 1968.
7. Lee, C. Y., "Some Properties of Nonbinary Error-Correcting Codes," IRE Trans. Inform. Theory, *IT-4*, No. 2 (June 1958), pp. 77–82.
8. Wyner, A. D., unpublished work.
9. Stein, S. K., "Factoring by Subsets," Pacific J. Math., *22*, No. 3 (September 1967), pp. 523–541.

Synthesis of Pulse-Shaping Networks in the Time Domain

By DAVID A. SPAULDING

(Manuscript received February 11, 1969)

A fundamental problem in the design of data transmission systems is the synthesis of pulse-shaping networks which satisfy specifications in both the time and frequency domains. This paper considers the problem of designing a network to shape an arbitrary input pulse into a band-limited pulse having minimum intersymbol interference. The design procedure uses the zeros of the network transfer function to achieve the band-limiting properties (using a modified Temes and Gyi constraint) while the transfer function poles are optimized with a computer to give the desired time response.

By limiting the specifications on the shaped pulse to an absolute minimum, very accurate results are achieved with simple networks. Some sample designs and experimental results are included. For example, an 11th order transfer function is designed to shape rectangular pulses for a synchronous baseband pulse amplitude modulation system. The shaped pulses have a bandwidth 20 percent in excess of the Nyquist bandwidth and a theoretical worst-case distortion of 2.1 percent. An active realization of this transfer function achieved a worst-case distortion of about 2.5 percent.

I. INTRODUTION

A fundamental problem in the design of data transmission systems is the synthesis of pulse-shaping networks which meet both time and frequency domain specifications. This paper considers the problem of designing, for a synchronous system, a network whose response to an arbitrary input pulse is a band-limited pulse with minimum inter-symbol interference.[1] The design procedure uses a slightly modified Temes and Gyi procedure to keep the pulses band-limited;[2] the time response is optimized by using a computer. By focusing attention only on the important instants of time, very efficient and accurate designs result. We include some sample designs and experimental results.

II. PROBLEM DESCRIPTION

Consider a baseband pulse-amplitude modulation system in which information is coded as the amplitude values of a single pulse shape $x(t)$. Assume that a pulse is transmitted every T seconds over a channel, which for the present is ideal. The received signal, $s(t)$, is

$$s(t) = \sum_{n=-\infty}^{\infty} a_n x(t - nT),$$

where a_n is the amplitude of the nth pulse. The receiver samples $s(t)$ at T second intervals to determine the a_n. If one requires that the amplitude of any particular transmitted pulse can be determined by a single sample of the received signal, that is, for all integers m,

$$s(mT + \tau) = \sum_{n=-\infty}^{\infty} a_n x(mT + \tau - nT) = a_m ,$$

then $x(t)$ must be a pulse with zero intersymbol interference; that is,

$$x(nT + \tau) = \delta_{n0} , \tag{1}$$

where τ is some appropriate reference time and δ_{mn} is the Kroneker delta function.

Insofar as detecting the transmitted amplitude is concerned, no other specification on $x(t)$ is required. However, in most situations it is desirable, if not mandatory, to band-limit the spectrum of $x(t)$ to frequencies less than some cutoff ω_c. Of course, the smallest allowable value for ω_c is π/T, the Nyquist frequency. Such a band-limiting constraint might result from a requirement to limit adjacent channel interference. In many cases band-limiting is the only frequency domain specification which is required. These simple time and frequency domain specifications represent the minimum requirements that a pulse-shaping network must meet in order to be useful in many pulse amplitude modulation systems (see Fig. 1). It is important to observe that such specifications do not uniquely define $x(t)$ except for the case where $\omega_c = \pi/T$.

Given these specifications, the problem now becomes that of generating a realizable rational transfer function which can achieve the specifications for a given input; that is, the approximation problem must be solved. With this problem solved, the physical network can be constructed using known techniques.

There are numerous ways of solving the approximation problem both in the time domain and the frequency domain. However, a more complete
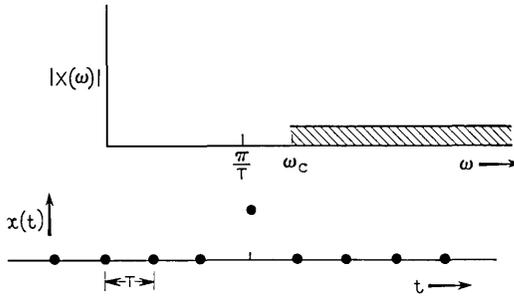
Fig. 1 — Minimum time and frequency domain specifications.

specification is generally required to use these techniques. For example, a standard frequency-domain approach is to completely specify a satisfactory $x(t)$ and to form the ideal transfer function of the network as $X(\omega)/Y(\omega)$, where $Y(\omega)$ is the Fourier transform of the network input. This transfer function is approximated by a rational function. The disadvantages of this straightforward approach are (*i*) whatever frequency domain measure of approximation accuracy is used, errors in the frequency domain are not easily related to errors at the sampling instants in the time domain, and (*ii*) completely specifying $x(t)$ requires the network to perform more shaping than is actually necessary. A particular selected $x(t)$ might give a transfer function which is more difficult to approximate than some other equally acceptable $x(t)$. Since the $x(t)$ most easily approximated is not known, specifying a particular $x(t)$ may require a transfer function of unnecessarily high order to achieve acceptable results.

Solving the approximation problem in the time domain permits more direct control of time domain errors. Ulstad has achieved good results in this manner; however, he completely specified $x(t)$.[3] In general, time-domain approximation procedures provide no direct control of the band-limiting properties of the network and one msut rely upon an accurate approximation of a completely specified $x(t)$ to achieve the band-limiting. Furthermore, when added weight in the approximation procedure is put at the sample times, the band-limiting properties of the network become increasingly difficult to control. This conflict, between approximating in the time domain with stress on sample times and achieving given band-limiting properties, seems to be common to most time-domain approximation techniques.

Jess and Schüssler considered the optimization of pulse-forming networks simultaneously in the time and frequency domains.[4] Their

approach, although a step in the right direction, minimizes the tails of the pulse; this does not necessarily give an acceptable value of intersymbol interference when the rate at which pulses are transmitted is a significant percentage of the Nyquist rate for the bandwidth available.

What is required is a method of approximating in the time domain which constrains the frequency domain behavior to be band-limited. Temes and Gyi show how to develop transfer functions which have band-limited impulse responses.[2] These ideas can be applied, with some modification, to give a useful solution to the problem of pulse shaping.

III. PULSE SHAPING USING THE TEMES AND GYI CONSTRAINT

For convenience, Appendix A reviews the manner in which Temes and Gyi develop a low-pass transfer function which has an equal-ripple stopband behavior. This is accomplished by expressing the transfer function in partial fraction form and constraining the residues to depend on the poles in a particular manner. If $G(s)$ is a transfer function with one zero at infinity, it is expressed as

$$G(s) = \sum_{i=1}^{N} \frac{R_i}{s - s_i},$$ (2a)

where

$$R_i = K z_i \prod_{\substack{j=1 \\ \neq i}}^{N} \frac{z_i + z_j}{s_i - s_j}$$ (2b)

and $z_i^2 = s_i^2 + \omega_c^2$, Re $z_i \geqq 0$; ω_c is the low-frequency edge of the stopband and K is the maximum gain in the stopband. It is important that this transfer function has all its zeros on the $j\omega$ axis and is therefore minimum phase.

In general, and specifically for the case where the input to the pulse-shaping network is a rectangular pulse, a minimum-phase transfer function does not have enough freedom to shape a pulse into a Nyquist pulse. (If a Nyquist pulse has a bandwidth of $(1 + \alpha)\pi/T$, where $0 < \alpha < 1$, then its Fourier transform must have linear phase over the frequency interval from zero to $(1 - \alpha)\pi/T$. To see this, compute the Fourier transform of the sampled pulse. This linear phase condition cannot be achieved, in general, with a minimum-phase shaping network.) To remedy this, the transfer function of equation (2) is multi-

plied by an all-pass transfer function which has the form,

$$H(s) = \prod_{i=N+1}^{N+L} \frac{-s - s_i}{s - s_i} ; \qquad \text{Re } s_i < 0. \tag{3}$$

The transfer function resulting from the product of equations (2) and (3) is band-limited and has arbitrary low-pass gain and phase characteristics. Observe that all the zeros of the transfer function $G(s)$ $H(s)$ (half of the available degrees of freedom) are constrained to be functions of the poles in order to get the band-limiting behavior. The poles (the remaining degrees of freedom) can now be used to optimize the time behavior of the pulse.

Assume for the present that the pulse to be shaped is rectangular, that is,

$$y(t) = u(t) - u(t - T_o), \tag{4}$$

where $u(t)$ is the unit step function and T_o the pulse width. From equations (2), (3), and (4), the output pulse $x(t)$ is

$$x(t) = \sum_{i=1}^{N+L+1} \mathcal{R}_i \{u(t) \exp (s_i t) - u(t - T_o) \exp [s_i(t - T_o)]\}, \tag{5a}$$

where

$$\mathcal{R}_i = \begin{cases} \dfrac{R_i H(s_i)}{s_i} , & 1 \leq i \leq N \\[3mm] \dfrac{G(s_i)}{s_i} \displaystyle\prod_{\substack{i=N+1 \\ \neq i}}^{N+L} \dfrac{-(s_i + s_i)}{(s_i - s_i)} , & N + 1 \leq i \leq N + L \\[3mm] G(0), & i = N + L + 1 \ (s_{N+L+1} = 0) \end{cases} \tag{5b}$$

and only simple poles are assumed to occur.

Equation (5) gives the output pulse in terms of the network poles. The pulse can now be optimized in the time domain using a digital computer and an appropriate optimization technique. For the particular application considered here, only the values of $x(t)$ at equally spaced intervals of time are important, that is, $t = kT + t_o$, where k is a positive integer, T is the sampling interval, and $-T < t_o \leq 0$. By concentrating on these instants of time rather than on the entire pulse waveform, excellent pulse-shaping networks can be designed which are not excessively complex.

The pulse shaping networks given in the following examples were designed by using a general purpose optimization program written

by Mrs. J. M. Schilling. The program used a steepest descent minimization technique. Typical running times on an IBM 7094 were about three to four minutes.

Figures 2 through 12 show the results of two sample designs and some experimental measurements. The first example (Figs. 2 through 6) is a seventh-order network which shapes a rectangular pulse into a Nyquist pulse with 50 percent excess bandwidth.* The network stopband rejection is 40 dB and the output pulse has a worst-case distortion of 0.38 percent.† The second example (Figs. 7 through 12)
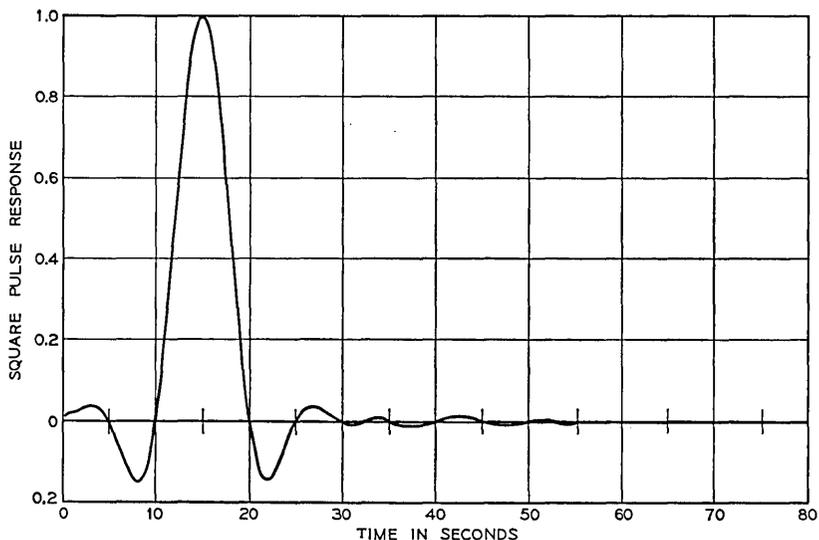
Fig. 2—Time response of a seventh order pulse-shaping network (one all-pass section) resulting from a rectangular input pulse of unit amplitude and of five seconds duration. The network has 40 dB stopband rejection, 50 percent excess bandwidth, and worst-case distortion of 0.38 percent. The Nyquist frequency is 0.1 Hz ($T = 5$ seconds).

shapes a rectangular pulse into a Nyquist pulse with 20 percent excess bandwidth. The worst-case distortion is 2.1 percent. For this example experimental measurements are shown for an active network realization using Tow's technique.[5]

---

* For a 50 percent excess bandwidth pulse, the low frequency edge of the stopband is at $(1.5/2T)$ Hz.

† Worst-case distortion is defined as $\sum\limits_{\substack{k=-\infty \\ \neq k_o}}^{\infty} | x(kT + t_o) |$, where $x(k_oT + t_o) = $ **1.**
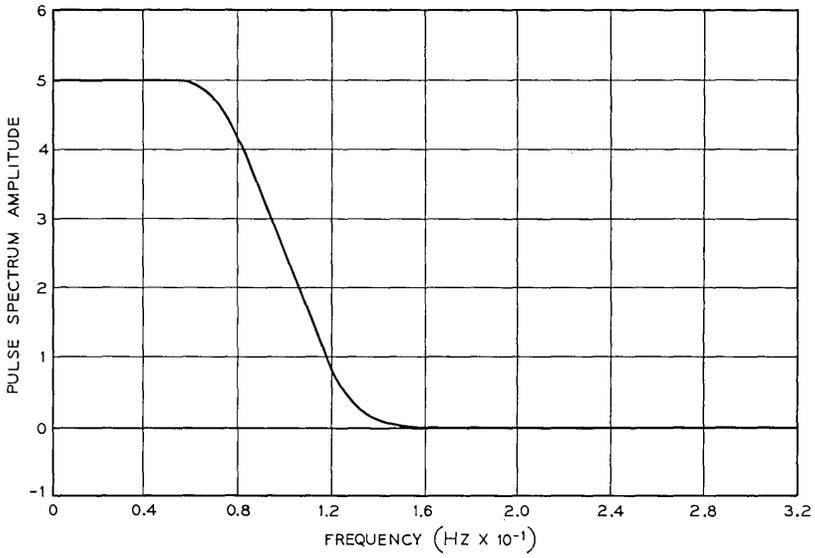
Fig. 3 — Pulse spectrum amplitude for time response shown in Fig. 2.
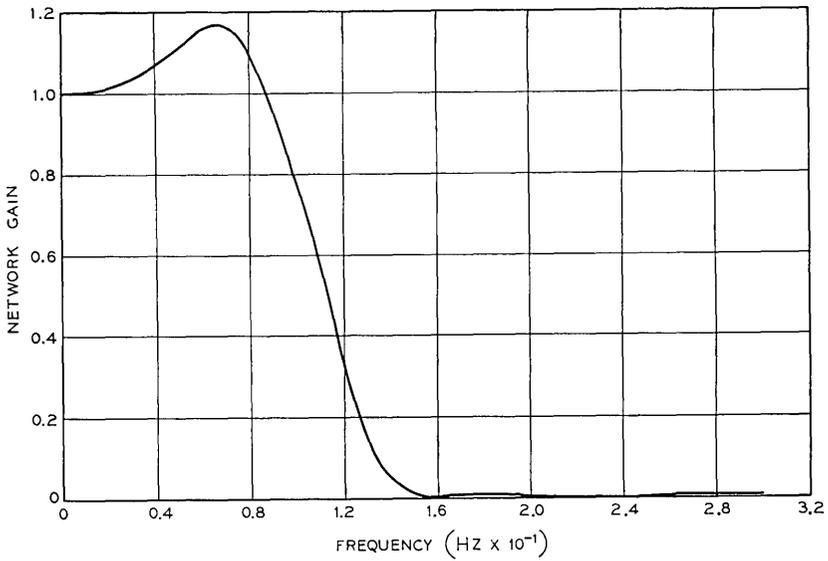
Fig. 4 — Gain of the network giving the output pulse shown in Fig. 2.
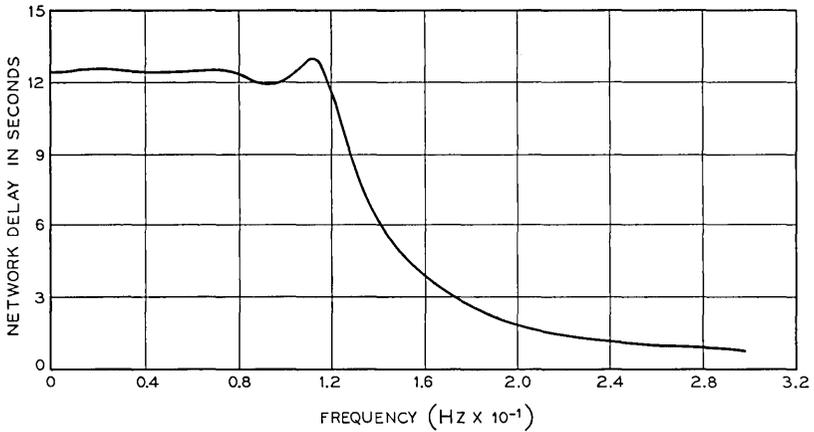
Fig. 5 — Delay of the network giving the output pulse shown in Fig. 2.

THE NUMERATOR POLYNOMIAL IS

+ 3.996795E-02S** 6

+-2.709413E-02S** 5

+ 1.275938E-01S** 4

+-8.224750E-02S** 3

+ 9.840381E-02S** 2

+-5.381433E-02S** 1

+ 1.244372E-02S** 0

THE DENOMINATOR POLYNOMIAL IS

+ 1.000000E+00S** 7

+ 1.835959E+00S** 6

+ 2.200246E+00S** 5

+ 1.799349E+00S** 4

+ 1.018142E+00S** 3

+ 4.043381E-01S** 2

+ 1.012517E-01S** 1

+ 1.244925E-02S** 0

THE POLES ARE

-1.999273E-01 ± 4.918577E-01J

-1.220059E-01 ± 7.300850E-01J

-3.389429E-01 ± 2.046175E-01J

-5.142068E-01   0.   J

THE ZEROS ARE

3.389439E-01 ± 2.046168E-01J

0.   ± 9.768803E-01J

0.   ± 1.442686E+00J

Fig. 6 — Transfer function data for the network of Figs. 2 through 5.
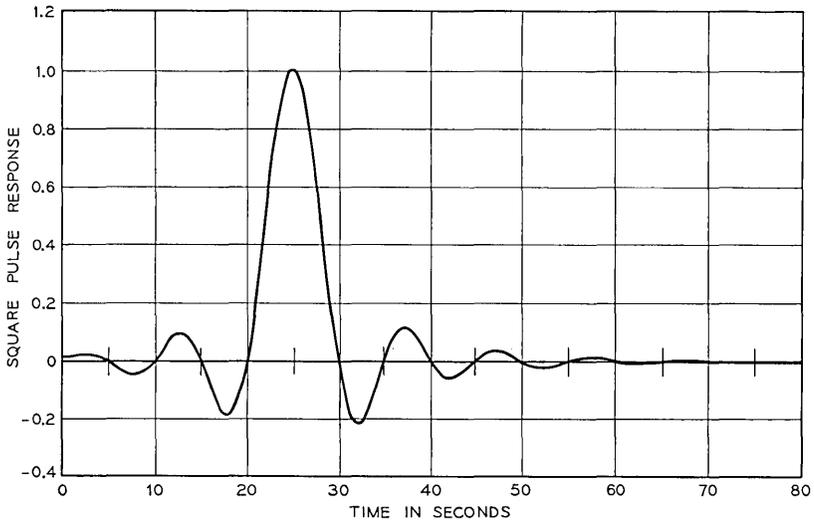
Fig. 7 — Time response of an eleventh order pulse-shaping network (two all-pass sections) resulting from a rectangular input pulse of unit amplitude and of five seconds duration. The network has 35 dB stopband rejection, 20 percent excess bandwidth, and worst-case distortion of 2.1 percent. The Nyquist frequency is 0.1 Hz ($T = 5$ seconds). Note: 35 dB network stopband rejection gives 40 dB or better rejection of signal energy when a rectangular pulse of five seconds duration is the network input.
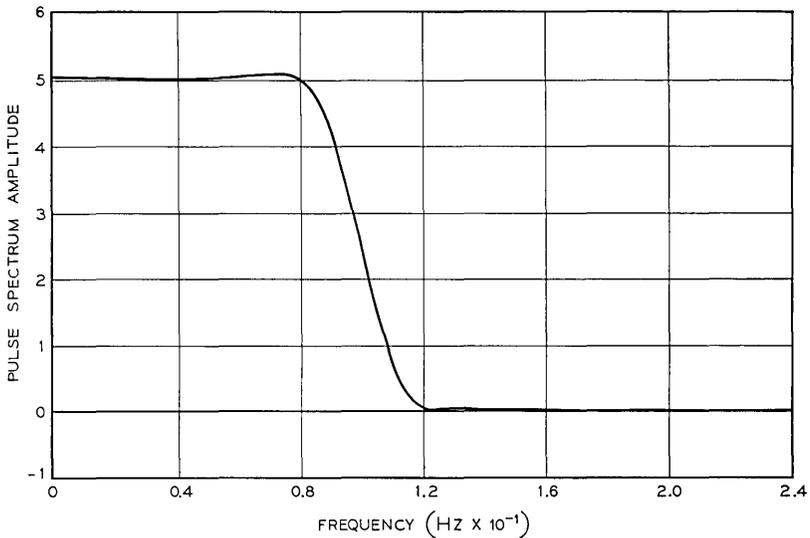
Fig. 8 — Pulse spectrum amplitude for time response shown in Fig. 7.
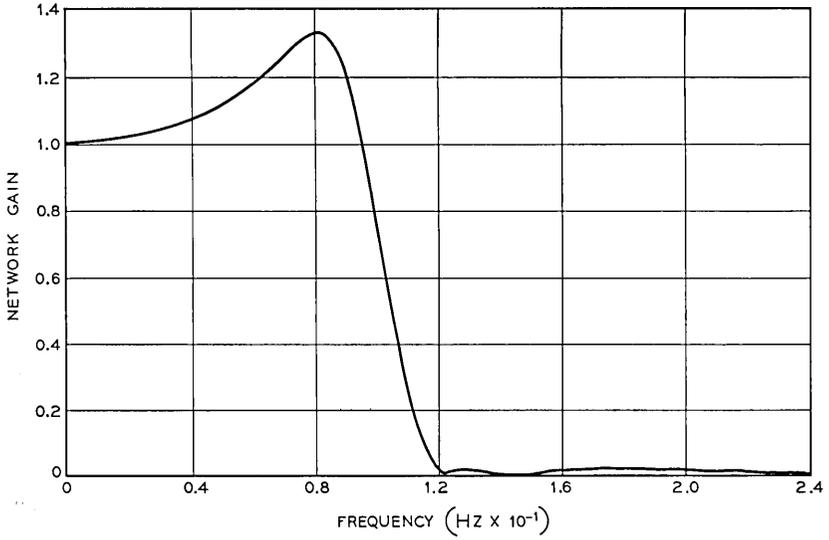
Fig. 9 — Gain of network giving the output pulse shown in Fig. 7.
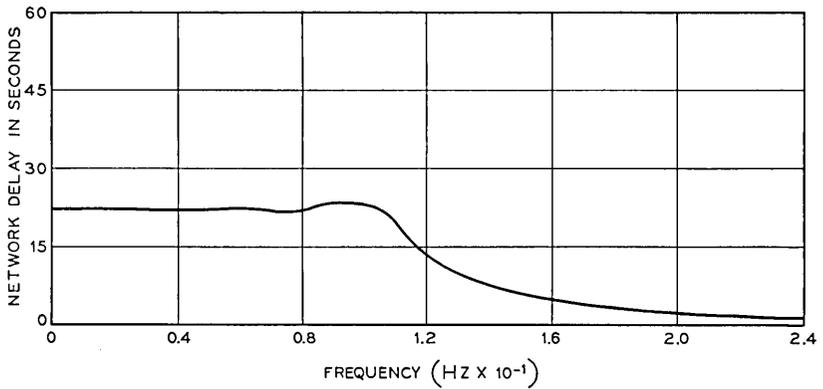
Fig. 10 — Delay of the network giving the output pulse shown in Fig. 7.

THE NUMERATØR PØLYNØMIAL IS

+ 9.010491E-02S**10

+-7.116008E-02S** 9

+ 3.996278E-01S** 8

+-2.963901E-01S** 7

+ 5.232092E-01S** 6

+-3.347661E-01S** 5

+ 2.679873E-01S** 4

+-1.276737E-01S** 3

+ 4.917976E-02S** 2

+-1.105460E-02S** 1

+ 1.200169E-03S** 0

THE PØLES ARE

-4.895659E-01 ±4.424214E-01J

-1.116714E-01 ±5.789655E-01J

-8.819848E-02 ±6.881603E-01J

-2.081902E-01 ±1.237181E-01J

-1.866762E-01 ±3.768048E-01J

-1.580778E+00   0.   J

THE DENØMINATØR PØLYNØMIAL IS

+ 1.000000E+00S**11

+ 3.749382E+00S**10

+ 6.603178E+00S** 9

+ 8.185644E+00S** 8

+ 7.463441E+00S** 7

+ 5.326736E+00S** 6

+ 2.948257E+00S** 5

+ 1.289339E+00S** 4

+ 4.230783E-01S** 3

+ 1.019484E-01S** 2

+ 1.565021E-02S** 1

+ 1.194575E-03S** 0

THE ZERØS ARE

  1.866781E-01 ±3.768034E-01J

  2.081920E-01 ±1.237177E-01J

  0.   ±7.677332E-01J

  0.   ±9.111657E-01J

  0.   ±1.620049E+00J

Fig. 11 — Transfer function data for the network of Figs. 7 through 10 and 12.
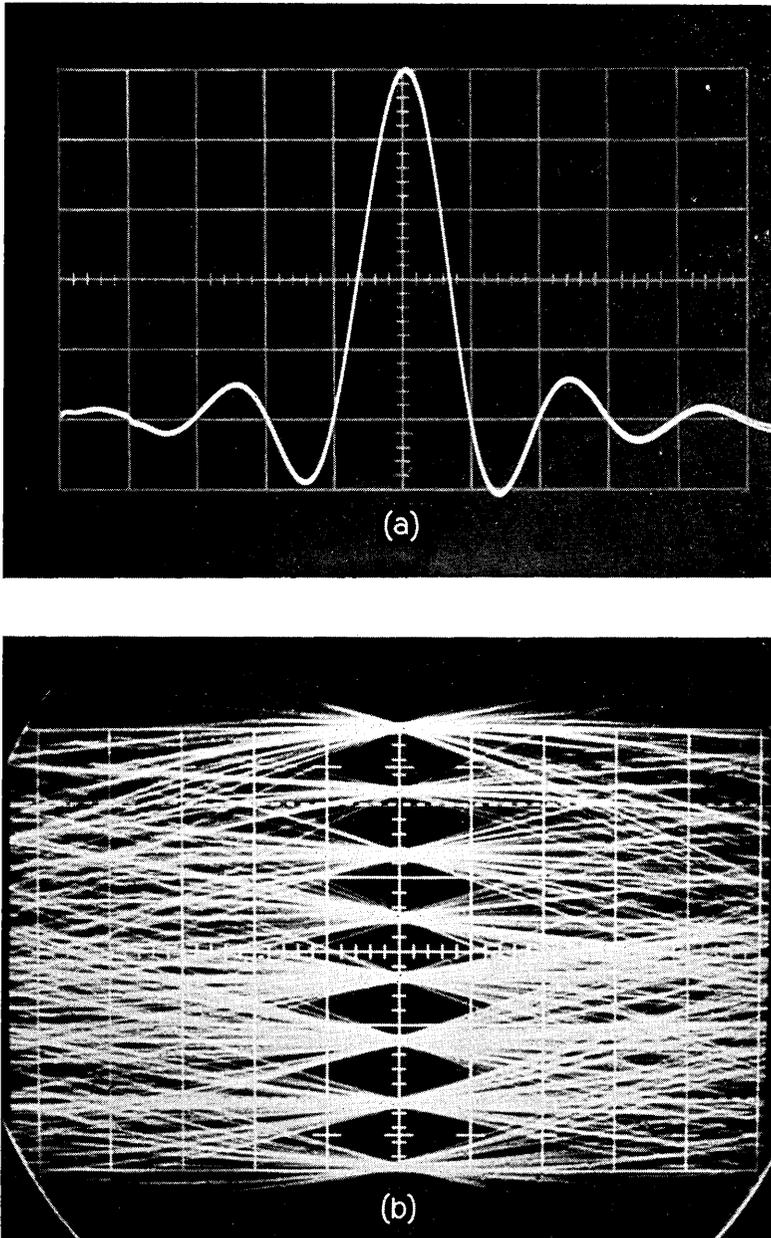
Fig. 12 — Experimental data on an active network realization of the transfer function given in Fig. 11. (a) Response of the network to a rectangular input pulse. The Nyquist frequency for the filter was 1800 Hz. (b) Eight level eye patterns generated by a random sequence of rectangular pulses with eight distinct amplitudes.

It should be stressed that although the design procedure used here gives excellent results, they are not necessarily optimum in any particular sense. It is clear from the complex way in which the poles enter into the time response of the output pulse that the error between the pulse samples, realized by equation (5) and the desired pulse samples, may not have a unique minimum; therefore, the computer program used to optimize the pole locations may actually converge to a local minimum. However, with a little experience the initial pole positions can be selected to give very satisfactory results.

IV. EXTENSIONS TO SHAPING ARBITRARY INPUTS

So far, only the shaping of a rectangular pulse has been considered. There are many situations where nonrectangular pulses must be shaped. For example, consider the case where the network is to shape a rectangular pulse to be transmitted over a channel which is no longer ideal as has been assumed so far; now the channel is assumed to introduce a known, fixed amount of amplitude and delay distortion. In this case, the pulse at the receiver is unchanged if the pulse-shaping network and the channel are interchanged (see Figs. 13a and b). Now the pulses presented to the shaping network from the channel are no longer rectangular. By having the network shape the channel output into a Nyquist pulse, the overall cascade connection of the pulse-shaping network and the channel shape a rectangular pulse into a Nyquist pulse. It is assumed that a solution to this problem is theoretically possible. A case which does not have a solution occurs when the channel is band-limited to less than $(1/2T)$ Hz.

For this example one might ask why the design process is rearranged in this manner. A more straightforward approach is to calculate the channel input required to give a particular Nyquist pulse at the channel output. The channel input is then approximated by the output of the shaping network. This approach has two disadvantages: (i) the channel output is over-specified, and (ii) the shaping network must approximate the channel input at more time points than is necessary in the other case.

In order to determine the output of the shaping network, $x(t)$, resulting from an arbitrary input, $y(t)$, one must perform a convolution. In general, a convolution is a very time-consuming calculation to carry out on a digital computer;* however, because the pulse-shaping net-

---

* A convolution must be performed *many* times when the poles of the pulse-shaping network are optimized by using a digital computer.

work is band-limited and we are interested in only equally spaced samples of the output pulse, this convolution can be made very efficient even without resorting to fast Fourier transform methods.

Since the pulse-shaping network is band-limited, an ideal band-limiting filter with the same bandwidth can be placed in front of it without appreciably affecting the shape of the output pulse (see Fig. 13b and c). Some effect occurs because the pulse-shaping network is not ideally band-limiting; this effect is small for reasonable stopband rejection levels. Now the input to the pulse-shaping network is band-limited. Since this is the case, the convolution can be performed using samples spaced at intervals of (π/ω_c) seconds or less, where ω_c is the cutoff frequency of the pulse-shaping network (see Figs. 13c and d). There is some aliasing error because the pulse-shaping network is not ideally band-limited; but this can be made small.

The bandwidth, ω_c, of the pulse-shaping network in all cases is greater than (π/T) radians per second (T is the time between successive pulses) and is usually less than $(2\pi/T)$ radians per second. For this situ-
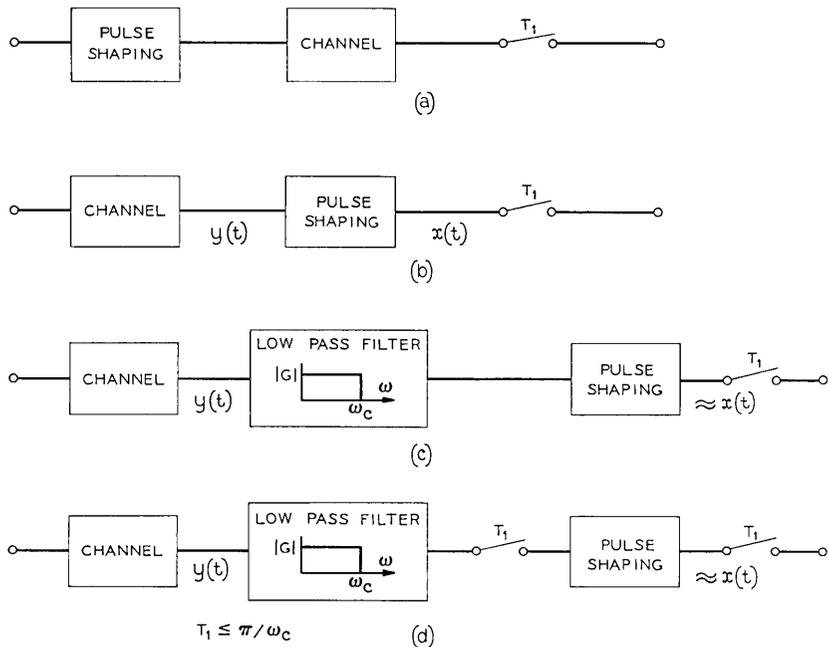
Fig. 13 — Approximately equivalent systems.

ation the output of the network can be found most conveniently by performing the convolution with samples spaced at intervals of length $T_1 = T/2$. Although a larger $T_1 (T_1 \leqq \pi/\omega_c)$ could be used, it requires interpolation to find the output at required sample times.

Figure 14 shows the results of the design of a pulse-shaping network to shape a nonrectangular pulse. The filters of a vestigial-sideband data transmission system were designed using a standard frequency-domain approach. The system was then simulated on a digital computer (assuming an ideal channel) and a binary eye pattern generated as Fig. 14a shows. The system was not ideal, as the figure indicates, because of errors introduced by the filters. The worst-case distortion was 64 percent.

The low-pass filter which follows the demodulator of the system was then designed, using the time-domain procedure described here. The order of the filter was kept the same. The portions of the data transmission system preceding the low-pass filter assumed the function of the channel as shown in Fig. 13. Figure 14b shows a binary eye pattern generated by a computer simulation of the data transmission system which incorporates the filter designed in the time domain. The worst-case distortion was 16 percent. The results in Fig. 14 occur without the aid of an automatic transversal equalizer.[6] When such an equalizer is used the results for both cases improve significantly and the advantage offered by the network designed in the time domain is reduced depending, of course, on the number of taps on the equalizer.

V. CONCLUSION

This paper has discussed a method of designing networks to shape arbitrary input pulses into band-limited Nyquist pulses. A modified Temes and Gyi constraint is used to keep the shaped pulses band-limited; the time responses are then optimized only at those time instants of interest. The resulting networks accurately realize both time and frequency domain specifications with minimum network complexity.

VI. ACKNOWLEDGMENTS

(a)

(b)
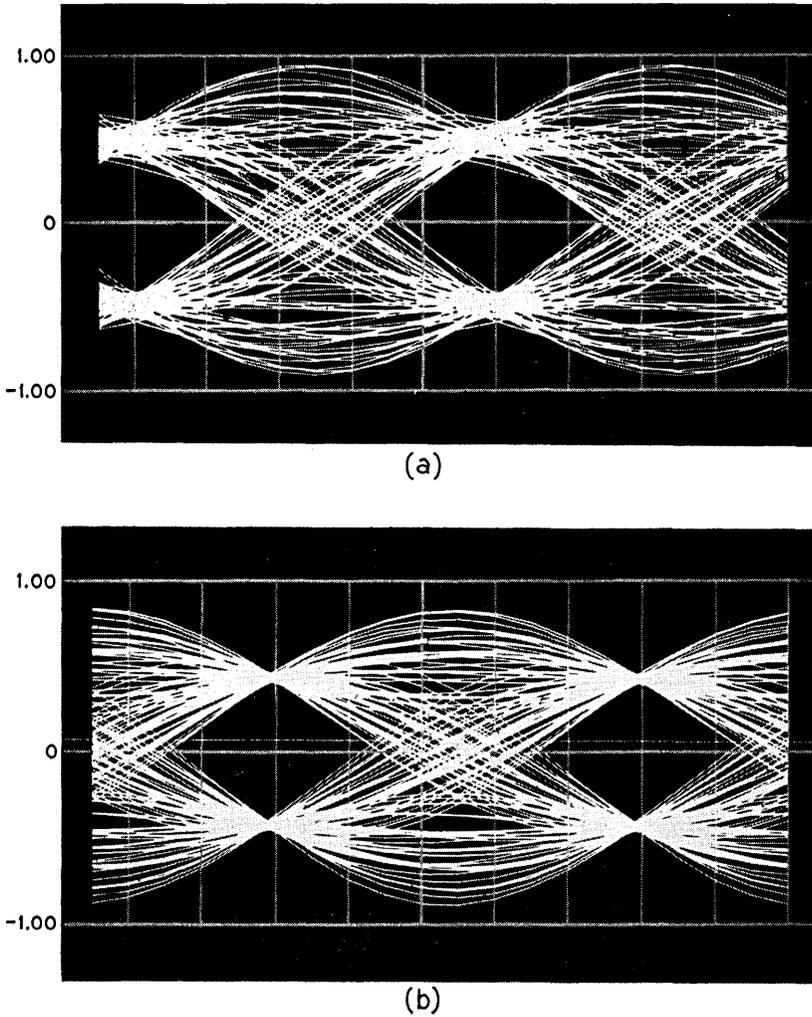
Fig. 14 — Computer generated binary eye patterns for a vestigial-sideband data transmission system. (a) Results occurring when the filters are designed using frequency domain techniques. (b) Results occurring when the low-pass filter following the demodulator is designed by the technique described in Section 4.

APPENDIX

Arbitrary Passband, Equal-Ripple Stopband Transfer Function of Temes and Gyi[2]

This appendix explains the procedure used by Temes and Gyi to develop $G(s)$, a low-pass, equal-ripple stopband, arbitrary passband transfer function. A rational $G(s)$ should (*i*) realize a gain less than or equal to some constant K for frequencies in the stopband, $| \omega | \geq \omega_c$ and (*ii*) have an arbitrary gain in the passband, $| \omega | < \omega_c$. This is achieved basically by using the poles of $G(s)$ to give the desired passband gain and the zeros of $G(s)$ to give the desired equal-ripple stopband gain.

To develop a transfer function with the desired gain properties, we consider the function $G(s)G(-s)$. For $s = j\omega$, $G(s)G(-s)$ equals the magnitude squared of $G(j\omega)$. Now consider the mapping of equation (6) which maps the s-plane to the z-plane:

$$z^2 = s^2 + \omega_c^2, \qquad \text{Re } (z) \geq 0, \qquad z = x + jy. \tag{6}$$

This mapping causes the stopband portion of $j\omega$ axis in the s-plane to correspond to the entire jy axis of the z-plane and the passband portion of the $j\omega$ axis in the s-plane to correspond to a portion of the x axis of the z-plane. The function $G(s)G(-s)$ can be transformed by equation (6) to the z-plane and, as will be shown, can be made to have equal-ripple stopband behavior by giving it the form of $H(z)$ in equation (7), where

$$H(z) = \frac{K^2}{1 + R(z)R(-z)} \tag{7a}$$

and

$$R(z) = zF(z)/E(z). \tag{7b}$$

$R(z)$ is a z-plane reactance function, and $E(z)$ and $F(z)$ are even functions. By transforming $H(z)$ to the s-plane and properly factoring it into $G(s)G(-s)$, the equal-ripple stopband transfer function is generated.

The z-plane reactance function in equation (7b) is written as an odd function over an even function. The reactance function could be the reciprocal of equation (7b); but, this form would not yield a rational $G(s)$. Since a reactance function has alternating poles and zeros on the jy axis and is pure imaginary there, $H(jy)$ has equal-ripple behavior ranging between K^2 when jy is a zero of $R(z)$ and zero when jy is a pole of $R(z)$. If $F(z)$ and $E(z)$ are the same order, $R(z)$ has a

pole at infinity; therefore, $G(s)$ has a zero at infinity. If $F(z)$ is of order two less than $E(z)$, $R(z)$ has a zero at infinity and $G(s)$ is not zero at infinity.

To determine $G(s)$ explicity, $H(z)$ is transformed into the s-plane and factored. To do this, $H(z)$ is written, using equation (7), as

$$H(z) = \frac{K^2 E^2(z)}{E^2(z) - z^2 F^2(z)} = \frac{K^2 E^2(z)}{[E(z) - zF(z)][E(z) + zF(z)]}. \tag{8}$$

Since $zF(z)/E(z)$ is a reactance function, $E(z) + zF(z)$ has roots in the left-half z-plane and $E(z) - zF(z)$ has roots in the right-half z-plane. Assuming that the polynomial $E(z) + zF(z)$ is nth order and the coefficient of z^n is unity, equation (8) can be factored into

$$H(z) = \frac{K^2 E^2(z)}{\left[\prod_{i=1}^{n} (z_i - z)\right]\left[\prod_{i=1}^{n} (z_i + z)\right]} = \frac{K^2 E^2(z)}{\prod_{i=1}^{n} [z_i^2 - z^2]}$$

where z_i are the roots of $E(z) - zF(z) = 0$. Using equation (6) and $s_i^2 = z_i^2 - \omega_c^2$, the s-plane version of $H(z)$ becomes

$$H(s) = \frac{K^2 E^2(z) \mid_{z^2 = s^2 + \omega_c^2}}{\prod_{i=1}^{n} (s_i^2 - s^2)} = \left[\frac{KE(z) \mid_{z^2 = s^2 + \omega_c^2}}{\prod_{i=1}^{n} (s - s_i)}\right]\left[\frac{KE(z) \mid_{z^2 = s^2 + \omega_c^2}}{\prod_{i=1}^{n} (-s - s_i)}\right]$$

$$= G(s)G(-s).$$

The s_i are the left-half plane images of the z_i. Therefore,

$$G(s) = \frac{KE(z) \mid_{z^2 = s^2 + \omega_c^2}}{\prod_{i=1}^{n} (s - s_i)} \tag{9}$$

is a realizable, rational transfer function. Note that $E(z)$ is an even function of z and thus is a rational function of s. Also all the zeros of $G(s)$ lie on the $j\omega$ axis in the stopband.

The construction of $G(s)$ is such that the poles, s_i, can be arbitrary (of course constrained to occur in comples conjugate pairs in the left-half plane). The numerator of $G(s)$ is found as a function of the poles by computing the polynomial

$$E(z) - zF(z) = \prod_{i=1}^{n} (z_i - z),$$

where $z_i^2 = s_i^2 + 1$, $\text{Re} (z_i) > 0$. The even part is taken and transformed back to the s-plane.

$G(s)$ can conveniently be expressed in partial fraction form as

$$G(s) = \sum_{i=1}^{n} \frac{R_i}{s - s_i}, \tag{10a}$$

$$R_i = \frac{KE(z_i)}{\prod_{\substack{k=1 \\ \neq i}}^{n} (s_i - s_k)}, \tag{10b}$$

where all the poles are assumed to be distinct and n is odd, so that $G(s)$ has a zero at infinity. $E(z_i)$ can be simplified:

$$E(z) + zF(z) = \prod_{k=1}^{n} (z_k + z),$$

$$E(z_i) + z_i F(z_i) = 2z_i \prod_{\substack{k=1 \\ \neq i}}^{n} (z_k + z_i),$$

and

$$E(-z_i) - z_i F(-z_i) = 0 = E(z_i) - z_i F(z_i).$$

The last equation is true since $E(z)$ and $F(z)$ are even functions of z. Adding the last two equations gives

$$E(z_i) = z_i \prod_{\substack{k=1 \\ \neq i}}^{n} (z_k + z_i),$$

which results in

$$R_i = Kz_i \prod_{\substack{k=1 \\ \neq i}}^{n} \frac{z_i + z_k}{s_i - s_k}. \tag{10c}$$

Thus, if the poles s_i are given, the residues R_i found from equation (10c) give a transfer function with equal-ripple stopband behavior.

Using this result the impulse response of $G(s)$ becomes, for odd n,

$$g(t) = R_n \exp(s_n t) + \sum_{i=1}^{\frac{n-1}{2}} \exp[\operatorname{Re}(s_i)t]$$

$$\cdot [2 \operatorname{Re}(R_i) \cos\{\operatorname{Im}(s_i)t\} - 2 \operatorname{Im}(R_i) \sin\{\operatorname{Im}(s_i)t\}]. \tag{11}$$

The real pole is s_n.

REFERENCES

1. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication,* New York: McGraw-Hill, 1968, Chapter 4.
2. Temes, G., C., and Gyi, M., "Design of Filters with Arbitrary Passband and Chebyshev Stopband Attenuation," 1967 IEEE Int. Conv. Digest, March 20-23, 1967, Paper 23.1, pp. 184–185.
3. Ulstad, M. S., "Time Domain Approximations and An Active Network Realization of Transfer Functions Derived From Ideal Filters," IEEE Trans. Circuit Theory, *CT-15,* No. 3 (September 1968), pp. 205–210.
4. Jess, J., and Schüssler, H. W., "On the Design of Pulse-Forming Networks," IEEE Trans. Circuit Theory, *CT-13,* No. 3 (September 1965), pp. 393–400.
5. Tow, J., "Active RC Filters—A State-Space Approach," Proc. IEEE, *56,* No. 6 (June 1968), pp. 1137–1139.
6. Lucky, R. W., "Techniques for Adaptive Equalization of Digital Communication," B.S.T.J., *45,* No. 2 (February 1966), pp. 255–286.

A Model for Generating On-Off Speech Patterns in Two-Way Conversation

By PAUL T. BRADY

This paper describes a model that generates on-off speech patterns representative of those in experimental two-way telephone conversations. The model assumes a conversant to occupy one of three speaking or one of three silent states. Transitions among the states are determined by Poisson processes governed by six parameters (one for each state). The validity of the model is tested by comparing the model computer simulation of 16 conversations with 16 real conversations. Cumulative distribution functions are compared for ten events (such as talkspurts, pauses, mutual silences, and so on) defined on the speech patterns. The model yields good fits to all events except "speech before interruption;" when an interruption occurs, a model speaker tends to interrupt the other's talkspurt later than a real speaker does.

Theoretical behavior of the model is also studied. All events consist of concatenations of exponentially distributed "state durations," even though most events are not themselves exponential. For some purposes, the exponential distribution is a satisfactory empirical fit to talkspurts, but not to pauses. Possible applications of the model include studying people's motivations to talk and fall silent on different circuits, and predicting statistical behavior of voice operated devices on the circuits.

I. INTRODUCTION

1.1 *Applications of the Model*

A model for generating on-off speech patterns in two-way conversations may have two uses:

(*i*) It may provide insight on the dynamic processes which determine when a person talks or is silent. For example, the model proposed here allows a person to be in one of six states, depending on whether he is talking, listening, or both conversants are talking, and so on. Each state is associated with a parameter which could be in-

terpreted as a "motivation" for either starting to talk or falling silent. As a subject talks over different experimental conditions, changes in the "motivation parameters" might be correlated with changes in subjective opinion of the circuit.

(ii) The model may predict the statistical behavior of voice-operated devices (such as echo suppressors, voice-switched amplifiers) as the circuits are changed. One alternative to using a model is to have people talk over different circuits and study the circuit behavior. This is often unsatisfactory because too much data may be required to isolate the effects of a particular circuit change. Another alternative to a model is to record an experimental prototype conversation and then play it over different circuits. This is also usually unsatisfactory because the conversants cannot react to circuit changes; their behavior remains the same. A model has the advantage of keeping the statistical structure of the "conversants" unchanged while allowing them to react as the circuit parameters are varied.

A model of on-off speaking patterns is not a new concept. The design of Time Assignment Speech Interpolation* was aided by the use of a number of one-way (that is, single speaker) models in parallel to simulate speech from many subscribers.[1] Jaffe, and others, have proposed a simple two-way Markovian model intended to study the speech behavior of psychiatric patients.[2] H. W. Gustafson of Bell Telephone Laboratories has proposed some improvements on the Markovian model to allow better prediction of speaker alternations.[3] The author has twice suggested a model; the first, with Mrs. N. W. Shrimpton (unpublished work), suggested a simple exponential fit to basic events such as talkspurts, and the second used a queueing system of "ideas" and "utterances" to yield a more complex model for talkspurts.[4]

The model proposed in this paper was developed after considering a large body of data from experimental two-way conversations conducted on telephone quality circuits containing no transmission delay or other degradations (See Table II footnote and Ref. 5).† To evalu-

---

* TASI is essentially a bank of voice-operated switches which may disconnect a subscriber from a channel when he is not talking to permit a talking subscriber to use the channel.

† Reference 5 describes an extensive statistical analysis of speech patterns in 16 conversations, and defines many "events," such as "talkspurt," "alternation silence," "pause in isolation," and so on. Average and median lengths for the events are tabulated, and cumulative distribution functions are included. The present paper assumes prior knowledge of Ref. 5. Notice that "event" is used to mean an interval of time, such as the interval of a talkspurt, and does not mean the occurrence of a probabilistic phenomenon such as the arrival of a pulse.

ate the model, we shall compare its simulation of the 16 conversations with data from the real conversations.

1.2 Relation of Model to Speech Detector

A speech detector is a rule which transforms speech into on-off patterns. Speech detectors are usually designed for specific needs, and vary considerably in their specifications. If a model is fit to one speech detector's output, then the model cannot be expected to be valid for all other detectors; but with minor changes, it may be adaptable to many of them.

The author's speech detector has previously been documented, and is described briefly here.[5,6] An initial hardware detector, with virtually no "pickup" and "hangover," yields a pattern of "spurts" and "gaps," after segmenting the speech into 5 ms intervals. All spurts $\leq$ 15 msec are presumed to be noise and are rejected (for throwaway); then all gaps $\leq$ 200 ms are filled in, as they were probably stop consonants or other minor breaks in continuous speech. The final on-off pattern contains, by definition, "talkspurts" and "pauses." No talkspurt can be $\leq$ 15 ms; no pause can be $\leq$ 200 ms. The model described here therefore generates talkspurts $\geq$ 20 ms and pauses $\geq$ 205 ms.

The speech patterns from a speech detector are strongly influenced by choice of threshold. In this study, the Ref. 5 data taken with a -40 dBm threshold were used as a basis for the simulated conversations.

1.3 Goals of This Paper

The remainder of this paper is divided into two main parts. Section II describes the model and illustrates its empirical behavior by comparing its output with real conversations. The question considered is: "Can this model generate patterns statistically similar to those of a randomly selected conversation?" We do not present data on applications such as determining differences among speakers or studying the behavior of a single speaker as he engages in various tasks. Future work is planned to investigate these problems.

Section III is a mathematical analysis of the model's behavior. From this analysis, one can gain an intuitive feeling of the model behavior, and acquire insight into the manner in which the two speakers interact. For a basic treatment of the model, however, Section III may be omitted. Section II assumes an elementary knowledge of probability theory; Section III requires some background in stochastic processes.

II. MODEL DEFINITION AND EMPIRICAL BEHAVIOR

2.1 *The Model*

2.1.1 *One-Port versus Many-Port Model*

Consider speakers A and B to be engaged in conversation. We shall model only speaker A's behavior and make no attempt to include B's behavior in the formulation. That is, speaker B's patterns are regarded only as they appear to A. It may be that B is really talking, but A does not receive him because of a blocking on the transmission line. Or, B may be delayed, and A may be receiving B's previous speech when in fact B is presently silent. We shall designate our model as a "one-port" model, since only one port, that is, A's side, is formulated. To use the model, it could be connected to anything, such as another one-port model, or a one-port model connected via a transmission delay, or several one-port models as in a conference circuit. (It may be invalid to assume that speaker A can be modeled the same way in a conference as in conversation with a single other speaker, but the model does at least allow such a connection to be formulated.)

In a many-port model, the entire system is modeled, with the drawback that a separate structure is required when special circuits are inserted between speakers. In addition, a one-port model leads to a description of each speaker, while if a many-port model is used, and a real conversation between A and B differs from one between A and C, it may not be possible to ascertain the change in speaker A's behavior. All we know is that the pair A-B is different from the pair A-C.

2.1.2 *Description of the Model*

Speaker A is either talking or silent, and he views B as either talking or silent. As Fig. 1 shows, in the simplest case four states occur at A's side. A is talking in the upper (shaded) half; B is talking in the right half. In considering transitions from state to state, as shown by the arrows, we apply the restriction that the two speakers cannot change their states at precisely the same time. Thus, in Fig. 1, diagonal crossings are prohibited.

Preliminary work with the Fig. 1 model showed that it was inadequate, especially in predicting events surrounding double talk. A natural extension of Fig. 1 is to expand each state into two states, the dichotomy decided by the previous state. Figure 2 illustrates the resulting 8-state model. Consider for example "A talks, solitary": to
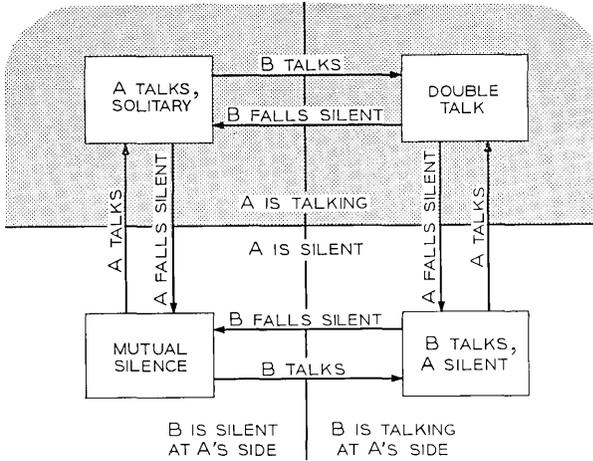
Fig. 1 — A four-state speech pattern model for speaker A. The shaded area indicates A is talking.

get to this state, either both speakers previously were silent or both were talking.

Figure 3, which is a reduction of Fig. 2, shows the model that the author has chosen to use. The upper left and lower right quadrants have been collapsed back to one state; simplicity has been gained at the expense of some loss of precision in modeling speech patterns.

Allowable state transitions are indicated on Fig. 3. There is no attempt to control B's behavior; he starts and stops talking in his own manner. Notice that his state changes cause horizontal transitions.

Vertical transitions are determined by A. If he is talking, he stops when a "fall silent pulse" occurs (Gustafson's terminology), and if silent he starts when a "start talking pulse" occurs. We call these $\beta-$ and α-pulses, respectively. These pulses are a result of Poisson processes,* so that, for example, if A is talking and B is silent (A solitary talk state), he stops talking in the next dt sec with probability $\beta_{sol}^A \cdot dt$.

For notation, the subscript on β, the fall silent parameter, describes the present state, while the subscript on α, the start talking parameter, denotes the event that will occur if the pulse occurs. The superscript refers to A or B. The six values for β and α are denoted β_{sol}, β_{ted}, β_{tor}, α_{pse}, α_{alt}, and α_{int} (See Fig. 3) in which the abbreviations mean solitary, interrupted, interruptor, pause, alternate, and interrupt.

---

* Poisson processes are tutorially discussed by Cox and Smith.[7]
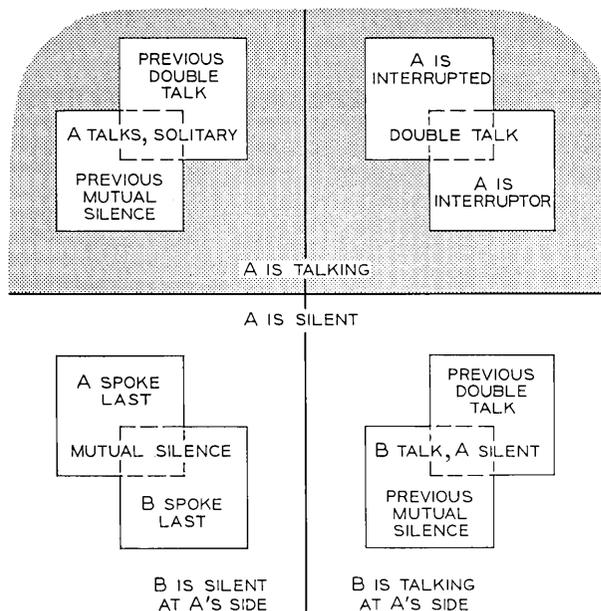
Fig. 2 — An eight-state model in which each state of Fig. 1 is divided into two states.

It is very important to understand the nature of the α or β parameters. They are not probabilities. However, if any α or β is multiplied by dt (for example, $dt = 0.005$ s), then αdt is the probability that A will leave the corresponding silence state "of his own volition" during the next dt seconds. (He may of course also be forced out of the state by B's action.) The αdt's and βdt's are "transitional" probabilities and do not represent the probability of being in each state. These "state" probabilities must be solved for, and can at times be difficult to obtain; they must consider the interaction of speaker A with his correspondent B. This is more fully discussed in Section III.

The α' and β's have a more appealing physical interpretation than just probability parameters. If some $\alpha = 2.5$, this implies that there is an "input stream" of α-pulses trying to drive A out of his state; the pulses occur at random times but at an average rate of 2.5 pulses per s, or with an average between-pulse interval of $1/2.5$ s. The units of α and β are pulses per second.

In general, none of the α's or β's is time dependent, so that the duration of occupation of a state has no effect on the value of that state's

α or β. An exception is that when A becomes silent, all α's are zero for 205 ms (so that only horizontal transitions can occur), after which time they resume their model values, and when A starts to talk all β's are zero for 20 ms. This guarantees that all silences are > 200 ms, and talkspurts are > 15 ms. (If an α-pulse occurs at the 210th ms, a 205 ms interval has occurred for that state, and the remaining 5 ms are assigned to the new state.)

A summary of the assumptions made in the model is:

(*i*) At any instant of time, A exists in one of six possible states.

(*ii*) A's talk-silence behavior is governed by Poisson processes, whose parameters are functions of the state A is in, but not of the length of time in the state (except for the previously noted minimum event length requirement).
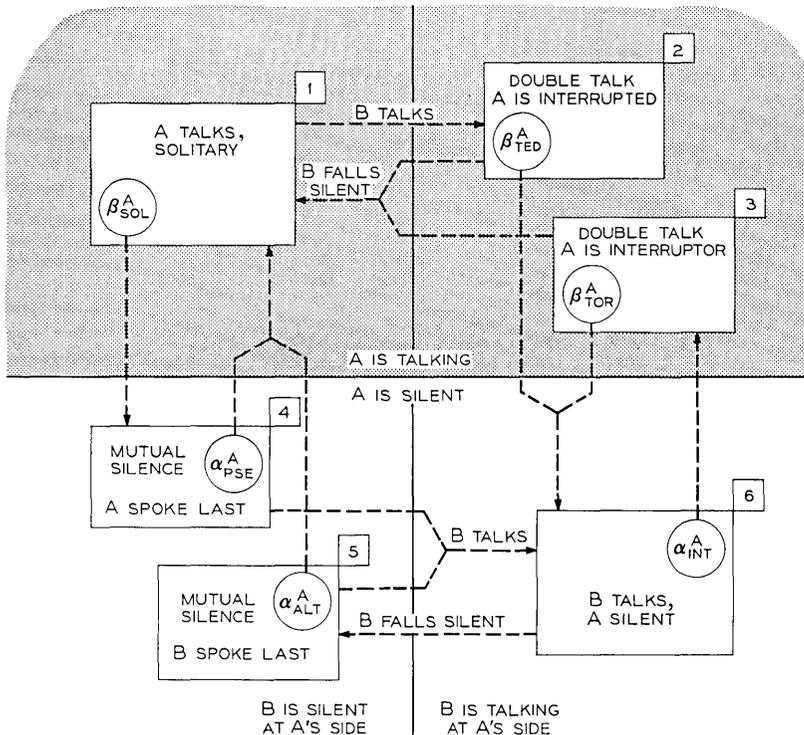
Fig. 3 — The six-state model used in this study. Vertical transitions are a result of Poisson processes at A's side. Horizontal transitions, resulting from B, are in A's external environment and are not generated by A's model.

(*iii*) The speakers cannot both change their speaking status at precisely the same instant of time.[†]

2.2 *Extracting the Model Parameters*

The six parameters for each of 32 speakers engaged in 16 conversations were derived in a very simple way: transition frequencies from state to state of the Fig. 1 model were counted by a brute-force stepping through each conversation.

To illustrate the process, recall that each person's speech is coded into 5 ms on-off intervals. Say that speaker A is in the solitary talk state (state 1, Fig. 3). β_{sol}^A can be found from a frequency count of A's falling silent from the state. Thus,

$$\beta_{sol}^A \cdot dt = \beta_{so:}^A \cdot (0.005) = \frac{\text{number of times } A \text{ falls silent from state}}{\text{number of times } A \text{ is in state, including numerator of this fraction}}. \qquad (1)$$

Whenever A is in a state, his behavior can be regarded as a succession of Bernoulli trials, in which case the above ratio is a best unbiased estimator for $\beta_{sol}^A \cdot (0.005)$, and hence of β_{sol}^A.

There are certain "trials" or 5 ms intervals which are not included in the frequency count. If A just begins to talk, he cannot leave the state until the talkspurt > 20 ms, or there are four intervals of 0.005 s; therefore, the first four intervals are not included. In silences, the first 21 intervals (205 ms) are not included. Also, if B's behavior produces a horizontal transition, this interval is not included, although the intervals up to that one are counted. The rare intervals containing both a horizontal and vertical transition are counted as vertical transitions. Table I is a list of the α and β values for all 32 speakers in 16 conversations.

2.3 *Testing the Model*

2.3.1 *Method of Testing*

To investigate the behavior of the model, a Monte Carlo simulator generated a model conversation of any desired length in a form which could be analyzed by Mrs. N. W. Shrimpton's speech analysis

---

[†] In the author's simulation, they cannot both change status in the same 5 ms time slot. This does occur in the author's data from the speech detectors, but it is very rare.

TABLE I—α AND β VALUES FOR 32 SPEAKERS, 16 CONVERSATIONS

β_sol	β_ted	β_tor	Conversation		Speaker	α_pse	α_alt	α_int
1.261455	3.411029	2.116402	1		1	1.831701	1.818579	0.563380
0.996458	2.225755	1.921885	1		2	2.457002	1.285416	0.588403
0.836082	2.018229	2.496434	2	FEMALE	1	3.610856	1.586914	0.319719
0.614610	1.817570	1.236264	2		2	3.177857	0.735383	0.193737
0.755323	2.134756	1.428163	3		1	1.391631	0.913121	0.354644
1.334627	1.687075	1.746107	3		2	1.701480	1.061364	0.426932
0.963705	2.773246	3.626714	4		1	1.497570	0.840841	0.281668
0.920078	1.170079	3.011515	4		2	1.514693	0.575869	0.139994
0.494011	1.156069	2.501303	5		1	2.033156	1.644157	0.107033
0.229878	1.290323	2.326664	5		2	3.231441	0.812301	0.183579
0.487893	1.976664	2.629389	6		1	1.880036	1.282691	0.406757
0.606033	1.421352	0.921261	6		2	3.042993	2.743970	0.688970
1.148240	2.318487	2.191781	7		1	1.442522	0.990402	0.123400
0.945180	3.740648	2.219321	7		2	1.271811	0.805795	0.205679
0.733397	2.586904	2.464066	8	MALE	1	2.099086	0.511809	0.088622
0.710458	1.932367	3.339192	8		2	2.207059	0.588446	0.199562
0.773362	1.574803	2.679831	9		1	1.606426	0.925181	0.230204
0.660433	1.994681	2.577710	9		2	1.742712	0.873908	0.217066
0.823098	4.733728	1.898734	10		1	1.790183	0.746016	0.118188
0.670252	6.010929	4.713805	10		2	1.646938	0.462535	0.103896
0.444459	1.107595	3.638569	11		1	2.203568	2.091714	0.417177
0.372235	1.805869	2.231405	11		2	3.431840	0.867276	0.275168
0.569062	1.234167	0.665083	12		1	1.260504	0.910657	0.271548
0.755803	1.996370	1.760921	12		2	1.784675	0.525237	0.239636
0.797666	1.438849	1.886792	13		1	1.798942	0.988468	0.275441
0.693121	1.790580	2.906574	13		2	3.064182	1.193967	0.260818
0.768697	1.940035	1.702128	14		1	1.563188	1.446204	0.356520
0.779857	2.012072	1.672862	14	FEMALE	2	2.225986	0.989827	0.116141
1.075338	1.115880	2.577320	15		1	2.040816	0.521993	0.206693
0.866694	1.822600	1.962533	15		2	2.952029	0.979129	0.474661
0.652447	2.364865	3.157895	16		1	2.615694	0.292312	0.134417
1.671175	2.427184	2.173913	16		2	4.990944	1.389631	0.343234

program. (The output of the program is illustrated in Ref. 5 and some of it is shown here.) The general procedure was to extract parameters from a real conversation and then simulate a conversation of 20 minutes duration. The original conversations were between 7 and 10 minutes long, but the simulated ones were longer to better estimate the true theoretical behavior. (Economic considerations prohibited simulations significantly longer than 20 minutes.)

If we could regard the two conversations of a real-simulated conversation pair as independent samples from two populations (or the same population), then classical statistical tests (such as t-test on means) would be appropriate. Unfortunately, the simulated conversations were derived from measurements of the real conversations, and standard tests no longer apply. For example, say that the talkspurt average lengths were very close for real and simulated conversations. With independent samples, this would suggest a good fit, but it may be that we have forced a good fit by setting simulated parameters equal to measured parameters of real speech.

Instead of using statistical tests, we define a "fit parameter," or FP, to indicate the correspondence between real and simulated events. This correspondence is examined for three quantities: average lengths of the events, cumulative distribution functions (cdfs) of the events, and rate of occurrence (for example, number of talkspurts per second). These three quantities are not independent; for example, a good fit of the cumulative distribution function (cdf) implies a good fit to the average (but the converse is not true). In assessing a good or bad fit of the model to the speech data, the fit parameters are not treated as yielding three independent pieces of information, but rather as representing three viewpoints of the goodness of a fit problem.

Table II is a list of the ten events. Two events, double talks (3) and mutual silences (4), merit a brief comment. In the experimental conversations, with no circuit degradation or delay, these events are identical for both speakers. However, for consistency with the other events, comparisons of the fit parameter are made twice, once for each speaker. This causes some redundancy in the tabulated comparisons of events 3 and 4 in Tables III through V.

2.3.2 Average Lengths

For average lengths, we define

$$FP_{(x)} \text{ (fit parameter)} \equiv \frac{\langle x \rangle_{real} - \langle x \rangle_{sim}}{\left(\dfrac{\sigma_{real}^2}{n_{real}} + \dfrac{\sigma_{sim}^2}{n_{sim}} \right)^{\frac{1}{2}}}, \qquad (2)$$

TABLE II—CATEGORIZED SPEECH EVENTS

Number	Event*
1	Talkspurt
2	Pause
3	Double talk
4	Mutual silence
5	Alternation silence†
6	Pause in isolation
7	Solitary talkspurt
8	Interruption
9	Speech after interruption
10	Speech before interruption

* For definition of "event" see Ref. 5.
† In Fig. 6 of Ref. 5, the alternation silences illustrated in the sample patterns are all incorrectly labeled. The A's and B's are transposed.

that is, the normalized difference between the means. If the observations were independent and from the same population, $FP_{(x)}$ would be normal, $\mu = 0$, $\sigma^2 = 1$. Independence is violated here, but we still can regard FP as an indication of similarity, and arbitrarily regard the fit as "bad" if $|FP_{(x)}| > 1.96$. Table III lists $FP_{(x)}$ for 10 events, 32 speakers. (See Ref. 5 for tabulated average lengths of real speech events.)

2.3.3 Cumulative Distribution Functions

In a two-sample Kolmogorov-Smirnov test, in which n_1 observations are made for one sample and n_2 for the other, the test statistic D is the maximum vertical discrepancy (absolute value) between the cumulative distribution functions for the two samples. If both n_1 and n_2 exceed 40, the identical population hypothesis is rejected at 0.05 level (see p. 131 of Ref. 8) if

$$D > 1.36 \left(\frac{n_1 + n_2}{n_1 n_2}\right)^{\frac{1}{2}}. \tag{3}$$

Again, in our data the samples are not independent, and although n_{sim} almost always exceeds 40, n_{real} often does not exceed 40 for those events surrounding interruptions. Nevertheless, we define

$$FP_{cdf} \equiv \frac{D}{1.36} \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}}. \tag{4}$$

If $FP_{cdf} > 1$, the fit will be considered bad. Table IV lists FP_{cdf} for 10 events and 32 speakers.

Comparative plots of cdf_{real} versus cdf_{sim} for all 10 events and all 32 speakers were generated. The curves for events 1 and 10 for speaker

TABLE III—$FP_{\bar{x}}$ FOR 32 SPEAKERS IN 16 CONVERSATIONS

Conversation	Speaker	Event 1	2	3	4	5	6	7	8	9	10
1	1	0.81	1.22	0.42	0.18	1.05	0.49	1.66	0.12	−0.24	1.50
	2	−0.06	−0.57	0.42	0.18	−0.35	−0.41	−0.34	0.50	0.87	−2.44*
2	1	0.54	0.65	0.14	0.44	0.81	−0.86	0.66	−0.40	−0.40	−1.07
	2	0.23	−0.38	0.14	0.44	1.25	−0.54	2.12*	−1.91	0.01	0.17
3	1	1.60	−1.12	1.28	−0.23	1.88	−1.05	3.67*	0.87	0.94	−2.58*
	2	0.67	1.26	1.28	−0.23	1.37	−2.50*	0.96	0.63	−0.38	−2.83*
4	1	−0.39	0.79	0.10	1.35	2.56*	−0.60	−0.03	0.54	−1.67	−2.62*
	2	−0.04	−0.42	0.10	1.35	1.08	−0.96	1.23	0.61	0.24	−2.56*
5	1	−1.53	−0.24	−0.15	0.46	0.26	−0.73	−2.35*	−0.48	−0.40	−1.20
	2	−0.63	−1.07	−0.15	0.46	0.19	0.56	−0.32	0.06	0.70	−2.25*
6	1	0.40	0.18	−0.04	0.34	−0.54	−0.50	1.67	−1.33	−1.58	−3.04*
	2	−0.44	0.37	−0.04	0.34	0.66	0.48	2.24*	−0.06	1.23	−0.64
7	1	−0.82	0.35	−0.44	1.06	1.97*	1.04	−.53	0.57	−1.17	−1.42
	2	0.24	−0.17	−0.44	1.06	0.30	−1.96*	1.31	−1.38	0.22	−1.19
8	1	−0.43	0.04	0.53	0.72	0.96	−0.15	0.58	−1.38	0.97	−2.63*
	2	1.30	0.18	0.53	0.72	0.26	0.18	2.02*	−1.13	−0.05	−0.07
9	1	0.58	0.62	0.39	0.03	−0.60	0.14	1.32	0.66	0.02	−0.47
	2	0.89	−0.27	0.39	0.03	−1.47	1.20	1.45	−1.33	−1.85	−0.23
10	1	0.39	0.25	0.15	0.93	−0.13	0.71	0.88	−0.17	0.19	0.87
	2	0.19	−0.08	0.15	0.93	−0.70	1.05	−0.23	0.16	−1.01	−0.39
11	1	0.81	−0.15	0.13	0.82	1.80	−0.67	2.73*	−0.84	−0.28	−0.50
	2	−0.52	0.26	0.13	0.82	0.96	−1.43	0.71	−0.53	−0.68	−2.47*
12	1	1.20	0.28	0.61	0.21	−0.77	−1.39	1.80	1.25	−0.03	−0.35
	2	−0.01	0.04	0.61	0.21	2.39*	0.43	−0.58	1.06	−0.87	−1.12
13	1	0.57	−0.78	−0.05	0.63	1.52	0.12	1.52	−0.61	−1.05	−1.66
	2	0.24	0.57	−0.05	0.63	−0.17	−0.23	1.62	−0.79	−0.21	−1.29
14	1	0.58	0.05	−0.64	0.23	0.75	−0.54	0.92	0.17	0.47	−3.12*
	2	0.39	−0.07	−0.64	0.23	0.45	−0.70	1.41	−0.87	0.66	−1.84
15	1	0.19	0.14	0.69	0.09	1.10	−1.16	0.55	−0.24	−0.35	−0.35
	2	0.33	−0.20	0.69	0.09	0.98	−0.85	0.96	−0.57	0.40	−1.60
16	1	0.05	−0.42	−1.26	−0.20	2.05*	−0.81	1.21	−0.39	1.47	−2.42*
	2	0.34	0.24	−1.26	−0.20	−0.05	−1.31	−0.00	−1.20	−1.49	−1.37
Total number of "bad" fits		0	0	0	0	4	2	6	0	0	11

$$FP_{\bar{X}} \equiv \frac{\bar{x}_{real} - \bar{x}_{sim}}{\left(\dfrac{\sigma_{real}^2}{n_{real}} + \dfrac{\sigma_{sim}^2}{n_{sim}}\right)^{\frac{1}{2}}}$$

The fit is considered "bad" if $|FP| \geq 1.96$; however, this must not be regarded as a statistical test. Bad fits are marked with asterisks. For typical values of $\bar{x}_{real}$ see Table IV of Ref. 5. Conversations 5 through 12 involve men, the rest involve women.

TABLE IV—FP_{cdf} FOR 32 SPEAKERS IN 16 CONVERSATIONS

Conversation	Speaker	Event 1	2	3	4	5	6	7	8	9	10
1	1	0.53	0.56	0.46	0.63	0.68	0.46	0.89	0.66	0.39	0.89
	2	0.59	1.44*	0.46	0.63	0.75	0.59	0.33	0.40	0.41	1.47*
2	1	0.87	0.74	0.47	0.51	0.53	0.78	0.93	0.34	0.49	1.07*
	2	0.39	1.17*	0.47	0.51	0.75	0.58	0.98	0.55	0.34	0.45
3	1	0.85	0.84	0.72	0.62	1.01*	1.15*	1.24*	0.37	0.49	1.63*
	2	0.97	0.74	0.72	0.62	0.65	1.02*	0.76	0.53	0.87	1.17*
4	1	0.80	0.97	0.56	0.69	1.28*	1.05*	0.67	0.52	0.80	1.16*
	2	0.40	1.15*	0.56	0.69	0.76	1.24*	0.66	0.23	0.56	0.91
5	1	0.61	1.16*	0.64	0.62	0.81	0.86	0.67	0.38	0.59	1.20*
	2	1.21*	0.69	0.64	0.62	0.52	0.47	1.13*	0.64	0.34	1.03*
6	1	0.77	1.11*	0.62	0.66	0.76	0.63	0.70	0.59	0.94	1.23*
	2	0.78	1.12*	0.62	0.66	0.68	0.62	1.06*	0.37	0.48	1.23*
7	1	0.76	0.48	0.38	0.74	0.84	0.85	0.50	0.53	0.46	0.69
	2	0.61	0.66	0.38	0.74	1.26*	0.90	0.76	0.72	0.47	0.89
8	1	0.90	0.63	0.47	0.65	0.99	0.70	0.70	0.54	0.41	1.14*
	2	1.38*	0.89	0.47	0.65	0.97	0.88	1.69*	0.37	0.42	1.15*
9	1	0.42	0.77	0.85	0.48	0.53	0.30	0.86	0.38	0.63	0.85
	2	0.52	0.84	0.85	0.48	0.76	0.48	0.78	0.58	0.82	0.86
10	1	0.90	0.89	0.45	0.54	0.69	0.79	0.98	0.62	0.40	0.58
	2	0.72	0.57	0.45	0.54	0.66	0.44	0.61	0.63	0.57	0.52
11	1	0.81	0.42	0.84	1.08*	0.92	0.49	1.25*	0.56	0.33	0.85
	2	0.88	0.75	0.84	1.08*	1.33*	0.60	0.46	0.91	0.88	0.79
12	1	0.78	0.63	0.52	0.55	0.34	0.72	1.25*	0.62	0.54	1.07*
	2	0.27	0.64	0.52	0.55	0.97	0.42	0.34	0.57	0.64	1.06*
13	1	0.60	0.78	0.56	0.39	0.54	0.62	0.78	0.49	0.39	1.10*
	2	0.52	0.41	0.56	0.39	0.46	0.45	0.56	0.51	0.48	0.50
14	1	1.24*	1.15*	0.59	1.08*	0.92	0.42	1.37*	0.28	0.50	0.95
	2	0.57	0.71	0.59	1.08*	1.22*	1.11*	0.64	0.32	0.77	0.77
15	1	0.50	0.96	0.83	0.57	0.59	1.03*	0.42	0.42	0.45	0.83
	2	0.53	0.58	0.83	0.57	0.96	0.60	0.66	0.70	0.40	1.08*
16	1	0.62	0.86	0.54	0.48	0.85	0.55	0.93	0.41	0.74	0.77
	2	0.65	0.68	0.54	0.48	0.33	0.71	0.72	0.56	0.63	0.78
Total number of "bad" fits		3	7	0	4	5	6	7	0	0	15

$$FP_{cdf} \equiv \frac{D}{1.36} \left(\frac{n_{real} m_{sim}}{n_{real} + n_{sim}} \right)^{\frac{1}{2}},$$

where D is the maximum absolute vertical distance between the two cumulative distribution functions. A "bad" fit occurs if $FP_{cdf} \geq 1.0$, as marked by asterisks. Cumulative distribution functions for events for all speakers collectively (for example, all talkspurts lumped together) are shown in Ref. 5.

TABLE V—VALUES FOR $FP_n \equiv \dfrac{\text{RATE OF OCCURRENCE OF SIMULATED EVENT}}{\text{RATE OF OCCURRENCE OF EVENT IN REAL CONVERSATION}}$.

Conversation	Speaker	Events							
		1	2	3	4	5	6	7	8
1	1	1.072	1.074	1.039	1.029	1.016	1.183	1.091	1.047
1	2	0.976	0.978	1.039	1.029	0.966	0.955	0.832	1.036
2	1	1.065	1.065	1.012	1.041	1.066	1.104	1.082	0.974
2	2	0.983	0.986	1.012	1.041	1.016	0.898	0.920	1.061
3	1	0.990	0.990	0.972	1.080	0.980	0.960	0.960	1.065
3	2	1.101	1.099	0.972	1.080	0.992	1.305	1.197	0.910
4	1	1.040	1.042	0.924	1.022	1.011	1.128	1.068	0.891
4	2	0.959	0.957	0.924	1.022	0.968	0.950	0.966	1.008
5	1	0.912	0.908	0.983	0.862	0.890	0.864	0.771	1.054
5	2	0.879	0.884	0.983	0.862	0.859	0.877	0.681	0.941
6	1	1.028	1.026	1.046	0.986	1.023	1.196	1.148	1.045
6	2	0.996	0.996	1.046	0.986	0.926	1.001	0.911	1.061
7	1	1.006	1.006	0.931	1.012	1.029	1.031	1.011	0.833
7	2	0.999	0.996	0.931	1.012	0.960	1.039	1.034	0.994
8	1	0.982	0.979	0.957	1.052	1.008	1.040	0.976	0.702
8	2	1.079	1.079	0.957	1.052	1.042	1.088	1.106	1.117
9	1	1.082	1.082	0.943	1.087	1.021	1.266	1.188	0.849
9	2	1.018	1.020	0.943	1.087	1.022	1.031	1.005	1.056
10	1	1.039	1.037	0.919	1.053	1.105	1.046	1.036	0.892
10	2	1.023	1.026	0.919	1.053	1.128	0.998	1.074	0.945
11	1	1.037	1.037	1.021	1.024	1.039	0.968	0.812	1.119
11	2	1.005	1.005	1.021	1.024	1.027	1.182	1.025	0.905
12	1	1.075	1.075	1.020	1.066	1.041	1.160	1.119	1.021
12	2	1.022	1.022	1.020	1.066	1.001	1.044	0.994	1.042
13	1	1.016	1.014	1.019	1.050	1.013	1.047	1.003	0.999
13	2	1.058	1.060	1.019	1.050	1.019	1.112	1.036	1.046
14	1	1.030	1.028	0.839	1.080	1.131	1.084	1.086	0.856
14	2	1.018	1.018	0.839	1.080	1.080	1.036	1.082	0.801
15	1	1.027	1.027	0.976	1.028	0.999	1.131	1.140	0.972
15	2	1.009	1.009	0.976	1.028	1.037	1.008	0.984	1.009
16	1	0.921	0.917	0.822	1.056	1.005	0.949	1.034	0.846
16	2	1.053	1.054	0.822	1.056	1.074	1.079	1.074	0.802

Events 8, 9, and 10 occurred an equal number of times and have equal FP_n.

2 of conversation 12 were arbitrarily selected for inclusion in this paper as Figs. 4 and 5. They illustrate a good and bad fit of the cumulative distribution functions, respectively. The plotted points are not data points; they represent category intervals of 15, 20, 30, $\cdots$ 200 ms, and 1, 2, 3 s, and so on. Thus, the number of asterisks in the cumulative distribution functions of real speech, or of breakpoints in the connected curves of cumulative distribution functions of simulated speech, do not equal n_{real} and n_{sim}, respectively.

2.3.4 *Rate of Occurrence*

To compare rate of occurrence of events

$$FP_n \equiv \frac{n_{sim}/\text{length of sim conversation}}{n_{real}/\text{length of real conversation}}. \qquad (5)$$

For a good fit, FP_n should be close to 1.0. When either n is small, FP_n may be changed considerably by the addition or subtraction of even one event, and unfortunately, FP_n does not consider the absolute values of the n's in the comparison, as do the other two FP's. In addition, we have not found a statistical test which is suitable for comparing rates of occurrences of events such as our speech events. Table V therefore is included only as a listing of the values of FP_n for eight events, 32 speakers without specifying good or bad fits. (Events 8, 9, and 10 occur an equal number of times; thus FP_n is equal for the three events.)
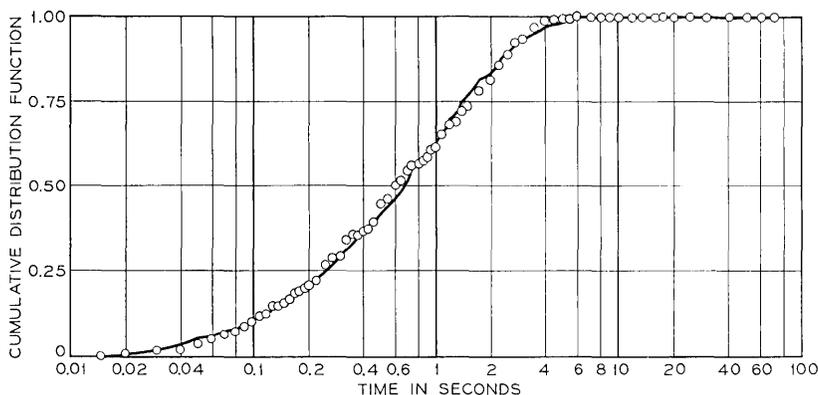

Fig. 4 — Real and simulated talkspurt distributions for speaker 2, conversation 12, illustrating a good fit. Circles are not data points; they occur at arbitrary category intervals. Circles represent real speech; connected curve, simulated patterns.
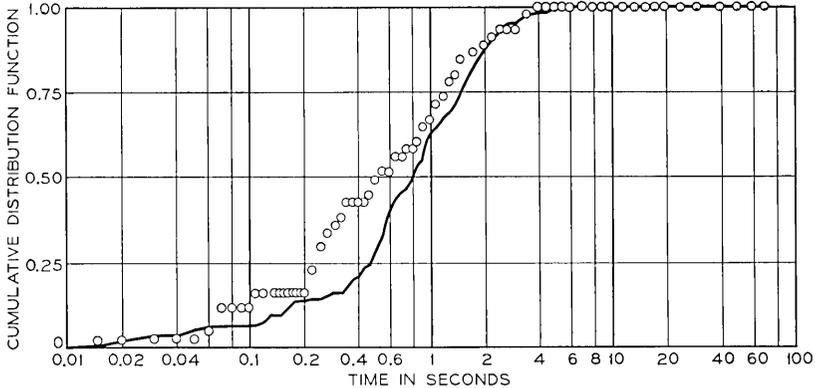
Fig. 5 — Real and simulated speech before interruption distributions, speaker 2, conversation 12, illustrating a bad fit.

2.4 Discussion

2.4.1 Goodness of Fit

In this study, the model is regarded as successful if it can match the distributions and rates of occurrence of the ten events listed in Table II. To see how well this criterion is met, the three FP's are considered separately.

First, notice in Table III that several columns (events) have no "bad" fits for average lengths. Now, in 32 trials of a legitimate 0.05 level test, we would expect about 1.5 failures; the six columns with no failures represent 160 trials (192 trials minus the 32 redundant trials for events 3 and 4) with eight expected failures. The lack of failures tends to rule out the $N(0, 1)$ distribution of entries in a column. Further, inspection of column 2, for example, reveals a variable which is apparently not $N(0, 1)$; 28 out of 32 (87.5 percent) of observations are within ± 1, as opposed to 68 percent for $N(0, 1)$. This substantiates our earlier statement that use of the FP does not constitute a legitimate statistical test.

Without regard to statistics, however, the Table III data do show a "good" correspondence between real and simulated (that is, model predicted) averages for all events except 10 (speech before interruption) and possibly 7 (solitary talkspurts). Looking now at Table IV, the model is clearly inadequate for event 10, but event 7 is not much worse than the rest. The Kolmogorov-Smirnov test (using $FP_{cdf} \geqq 1$ as failure criterion) is powerful and would be a severe test if statistical tests were valid; for this reason, we regard the cumulative distribution functions fits as generally successful (except for event 10), but with room for improvement. Further, some fits are remarkably close, such as the talkspurt

cumulative distribution functions for conversation 12 speaker 2 (see Fig. 4).

The rate of occurrence ratios in Table V are generally close to 1.0; there are a few scattered discrepancies but the model does not appear to have serious problems in generating a realistic number of events.

Regarding individual conversations, a study of Tables III and IV shows no tendency for the model to fail on particular speakers. The speakers in conversation 3 have a few more failures than seems normal, but even here the model exhibits "acceptable" fits for most events. The model appears equally valid for men (conversations 5 through 12) as for women (1 through 4 and 13 through 16).

2.4.2 Conversational Behavior

The failure of the model in predicting speech-before-interruption intervals may shed some light on the behavior of the subjects. Table III shows that the simulated intervals are too long; real people tend to interrupt sooner than predicted by simulation. This may be a question of reaction time. The model assumes that the instant A (who is silent) hears B begin to speak, A is immediately in a "listen to B" state and will speak only if he wishes to interrupt. In reality, A may require some time—perhaps 200 ms—before he adjusts to the presence of B's speech; in the meantime A's speech may not be intended as an interruption. A more sophisticated model might in fact assume the existence of a short delay in A's reception of B.

The numerical values of the α's and β's (Table I) also provide clues to behavior. The absolute values are a little hard to interpret since they are so closely related to the design of the author's speech detector. But notice that α_{alt} is less than α_{pse} for each speaker, confirming our intuitive belief that a person is more likely to resume talking after a pause he generates than after a pause the other party generates. This also justifies having two different states for A in which B is silent; the model would certainly deteriorate if the states were merged to one with an "averaged" α parameter.

Considering β_{tor} and β_{ted}, there is no consistent difference; $\beta_{ted} > \beta_{tor}$ for 15 of the 32 speakers. Thus, 15 (about half) of the subjects are more likely to terminate double talking if interrupted than if they are interruptors. A simpler model might merge these states, but serious errors might result for some subjects, since β_{ted} is often considerably different from β_{tor}.

The lowest of the α's is predictably α_{int}. A person is less likely to start talking when his correspondent is talking than when he is silent.

The precision of the α's and β's merits some attention. Because these quantities were measured over a person's entire conversation, they are not statistical estimators, but exact measures, correct to six figures. If you wish to regard the conversation as a sample of a larger population, however, you could regard an estimated value of α or β as a measure of a population α or β and establish confidence limits. The α's and β's were measured from Bernoulli trials where n varied from about 1000 to 40,000, depending on the conversation and parameter to be measured. Although the n's were large, the p values ($\alpha\,dt$) were generally very small, typically about 0.005; standard deviations of α or β estimates could equal about 0.1, with resulting 95 percent confidence limits of about ± 0.2.

2.4.3 Scope of the Model

Telephone conversations usually begin with a brief but rapid interchange of short words ("hello," and so on). In many calls the calling party then assumes dominance, and then possibly the other party may dominate. Our model attempts to duplicate speech patterns using six time-invariant parameters for each speaker and cannot, except by chance, generate the alternation of dominance which often occurs in real conversations.

The model is, however, a very simple one. With only six states we are attempting to simulate the utterance patterns of a person, who is certainly not a six-state device. Simplicity is also achieved by the Markovian technique of having a person leave a state with a time-invariant probability, independent of the duration of state occupation. (The minimum pause and talkspurt lengths constitute minor violations of this philosophy, but add little to the complexity of the model.)

The real issue here is not whether such a simple model can duplicate *all* aspects of conversation behavior, but rather whether such a model is useful on its own. The author plans to test it by using it to investigate speech behavior on circuits with transmission delay; another group at Bell Laboratories is studying its applicability to circuits with switched-gain amplifiers. The ease with which the model can be simulated, plus its success in matching overall patterns, gives it the potential of becoming an important tool in the study of conversational dynamics.

It may eventually prove worthwhile to extend the model and try to get a closer match to the dynamics of conversation. One way to do this would be to increase the number of states. This might improve the fit to the "total pattern" distribution, but might require a huge

number of states before a realistic "dominance alternation" occurs. Another way would be to introduce time-varying α and β parameters in the present six-state model. It appears that the development of either of these extended models (or a combination of them) would require an intensive amount of additional research.

III. MATHEMATICAL ANALYSIS

The principal goal of this section is to find theoretical distribution functions of the ten speech events in Table II. A complete analysis of the Fig. 3 model is not possible, but it is possible to analyze a simplified model and extend the results. For analysis, the model must be connected to another speaker. Section 3.1 considers speakers A and B to be directly connected with no minimum pause and talkspurt restrictions. Section 3.2 introduces these restrictions, to make the model match the author's speech detector. Section 3.3 considers an exponential approximation to talkspurts and pauses, and Section 3.4 discusses the effects on the analysis of introducing special circuits between subjects (transmission delay, echo suppressors).

3.1 *Direct Connection of Two Speakers*

Let the speech pattern model (Fig. 3) for speaker A be directly connected with one for speaker B. The entire A-B system thus exists in six states, since each state for A can be shown to correspond to a unique state for B. If all α's are forced to zero in the first 200 ms of silence, then a 200 ms minimum pause restriction is achieved; if β's are zero for 15 ms of talking, a 15 ms minimum talkspurt is achieved. In this section, we do not use these restrictions; we regard all α's and β's as time invariant. Because each state is terminated by a Poisson pulse from either A or B, the entire system is Markovian and the duration of each state has an exponential distribution.

This is illustrated, for example, by state 5. A will leave state 5 of his own volition in dt seconds with probability $\alpha^A_{alt} \cdot dt$. State 5 for A corresponds to state 4 for B; hence, B causes A to leave state 5 with probability $\alpha^B_{pse} \cdot dt$. A remains in state 5 with probability $1 - \alpha^A_{alt} \cdot dt - \alpha^B_{pse} \cdot dt$.* State 5 is thus terminated by a Poisson process with parameter $(\alpha^A_{alt} + \alpha^B_{pse})$; its duration is exponentially distributed with that parameter (see p. 154 of Ref. 9).

The appendix shows that even if only those events are considered in which A happens to terminate a state, these events are also exponential

---

* Cox and Smith give an expository treatment of this kind of analysis.[7]

with the parameter equal to the sum of the A and B "exit" parameters. For example, a "solitary talkspurt," in which A generates a talkspurt entirely within B's silence, is terminated when A leaves state 1 because of a β_{sol}^A - pulse. Nevertheless, A's solitary talkspurt is exponential with parameter $(\beta_{sol}^A + \alpha_{int}^B)$ and therefore has an average length of $1/(\beta_{sol}^A + \alpha_{int}^B)$. (State 1 at A's side corresponds to state 6 at B's side.)

This prediction for solitary talkspurt average lengths is well supported by simulation and is in fair agreement with actual speech data. Table VI compares the predicted average talkspurt lengths for 32 speakers with the measured averages from simulation. Only 2 out of 32 fail a 5 percent level test, which indicates that the simulator (that is, model) behaves as predicted.

Table VI also shows data from real speech. It is more appropriate to compare the real speech averages with simulated averages than with theoretical predictions, since the simulator contained the 15 ms and 200 ms minimum talkspurt and pause restrictions. Table III showed that 6 of the 32 average lengths of simulated solitary talkspurts were judged to be "bad" fits to empirical averages. In addition, a product-moment correlation of 0.91 is found for the two columns of average lengths in Table VI. A reasonably good fit is thus suggested; but Table VI shows that the real speech average exceeds the simulated average in 25 of the 32 cases. There is therefore a definite but mild tendency for the model (that is, simulator) to predict solitary talkspurts which are too short. This in no way refutes the result of the appendix, which is related only to the theoretical model.

In summary, the six-state Markovian system in this section may be solved by standard techniques. The following conclusions seem most relevant to speech analysis.

(*i*) A solution of the steady state probabilities of being in each of six states (that is, percent time in each state) may be obtained by routine solution of Markovian transition equations. This solution is not presented here because it is cumbersome, and it is not required for finding the distributions of durations of many of the states.

(*ii*) The distribution of the duration of A's being in any one of the six states is exponential with its parameter equal to the sum of the A and B parameters for leaving the state.

(*iii*) The distribution of three speech events may be immediately deduced. The events are:

(*a*) Alternation silence from B to A, in which B stops talking, there is a mutual silence, and A starts. This is distributed as the duration of state 5: exponential $(\alpha_{alt}^A + \alpha_{pss}^B)$.

TABLE VI—PREDICTED AND EMPIRICAL AVERAGE
SOLITARY TALKSPURT LENGTHS

Conversation	Speaker	Predicted (s)	Simulated		Real Speech	
			n	Ave (s)	n	Ave (s)
1	1	0.541	342	0.556	158	0.655
	2	0.641	246	0.628	149	0.606
2	1	0.971	377	0.983	183	1.035
	2	1.070	149	0.995	85	1.343
3	1	0.846	213	0.773	116	1.172
	2	0.592	380	0.566	166	0.621
4	1	0.906	374	0.969	208	0.966
	2	0.832	231	0.834	142	0.952
5	1	1.476	122	2.225*	87	1.631
	2	2.968	73	3.282	59	3.045
6	1	0.850	79	0.729	38	1.135
	2	0.987	175	0.957	106	1.259
7	1	0.739	363	0.832*	181	0.796
	2	0.936	320	0.942	156	1.064
8	1	1.072	222	1.114	117	1.192
	2	1.251	333	1.160	155	1.380
9	1	1.010	280	1.013	82	1.203
	2	1.123	231	1.025	80	1.242
10	1	1.079	356	1.071	116	1.157
	2	1.268	226	1.271	71	1.234
11	1	1.390	144	1.290	51	1.879
	2	1.267	82	1.214	23	1.467
12	1	1.237	226	1.130	77	1.405
	2	0.973	172	1.027	66	0.948
13	1	0.945	214	0.880	76	1.087
	2	1.032	282	1.012	97	1.238
14	1	1.130	336	1.121	118	1.221
	2	0.880	244	0.860	86	1.047
15	1	0.645	203	0.629	56	0.684
	2	0.932	388	0.893	124	0.992
16	1	1.004	116	0.958	18	1.335
	2	0.554	864	0.549	129	0.549

Predicted averages for A speakers (speakers 1) $= 1/(\beta_{sol}{}^{A} + \alpha_{int}{}^{B})$, for B speakers $= 1/(\beta_{sol}{}^{A} + \alpha_{int}{}^{A})$. Values for α's and β's were obtained from Table I. This prediction is slightly in error because of the 200 ms minimum pause requirement, as explained in Section 3.2. Significance (marked by asterisks) is at 0.05 level; $\bar{x}$ is assumed normal with mean = predicted average, $\sigma = \text{mean}/(n)^{\frac{1}{2}}$, since for a single observation from exponential distribution, $\sigma = \mu$. (Simulated and real speech n's are considerably different because lengths of conversations are different.) Product-moment correlation of simulated and real averages = 0.91.

(b) Pause in isolation, which has the distribution of state 4: exponential $(\alpha_{pse}^{A} + \alpha_{alt}^{B})$.

(c) Solitary talkspurt which is exponential with parameter $(\beta_{sol}^{A} + \alpha_{int}^{B})$. (State 1 also has this distribution; but A's being in state 1 does not imply a solitary talkspurt, since state 1 can be entered from double talking.)

(iv) Two distributions are a little more difficult, but straightforward.

(a) Double talk, in which states 2 and 3 are each exponential, but

with different parameters. The double talk density function is an average of the two exponential density functions, each weighted by the steady state probabilities of the states 2 and 3, respectively. The resulting distribution probably resembles an exponential, but is in general not strictly exponential unless states 2 and 3 are identically distributed.*

(b) Mutual Silence, which is the same as in the case of double talk, but with states 4 and 5, which are each exponential.

(v) The distributions of the remaining events of Table II are very difficult to derive. For example, we notice that a talkspurt can consist of an infinite possible sequence of states 1, 2, and 3.† Although there are techniques for handling problems of this type, they are complicated and in this case may yield formidable analytic expressions.

Notice that for this completely Markovian system of the ten speech events of Table II, only three—alternation silence, pause in isolation, and solitary talkspurt—are strictly exponentially distributed. But all events consist of concatenations of the six states, which in turn are exponential. We could think of these states as exponential "building blocks" with which the speech events are constructed.

3.2 *Effect of Minimum Pause and Talkspurt Length*

The introduction of time-varying parameters to obtain minimum lengths for pauses and talkspurts ruins the Markovian structure of the model, and standard techniques are not applicable for solution. However, certain results are still obtainable.

First of all, in the speech model, the 15 ms minimum talkspurt requirement is included because the author's speech detector, used to collect the speech data to test the model, uses a 15 ms throwaway for noise rejection in the raw speech data; hence all measured talkspurts exceed 15 ms. Even without the 15 ms restriction in the model, most simulated speech events are much longer than 15 ms, and we can anticipate only minor errors by ignoring the minimum length in the analysis.

The 0.2 s minimum pause is harder to deal with; it is long enough to affect the results. In general, constant lengths of 0.2 s are added to exponentially distributed silent state durations. We refer to the resulting distribution as constant-plus-exponential.

In some cases, one of the state exit parameters (say β) may be zero for 0.2 s, while the other (α) remains at its usual value. Then, the first

---

* Averaging two density functions is not equivalent to averaging two independent exponential random variables. The latter operation yields a gamma distribution.

† Certain sequences are not possible, such as 1, 3, 2; but there are still infinitely many allowable ways A can wander among the three states before falling silent.

0.2 s is exponential with parameter α, and the probability that the interval will extend beyond 0.2 s is $e^{-0.2\alpha}$. Intervals beyond 0.2 s are constant-plus-exponential distributed, with constant = 0.2 s and exponential parameter = $(\alpha + \beta)$. Since the relative fraction of less-than versus greater-than 0.2 s intervals is known, the total distribution can be found by combining the pre- and post-0.2-s exponential segments.

These results can be used to draw the following conclusions regarding event distributions.

(*i*) Alternation silence (state 5): assuming A has not talked for 200 ms prior to the state entry;* this is exponential (α_{alt}^A) for 0.2 s, and then exponential ($\alpha_{alt}^A + \alpha_{pse}^B$).

(*ii*) Pause in isolation (state 4): these must be at least 0.2 s long, since A cannot terminate state 4 until that time. Hence, these are constant-plus-exponential distributed; constant = 0.2 s, exponential parameter = $(\alpha_{pse}^A + \alpha_{alt}^B)$.

(*iii*) Solitary talkspurt tends to be exponential ($\beta_{sol}^A + \alpha_{ini}^B$); most state 1 durations are unaffected by the minimum pause requirement.†

(*iv*) Double talk: states 2 and 3 distributions are completely unaffected by the minimum pause requirement, but their relative steady state probabilities may be changed somewhat, thus affecting the blend of the two density functions. The effect is probably slight, however, and the general shape of the distribution still looks very much as it did without the minimum pause length. This has the appearance of an exponential distribution, although not precisely exponential.

(*v*) Mutual Silence: this distribution was predictable without the minimum pause requirement, but it now appears to be very complex and strongly affected by the 200 ms constant. All mutual silences which are "pauses in isolation" are at least 200 ms long, and those which are "alternation silences" usually start exponentially with parameter α_{alt}^A, and after 200 ms they become exponential with parameter $(\alpha_{alt}^A + \alpha_{pse}^B)$. Figure 5 of Ref. 5 clearly shows the importance of the 200 ms constant in mutual silences.

(*vi*) Remaining events are too complex to predict. Certainly, however, all talkspurts start with a 15 ms constant duration, and pauses start with a 200 ms constant duration.

3.3 *Exponential Approximation to Talkspurts and Pauses*

Exponential and constant-plus-exponential events are easy to simu-

---

* Ref. 5 data suggest that less than 10 percent of state 5 intervals begin within 200 ms of A's speech.

† Based on data from Ref. 5, we estimate that only about 6 or 7 percent of state 1 intervals begin with 200 ms of B's speech.

late. If one wanted to generate artificial talkspurts and pauses and was unconcerned with speaker interaction, could he use such a simplified model? We tried such a fit to the empirical talkspurt and pause distributions of the conversations described in Ref. 5.

Talkspurts were fit by a straight exponential distribution, without a 15 ms constant, in which the exponential parameter was deduced from the average event length. That is, for a particular speaker let

$$\beta_{ts} \equiv 1/\text{average talkspurt length;} \tag{6}$$

then

$$\text{Prob } (T \leqq t) \text{ (cumulative function)} = 1 - \exp (-\beta_{ts}t). \tag{7}$$

For pauses, we used a constant-plus-exponential. Let

$$\alpha_{pse} \equiv 1/(\text{average pause length } -0.2), \tag{8}$$

that is, the reciprocal average of the above 200 ms part of all pauses. Then

$$\Pr (T \leqq t) = \begin{cases} 0 & \text{for } 0 \leqq t \leqq 0.2 \\ 1 - \exp [-\alpha_{pse}(t - 0.2)] & \text{for } t > 0.2. \end{cases} \tag{9}$$

For comparing distribution functions, a Kolmogorov-Smirnov test was used to see if the empirical distribution function came from the particular exponential function based on β_{ts} or α_{pse}.[10] Once again, the statistical test is not strictly appropriate, since the mean of the exponential function is forced equal to the sample average. It still appears, however, to be a reasonable heuristic method to determine if the "shape" of the curve is exponential. Only four out of 32 sets of talkspurts fail the test, suggesting that the exponential model is a good approximation for talkspurts. This is in agreement with the findings of Jaffe and others.[2] None of the pauses fit constant-plus-exponential. This probably results from trying to fit one distribution to two distinctly different kinds of pause: pause in isolation, which occurs between words and is short, and the long silence which occurs when listening to the other speaker.

The good exponential fit to talkspurts might cause one to feel that the talkspurts could be modeled by a single parameter Poisson process. This would be achieved by having a single "talk" state, instead of three; once the state is entered, the speaker would ignore the other's speech and stop talking when his single parameter β-pulse occurred. Although a reasonable talkspurt fit would be achieved, other speech events, such as double talk and interruptions, would be poorly

matched for most speakers. This is true because the measured values of the three β's of Fig. 3 are generally quite different, with the two double talking β's often different from each other and typically at least twice β_{sol} (see Table I). A single parameter Poisson process would incorrectly assume these β's to be equal to each other.

Why, then, do we get a good exponential fit to the general talk-spurt distribution? Table II of Ref. 5, for -40 dBm threshold, shows that state 1 accounts for about 88 percent of A's talking time, so that the different β's during double talking exert only a minor effect upon the predominant state 1 single parameter Poisson process.* That is, the long and frequent state 1 intervals tend to obliterate the fine structure of the double talks.

3.4 *Connection of Two Models Over Special Circuits*

When A and B are directly connected, equations at A's side are easily written because knowledge of A's state at a random instant implies knowledge of B's state. (Once again, for simplicity, assume a Markovian model with no minimum event length requirements.) Analysis becomes very difficult when the circuit prohibits such knowledge. Two such circuits are considered here: Circuits with transmission delay and with echo suppressors.

3.4.1 *Delay*

The feasibility of transmitting two-way telephone calls over satellite circuits has generated widespread interest in the effects of transmission delay on the behavior of the conversants. We have previously dealt with a system which connected two three-state Markovian devices over a channel with transmission delay.[4,11] The following conclusions are of interest here.

(i) If the delay is "short" (in the order of average pause lengths or less, as occurs in cases of practical interest, where $D \leqq 1200$ ms), an exact analysis has not yet been found, and approximations are required to solve even the simple three-state system.

(ii) For very long delays, asymptotic system behavior of the model is obtainable; but the model is of doubtful validity since an entirely different kind of speech behavior might result from excessively long delays.

---

* One or the other speaker, but not both, talks for 100-24.99 (mutual silence) — 4.62 (double talk) $=$ 70.39 percent of the time; this accounts for states 1 and 6 at A's side. State 1 is occupied about half this time, or 35.20 percent. This is 88 percent of 35.20 $+$ 4.62, which is A's talking time.

3.4.2 *Echo Suppressors*

For our purposes, echo suppressors are devices which occasionally block the A to B or B to A (or both) transmission paths, at times depending on the interaction of the A and B speech patterns. (For further details on echo suppressors see Ref. 12.) There may also be delay, but even without delay the time dependency and uncertainty in the system is apparent and virtually prohibits formal analysis.

For both delay and echo suppressors, simulation is not difficult (the author's simulator already incorporates delay) and provides at present the only means of assessing the performance of the model.

3.5 *Summary*

The six-state model described by the author contains time dependencies which prevent formal Markovian analysis, but there is a tendency for the speech events to be formed from exponential, and in some cases, constant-plus-exponential "building blocks." Practically all of the exponential blocks or exponential parts of the constant-plus-exponential blocks have distributions with parameters equal to the sum of the A and B "exit probability" parameters; and even those events which seem exclusively a result of one speaker (such as solitary talkspurts) are in fact influenced by both speakers in a predictable way.

Although several theoretical results are obtainable, one is forced to turn to simulation for complete quantitative results. The ease by which the model is simulated helps compensate for the numerous computer runs required for studying model behavior as a function of parameter or circuit changes.

IV. ACKNOWLEDGMENT

APPENDIX

Distribution of a State Terminated By a Particular Speaker

This appendix is a derivation of the result stated in Section 3.1, that if, for example, one considers only those state 1 intervals terminated

by A, these will have an exponential distribution with parameter $(\beta_{sol}^A + \alpha_{int}^B)$. For shorthand, we call the parameters β and α. Let state 1 begin at time $t = 0$. The joint probability that it is t s long and terminated by A is:

$$\Pr \text{ (terminated by } A \text{ in } t, t + dt) = e^{-(\alpha+\beta)t \cdot} \beta \, dt; \qquad (10)$$

that is, neither an $\alpha-$ or $\beta-$ pulse can occur in $(0, t)$, and one β-pulse must occur in $(t, t + dt)$. Integrating equation (10) over all t,

$\Pr$ (state terminated by A at any time)

$$= \int_0^\infty (10) \, dt = \beta/(\alpha + \beta), \qquad (11)$$

as it should. We desire the conditional probability that state 1 ends in $(t, t + dt)$ given that it is terminated by A. By Bayes' rule,

$\Pr$ (state ends in $(t, t + dt)$ | terminated by A)

$$= \frac{\text{joint } \Pr \text{ (state ends in } (t, t + dt) \text{ and is terminated by } A)}{\Pr \text{ (terminated by } A)}$$

$$= \text{equation (10)/equation (11)} = e^{-(\alpha+\beta)t}(\alpha + \beta) \, dt, \qquad (12)$$

which is recognized as an exponential density function with parameter $(\alpha + \beta)$.

REFERENCES

1. Fraser, J. M., Bullock, D. B., and Long, N. G., "Overall Characteristics of a TASI System," B.S.T.J., *41*, No. 4 (July 1962), pp. 1439–1454.
2. Jaffe, J., Cassotta, L., and Feldstein, S., "Markovian Model of Time Patterns of Speech," Science, *144*, No. 3620 (May 15, 1964), pp. 884–886.
3. Gustafson, H. W., "Model for the Analysis of Talkspurt and Silence Durations in Conversational Interaction," Proc. 77th Annual Conv. Amer. Psychological Assn., *44*, Part I (1969), pp. 43–44.
4. Brady, P. T., "Queuing and Interference Among Messages in a Communication System with Transmission Delay," Ph.D. Thesis, Department of Electrical Engineering, New York University, June 1966.
5. Brady, P. T., "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., *47*, No. 1 (January 1968), pp. 73–91.
6. Brady, P. T., "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44*, No. 1 (January 1965), pp. 1–22.
7. Cox, D. R., and Smith, W. L., *Queues*, Methuen: London, 1961.
8. Siegal, S., *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill, 1956.
9. Cox, D. R., and Miller, H. D., *The Theory of Stochastic Processes*, New York: Wiley, 1965.

10. Hoel, P. G., *Introduction to Mathematical Statistics,* New York: Wiley, 1962.
11. Brady, P. T., "A Stochastic Model of Message Interchange on a Channel With Transmission Delay," IEEE Trans Commun. Techniques, *15,* No. 3 (June 1967), pp. 405–412.
12. Unrue, J. E., "Echo Suppressor Design Considerations," IEEE Trans. Commun. Techniques, *16,* No. 4 (August 1968), pp. 616–624.

Mobile Radio Diversity Reception

By E. N. GILBERT

This paper examines a particular kind of diversity system, under conditions of multipath fading, when there is interference from either random noise or from an unwanted station. The transmitter sends a pilot wave along with the modulated signal. The receiver's mixer stage heterodynes the signal with the pilot (instead of with a locally generated tone). Doppler phase distortion, which affects the signal and pilot in nearly the same way, cancels out during mixing. The diversity system with N antennas adds the outputs from N such mixers. This kind of diversity tends to add the N signal outputs in phase, while random noise components as well as certain other interferences add powerwise. In the presence of an interfering station, diversity smooths out amplitude fluctuations. It thereby reduces the probability that the interference will override the desired station.

I. INTRODUCTION

D. O. Reudink, in an unpublished work, has suggested a diversity system especially suited for mobile radio. In his system the transmitter sends a pilot wave along with the modulated signal. The receiver's mixer stage beats the signal against the received pilot (instead of against a locally generated tone). Doppler distortion, which affects the signal and pilot in nearly the same way, cancels out during mixing. The diversity system with N antennas adds the outputs of N such mixers and demodulates the sum by means of an ordinary AM or FM detector.

The receiver obtains a signal-to-noise advantage by adding signal components from the N mixers in phase while adding most interference terms powerwise. To obtain this advantage under multipath propagation conditions, the receiver's IF (that is, the difference f between the signal and pilot frequencies) must be chosen small enough. It suffices to make f so small that the propagation times along the different paths all agree to within a small fraction of $1/f$ (see Section

2473

2.2). The analysis presented in this paper is only valid for the situations in which signal components add in phase.

The effectiveness of these receivers is most clearly seen by examining the signal and noise levels at their outputs. Here the noise in question may be either random noise or an unwanted beat from an interfering station. Several kinds of signal-to-noise ratios can be defined because the signal and noise levels fluctuate as the receiver moves. The ratio snr of output signal power to output noise power depends on the receiver's position. Here snr is regarded as a random variable and its probability distribution function is derived. A simpler ratio, called SNR, is obtained by dividing the mean output signal power by the mean output noise power. SNR is simply a fixed number but it gives less information about receiver failure than the distribution of snr does.

The probability distribution of snr is derived for cases in which the signal experiences rayleigh fading. The rayleigh fading model is known to agree well with experiment within small areas, say ten wavelengths across, although it cannot account for largescale effects like shadowing by buildings and hills.[1] SNR is derived without assuming rayleigh fading.

Table I gives excerpts from more complete tables which follow. It compares receivers under rayleigh fading conditions by giving transmitter powers needed to keep snr above 3 dB or 10 dB with probability 0.99. The transmitter powers are given in decibels above a common level which need not be specified at this point. Of course the required powers depend on the interference power and on the propagation losses, but these terms are the same in all cases; they contribute a constant number of decibels to all the tabulated values. Only differences in decibel values need be considered when comparing receivers.

The table considers four kinds of interference and gives the signal power needed to keep snr at the given level for each separately. Random interference is supposed to be gaussian noise. In diversity receivers an interfering station produces three noise signals having different properties. These are called $2PS'$, $2P'S$, $2P'S'$, the letters denoting the components which beat to produce the noise. Thus $2P'S$ is a beat between interfering Pilot and desired Signal. For comparison, the conventional receiver has only one kind of output noise. Notice that the relative strengths of the three noises in the diversity receiver, and hence the character of the combined noise, depends both on N and on the signal level. Even a two-antenna diversity system has a noise

TABLE I—RELATIVE TRANSMITTER POWERS (DB) REQUIRED FOR
0.01 PROBABILITY OF snr $\geqq$ 3 DB OR 10 DB

snr (dB)	Interference		Diversity Receivers				Conventional Receiver
			$N = 1$	2	4	8	
3							
	random		26.0	14.3	6.6	1.4	20.0
	station	$2PS'$	23.0	12.5	6.3	1.9	
		$2P'S$	23.0	12.5	6.3	1.9	23.0
		$2P'S'$	21.5	13.5	9.3	6.8	
10							
	random		36.0	24.3	16.6	11.4	30.0
	station	$2PS'$	30.0	19.5	13.3	8.9	
		$2P'S$	30.0	19.5	13.3	8.9	33.0
		$2P'S'$	25.0	17.0	12.8	10.3	

advantage over the conventional system and has immunity to doppler distortion too.

II. THE DIVERSITY RECEIVER

The transmitter sends a pilot tone $A \cos 2\pi F t$ along with the modulated signal $AB \cos[2\pi(F + f)t + \theta]$. Here f is an intermediate frequency, small compared with F but large enough so that the signal spectrum does not overlap the pilot. B and θ are an amplitude and a phase, either one of which may be varied slowly to represent the modulating signal. The receiver (see the block diagram, Fig. 1), contains elements SQ which square received antenna voltages. Each square contains a component at frequency f which results from a beat between the pilot and the modulated signal. This component contains the modulation, AM or FM, of the original transmission. The N squares are added and the sum is filtered to remove other components at frequencies far from f. The filtered sum is an IF signal to be demodulated in the usual way.

2.1 Single Path In Phase Addition

In effect the transmitted pilot tone replaces the local oscillator tone which a conventional receiver generates internally. The advantage is that any doppler distortion affects the pilot as well as the modulated signal. As a result, the circuit of Fig. 1 tends to add IF components in phase if f is small. This may be seen as follows.

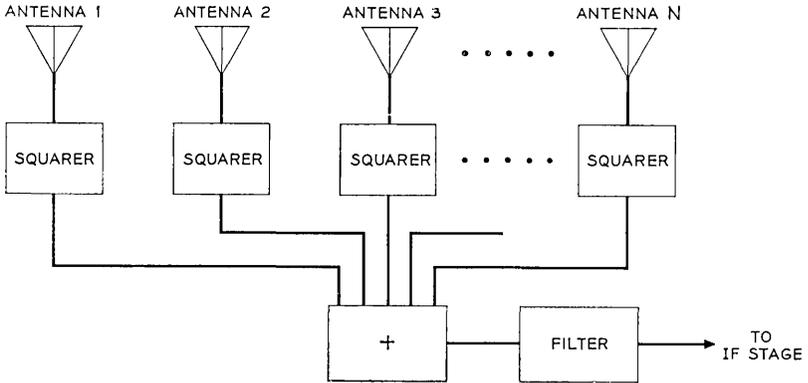Figure 2 shows N antennas receiving a signal which arrives from

Fig. 1 — Diversity receiver.

the direction indicated by the arrow. Suppose for the moment that this is the only incident signal (no multipath effects). Now consider two typical antennas, say 1 and 2. Let the difference between the lengths of the paths from 1 and 2 to the transmitter be called s.

If the voltage in antenna 1 is

$$A \cos (2\pi F t + \varphi) + AB \cos [2\pi (F + f)t + \psi], \qquad (1)$$

then the voltage in antenna 2 is

$$A \cos [2\pi F(t - s/c) + \varphi] + AB \cos [2\pi (F + f)(t - s/c) + \psi], \qquad (2)$$

where c is the velocity of light. After squaring, the IF components are $\frac{1}{2}A^2 B \cos (2\pi f t + \psi - \varphi)$ from antenna 1, and $\frac{1}{2}A^2 B \cos (2\pi f t + \psi - \varphi - 2\pi f s/c)$ from antenna 2.
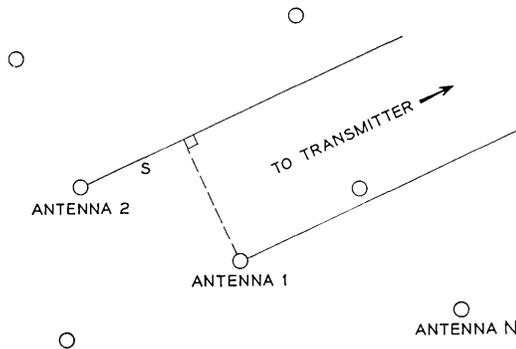
Fig. 2 — Reception by N antennas.

These two components differ in phase by $2\pi fs/c$ radians. To keep this angle small, s must be a small fraction of c/f, the wavelength at IF. For instance if the IF is $f \leqq 1$ MHz and if no two antennas are more than ten feet apart, then s is less than 0.01 wavelength and the N beat components are in phase to within 3.6°.

2.2 *Multipath Inphase Addition*

Under multipath conditions cross beats occur between pilots and modulated signals received via different paths. This section derives a more stringent sufficient condition for inphase addition. Now the lengths of all major propagation paths from transmitter to receiving antennas must agree within a small fraction of the IF wavelength. For example, if the IF were 100 kHz, the wavelength in question would be 3000 meters. Path differences of hundreds of feet would still permit nearly inphase addition. Path differences of this size might occur if only nearby buildings serve as reflectors. The data which W. R. Young took in New York City shows that some longer path dierences can be expected there.[1]

The voltages in antennas 1 and 2 of Fig. 2 are now sums of voltages received over different paths. The kth path contributes terms like (1) and (2) but with parameters A_k, ϕ_k, ψ_k, and s_k which depend on k. Suppose the kth path has length L_k. Then ϕ_k is a sum of phase shifts at reflections plus a propagation term $-2\pi FL_k/c$. Likewise ψ_k is a sum of the same phase shifts at reflections, a propagation term $-2\pi(F + f)L_k/c$, and the modulation angle θ. Then $\psi_k = \phi_k + \theta - 2\pi fL_k/c$. At antenna 2 the kth pilot is $P_k = A_k \cos(2\pi Ft + \phi_k - 2\pi Fs_k/c)$ and the kth modulated signal is $S_k = A_k B \cos[2\pi(F + f)t + \theta + \phi_k - 2\pi fL_k/c - 2\pi(F + f)s_k/c]$. At antenna 1 the kth path produces voltages of the same form but with $s_k = 0$.

When the antenna 2 voltage is squared, cross beats between the jth and kth paths occur. The IF part of P_kS_j is

$$P_kS_j : \tfrac{1}{2}A_kA_jB \cos[2\pi ft + \theta + \phi_j - \phi_k$$
$$- 2\pi fL_j/c - 2\pi(F + f)s_j/c + 2\pi Fs_k/c].$$

There is also a P_jS_k beat, and the sum of the two beats contains the IF component

$$P_kS_j + P_jS_k : A_kA_jB \cos[2\pi ft + \theta - \pi f(L_k + L_j + s_k + s_j)/c]$$
$$\cdot \cos[\phi_j - \phi_k - \pi f(L_j - L_k + s_j - s_k)/c - 2\pi F(s_j - s_k)/c].$$

The same expression gives the IF component of $P_k S_j + P_j S_k$ at antenna 1 when s_j and s_k are replaced by zero. In this expression the first cosine contains the time dependence while the second cosine is purely an amplitude factor.

Now suppose, as in Section 2.1, that s_1, s_2, $\cdots$, are all so small that the terms $\pi f s_k / c$ are small angles. Then the first cosine in the $P_k S_j + P_j S_k$ contribution is nearly the same at antenna 2 as it is at antenna 1. However the second cosine contains the large angle $2\pi F(s_j - s_k)/c$ at antenna 2 only. Indeed one can construct numerical examples to show that further assumptions are needed to make the total IF outputs of the two squarers be inphase. It will suffice to assume that the path lengths L_1, L_2, $\cdots$, are nearly equal, differing from one another by only a small fraction of c/f. Under this extra condition, the first cosine factor is approximately $\cos(2\pi f t + \theta - 2\pi L_1/c)$ for all k, j and at both antennas. For a given k, j the second cosine factor can still have opposite signs at the two antennas. However, when all beats are combined, the amplitude at antenna 2 is approximately

$$\tfrac{1}{2} \sum_{k,i} A_k A_j B \cos [\phi_i - \phi_k - 2\pi F(s_i - s_k)/c]$$

$$= \tfrac{1}{2} B \operatorname{Re} \sum_{k,i} A_k A_j \exp i[\phi_i - \phi_k - 2\pi F(s_i - s_k)/c]$$

$$= \tfrac{1}{2} B \operatorname{Re} \mid \sum_i A_j \exp i[\phi_i - 2\pi F s_i/c] \mid^2,$$

which is positive. The same argument with $s_i = 0$ gives a positive amplitude at antenna 1; the two sums are inphase.

In New York City large path differences are observed. There it may be difficult to make f small enough to satisfy always the condition just derived. However if the total number K of paths is small, there is still some tendency for the phases from squarers 1 and 2 to be close. For although the $P_k S_j$ contributions from antennas 1 and 2 differ in the $K(K - 1)$ cases with $j \neq k$, the argument of Section 2.1 shows that the two antennas give equal contributions in the K cases with $j = k$. One can analyze simple models in which L_k and other parameters are randomly chosen and still conclude that the IF outputs from the two squarers are correlated, but to an extent that decreases as K increases. However I omit those details and assume from now on that signal outputs from the squarers add inphase. I also assume that F is large enough, say about 1000 MHz, so that the phases of noise received in antennas placed a few feet apart can be considered independent.

III. RESPONSE TO RANDOM NOISE

This section considers the effect of random noise on diversity reception and gives expressions (16), (17), and (18) for output noise spectra. Multipath fading effects make the output signal to noise ratio, snr, depend on the position of the receiver. A single mathematically convenient figure of merit is the ratio of expected signal power to expected noise power. This ratio is called SNR here. Before the mathematical details begin, some of the results will be summarized.

SNR increases linearly with the number N of antennas [equation (20)]. For a given amount of total transmitter power, the largest output signal power is obtained by transmitting equal amounts of power in the pilot and modulated signal. The diversity system will be compared with a conventional system using the same transmitter power. If N is small, the conventional system has a slight noise advantage because it uses the full transmitter power for the modulated signal (the pilot is generated in the receiver). The diversity system with $N = 3$ has about the same SNR as a conventional system. However, the probability distributions of snr for these receivers are very different; the one for the diversity receiver is more sharply peaked. As a result a diversity system, even with $N = 2$, produces a small snr less often than the conventional system (compare with Table I).

When making SNR comparisons one must also recognize qualitative differences between the output noises from different receivers. The conventional receiver has a steady noise output resulting from input noise beating against the steady local oscillator signal. In the diversity system the output noise results largely from input noise beating against fluctuating pilot and modulated signals. During fades the output noise from the diversity receiver also fades while the noise from the conventional receiver does not. Thus, the diversity receiver has acceptable snr more often than a conventional receiver with the same output SNR.

3.1 *Noise Spectra*

The mathematical treatment will begin with the case $N = 1$; the extension to more antennas will be easy. The input to the squarer is the sum of three voltages:

$$\text{Pilot} \qquad P(t) = A \cos (2\pi F t + \varphi), \tag{3}$$

$$\text{Signal} \qquad S(t) = AB \cos [2\pi(F + f)t + \psi], \tag{4}$$

$$\text{Noise} \qquad n(t) \;=\; \sum n_i \cos\,(2\pi f_i t + \xi_i). \tag{5}$$

Here the noise is represented, as by S. O. Rice,[2] as a sum of sinusoids with random phases ξ_i and amplitudes n_i. Rice studied the effect of squaring a random noise; this section adapts his work to the present problem.

The received pilot power is $\tfrac12 A^2$ (into a one ohm load); likewise the signal has power $\tfrac12 A^2 B^2$. The noise has a one-sided power spectrum function $w(\nu)$ such that

$$w(\nu)\,\triangle\nu \;=\; \tfrac12 \sum_{\nu < f_i < \nu + \triangle\nu} n_i^2$$

represents the noise power in the frequency band from ν to $\nu + \triangle\nu$. The shape of the function $w(\nu)$ is determined by the tuned circuits (not shown in Fig. 1) which filter the antenna signal before squaring. Figure 3 shows a typical case

$$w(\nu) \;=\; \begin{cases} N_o\,, & F - b \leqq \nu \leqq F + f + a, \\[4pt] 0, & \text{otherwise,} \end{cases} \tag{6}$$

which uses a band-pass filter slightly wider than necessary to pass the pilot and signal at frequencies F and $F + f$.

Squaring $P + S + n$ produces six terms; P^2, S^2, n^2, $2PS$, $2Pn$, $2Sn$. P^2 and S^2 contribute nothing to the output after the output filter removes components remote from frequency f. The other contributions are

$$A^2 B \cos\,(2\pi f t + \psi - \varphi) \quad \text{from} \quad 2PS, \tag{7}$$

$$A \sum n_i \cos\,[2\pi(f_i - F)t + \xi_i - \varphi] \qquad \text{from} \quad 2Pn, \tag{8}$$

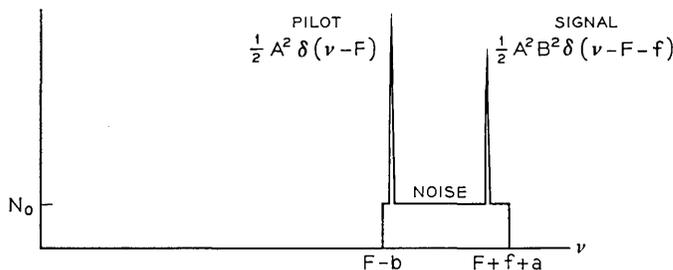$$AB \sum n_i \cos\,[2\pi(f_i - F - f)t + \xi_i - \psi] \quad \text{from} \quad 2Sn, \tag{9}$$

Fig. 3 — Power spectra at the input to a squarer.

$$\sum_{j<i} n_i n_j \cos\left[2\pi(f_i - f_j)t + \xi_i - \xi_j\right] \qquad \text{from} \quad n^2. \qquad (10)$$

The $2PS$ contribution is the desired output; its power is $\frac{1}{2}A^4B^2$. The spectra of the other contributions appear in Fig. 4. The spectral density functions are

$$A^2 w(\nu + F) \qquad \text{from} \quad 2Pn, \qquad (11)$$

$$A^2 B^2 w(F + f + \nu) \qquad \text{from} \quad 2Sn, \qquad (12)$$

$$2 \int_0^\infty w(x)w(\nu + x)\, dx \quad \text{from} \quad n^2. \qquad (13)$$

Functions (11), (12), and (13) assign some power to negative values of ν; these are to be aliased to positive frequencies. This aliasing accounts for the peculiar discontinuities in the spectra at low frequencies. The dotted lines show functions (11), (12), and (13) before aliasing. The values of a and b will be assumed smaller than f so that, as in Fig. 4, the noise power densities at frequency f are $A^2 N_o$ for Pn noise and $A^2 B^2 N_o$ for Sn noise.

In the case of gaussian noise, the phases ξ_i in functions (8), (9), and (10) are independent. It then follows that the three kinds of output noise components at a given frequency ν are uncorrelated. Then these noises add powerwise and the total noise spectral density is the sum of functions (11), (12), and (13).

3.2 Noise in Diversity System

In a diversity system the same kind of analysis applies for each of N antennas. The amplitudes and phases would now be written as A_k, n_{ik}, ψ_k, φ_k, and ξ_{ik} where the subscript k ($k = 1, \cdots, N$) specifies the antenna. All these random variables are independent of one another except for ψ_k and φ_k which satisfy $\psi_1 - \varphi_1 = \psi_2 - \varphi_2 = \cdots = \psi_N - \varphi_N = \theta$ because, as discussed in Section II, the $2PS$ terms have a common phase angle θ. Thus the N signal components add voltage-wise and the expected signal power at the output is

$$\tfrac{1}{2}B^2 E\left(\sum A_k^2\right)^2 = \tfrac{1}{2}B^2 E\{NE(A^4) + N(N - 1)[E(A^2)]^2\}.$$

Let k_o denote the (dimensionless) ratio

$$k_o = E(A^4)/[E(A^2)]^2. \qquad (14)$$

For rayleigh fading, $k_o = 2$. For no fading $k_o = 1$. The expected output signal power is
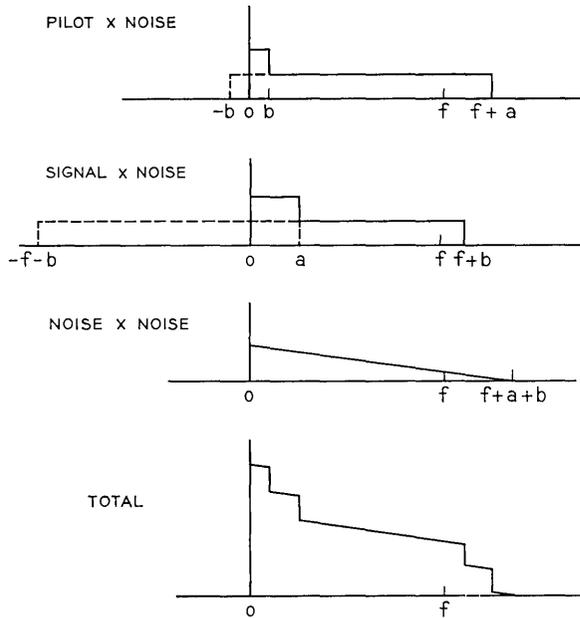
Fig. 4 — Output noise spectra.

$$E(\text{Signal Power Out}) = \tfrac{1}{2}N(N + k_o - 1)[E(A^2)]^2 B^2$$

$$= \tfrac{1}{2}N(N + 1)[E(A^2)]^2 B^2(\text{rayleigh}). \qquad (15)$$

According to equations (3) and (4), $\tfrac{1}{2}E(A^2)$ and $\tfrac{1}{2}B^2 E(A^2)$ are the expected received powers of the pilot and signal. With fixed transmitter power [fixed $\tfrac{1}{2}(1 + B^2)E(A^2) = P_o$], the output signal, equation (15), is maximized when the transmitted power is divided equally between pilot and signal [$B = 1$, $E(A^2) = P_o$]. Then equation (15) becomes $E(\text{Signal Power Out}) = \tfrac{1}{2}N(N + k_o - 1)P_o^2$.

The noise terms (11), (12), and (13) for the N antennas add power-wise and the expected output noise power spectrum is a sum of three terms

$$NE(A^2)w(\nu + F) \qquad \text{from} \qquad 2Pn, \qquad (16)$$

$$NB^2 E(A^2)w(F + f + \nu) \qquad \text{from} \qquad 2Sn, \qquad (17)$$

and

$$2N \int_0^\infty w(x)w(\nu + x)\,dx \quad \text{from} \quad n^2. \qquad (18)$$

.3 SNR Formulas

For a typical case, suppose $w(\nu)$ is the function (6) with $a < f$ and $b < f$. Suppose also that the output filter in Fig. 1 has a narrow rectangular transfer function with bandwidth Δf about frequency f. Then the expected output noise power is

$$E(\text{Noise Power Out}) = 2NN_o[P_o + N_o(a + b)]\Delta f \qquad (19)$$

where again $P_o = \frac{1}{2}(1 + B^2)E(A^2)$ is the total expected power which an antenna receives from pilot and signal. In this case the output noise power does not depend on B and the choice $B = 1$ maximizes not only the output signal power but the output signal to noise ratio as well. With $B = 1$, equations (15) and (19) combine to give

$$\text{SNR} = \frac{N + k_o - 1}{4} \frac{P_o/(N_o \Delta f)}{1 + (a + b)N_o/P_o}, \qquad (20)$$

where k_o is given by equation (14) ($k_o = 2$ for Rayleigh fading).

If the input noise spectrum is not flat as in Fig. 3, the output noise contributions (16) and (17) do not combine into the term $2NN_oP_o\Delta f$ which appears in equation (19). In that case the value of B which gives the best output SNR may not be one but will depend on the input noise power densities at F and $F + f$.

Formulas (11), (12), and (13) also apply, with slight reinterpretations, to the conventional receiver without diversity. $\frac{1}{2}A^2$ is the power of a local oscillator and $\frac{1}{2}A^2B^2$ is the received signal power. Then A has a well determined value, but B is a random variable having perhaps a Rayleigh distribution. Now P_o is $E(\frac{1}{2}A^2B^2)$. The desired output signal has amplitude A^2B and so has expected power $E(\text{Signal Power Out}) = \frac{1}{2}A^4E(B^2) = A^2P_o$. The local oscillator is deliberately made much stronger than the incoming signal or noise; then the output noise components (12) and (13) are negligible compared with formula (11). For the output filter of bandwidth Δf, $E(\text{Noise Power Out}) = A^2w(F + f)\Delta f$. When $w(\nu)$ is the function (6) again,

$$\text{SNR} = P_o/(N_o\Delta f). \qquad (21)$$

The output signal to noise ratio in equations (20) and (21) differ by a factor

$$(\text{SNR})_{\text{conventional}}/(\text{SNR})_{\text{diversity}}$$

$$= 4[1 + (a + b)N_o/P_o]/(N + k_o - 1). \qquad (22)$$

The term $(a + b)N_o/P_o$ represents that part of the input noise to signal ratio which results from noise arriving outside the band $F \leqq \nu \leqq F + f$. Then this term will be small in any useful case. The remaining factor $4/(N + k_o - 1)$ gives the conventional system the advantage unless $N \geqq 5 - k_o$. When Rayleigh fading holds, a three-antenna diversity system has the same output SNR as the conventional system.

3.4 snr Distribution

As mentioned in Section 3.3, the signal and noise levels of conventional and diversity receivers fluctuate differently as the receiver moves. In the case of Rayleigh fading one can obtain the probability distribution functions for snr = (Signal Power Out/Noise Power Out) for the two receivers. Again take the simple input noise spectrum of equation (6) with small values of a and b.

Expressions (7), (11), and (12) show that snr $= (4N_o \triangle f)^{-1} \sum A_k^2$ for the diversity receiver ($B = 1$). Each Rayleigh amplitude A_k may be expressed in terms of independent gaussian variables x_k, y_k of mean zero and unit variance by means of $A_k^2 = \frac{1}{2} P_o(x_k^2 + y_k^2)$.

Then snr $= (\chi_{2N}^2/8)(P_o/N_o \triangle f)$ where $\chi_{2N}^2 = X_1^2 + \cdots + X_N^2 + Y_1^2 + \cdots + Y_N^2$ has the chi-squared probability distribution with $2N$ degrees of freedom. The same result might be obtained by interpreting the receiver as a maximal ratio combiner.[3]

Expressions (7), (11), and (12) also apply to the conventional receiver if, as explained above, A is a fixed number while B is a small rayleigh variable. Only $2Pn$ noise need be considered; then snr $= (2N_o \triangle f)^{-1}A^2B^2 = \frac{1}{2}\chi_2^2(P_o/N_o \triangle f)$, where again χ_2^2 has the chi-squared distribution, now with two degrees of freedom.

Suppose the system fails when snr is below some known critical value. Suppose such failure can be tolerated only a small fraction Q of the time. The given value of Q is reached at some χ^2 value which can be read from probability tables. To achieve the desired small failure probability the ratio $P_o/(N_o \triangle f)$ (a kind of input SNR) must be

$$P_o/(N_o \triangle f) = \begin{cases} (8/\chi_{2N}^2) \text{ snr} & \text{(diversity)} \\ \frac{1}{4}(8/\chi_2^2) \text{ snr} & \text{(conventional)} \end{cases} .$$

Table II gives 10 log $(8/\chi_{2N}^2)$ as a function of Q. Thus for a 0.01 probability of failure, $P_o/(N_o \triangle f)$ must exceed the critical snr by 26.0 dB, 14.3 dB, 6.87 dB, and 1.39 dB for diversity systems of one, two, four, and eight antennas. The conventional receiver requires $26.0 - 6.0 = 20.0$ dB and so is intermediate between diversity systems with $N = 1$ and 2.

TABLE II—VALUES OF 10 LOG $(8/\chi_{2N}^2)$ FOR WHICH PROBABILITY
OF FAILURE = Q

Number of Antennas	Q					
	0.001	0.005	0.01	0.025	0.05	0.1
1	36.0	29.0	26.0	22.0	18.9	15.8
2	19.5	15.9	14.3	12.1	10.5	8.78
3	13.3	10.7	9.62	8.01	6.89	5.60
4	9.74	7.75	6.57	5.65	4.67	3.60
5	7.36	5.70	4.95	3.92	3.08	2.16
6	5.61	4.15	3.50	2.59	1.85	1.03
7	4.23	2.96	2.35	1.53	0.86	0.12
8	3.15	1.92	1.39	0.64	0.02	−0.63

IV. INTERFERENCE FROM A SECOND STATION

Suppose a diversity system tries to receive a desired signal while another station uses the same channel. The pilots and modulated signals of the two stations produce a variety of beat components, three of which cause interference at IF [functions (23), (24), and (25) below]. Two sound like doppler-distorted versions of the modulated signals from the desired station and its competitor. The third is an undistorted copy of the modulated signal from the competing station.

Under multipath conditions, the two doppler-distorted beats have phases which are uncorrelated from antenna to antenna. The output SNR's for these noises grow linearly with the number N of antennas [equation (26)]. The third components from the separate squarers add in phase. Then the SNR for this interference is not reduced by increasing N [equation (27)]. However, increasing N reduces the variability of the power levels of the output signal and noise. Thus, if the desired station is a few decibels stronger than the competing station, increasing N reduces the chance that multipath fading will allow the competing station to override the desired station (Table III).

One may reuse much of the formalism of Section III. A single antenna again receives a pilot [equation (3)], modulated signal [equation (4)], and a noise which is a special case of equation (5). The noise now has only two components. One is a pilot $P'(t)$ of frequency F, phase φ', and amplitude A'. The other is a modulated signal $S'(t)$ of frequency $F + f$, phase ψ', and amplitude $A'B'$.

Squaring produces IF components which are obtainable from func-

TABLE III—VALUES OF 10 LOG F SUCH THAT PROBABILITY
OF FAILURE $= Q$ (2P'S' NOISE)

Number of Antennas	Q							
	0.001	0.005	0.01	0.025	0.05	0.1	0.25	0.50
1	30.0	23.0	20.0	15.9	12.8	9.54	4.77	0
2	17.3	13.6	12.0	9.82	8.05	6.14	3.14	0
3	13.0	10.4	9.27	7.64	6.31	4.84	2.50	0
4	10.8	8.74	7.80	6.46	5.36	4.13	2.14	0
5	9.56	7.70	6.88	5.71	4.79	3.67	1.90	0
6	8.35	6.92	6.19	5.16	4.29	3.32	1.73	0
7	7.73	6.33	5.70	4.75	3.94	3.04	1.58	0
8	7.17	5.86	5.29	4.42	3.69	2.88	1.49	0

tions (7), (8), (9), and (10). The desired signal component is func-
tion (7) again. The $2Pn$ component, function (8), has two parts, one
of which $[P(t)$ beating against $P'(t)]$ contributes nothing. The re-
maining IF contribution from function (8) is

$$AA'B' \cos (2\pi ft + \psi' - \varphi) \quad \text{from} \quad 2PS'. \tag{23}$$

Likewise functions (9) and (10) contribute only

$$AA'B \cos (2\pi ft + \psi - \varphi') \quad \text{from} \quad 2SP', \tag{24}$$

and

$$A'^2B' \cos (2\pi ft + \psi' - \varphi') \quad \text{from} \quad 2P'S'; \tag{25}$$

the $2SS'$ and S'^2 terms do not contribute at IF.

The three interference terms (23), (24), and (25) have different
characteristics. The $2PS'$ and $2P'S'$ components carry the modulation
(AM or FM) of $S'(t)$ and act like interfering stations at IF. Likewise
the $2SP'$ term sounds like a station with the desired modulation of
$S(t)$. As the receiver moves, the two angles ψ', and φ undergo different
doppler shifts. Then the $2PS'$ component contains a residual doppler
distortion. Likewise the $2SP'$ component is doppler distorted and so
will be considered a noise. By contrast, as in the $2PS$ term, the doppler
shifts in the $2P'S'$ term cancel out leaving an undistorted interfering
signal.

Because the $2P'S$ component has both the desired modulation and
doppler distortion it is not clear whether it should be treated as a
signal term or as a noise term. If it were counted as part of the sig-
nal, the $2P'S$ term would be a source of fluctuation of the output sig-
nal level (it differs in phase from the $2PS$ term by a random amount).

To call the $2P'S$ term a kind of noise is probably overconservative if the system uses FM of index high enough to make the doppler distortion unimportant. It turns out that the power levels of the $2PS'$ and $2P'S$ terms have the same probability distribution. Thus, whenever other interference terms are small it does not matter much whether $2P'S$ components are treated as signal or as noise.

4.1 SNR Formulas

As in Section 3.3 one can compute an SNR, defined as E(Signal Power Out)$/E$(Noise Power Out), for each of the three interferences. Again multipath fading conditions will be assumed so that pilot amplitudes and phases from the N antennas are independent variables. The conditions $\psi_1 - \varphi_1 = \psi_2 - \varphi_2 = \cdots = \psi_N - \varphi_N = \theta$ and $\psi'_1 - \varphi'_1 = \psi'_2 - \varphi'_2 = \cdots = \psi'_N - \varphi'_N = \theta'$ relate the signal phases to the pilot phases.

The expected signal output power is given by equation (15) as before. The expected power from the N terms of type $2PS'$ is $E(\sum \frac{1}{2} A_k^2 A_k'^2 B'^2)$ $= \frac{1}{2} N B'^2 E(A^2) E(A'^2)$. Likewise the $2SP'$ power has expected value $\frac{1}{2} N B^2 E(A^2) E(A'^2)$. The SNR's are SNR $= (N + k_o - 1)(B/B')^2 E(A^2)/E(A'^2)$ for $2PS'$ interference and SNR $= (N + k_o - 1)E(A^2)/E(A'^2)$ for $2SP'$ interference [recall the definition of k_o given by equation (14)]. When $B = B' = 1$ and the expected received powers from the two stations are P_o and P'_o, both interferences have

$$\text{SNR} = (N + k_o - 1)P_o/P'_o. \tag{26}$$

The expected power of $2P'S'$ interference is given by equation (15) with A' and B' replacing A and B. Then, if $B' = B$, the SNR for $2P'S'$ is

$$\text{SNR} = (P_o/P'_o)^2. \tag{27}$$

The two expressions (26) and (27) have interesting differences. They depend on N in different ways because the $2SP'$ and $2S'P$ components from separate antennas add with random phases while the $2S'P'$ components add in phase. The input signal to noise ratio P_o/P'_o appears with different exponents in equations (26) and (27) because equation (26) relates to beats between the desired station and the interfering one, while equation (27) relates to beats of the interfering station with itself.

Because of these differences, either kind of output noise can be the more serious one, depending on the situation. For a given number of antennas, the $2S'P$ and $2SP'$ noises are stronger than the $2S'P'$ noise when P_o/P'_o is large. As P_o/P'_o becomes smaller, all noises increase and, at $P_o/P'_o = N + k_o - 1$, they have equal powers. When P_o/P'_o is still

smaller, the $2S'P'$ noise (undistorted copy of the interfering signal) predominates. With Rayleigh fading and $N = 4$ antennas, the $2S'P'$ noise predominates at input signal to noise ratios of 7 dB or less.

In conventional systems, an interfering station produces only one output noise component. It has

$$\text{SNR} = P_o/P_o' . \tag{28}$$

None of the noise components of the diversity system are as bad as this unless the interfering station is stronger than the desired one.

4.2 snr Distributions, $2P'S'$ Noise

Equation (27) shows that adding more antennas does not improve the SNR for $2P'S'$ noise. However, diversity helps by reducing the chance that a large fluctuation of the interfering signal level will cause the system to fail. To study this effect let $A_1 , \cdots , A_N$ be signal amplitudes, as in expressions (7) and (8), received by the N antennas. Likewise let these antennas receive $A_1' , \cdots , A_N'$ from the interfering station. Under severe multipath conditions these $2N$ amplitudes may be regarded as independent random variables. Again take $B = B' = 1$ so that $E(A_k^2) = P_o , E(A_k'^2) = P_o'$. The desired and interfering stations produce output signals with amplitudes $\sum A_k^2$ and $\sum A_k'^2$. Then

$$\text{snr} = (\sum A_k^2 / \sum A_k'^2)^2 \tag{29}$$

is the random variable which must be studied.

The probability distribution function for snr can be obtained easily in the case of rayleigh fading. Each Rayleigh amplitude A may be represented by the formula $A^2 = X^2 + Y^2$ where X and Y are independent gaussian variables with variance $E(X^2) = E(Y^2) = \frac{1}{2}P_o$.

In these terms, the quantity

$$F = \frac{(X_1'^2 + Y_1'^2 + X_2'^2 + \cdots + Y_N'^2)/(\frac{1}{2}P_o')}{(X_1^2 + Y_1^2 + X_2^2 + \cdots + Y_N^2)/(\frac{1}{2}P_o)}$$
$$F = (P_o/P_o') \, \text{snr}^{-\frac{1}{2}}, \tag{30}$$

is the ratio of two sums of $2N$ independent squares of gaussian variables of unit variance. Statisticians use such ratios frequently and have tabulated their probability distributions. Abramowitz and Stegun give such a table.[4] In their notation the cumulative probability function for F is $P(F \mid 2N, 2N)$, a special case of their $P(F \mid \nu_1 , \nu_2)$. Their Table 26.9 gives $Q(F \mid \nu_1 , \nu_2) = 1 - P(F \mid \nu_1 , \nu_2)$, so that snr has the distribution function

$$\text{Prob } \{\text{snr} \le (P_o/P_o')^2 F^{-2}\} = Q(F \mid 2N, 2N). \tag{31}$$

Table III reproduces part of Abramowitz and Stegun's table after converting F values to decibels. The numbers tabulated are values 10 $\log_{10}F$ which are needed to make the probability of equation (31) a small value $Q = 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25$, or 0.5. To use Table III one must first know how small snr can become before the system will fail; one also decides on an acceptable probability Q of failure. The table gives a corresponding value of F and the conditions for not failing are met as long as the input signal to noise ratio P_o/P'_o satisfies

$$F \, \mathrm{snr}^{1/2} \leqq P_o/P'_o . \tag{32}$$

For example, suppose the system fails if snr becomes as small as 3 dB. Suppose failure can be tolerated only 1 percent of the time. The tabulated values of F for $Q = 0.01$ and $N = 1, 4, 8$ are 20.0, 7.80, and 5.29 dB. Then inequality (32) requires the input signal to noise ratio to be

$$
\begin{array}{lll}
20.0 + 1.50 = 21.5 \text{ dB} & \quad \text{for one antenna,} \\
7.80 + 1.50 = 9.30 \text{ dB} & \quad \text{for four antennas,} \\
5.29 + 1.50 = 6.79 \text{ dB} & \quad \text{for eight antennas.}
\end{array}
$$

In the case of one and four antennas at these signal levels, equations (26) and (27) show that the other noise components $2SP'$ and $2PS'$ are stronger than the $2P'S'$ component. Thus the snr for $2SP'$ and $2PS'$ noises must be considered later.

To show the advantage of diversity over a conventional system, one may examine the probability distribution function for the conventional snr. This function is not just equation (31) with $N = 1$. A conventional system has snr $= (AB)^2/(A'B')^2 = (X^2 + Y^2)/(X'^2 + Y'^2)$ instead of equation (29). To get a ratio of sums of squares of gaussian variables with unit variance, one must now define $F = (P_o/P'_o)/\mathrm{snr}$ instead of equation (30). The value of F for a given failure probability Q is again obtained from Table III with $N = 1$. The input signal to noise ratio P_o/P'_o must then satisfy

$$F \, \mathrm{snr} \leqq P_o/P'_o \tag{33}$$

instead of inequality (32). To have snr as low as 3 dB for only a fraction $Q = 0.01$ of the time, the input signal to noise ratio must now be 23 dB or more.

4.3 snr Distributions, $2SP'$ and $2P'$ Noises

The SNR calculation showed that $2SP'$ and $2PS'$ components are apt to be the strongest noises when N is small. The distribution functions

for their snr may also be derived. Again rayleigh fading is assumed and $B' = B = 1$. The latter assumption makes the $2PS'$ and $2SP'$ components have the same snr distribution [compare expressions (23) and (24)].

It is convenient to rewrite the $2PS'$ component (23) in terms of cosine and sine amplitudes

$$X' = A' \cos (\psi' - \varphi), \qquad Y' = -A' \sin (\psi' - \varphi). \qquad (34)$$

Then expression (23) becomes $AX' \cos 2\pi ft + AY' \sin 2\pi ft$. Now X' and Y' are independent gaussian random variables with mean zero and variance $\frac{1}{2}P_o'$. When there are N antennas, equations (34) give amplitudes X_k' and Y_k' for the kth antenna. The kind of argument that produced equations (28) and (29) now leads to

$$\text{snr} = \frac{(\sum A_k^2)^2}{(\sum A_k X_k')^2 + (\sum A_k Y_k')^2}. \qquad (35)$$

It is possible to transform equation (35) into a form to which an F-distribution again applies. As a first step, introduce two new random variables

$$x' = \sum A_k X_k'/(\tfrac{1}{2}P_o' \sum A_i^2)^{\frac{1}{2}}, \qquad y' = \sum A_k Y_k'/(\tfrac{1}{2}P_o' \sum A_i^2)^{\frac{1}{2}}.$$

For any $A_1, \cdots, A_N$, x' and y' are independent gaussian variables of mean zero and variance 1. Now equation (35) becomes

$$\text{snr} = 2 \sum A_k^2/[P_o'(x'^2 + y'^2)]. \qquad (36)$$

Next one can express the pilot $P(t)$ in terms of cosine and sine amplitudes. In this way one obtains $A_k^2 = \frac{1}{2}P_o(x_k^2 + y_k^2)$, where x_k and y_k are independent gaussian random variables of mean zero and variance 1. Finally equation (36) becomes

$$\text{snr} = (P_o/P_0')/G, \qquad (37)$$

where $G = (x'^2 + y'^2)/ \sum (x_k^2 + y_k^2)$.

Again the snr involves a ratio G of sums of squares of gaussian variables and formulas for a suitable F-distribution are applicable. This time the numerator and denominator of the ratio contain unequal numbers of terms; the appropriate definition of F is $F = NG$. In the notation of Abramowitz and Stegun[4], the cumulative probability function for F is $1 - Q(F \mid 2, 2N)$. From their table, I obtain Table IV which gives values of $10 \log G$ which may be used with equation (37). Thus if

TABLE IV—VALUES OF 10 LOG G SUCH THAT PROBABILITY
OF FAILURE = Q ($2SP'$ AND $2PS'$ NOISES)

Number of Antennas	Q							
	0.001	0.005	0.01	0.025	0.05	0.1	0.25	0.50
1	30.0	23.0	20.0	15.9	12.8	9.54	4.77	0
2	14.9	11.2	9.54	7.26	5.40	3.34	0	−3.83
3	9.54	6.86	5.61	3.84	2.33	.61	−2.32	−5.85
4	6.64	4.41	3.34	1.79	.45	−1.09	−3.82	−7.24
5	4.74	2.76	1.79	.37	−.86	−2.34	−4.95	−8.28
6	3.34	1.52	0.61	−.70	−1.88	−3.31	−5.85	−9.12
7	2.25	0.53	−.32	−1.59	−2.72	−4.09	−6.64	−9.83
8	1.37	−.27	−1.08	−2.32	−3.43	−4.76	−7.24	−10.5

a given output snr must be maintained for all but a fraction Q of the time, Table IV determines G. Then equation (37) determines th einput signal to noise ratio $P_o/P'_o = G$ snr.

Continuing the earlier example with snr = 3 dB, and Q = 1 percent, Table IV gives G values of 20.0, 3.34, and −1.08 dB for 1, 4, and 8 antennas. The required input signal to noise ratios are

$$20.0 + 3.0 = 23.0 \text{ dB} \quad \text{for one antenna,}$$
$$3.34 + 3.00 = 6.34 \text{ dB} \quad \text{for four antennas,}$$
$$-1.08 + 3.00 = 1.92 \text{ dB} \quad \text{for eight antennas.}$$

4.4 Transmission Path Lengths

Suppose that a vehicle receives a station D miles away while a second station D' miles away interferes. If the two stations radiate equal powers, the ratio P_o/P'_o is determined by the path losses to the two stations. For example, with isotropic antennas and inverse square law propagation $P_o/P'_o = D'^2/D^2$.

The numbers in Tables III and IV can be used to set limits on D'. For example, suppose snr must be above 3 dB with probability 0.99; then $2P'S'$ noise is the most serious one. P_o and P'_o must differ by at least 9.3 dB for a four-antenna diversity receiver or by 23 dB for a conventional receiver. If the inverse square law held, D' would have to be at least 2.9 D for four-antenna diversity reception and 14.1 D for conventional reception. While the inverse square law holds in free space, waves near the earth's surface attenuate more rapidly. Measurements by W. C. Jakes followed roughly an inverse fourth power law for ranges between 2 and 15 miles. Then, allowed values of D' can be as small as 1.7 for four-antenna diversity receivers and 3.8 for conventional receivers.

REFERENCES

1. Young, W. R., Jr., "Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc," B.S.T.J., *31*, No. 6 (November 1952), pp. 1068–1085.
2. Rice, S. O., Mathematical Analysis of Random Noise," B.S.T.J., *23*, No. 3 (July 1944), pp. 282–332, and *24*, No. 1 (January 1945), pp. 46–156.
3. Schwartz, M., Bennett, W. R., and Stein, S., *Communication Systems and Techniques,* New York: McGraw Hill, 1966.
4. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions,* National Bureau of Standards, AMS55 (1964).

An Analysis of Time Usage in Bell System Business Offices

By W. H. WILLIAMS and H. CHEN

The everyday contact with customers of the Bell System is carried out in approximately 2100 business offices. The assessment of such a large number of offices makes the continued improvement of formal office measurement schemes attractive. This paper describes an analysis of models for time usage in Bell System business offices. In addition, it was hoped that these models would be potentially useful for interoffice comparisons. A model for single offices is described first. This is followed by the development of a multioffice model which is constructed in such a way that it has good statistical characteristics and attempts to make the office comparisons as fair as possible.

The inputs to the multioffice model are: (i) the gross time used by each business office, (ii) the number of contacts that each office had with business and residence customers, (iii) the number of accounts carried by each office, and (iv) certain characteristics which were judged to reflect the nature of the exogenous demand put on the office, for example, percent of business main telephones.

I. TIME MEASUREMENT SCHEMES AND THE BELL SYSTEM BUSINESS OFFICES

The everyday contact with customers of the Bell System is carried out in approximately 2100 business offices. These offices have many functions. To specify just a few, most orders for telephone service, toll inquiries, and complaints of various kinds are handled by them. Consequently these offices are very important and need to be well run. However, their large number emphasizes the need for formal office measurement schemes which can be studied objectively.

While such schemes can be very useful, they can also contain very troublesome features. The first of these troubles relates to the operational definition of the word efficiency. It should not be so broad as to be meaningless or misleading, nor should it be overly narrow. The re-

sult of a narrow definition of efficiency is likely to be multiple meas-
ures which would be very unwieldy with a large number of offices. Ad-
ditionally, this definition must reflect the internal and external office
characteristics that are statistically associated with efficiency.

While the problems of meaning and measurement are obvious (but
are not necessarily easy to solve), the problems which come about as
a result of the influence of the measurement scheme on the office itself
are not usually so obvious. If a scheme is not carefully evaluated it
can modify the office itself in undesirable ways. On the other hand, the
influence of a measurement plan on the behavior of the office is po-
tentially useful for inducing desired objectives. However, to attempt
such inductions requires a good deal of knowledge about the offices.

A final, and perhaps tangential, difficulty with any analytic meas-
urement scheme is that it will not itself separate the offices into those
which are "efficient" and those which are "inefficient." Such a separa-
tion is usually achieved by a comparison with norms which may be
obtained from statistical studies or from theoretical considerations. At
some stage the separation always requires the judgment of manage-
ment.

In summary, a measurement scheme, to be useful to management,
must relate to and measure some understandable characteristics of
office work performance in such a way that it is informative, and not
potentially misleading. At the same time it must not interact with the
actual office procedure in such a way that it invites the offices to be-
come less efficient. It must allow the local managers to be flexible. It
naturally follows that if meaningful office measurements can be con-
structed, they would be very helpful to both the immediate office man-
agement and the higher staff personnel.

This analysis was performed in conjunction with studies by the
American Telephone and Telegraph Company and Bell Telephone
Laboratories.

II. THE DATA

All analysis was carried out on an "entity" basis. The entities are
groupings of office locations and "departments" such that each entity
carries out approximately the same set of work functions. For example,
some larger offices have part of the handling of telephone orders car-
ried out by separate groups and not by the service representatives;
these groups may even be at a different location but must be included
in any interoffice comparisons. There are other similar situations and

it was clearly necessary to construct groups as nearly alike in function as possible. While this grouping is necessary for consistent analysis, the details of it are not necessary for this paper. Finally, while the entities described above do not necessarily correspond to any other definition of a business office, they are referred to as offices in the remainder of this paper. No misunderstanding of this terminology should occur.

The basic data were of five different types: (*i*) daily counts of customer contacts, (*ii*) daily gross time data, (*iii*) daily work sampling observations, (*iv*) monthly numbers of accounts carried by the offices, (*v*) profile survey variables. Each of these types of data played an important role in the development of the statistical models; however, the final models do not use work sampling. Let us take a closer look at these data types.

(*i*) Much of the work of the business offices is generated by the customer on the telephone; some personal contact occurs in public offices, but relatively little. Most of these customer contacts are counted and classified. Eight of the categories are orders, toll inquiries, other billing inquiries, and miscellaneous contacts, each for business and residence customers. These eight are among the most important classifications, and account for most of the office working time.

(*ii*) The daily gross time spent on all categories of work is available as a normal accounting item. This gross time is the total work time for which commercial employees in an office are paid. It includes: (*a*) time spent for the previously mentioned eight classifications of customer contacts; (*b*) time spent in the company's public office; (*c*) time spent on treatment work; (*d*) time spent on teller work; (*e*) normally scheduled relief time and personal time; (*f*) idle time; (*g*) time spent on work classified as general activity; and (*h*) time spent on miscellaneous activities. These categories are listed mainly for information and understanding. The bulk of the analysis is dependent only upon the availability of gross time data. Time does not have to be available in subcategories.

(*iii*) Time slice work sampling studies were carried out in 46 offices of the System in 1964. This study gave daily estimates of the total time spent on the various work categories including the eight categories mentioned in item *i*. In 42 of these offices, data were gathered for a 13-week period from May through July, and in the remaining four categories, the study continued for seven months through November 1964.

(*iv*) The number of accounts carried by each of the 46 offices was

obtained for each month the office was in the study. The numbers used are totals of both business and residence accounts.

(v) A profile survey was made of all offices in the System to determine basic characteristics about each office and its environment. Data on over 200 exogenous variables were obtained, 55 of which were studied in detail. Only those used in the models described in this paper are explicitly introduced.

III. THE DEVELOPMENT OF THE SINGLE OFFICE MODEL

To construct a first model for daily time expenditure in a single office, assume that time is used up partly as a result of direct customer demands, and partly by overhead time, see equation (1).

$$\begin{bmatrix} \text{time spent on all} \\ \text{commercial office} \\ \text{work} \end{bmatrix} = [\text{overhead time}] + \begin{bmatrix} \text{time required for} \\ \text{customer generated} \\ \text{demands} \end{bmatrix}. \quad (1)$$

Next, suppose that the time required to carry out a single customer contact in the jth work category is a_j, and that it is performed F_{ij} times on day i. Then the total time spent that day on category j is $F_{ij}a_j$ and the right bracket of the right side of equation (1) could be written as $\sum_{j=1}^{k}F_{ij}a_j$, where k is the total number of work categories. Thus, if a_o denotes overhead time, equation (1) can be written as

$$T_i = a_o + \sum_{j=1}^{k} F_{ij}a_j, \quad (2)$$

where T_i is the total time expenditure on day i, $i = 1, 2, 3, \ldots, n$. Since it is doubtful that such exact relationships ever hold, the model given in equation (2) needs to be modified. Only the specific modifications used in this paper are discussed. For a more general discussion see Ref. 1.

The first modification was a transformation of all observations to logarithms. This transformation was performed because plots of the estimated daily time on each of the eight categories (using data from the work sampling study) against the corresponding daily contact frequencies showed that the two were related approximately logarithmically. Consequently such a transformation could be expected to improve the statistical characteristics of the models.

The second major modification in the model formulation was the reduction in the number of categories. This came about because multicollinearities among the independent variables led to an extensive study to find which work categories were the best predictors of time.

The result was that two categories, total number of business contacts and total number of residence contacts were found to be better statistical predictors of time than any other combination of single categories or groupings of categories. While these two modifications are important they are intermediate steps and so the details have not been presented.

Consequently, the model used for each single office had the functional form,

$$T_i = \beta_o F_{1i}^{\beta_1} F_{2i}^{\beta_2}, \qquad i = 1, 2, \cdots, n \tag{3}$$

where T_i is the gross time on commercial operation on day i,

F_{1i} is the total daily number of business contacts,

and F_{2i} is the total daily number of residence contacts.

At this point, log β_o is an estimate of overhead time in an additive model like equation (2), and β_1 and β_2 are estimates of the average time requirements on a log basis.

This model was applied to each of the 46 offices individually. The statistical details using the data from an individual office are presented in a paper which emphasizes the statistical development of this model.[2]

IV. THE DEVELOPMENT OF THE MULTIOFFICE MODEL

4.1 Selection of an Appropriate Model

The models used in Section III are of the general form $t = q(f_1, f_2, \ldots, f_k)$ which relates the demand put on an office and the time consumed by it. Possibly the most interesting use of these models is to give estimates of the time required by an office to carry out a given demand load. Such an estimate could then be compared with the actual time used to produce an efficiency factor. A natural way to do this is as a ratio, $E =$ allowed time/actual time. These factors could be computed monthly to follow the progress of an office.

The comparison of different offices is not so simple, however. There are a number of possible approaches. One of these is to obtain a model fit and an efficiency score, E, for each of the offices for a specified month and then to compare the office E scores. This is in effect fitting a model to all the offices in which each office is associated with an individual set of parameters $(\beta_0, \beta_1, \beta_2)$. But such a model has two major defects for use in interoffice efficiency comparisons.

The first is that the approach would be very cumbersome for use

with such a large number of offices, but the second is the most devastating. It is that comparisons of two offices by use of such a model gives an inefficient operation a time allowance which is based on its own inefficient procedures. Similarly, an efficient office would be hurt in the comparison by being given only a time allowance based on its own efficient organization. This is clearly what is not wanted.

This defect suggests fitting a three parameter model to all offices. This would allow all offices the same standard overhead time and the same standard times for business and residence contacts. Such a model would be tractable and would eliminate the defect of allowing each office a standard time based on its own procedures.

However, the proposal of a three-parameter model makes it very clear that there really may be valid reasons why one office should have different time allowances from another. Consequently, we seem to stand between two models, one which allows every office the same overhead and average time allowances, and one which gives every office different allowances based on their individual performances.

What is clearly needed at this stage is a method and a model which gives offices a fair time allowance, based on the factors which actually influence the performance times. Operationally, this means relating the estimates of $(\beta_0, \beta_1, \beta_2)$ for each office to the exogenous variables which were measured in the profile survey.

4.2 *The Adjustment for Overhead Time*

It has been pointed out that the log β_o can be interpreted as measures of overhead time.* It has also been pointed out that it does not seem reasonable for interoffice comparisons to allow each office its own overhead time. There are two reasons for this. One is that such a procedure allows an inefficient office a time credit based on its own inefficient procedures. The other is that one would expect that a well-run large office might have more overhead time associated with it than a poorly-run small office. This means that a measure of office size must be introduced to scale these estimates of overhead time. The one selected was A_i, the monthly number of accounts carried by the office. Figure 1 is a plot of log $\hat{\beta}_{oi}$ against log $(A_i/100)$. There is one point for each of the 46 offices. A linear regression model was fitted to these data. While the statistical details of the fit affect the decision to use A_i as a scale variable, they only indirectly affect the final model, and consequently are not presented.

---

* Overhead time, as used here, means time for which no frequency count can sensibly be made.
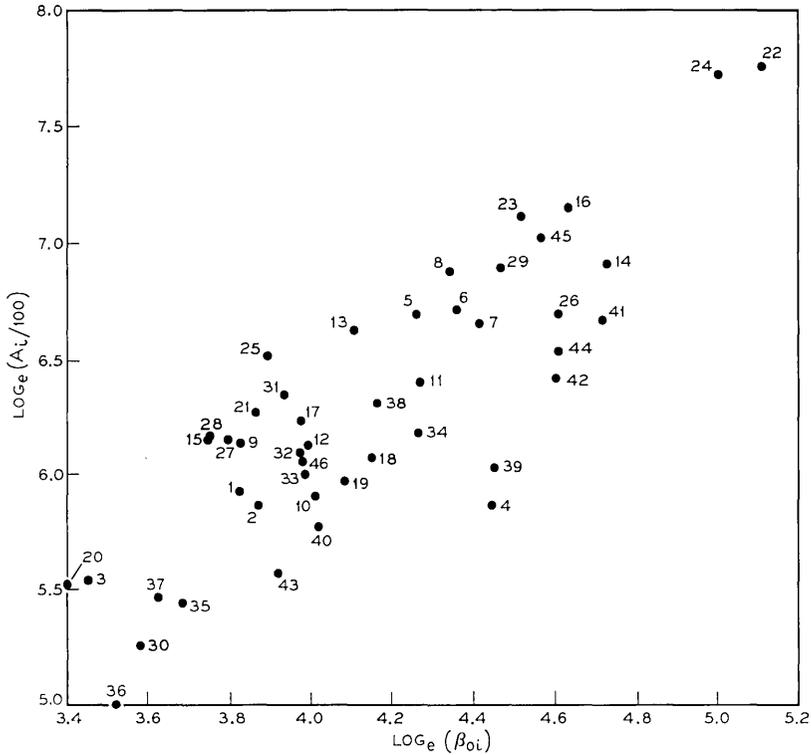
Fig. 1 — The number of office accounts versus the estimates of overhead time.

The good linear relationship between $\log \hat{\beta}_{oi}$ and $\log (A_i/100)$ suggests modifying the model by inserting $\alpha_o A_i^{\alpha_1}$ for the $\hat{\beta}_{oi}$. This modification gives rise to equation (4),

$$T_i = \alpha_o A_i^{\alpha_1} F_{1i}^{\beta_1} F_{2i}^{\beta_2}, \qquad i = 1, 2, \cdots, 46 \tag{4}$$

where the parameters α_o, α_1, β_1, β_2 are common to all offices. Again the statistical details of the fit are not included.

4.3 Adjustment for Contact Factors

As has been pointed out, the time that it takes an office to carry out a business or residence contact may well be influenced by outside factors. The hope was that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the business and residence contact times would be related to variables that were included among the profile variables. Consequently, a search of these variables was undertaken.

Perhaps surprisingly, a number of good relationships were found. More, in fact, than were used. Percent service representative losses and number of main business telephones were found to be nicely related to the business parameter, β_1. Percent business main and number of customer bills handled were found to be usefully related to the residence parameter, β_2. The relationships were approximately logarithmic. Consequently, the multioffice model was modified in a manner similar to that for overhead time. Specifically, the multioffice model was put in the form,

$$T = \alpha_o A^{\alpha_1} F_1^{\gamma_0 + \gamma_1 \log C_1} F_2^{\delta_0 + \delta_1 \log C_2}, \tag{5}$$

where C_1 and C_2 are the selected profile variables and the other variables, T, F_1 and F_2 remain as previously specified. In this form the time allowances (log basis), β_1 and β_2 have been modified so that each office's allowance is adjusted by the related profile variable.

So for example, if C_1 is percent service representative loss and C_2 is percent business main, the time allowance for an office would be made up of two components as originally specified.

$$\begin{bmatrix} \text{time} \\ \text{allowance} \end{bmatrix} = \begin{bmatrix} \text{overhead time} \\ \text{allowance} \end{bmatrix} + \begin{bmatrix} \text{allowance for time} \\ \text{generated by} \\ \text{customer demands} \end{bmatrix}. \tag{6}$$

Now, however, the overhead time allowance is based on the size of the offices as specified by the number of accounts it carries. In addition, the time allowed for customer generated demands is based on the (log) number of contacts multiplied by an allowed time per contact. The time per business contact is bigger for offices with higher service representative losses. For residence, however, the time per contact is higher for offices which have a higher business main percentage. Apparently in these cases the residence customer requires more time to handle.

Percent service representative losses and percent business main telephones are not the only factors that can be used successfully. As stated earlier, a number of other variables are nicely related to the parameters β_1 and β_2 and have approximately the same statistical efficiency. In addition, the inclusion of even more exogeneous variables can reduce the residual mean square error of the fit. For example, in Ref. 2 the average time required for a residence contact is effectively related to both percent business main telephones and the total number of main stations. Then the model takes the form,

$$T = \alpha_o A^{\alpha_1} F_1^{\gamma_0 + \gamma_1 \log C_1} F_2^{\delta_0 + \delta_1 \log C_2 + \delta_2 \log C_3}, \tag{7}$$

where C_3 is the total number of main stations.

TABLE I — ANALYSIS OF VARIANCE (LOG BASIS):
MODIFIED MODEL

Source	Degrees of Freedom	Sum of Squares (Fitted in order)	Mean Square
$\log \alpha_0$	1	96,482.4814	96,482.4814
α_1	1	781.4373	781.4373
γ_0	1	111.4103	111.4103
δ_0	1	38.6524	38.6524
γ_1	1	6.8262	6.8262
δ_1	1	24.7510	24.7510
Subtotal $\alpha_1\gamma_0\delta_0\gamma_1\delta_1$	5	963.0752	192.6150
Residual	2952	98.4815	0.0333
Total	2958	97,544.0410	

For illustration, Table I gives the details of the analysis of variance for the model of equation (5), where C_1 is percent service representative loss and C_2 is percent business main telephones. Table II gives the estimates of the parameters from the complete fit, along with their standard errors. Table III presents the correlations among the estimated parameters. One important aspect of such correlations is that

TABLE II — ESTIMATES OF PARAMETERS:
MODIFIED MODEL

Parameter	Estimate	Standard Error
$\log \alpha_0$	0.1763	0.0040
α_1	0.4400	0.0131
γ_0	0.1440	0.0074
δ_0	0.2935	0.0106
γ_1	0.0413	0.0027
δ_1	0.2507	0.0092

TABLE III — CORRELATION OF ESTIMATES OF PARAMETERS:
MODIFIED MODEL

	$\log \alpha_0$	α_1	γ_0	δ_0	γ_1	δ_1
$\log \alpha_0$	1.000	−0.622	0.386	0.027	−0.035	−0.370
α_1	−0.622	1.000	−0.458	−0.724	0.128	0.400
γ_0	0.386	−0.458	1.000	−0.090	−0.038	−0.663
δ_0	0.027	−0.724	−0.090	1.000	−0.235	−0.070
γ_1	−0.035	0.128	−0.038	−0.235	1.000	0.047
δ_1	−0.370	0.400	−0.663	−0.070	0.047	1.000

the actual values obtained as estimates of the parameters cannot be separated from the model used to obtain them. For example, the estimate of δ_1 is not the same in the models of equations (5) and (7).

Another feature of these exogeneous variables is that their selection and use in a measurement scheme will be heavily influenced by non-statistical factors. The reason is that the mere selection of variables to be included in a measurement scheme can influence the operation of the office. If not carefullly selected the measured variables may become ends in themselves and the office may operate in such a way that its objective is not performing the real work function, but rather getting credit for the measurement scheme. Such a situation could even prevent office reorganization. An office may not feel inclined to automate if such a modernization would eliminate items for which credit is given. These are undesirable results; but it is also true that this type of an interaction can be used to bring about more favorable ends. For example, if larger offices are thought to be desirable, the allotment of larger time credits to larger offices would probably create a movement towards consolidation.

It is interesting to ask how the inclusion of percent service representative losses in a measurement scheme would affect the offices. One answer is that it seems unlikely that a manager would or could try to remove employees in order to increase the turnover rate. He already has considerable pressure on him to keep these losses as small as possible. However whether this is an accurate statement or not, this example makes it clear that major management decisions are needed during the development of any measurement plan.

In summary, it seems clear that the decision to include any variable in a measurement plan should be influenced not only by the statistical characteristics of the variable but also by very careful management considerations.

V. SOME ACTUAL OFFICE COMPARISONS

The suggested measurement basis gives each office a time allotment based on the number of business and residence contacts handled and an adjusted (by the profile variables) standard time per contact, plus an allotment for overhead time based on the size (number of accounts) of the office. The formula is given in equation (8) using percent service representative losses and percent business main. This allotment is to be compared with the actual time consumed. Presumably this would be done each month.

It seems most natural to compare the allotted and actual times as

a percentage; see equation (9). Other comparisons would be possible, such as one based on the difference of the actual and allotted times, but the percentage seems preferable because of its more natural scaling. The formula used is

$$\begin{bmatrix} \text{alloted time} \\ \text{for office } i \end{bmatrix} = 1.193 A_i^{0.440} F_{1i}^{0.144+0.041C_1} F_{2i}^{0.293+0.251C_2} \tag{8}$$

where A_i is the monthly number of accounts carried by the office,

F_{1i} is the daily number of business contacts,

F_{2i} is the daily number of residence contacts,

C_1 is the monthly percent Service Representative loss,

C_2 is the percent business main telephones for the month.

$$E = \frac{\text{time allotment}}{\text{actual time reported}} \times 100. \tag{9}$$

Thus at the end of each month each office receives a rating which tells how it performed in relation to its own time standard. This allows two types of comparisons. The first is the month to month comparison of each office with itself; the second is the comparison of offices with each other on the basis of their percent efficiency. It is important to notice that these are different comparisons. It would not be impossible for an office to slip in comparison with itself from one month to the next but rank higher when compared with all other offices.

Based on the data of the three month study, the suggested procedure gives the rankings shown in Table IV. Notice that the rankings are relatively stable and that cases do occur in which the E number and the ranking go in opposite directions from one month to the next. For example, consider offices 4 and 34.

After these rankings were calculated, they were checked for obvious systematic behavior. None was found. The E values are not related to the gross time used by the office nor to any of the variables used as inputs to the estimated time. This means that the scheme does not seem to be favoring offices with special characteristics.

VI. PRINCIPAL STEPS IN FORMING THE MODEL

The key steps which lead to the final model formulation are:

(*i*) The formation of the entities. This allows analysis of comparable office groupings without which consistent statistical relationships would probably not have been found.

(*ii*) The recognition that the relationship between time consumed and demand load is nonlinear and that a log transformation allows

TABLE IV — OFFICE E NUMBERS AND RANKINGS

Office Designation	Month 1		Month 2		Month 3	
	Rank	E-Value	Rank	E Value	Rank	E-Value
1	26	96.11	37	93.98	27	99.42
2	41	83.52	34	94.69	32	95.40
3	15	105.77	16	107.29	20	104.57
4	18	98.87	20	102.67	36	92.44
5	11	111.60	8	118.58	3	123.10
6	34	93.18	21	102.63	14	108.40
7	29	94.40	32	96.07	38	92.26
8	20	98.41	22	102.06	21	103.27
9	8	115.29	12	114.08	11	115.09
10	27	95.57	39	92.94	41	85.66
11	42	80.50	45	75.73	43	78.75
12	39	89.58	23	101.13	22	101.21
13	10	112.04	11	115.40	8	119.22
14	46	70.79	46	74.32	45	73.34
15	3	123.27	4	121.93	5	121.50
16	19	98.61	24	99.38	24	100.47
17	4	120.53	7	119.10	10	116.77
18	33	93.25	35	94.46	31	95.78
19	21	98.01	38	93.97	23	100.65
20	7	116.60	5	120.60	6	120.83
21	9	114.94	6	120.53	2	129.15
22	36	91.23	36	94.07	34	93.67
23	13	109.49	15	111.11	9	117.14
24	23	97.31	26	99.15	33	94.12
25	1	159.53	1	138.73	7	120.68
26	5	118.60	13	111.73	16	107.15
27	6	118.48	3	128.52	4	121.97
28	2	130.53	2	130.76	1	132.68
29	35	92.62	25	99.17	35	93.26
30	38	89.81	30	96.48	29	98.34
31	28	94.80	19	103.05	12	112.83
32	14	109.22	10	116.47	19	106.14
33	37	90.41	27	99.00	30	96.08
34	31	93.69	33	94.94	17	106.60
35	32	93.51	31	96.11	37	92.33
36	22	98.00	40	91.44	40	86.92
37	24	97.12	28	97.94	15	107.29
38	25	96.12	17	105.94	25	99.95
39	30	93.78	29	97.41	28	98.43
40	17	102.32	14	111.26	26	99.92
41	45	74.47	43	77.79	42	80.03
42	43	79.23	41	81.59	39	90.83
43	40	84.41	44	77.15	44	76.83
44	44	75.79	42	78.32	46	72.17
45	16	104.34	18	104.60	18	106.50
46	12	110.14	9	116.77	13	112.13

simple and effective fitting. This transformation also has the important advantage of stabilizing the variances, thus making the spread of the resulting E estimates about the same for different classes of entities.

(*iii*) The recognition that the business-residence classification of customer contacts was more closely related to time usage than any other work categories. This grouping not only predicts time very well, but also requires substantially less data gathering than the more detailed classifications. In addition, since System offices tend to be organized according to the business-residence function, data gathering for this classification might possibly be completely automated.

(*iv*) The recognition that gross time can be predicted with more accuracy than the time associated with any subcategories. This means that work categories for which no frequency counts are available, are included in the analysis as "overhead" time. It also means that no work sampling is required.

(*v*) The introduction of the number of accounts as a measure of office size and its use in scaling the estimates of overhead time. Similarly, the use of the profile variables for adjusting the average time made the office comparisons more equitable.

VII. SUMMARY

This paper has described the development of statistical models for time usage in Bell System business offices. These experimental models have been designed so that they are good predictors of time, and can be used to give time allowances to different offices in an equitable way. The latter requirement means that suitable external variables have to be included. The manner in which this is carried out (see Section 4) is one of the key parts of the paper.

Finally it is pointed out that although work sampling may give very useful information in time studies, the use of a measurement scheme based on statistical models of the kind suggested in the paper would not require it. Data obtained by work sampling was used in the analysis but is not necessary for the general application of this approach.

VIII. ACKNOWLEDGMENTS

Many people have contributed significantly to this project. This is particularly true of persons in the Bell System operating companies.

There are far too many of them to be named. All of the actual observations and preliminary data tabulations were made by these people and without them no results of any kind would ever have been achieved. In addition, many of them patiently took us on guided tours of their business offices and carefully explained their operation to us. This document owes all of them a great deal.

Claire Gerity and his group at the American Telephone and Telegraph Company were very instrumental and cooperative in making arrangements and passing along their insight into the business office operations. In particular, David Macarthy of that group must be singled out for his valuable contributions.

REFERENCES

1. Williams, W. H., "A Linear Model Approach to Time and Cost Analysis," *Management Science, 12,* No. 6 (February 1966) pp. B216–223.
2. Williams, W. H., and Chen, H., unpublished work.

A Heterodyne Scanning System for Hologram Transmission[†]

By ARTHUR B. LARSEN

(Manuscript received November 19, 1968)

This paper describes the experimental realization of a recently proposed scanning reference beam technique for hologram transmission. The apparatus uses an extremely simple method for obtaining the two different but phase-locked optical frequencies necessary for the heterodyne mode of operation. The paper shows reconstructions obtained from transmitted holograms of two- and three-dimensional objects, analyzes the signal-to-noise ratio and resolution attainable with this technique, derives a new general theorem concerning the detectability of the interference between two arbitrary beams, and discusses the theorem's applications to this system.

I. INTRODUCTION

The transmission of holograms over electrical channels is of interest not only because of the three-dimensional images obtainable with such a system but also because of the possible advantages of the holographic process as a coding technique for subjectively more error-resistant transmission of two-dimensional material. Indeed, the transmission of thin holograms over conventional television systems presents no conceptual difficulties and has already been demonstrated with low resolution holograms.[1] However, the necessity of resolving the holographic carrier fringes, as well as other unnecessary spatial frequency components, results in the waste of $3/4$ of the resolution capability of the camera. Although recently devised techniques can avoid this waste, the use of these techniques in useful holographic transmission systems is still limited by the resolution of camera tubes.[2]

This paper describes the experimental realization of a scanning reference beam technique for hologram transmission recently published by Enloe, Jakes, and Rubinstein.[3] This technique not only eliminates

---

the camera tube but also requires minimum resolution from the optical scanner used in its place. Furthermore, the advantages of this system, because they accrue through the elimination of totally extraneous components present in all holograms, can be obtained while simultaneously using other bandwidth reduction schemes, such as those of Lin or Haines and Brumm.[4,5]

We briefly describe the system here, and give more specific apparatus information in Section IV. As Fig. 1 shows, the scanning system replaces the conventional reference beam by a focused spot which is optically scanned in a raster fashion over the surface of a large area photodetector. The detector provides an output current proportional to the integrated intensity of the total incident light. The time-varying interference between the stationary object beam and the constant amplitude scanning spot causes a variation in the detector output. This signal is amplified and transmitted electrically to a receiver, where it modulates the kinescope intensity. The hologram made by photographing the kinescope display is then used to reconstruct the original scene.

II. ANALYSIS OF HETERODYNE SCANNING

For mathematical simplicity and ease of understanding, the original analysis of the operation of this system as given by Enloe and others, was based on the assumption that the focused spot of the scanning reference beam could be represented by a delta function.[3] Actually, the reference beam cannot be focused to a mathematical point but is spread over a nonzero area. The shape and size of the limiting aperture in the system determine the nature and amount of this spreading and hence, the possible resolution. The limiting aperture in a real system is typically determined by the optical deflection system. We will here analyze the simple but practically important case of a focused spot formed from a uniform plane wave passing through a circular aperture.

Assuming the limiting aperture and center of deflection of the scanning beam to be located at the front focal point of the focusing lens, the distribution of $e_R(r, t)$, the electric field at a detector located in the rear focal plane of the lens, is given by

$$e_R(r, t) = \frac{2E_R J_1(\sigma r)}{\sigma r} \exp i(\omega_o t + \phi_R), \tag{1}$$

where $\sigma = 2\pi b/f\lambda$, b is the radius of the limiting aperture, f is the focal length of the focusing lens, r is the radial distance measured
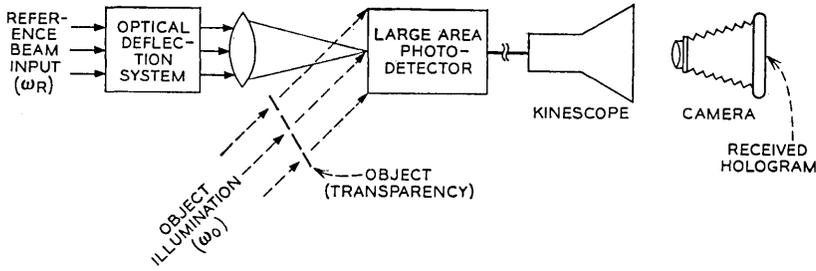
Fig. 1 — Simplified heterodyne scanning system diagram.

from the geometric center of the focused spot, and E_R, ω_R, ϕ_R, and λ are the amplitude, angular frequency, phase, and wavelength of the reference beam. The object field e_o is of the form

$$e_o(x, t) = E_o \exp i(\omega_o t + kx + \phi_o); \tag{2}$$

that is, a plane wave of amplitude E_o, angular frequency ω_o, and phase ϕ_o incident on the detector at an angle of $\theta = k\lambda/2\pi$ with respect to the normal. (Because they are virtually equal for all purposes of this derivation, no distinction is made between the wavelengths of the object and reference beams.)

With the scanning spot moving at a horizontal velocity u and a vertical velocity v, $I(r, \theta, t)$, the intensity at any point (r, θ) on the detector surface as measured from the geometric center $(x = ut, y = vt)$ of the focused spot, is given by

$$I(r, \theta, t) = \mid e_R(r, t) + e_o(r, t) \mid^2 = E_o^2 + \frac{4E_R^2 J_1^2(\sigma r)}{(\sigma r)^2} + \frac{4E_R E_o J_1(\sigma r)}{(\sigma r)}$$

$$\cdot \cos [k(ut + r \sin \theta) + (\omega_o - \omega_R)t + \phi_o - \phi_R]. \tag{3}$$

The detector output current $i(t)$ is proportional to the integral of this intensity over the detector surface of area A. Incorporating the necessary physical constants to allow writing an equality, and assuming that the detector intercepts all of the significant energy of the scanning beam so that the integrations over r can be extended to infinity, we have

$$i(t) = \frac{\eta e}{Z_o h\nu} \int I(r, \theta, t) \, dA$$

$$= \frac{\eta e}{Z_o h\nu} \left\{ E_o^2 A + \int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} \frac{4E_R^2 J_1^2(\sigma r)}{(\sigma r)^2} r \, d\theta \, dr \right.$$

$$+ \int_{r=0}^{\infty} \int_{\theta=0}^{\pi} \frac{4E_R E_o J_1(\sigma r)}{\sigma r} \cos \left[k(ut + r \sin \theta) \right.$$

$$\left. + (\omega_o - \omega_R)t + \phi_o - \phi_R \right] r \, d\theta \, dr \bigg\} , \tag{4}$$

where η is the detector quantum efficiency, e the electronic charge, $h\nu$ the photon energy, and Z_o the impedance of free space.

The first of the two integrals in equation (4) can be evaluated by a technique outlined by Born and Wolf (see p. 398 of Ref. 6); the total dc component of the detector current then becomes

$$i_{\mathrm{dc}} = \frac{\eta e}{Z_o h\nu} (E_o^2 A + 4\pi E_R^2 / \sigma^2). \tag{5}$$

The last integral of equation (4) gives the ac (signal) component of the detector output current and can be further simplified to

$$i_s(t) = \frac{\eta e}{Z_o h\nu} \frac{8\pi E_R E_o}{\sigma} \cos \left[(ku + \omega_o - \omega_R)t + \phi_o - \phi_R \right]$$

$$\cdot \int_{r=0}^{\infty} J_1(\sigma r) J_o(kr) \, dr. \tag{6}$$

Equation (6) can then be evaluated to yield[7]

$$i_s(t) = \begin{cases} \dfrac{\eta e}{Z_o h\nu} \dfrac{8\pi E_R E_o}{\sigma^2} \cos \left[(ku + \omega_o - \omega_R)t + \phi_o - \phi_R \right] \\ \qquad\qquad \text{for} \quad k < \sigma, \quad \text{that is,} \quad \theta < b/f \\ 0 \qquad\qquad \text{for} \quad k > \sigma, \quad \text{that is,} \quad \theta > b/f. \end{cases} \tag{7}$$

Thus, the scanning spot cannot resolve the phase variations in an off-axis plane wave unless $\theta < b/f$. In other words, the object beam must appear to come from within the active aperture of the lens used to form the scanning spot. Therefore, a beamsplitter of some type must be used in this system to recombine the object and reference beams.

If $\omega_o = \omega_R$, equation (7) shows the maximum electrical output frequency to be $ku/2\pi = u\theta/\lambda$. With a scanning system that provides a peak-to-peak angular beam deflection of Ω, the scan length in the focal plane of the scanning beam focusing lens is Ωf. At a horizontal scan velocity of u, this length will be traversed in $\Omega f/u$ seconds, during which time a maximum of $\Omega b/\lambda$ cycles will be generated. Thus, the maximum number of resolvable line pairs (or phase changes) is completely determined by the clear aperture and angular deflection of the optical scanner. Because the information to be modulated onto

the spatial carrier contains both positive and negative spatial frequencies, the maximum allowable modulating spatial frequency is just one-half of that determined above, giving a usable resolution of $\Omega b/2\lambda$ line-pairs per scan line. (A more detailed discussion of the resolution and bandwidth requirements for this and the following case may be found in Ref. 3.)

With $\omega_o \neq \omega_R$, equation (7) shows the ac output frequency to be limited only by $\omega_o - \omega_R$. The maximum value of k is still restricted as before, but if $\omega_o - \omega_R$ is chosen greater than uk_{max}, the entire range of k values can be used to contain a single sideband of object information, effectively doubling the scanner resolution. In this mode of operation, the transmitted holograms should yield image reconstructions having the same resolution that could be obtained by using the scanner as a flying spot image dissection system. (It may also be possible to obtain increased resolution in the case of $\omega_o = \omega_R$ by operating in a single-sideband mode; this has not yet been experimentally investigated.) The method of obtaining the two different, but phase-locked, optical frequencies needed for this maximum resolution heterodyne operation is described in Section 4.1. In both cases, the resolution in the direction parallel to the carrier fringes is determined only by the spot size.

III. SIGNAL-TO-NOISE RATIO

Under optimum conditions, the only significant noise source will be the shot noise generated in the photodetector by the dc component of the detected signal. Substituting i_{dc} as given by equation (5) into the conventional shot noise formula gives a mean-square noise current of

$$\bar{i}_n^2 = \frac{2e^2 B \eta}{Z_o h\nu} (E_o^2 A + 4\pi E_R^2/\sigma^2),\tag{8}$$

where B is the electrical bandwidth required by the system.

Using the rms signal current as given by equation (7), the signal-to-noise (power) ratio becomes

$$\frac{\bar{i}_s^2}{\bar{i}_n^2} = \frac{32\pi^2 \eta E_R^2 E_o^2}{Z_o h\nu\sigma^4 B} \bigg/ \left(E_o^2 A + \frac{4\pi E_R^2}{\sigma^2}\right).\tag{9}$$

As is usual with holography, E_R and E_o will be obtained from the same laser and will thus be subject to the constraint that

$$E_o^2 A + 4\pi E_R^2/\sigma^2 \leq Z_o P_o,\tag{10}$$

where P_o is the laser output power. [Equation (10) states that the sum of the object and reference beam powers cannot exceed the laser power.]

It can then be shown that when $E_o^2 A = 4\pi E_R^2/\sigma^2 = \frac{1}{2}Z_o P_o$, equation (9) has a maximum given by

$$\left(\frac{i_s^2}{i_n^2}\right)_{\max} = \frac{4\pi\eta P_o}{B\sigma^2 A h\nu}. \tag{11}$$

Defining an equivalent scanning spot radius r_{eq} such that a field of uniform intensity E_R incident on a circular area of radius r_{eq} provides the same photodetector current as the actual incident field, we have, using the appropriate term from equation (5)

$$\pi r_{eq}^2 E_R^2 = 4\pi E_R^2/\sigma^2 \quad \text{or} \quad r_{eq} = 2/\sigma. \tag{12}$$

With N_s, the number of resolvable spots, then given approximately by $N_s = A/\pi r_{eq}^2 = A\sigma^2/4\pi$, equation (11) becomes

$$\left(\frac{i_s^2}{i_n^2}\right)_{\max} = \frac{\eta N_p}{B N_s}, \tag{13}$$

where $N_p = P_o/h\nu$ is the number of photons per second incident on the detector. To transmit a hologram in a specified time, B will have to be increased as N_s increases, in which case the signal-to-noise ratio decreases with the square of the number of resolvable spots.

IV. EXPERIMENTAL APPARATUS AND PRELIMINARY SYSTEM OPERATION

Figure 2 is an overall diagram of the apparatus used for the experimental verification of the heterodyne scanning system.

4.1 Object and Reference Beam Generation

A Spectra-Physics Model 125 50-mw He-Ne laser operating at 6328 Å is used as the optical source. The portion of its output transmitted by beamsplitter 1 is used for object illumination; the remainder is reflected at normal incidence from the moving mirror M. This introduces a Doppler shift and is the method by which the two different but phase-locked optical frequencies are obtained. The large motions required (≈ 0.1 mm peak-to-peak) are readily obtained by using a modified loudspeaker assembly for the mirror driver. The Doppler-shifted light returns to the beamsplitter, where the transmitted portion proceeds to the optical deflection system.

4.2 Beam Deflection and Focusing

The first deflection (horizontal) is performed by an American Time Products type 44 optical scanner operating at 7.2 kHz. This unit has
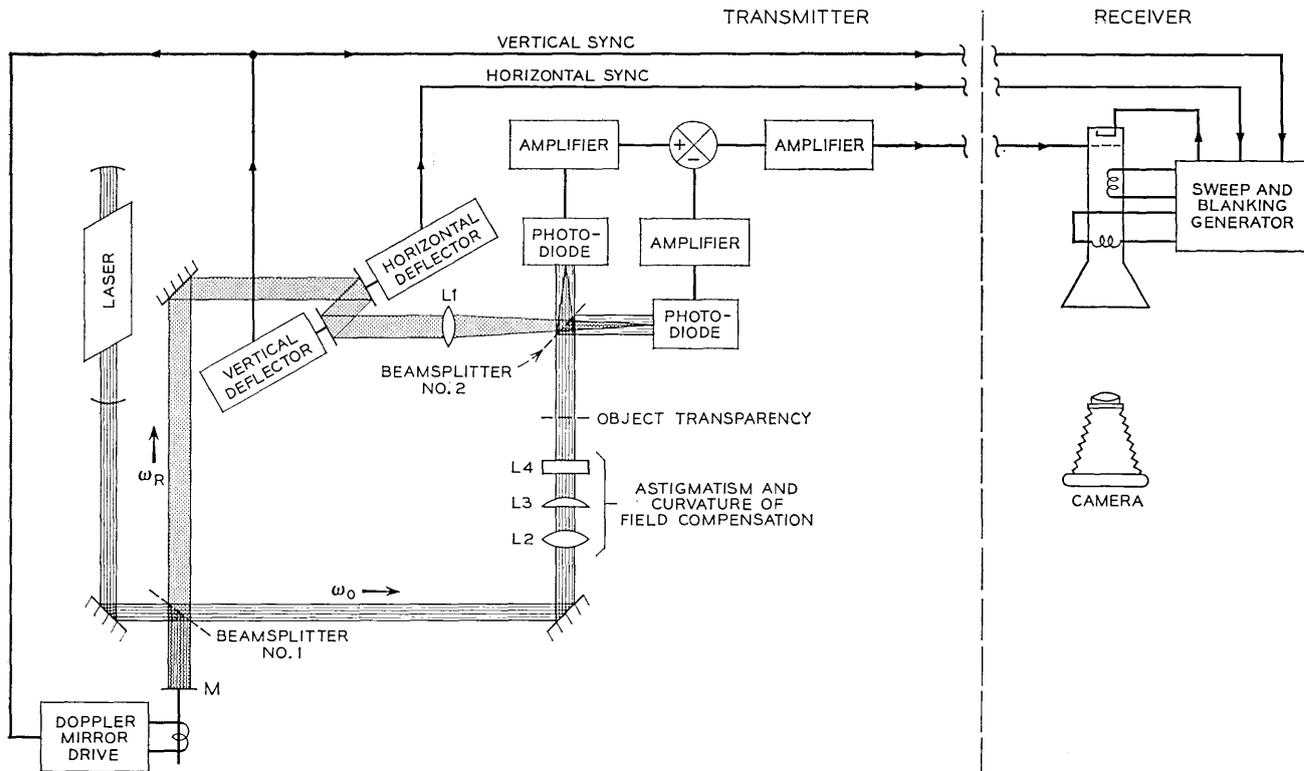
Fig. 2 — Heterodyne scanning system block diagram.

a clear aperture of 2 mm radius and provides a 6° peak-to-peak scan. The line scan thus formed is then deflected vertically by a second scanner that affords a clear aperture radius of 8 mm with a 15° peak-to-peak deflection at 60 Hz. Both scanners operate in torsional-mechanical resonance and hence provide only sinusoidal deflections.

Lens L1 transforms the angularly deflected beam into the focused scanning spot. The unavoidable physical separation of the two optical deflectors separates the horizontal and vertical centers of deflection, causing the locus of the waist of the scanning spot to be astigmatic. Compensation for this effect is provided by the inclusion of cylindrical lenses L3 and L4 in the object beam path. Spherical lens L2, in conjunction with L3 and L4, corrects for the curvature of field of the focusing lens L1 and expands the object beam to the size required to match the scanning spot raster.

4.3 *Detection*

Beamsplitter 2 is used to recombine the object and reference beams. As Section II shows, the use of a beamsplitter for this purpose is a necessity rather than a convenience. The actual detection was done using United Detector Technology type PIN-10 large area silicon photodiodes. The sensitive surface of the photodetector was originally placed at the locus of the scanning beam waist; it is still convenient to consider the detection process to occur there. However, it was experimentally observed that this particular detector location is not only unnecessary but undersirable. It is unnecessary because it can be shown that the detected beat signal is independent of the detector position provided the detector intercepts all of the area common to both beams. The special case where both beams are essentially plane waves has been long known and used by those engaged in optical heterodyne experiments, but to our knowledge, the general case has not. A derivation of this very useful result, modeled after one first given by H. Kogelnik, may be found in the appendix, which also contains other interesting applications of the general theorem.[8]

With no need to either carefully position the detector or require its surface to conform to the locus of the scanning beam waist, it can be located away from the focus, thereby reducing, by orders of magnitude, the peak instantaneous power densities at its surface. In addition, the effects of dust particles and other local anomalies of the detector surface are considerably reduced by an out-of-focus location. In all cases, the equivalent hologram is made at the locus of the scanning beam waist, independent of the actual detector position.

4.4 *Hologram Display and Recording*

The amplified and processed[†] output of the detector is used to intensity modulate a Westinghouse WX–30176P 10-inch high resolution kinescope. Synchronizing pulses from the optical deflectors are used to regenerate the sinusoidal sweeps necessary to match the kinescope sweeps to the optical ones. Because the horizontal and vertical optical deflection systems are both free-running oscillators, the kinescope display has random interlace. In conjunction with the several seconds of exposure required to record the kinescope output on Polaroid 46–L transparency film, this random interlace causes the scanning lines to be smeared together and undiscernible in the final hologram, eliminating the problem of diffraction by them.[9]

4.5 *Reconstruction*

The still limited resolution available with this system requires object-reference beam angles of less than 2°. To separate the real image from the direct beam and virtual image when reconstructing, the Fourier transform technique described by Enloe and others is used, the only modifications being the inclusion of cylindrical lenses in the final imaging process.[1] These permit compensation for astigmatic effects arising from both the oblique optical paths through the second beamsplitter and geometric distortions caused by disparities in the optical and electrical sweeps.

V. SIGNAL ENHANCEMENT TECHNIQUES

In addition to the desired signal [equation (7)] and shot noise [equation (8)], the detector output includes ac components due both to variations in the laser source output and position dependent modulations of the scanning beam. The largest of the source variations are periodic and result from plasma oscillations within the laser active medium. These are suppressed by the use of *rf* excitation of the discharge. The smaller, random, source fluctuations remaining, though comparable in amplitude with the desired signal, can be considerably attenuated by using two photodetectors in a balanced modulator configuration, as shown in Fig. 2. Source amplitude fluctuations, which produce in-phase variations in both photodetector outputs, are canceled in the following difference amplifier; the desired interference terms give rise to out-of-phase signals which are enhanced. As shown

---

† Various techniques for improving the signal-to-noise ratio which are used are discussed in the Section V.

in the appendix, this out-of-phase condition for the desired interference term can be assured only when using a lossless beamsplitter.

Position dependent modulations of the scanning beam are caused by dust and imperfections on the beamsplitter and detector surfaces as well as by multiple reflections. The poor impedance match between silicon and air causes particularly severe reflections at the detector surface. When conditions are right for this reflected light to be returned to the detector surface by a second reflection at some other optical surface, a modulation of the detector output in synchronism with the scanning spot position results. Either of these effects produce on the display kinescope a stationary pattern which can be distinguished from a true hologram by its presence in the absence of the object beam. Such position dependent modulations can be greatly attenuated by photographing the kinescope display with a double exposure technique: half of the necessary exposure is made in the usual way, while for the remainder the reference beam path is lengthened by a half wavelength and the gain of the final video amplifier is reversed in sign. This combination of optical and electrical phase shifts leaves unchanged those components of the video signal arising from interference between the object and reference beams, but reverses the polarity of the position dependent modulations described above. The position dependent modulations during the second exposure thus cancel those of the first, leaving only the desired object-reference beam interference terms.

The efficacy of these signal enhancement techniques is demonstrated in Fig. 3, which compares the outputs obtained using single detection, Fig. 3a, balanced detection, Fig. 3b, and balanced detection with double exposure, Fig. 3c. The presence of a significant amount of random noise, indicated by poor definition and contrast, is readily evident in Fig. 3a. The suppression of this noise by a balanced detection system yields considerable improvement, as shown by Fig. 3b. It is not possible, by inspection of Fig. 3b only, to determine whether or not any position dependent modulations are present. However, comparing it with Fig. 3c, in which they are suppressed, shows the nature and severity of their contribution.

VI. EXPERIMENTAL RESULTS

6.1 *Resolution*

When operated in the heterodyne mode, that is, with the off-axis reference beam simulated by a controlled frequency difference between
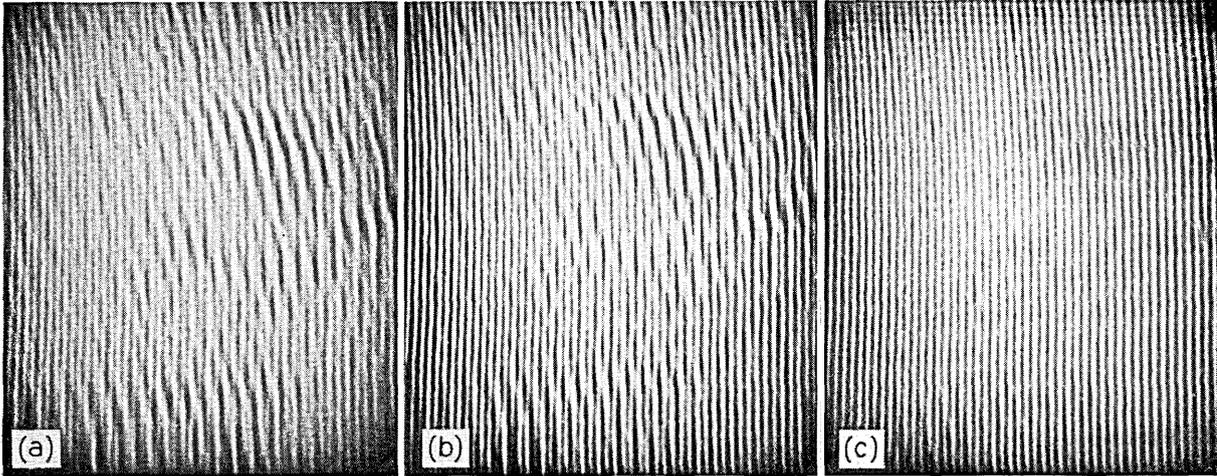
Fig. 3 — Effects of signal enhancement techniques on hologram quality:
(a) single detection, (b) balanced detection, (c) balanced detection with phase-reversal and double exposure.

the object and on-axis reference beams, this system should produce holograms that yield reconstructions with resolutions equal to those obtainable when using the same deflection system as a conventional flying-spot scanner. This assumes that the kinescope-camera portion of the system has a resolution capability of at least twice that of the optical scanner.

How well this prediction is met can be seen by comparing Fig. 4a, a photograph of the kinescope display taken when operating the system as a flying-spot scanner, with Fig. 4b, the real image reconstruction of the same object made from a transmitted hologram. Calculations, based on the parameters of the optical deflection system used, predict a resolution capability of 200 line-pairs; this value is reached in the flying-spot display of Fig. 4a. The measured limiting resolution of Fig. 4b, though only 160 line-pairs, is considerably in excess of the theoretical maximum of 100 obtainable with nonheterodyne scanning.

Figure 4a also shows the ability of the random interlace, when used in conjunction with a long exposure, to reduce the visibility of the scanning lines; the 60 lines per frame would otherwise produce a very coarse raster. Figure 5 indicates the subjective quality of the reconstructions obtainable with this system.

6.2 *Transmission of Three-Dimensional Images*

Heretofore it has been tacitly assumed but not really required that the subjects be two-dimensional. Actually this assumption may be dropped and more complicated objects considered. The next step in complexity, the simplest three-dimensional scene, consists of two (two-dimensional) transparencies separated longitudinally. The hologram transmitted for such a three-dimensional "object" (a vertical grating of period 0.5 mm located 5 cm behind a transparency portrait) is shown in Fig. 6.

The necessarily nondiffuse nature of both the illumination and the subject transparencies used in this experiment results in an extremely limited field of view, preventing the use for depth cues of not only binocular vision but also parallax. Demonstration of the three-dimensional nature of the reconstruction obtained from this hologram is therefore limited to showing the optimum focus for different portions of the reconstruction to lie in different planes. The real image reconstruction from Fig. 6, when taken in the plane of best focus for the grating, is shown in Fig. 7a. Figure 7b shows the corresponding re-
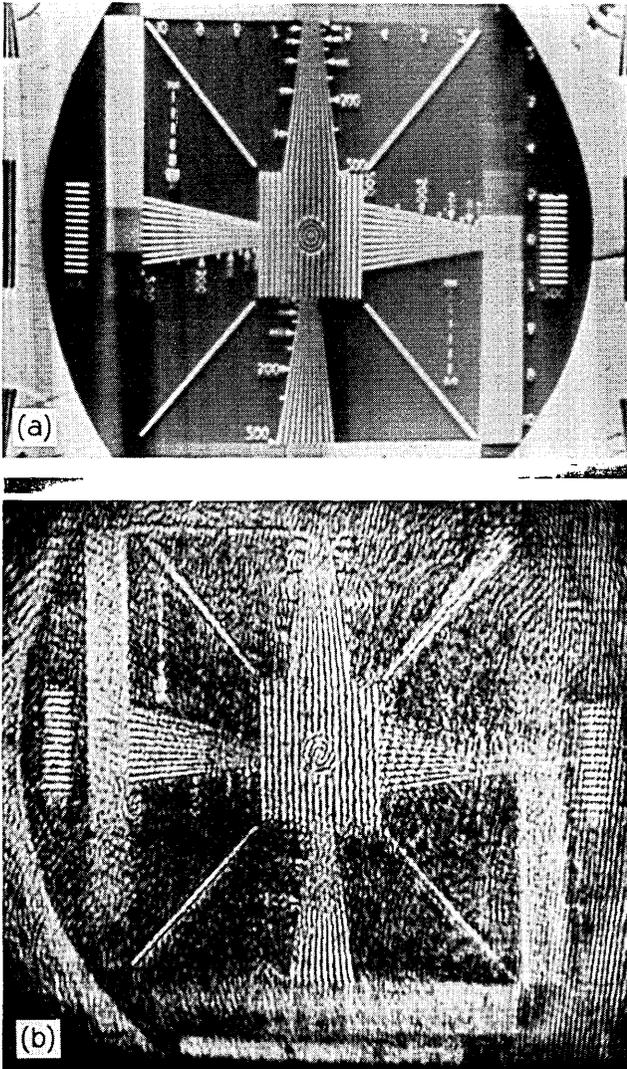
Fig. 4 — Comparison of image quality of flying-spot and heterodyne scanning transmissions: (a) flying spot, (b) heterodyne scanning.

Fig. 5 — Two-dimensional portrait reconstructed from transmitted hologram.

sult when the reconstructed image is recorded in that plane that pro-
vides optimum focus for the portrait, thus demonstrating the three-
dimensional nature of the reconstructions obtainable with this system.

VII. SUMMARY AND CONCLUDING REMARKS

A heterodyne scanning system for transmitting holograms, which
requires no camera tube and the theoretically minimum resolution
from the optical deflectors, has been constructed. This required the
development of a technique for obtaining two different but phase-
related optical frequencies. Analyses have been made for the signal-
to-noise ratio and resolution as functions of the system parameters,
and the resolution predictions verified experimentally. Several tech-
niques for improving the system signal-to-noise ratio have been im-
plemented. The use of random interlace and many-field exposures
avoided the problems of diffraction by the scanning lines when recon-
structing. Off-axis holograms of both two- and three-dimensional ob-
jects have been transmitted and reconstructed.

The kinescope-camera receiving system, though easily implemented

in the laboratory, is not only limited in its resolution capability, but is also unsuited for real-time operation. However, a receiver operating on the Eidophor principle would not only solve the real-time problem but the resulting phase holograms would also provide increased optical efficiency in reconstruction.[10]

Because of its compatibility with other bandwidth reduction schemes, the heterodyne scanning technique should find application wherever holographic information is to be transmitted over systems having limited resolution or bandwidth. The present edge in resolution held by conventional camera tubes over optical scanning devices is expected to disappear as a result of the considerable effort now being applied to optical deflection techniques.

The experimental observation and subsequent proof that the detector output is independent of the position of the scanning beam waist relative to the detector should prove to be important not only to the successful operation of this experiment but to the extension of flying-spot scanning techniques into areas where depth of focus problems have heretofore prevented their application.

Concurrent with the submission of this paper and the publication
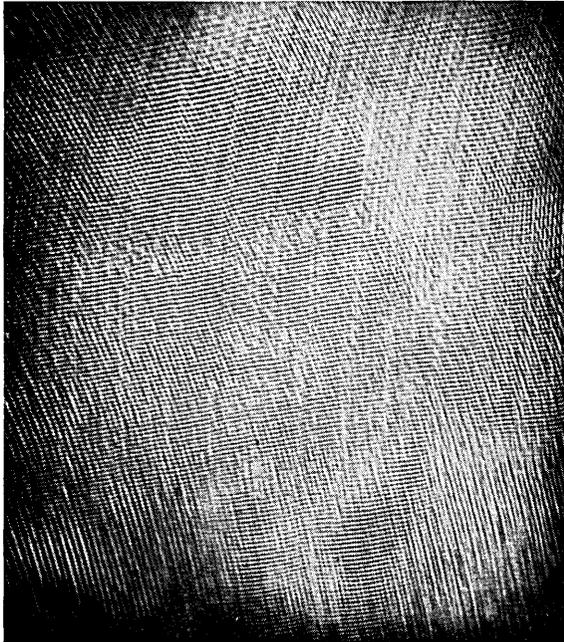
Fig. 6 — Hologram of three-dimensional scene.
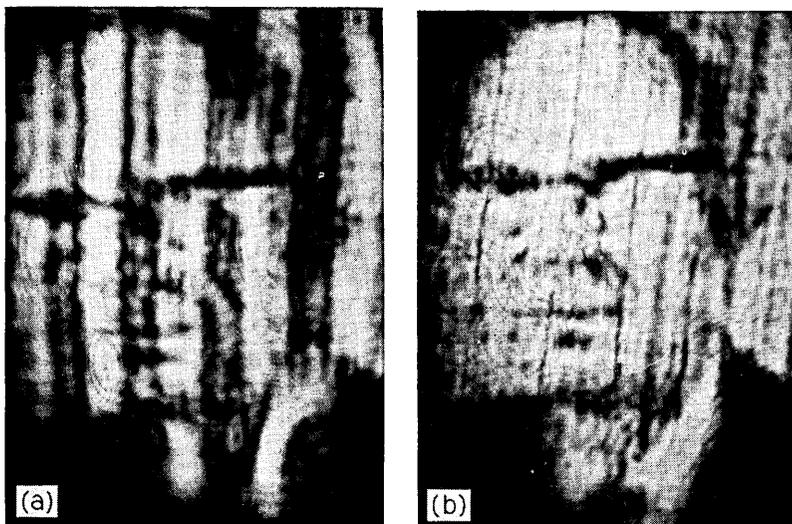
Fig. 7 — Demonstration of three-dimensional nature of image reconstructed from hologram shown in Fig. 6: (a) bars in optimum focus, (b) portrait in optimum focus.

of Ref. 3, Bertolotti and others published the results of their analysis and experiments on a one-dimensional holographic transmission system.[3,11] The transmitter described here, when operated in a non-heterodyne mode, is similar to theirs, but their analysis and proposed receiving system are different. The interested reader will find it worth-while to become acquainted with their approach to the problem of hologram transmission.

VIII. ACKNOWLEDGMENTS

The author acknowledges fruitful discussions with L. H. Enloe, H. Kogelnik, R. C. Brainard, and C. B. Rubinstein.

APPENDIX

Conservation of Beat Energy

A.1 *Derivation*

We consider the limitations and conditions under which the beat signal obtained by detecting the interference between two optical beams is independent of the detector location. The analysis is modeled after one proposed by H. Kogelnik.[8]

Consider a volume containing no optical sources or sinks and having a boundary that is everywhere in free space. Under these conditions, Poynting's theorem for any incremental volume in this region can be written

$$\nabla \cdot \mathbf{S} + \frac{\partial W}{\partial t} = 0, \tag{14}$$

where $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ and $W = \frac{1}{2}\epsilon E^2 + \frac{1}{2}\mu H^2$. $\mathbf{E}$ and $\mathbf{H}$, the resultant real electric and magnetic fields produced by the combination of the two beams, can be written as the sum of the single frequency real fields corresponding to each beam:

$$\mathbf{E} = \mathbf{E}_1 \exp i\omega_1 t + \mathbf{E}_1^* \exp -i\omega_1 t$$

$$+ \mathbf{E}_2 \exp i\omega_2 t + \mathbf{E}_2^* \exp -i\omega_2 t \tag{15}$$

$$\mathbf{H} = \mathbf{H}_1 \exp i\omega_1 t + \mathbf{H}_1^* \exp -i\omega_1 t$$

$$+ \mathbf{H}_2 \exp i\omega_2 t + \mathbf{H}_2^* \exp -i\omega_2 t,$$

where ω_1 and ω_2 may or may not be equal. Substituting the values of $\mathbf{E}$ and $\mathbf{H}$ from equation (15) into equation (14), and comparing beat terms varying as $\exp i(\omega_1 - \omega_2)t$, we have

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) = -i(\omega_1 - \omega_2)(\epsilon \mathbf{E}_1 \cdot \mathbf{E}_2^* + \mu \mathbf{H}_1 \cdot \mathbf{H}_2^*). \tag{16}$$

Equation (16) rewritten in the integral form gives

$$\oiint_C (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) \cdot d\mathbf{A}$$

$$= -i(\omega_1 - \omega_2) \iiint_V (\epsilon \mathbf{E}_1 \cdot \mathbf{E}_2^* + \mu \mathbf{H}_1 \cdot \mathbf{H}_2^*) \, dV, \tag{17}$$

where the surface C encloses the volume of integration V. A photodetector intercepting these fields will provide a beat signal current, I_b, having a complex amplitude given by

$$I_b = K \iint_D (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) \cdot d\mathbf{A}, \tag{18}$$

where D is the area of the detector and K incorporates several physical constants. The left side of equation (17) is now recognized as giving, within a constant multiplier, the response of a detector intercepting all of the beat energy crossing surface C. For the case where $\omega_1 = \omega_2$, the right side of equation (17) is identically zero. Under appropriate condi-

tions, defined in Section A.3, it is possible for the right side to be negligible even with $\omega_1 \neq \omega_2$. The following three important results, derivable from this theorem assume a zero right side.

A.2   Applications

A.2.1   Constancy of Detected Beat

Consider, as shown in Fig. 8a, a volume through which the combined beams are propagating. Let C_1 contain all of the surface C common to the two beams as they propagate into the volume, and C_2 contain all of C common to the beams as they leave. Then, by definition, the vector product is zero everywhere except over some regions of C_1 and C_2, and equation (17) can be written

$$\iint_{C_1} (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) \cdot d\mathbf{A}$$

$$+ \iint_{C_2} (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) \cdot d\mathbf{A} = 0. \qquad (19)$$

Taking into account the relative directions of the vector products and the surface normals at C_1 and C_2 leads to the conclusion that a detector intercepting the beams crossing C_2 yields identically the same output as a similar detector intercepting the beams crossing C_1. Since the separation of C_1 and C_2 is arbitrary, the detector output is independent of its location, provided all of the area common to both beams is intercepted.

A.2.2   Phase Relationships with a Lossless Beamsplitter

In the case shown in Fig. 8b, the surface C encloses a lossless beamsplitter which is combining two input beams into two output beams, each output containing a part of both inputs. C_1 and C_2 again contain all the portions of C that are common to both beams. As before, the cross-product is, by virtue of the definition of C_1 and C_2, zero everywhere except on C_1 and C_2, so that equation (19) still applies. However, when the relative directions of the beat energy flux and the surface normals are taken into account, we arrive at the result that the interaction at one detector must be the negative of that at the other, yielding detector outputs 180° out of phase.

A.2.3   Combining Beams without a Beamsplitter

The last case to be considered involves two otherwise separate beams brought in at appropriate angles so they overlap at the detector sur-
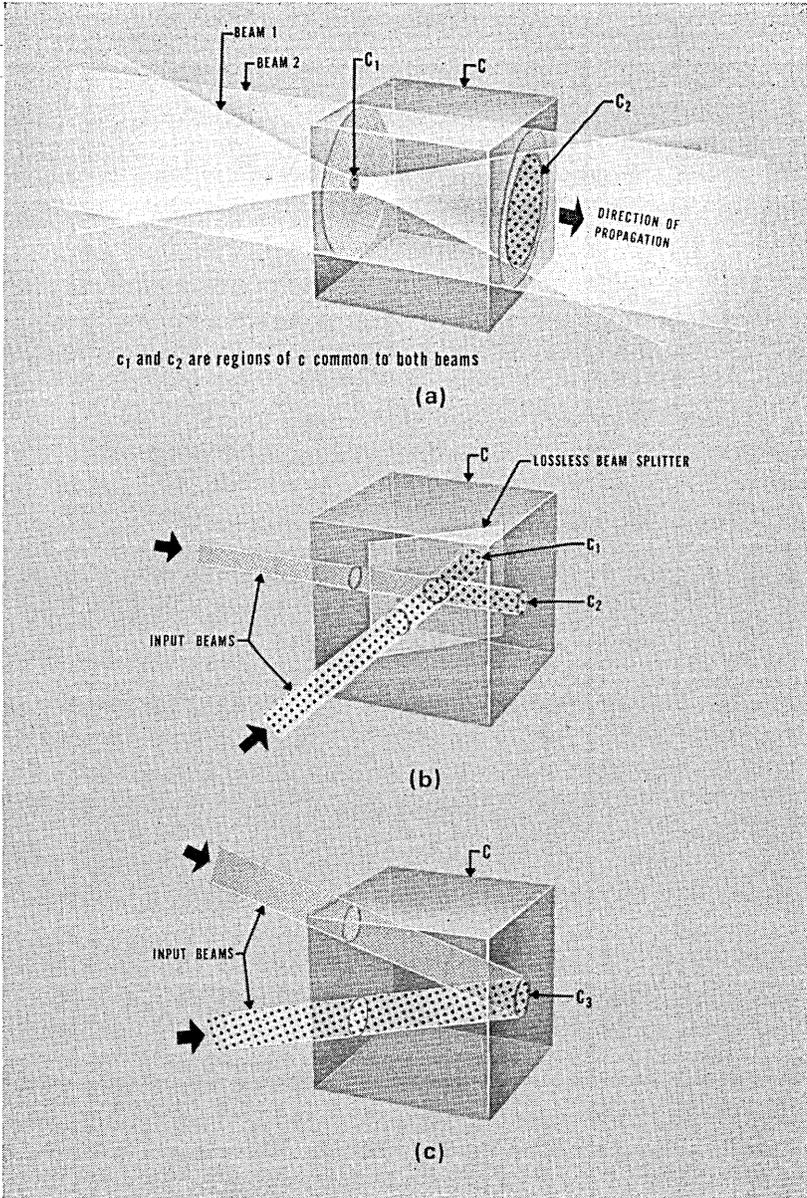
Fig. 8 — Applications of theorem on conservation of beat energy: (a) independence of detector position, (b) phase relationships with lossless beamsplitter, (c) combining beams without a beamsplitter.

face. As shown in Fig. 8c, the surface C is chosen in this case to pass just in front of the detector and to close in some region of space where the two beams are separate. Now the interaction term is everywhere zero on C except at C_3, where the two beams overlap. Equation (17) for this situation reduces to

$$\iint_{C_3} (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) \cdot d\mathbf{A} = 0, \tag{20}$$

which says that no interference is detected in this arrangement. This identical result was derived in a completely different fashion in the body of the paper in connection with the analysis of the system resolution.

A.3 *Extensions and Limitations*

Notice that these derivations require absolutely no assumptions as to the structure of either of the fields other than that they obey Maxwell's equations. Furthermore, there are no restrictions on any distances involved.

For the heterodyne case where $\omega_1 - \omega_2 \neq 0$, equation (17) has a nonzero right side. However, for small beat frequencies the above rules are still true. Although evaluation of the right side of equation (17) for the general heterodyne case is impractical, a simple example can be analyzed to indicate when it may reasonably be ignored.

Imagine a region of free space containing two plane waves with the same polarization but different frequencies, both propagating in the $+z$ direction. It can easily be shown that the single (z-directed) beat frequency component of the Poynting vector has an amplitude which is independent of z. However, its phase does vary with longitudinal position, changing by $\pi/2$ for every change in z of $\pi c/2(\omega_1 - \omega_2)$. For the 3-MHz beats observed in this experiment, this quarter-wave distance was 25 meters, much larger than the dimensions of the experimental apparatus. It is reasonable to assume that a similar phase variation would be found in the general case.

Although the amplitude of the interaction was constant even in the heterodyne case for the plane wave example, it can be shown to decrease with z for gaussian beams.[8] This amplitude change, however, typically occurs over distances orders of magnitude larger than that for the phase change; even for the extreme case of different wavelength gaussian beams focused by an $f/1$ optical system, significant phase variations of the detected beat signal occur with detector move-

ments an order of magnitude less than required for correspondingly significant amplitude changes.

REFERENCES

1. Enloe, L. H., Murphy, J. A., and Rubinstein, C. B., "Hologram Transmission via Television," B.S.T.J., *45*, No. 2 (February 1966), pp. 335–339.
2. Burckhardt, C. B. and Doherty, E. T., "Formation of Carrier Frequency Holograms with an On-Axis Reference Beam," Appl. Opt., *7*, No. 6 (June 1968), pp. 1191–1192.
3. Enloe, L. H., Jakes, W. C. Jr., and Rubinstein, C. B., "Hologram Heterodyne Scanners," B.S.T.J., *47*, No. 9 (November 1968), pp. 1875–1882.
4. Lin, L. H., "A Method of Hologram Information Reduction by Spatial Frequency Sampling," Appl. Opt., *7*, No. 3 (March 1968), pp. 545–548.
5. Haines, K. A. and Brumm, D. B., "Holographic Data Reduction," Appl. Opt., *7*, No. 6 (June 1968), pp. 1185–1189.
6. Born, M. and Wolf, E., *Principles of Optics,* New York: Pergamon Press, 1965, p. 396 and 398.
7. Gradshteyn, I. S. and Ryzhik, I. M., *Table of Integrals Series and Products,* New York: Academic Press, 1965, p. 667.
8. Kogelnik, H., unpublished work.
9. Klimenko, I. S. and Rukman, G. I., "Wavefront Reconstruction by Holograms Transmitted by Television," Zh. Tekh. Fiz., *37*, (August 1967), pp. 1532–1534. Translated in Sov. Phys.—Tech. Phys., *12*, No. 8 (February 1968), pp. 1115–1116.
10. Baumann, E., "The Fischer Large-Screen Projection System (Eidophor)," J. Soc. Motion Picture and Television Engineers, *60*, No. 4 (April 1953), pp. 344–356.
11. Bertolotti, M., Gori, F., Guattari, G., and Daino, B., "On a Method of Conversion and Reconversion of Spatial into Temporal Frequencies," Appl. Opt., *7*, No. 10 (October 1968), pp. 1961–1964.

Jump Criteria of Nonlinear Control Systems and the Validity of Statistical Linearization Approximation

By SANG H. KYONG

We study the conditions for the unique response in a class of nonlinear control systems subject to random inputs using statistical linearization approximation. As in the case of sinusoidal inputs, we show that jump phenomena may occur if the inverse vector locus of the linear part passes through certain regions on the complex plane, where the regions are defined by the characteristics of nonlinear part. Such jump phenomena regions for several typical nonlinearities are given; we also show that, among a restricted class of nonlinearities, the saturation and dead zone produce the largest jump phenomena regions.

A new result concerning the validity of statistical linearization approximation of nonlinear control systems is also presented. We show that the condition for the uniqueness of response to a given input in a nonlinear feedback system obtained through statistical linearization approximation is compatible with a related rigorous result, thus providing additional confidence in the applicability of statistical linearization.

I. INTRODUCTION

It is well known that jump resonance can occur in nonlinear control systems with attendant worsening of the control performance. In the case of periodic input signals, the rigorous conditions for the unique response,* or equivalently, for the absence of jump resonance, are available.[1] In addition, various authors have studied the conditions for the absence of jump resonance using the describing function method (see Refs. 2 and 3); the describing function method criteria

---

* Although the present terminology is widely used, a more precise term will be "unique solution to the equations arising from the steady state situation for a given input realization."

for jump resonance have been found for many common nonlinearities. For systems with random inputs, the exact condition for the unique response is not known, although a rigorous condition for the convergence of a successive approximation is available.[4]

A useful approximate technique for studying the performance of nonlinear feedback systems subject to random inputs is Booton's method of statistical linearization.[5] Although the method of statistical linearization has been widely used, the conditions for its validity are not fully known.

The first part of this paper concerns the determination of the criteria for unique response, in a class of nonlinear control systems subject to random inputs, using statistical linearization approximation. We present the statistical linearization criteria for unique response for several common nonlinearities. We also show that an idealized saturation and an idealized deadzone yield the limit jump phenomena regions among a restricted class of nonlinearities.

In view of the uncertainty concerning the conditions for the validity of statistical linearization approximation, it is of interest to compare the results of statistical linearization analysis with those of a rigorous analysis. The second part of this paper presents a result that provides new evidence on the validity of statistical linearization approximation. More specifically, the conditions for the unique response obtained in the first part on the basis of statistical linearization are compared with a related result of Holtzman,[4] which is a rigorous sufficient condition for the convergence of a successive approximation starting with the statistical linearization approximation. We show that the two conditions are "compatible"; that is, the satisfaction of the rigorous condition for the convergence of the successive approximation guarantees the satisfaction of the conditions for unique response based on statistical linearization. However, since Holtzman's rigorous condition guarantees only the convergence of a specific successive approximation but not necessarily a unique reponse, while the conditions derived from statistical linearization are for the globally unique response, the precise interpretation of the present comparison is largely open to debate. The present comparison lacks the finality of a similar comparison concerning the method of describing function in the case of periodic inputs.[6] Still, the comparison appears to provide some additional confidence in the validity of statistical linearization approximation.

Section II defines the class of nonlinear control systems to be studied and derives the conditions for the unique response based on the statistical linearization analysis. Section III presents such condi-

tions for the unique response for several typical nonlinearities. Section IV shows that if the conditions for the unique response are met by saturation and deadzone nonlinearities, then a large class of other nonlinearities will also meet the conditions. Section V shows that the statistical linearization conditions for the unique response are compatible with a related rigorous condition.

II. CONDITIONS FOR THE UNIQUENESS OF RESPONSE

Consider the nonlinear feedback system of Fig. 1. The nonlinear characteristic $f(\cdot)$ is assumed to be single-valued, odd, and piecewise continuously differentiable, and to satisfy

$$0 \leqq a \leqq f'(m) \leqq b \tag{1}$$

for all real m, where a and b are real. Concerning the linear element, it is assumed that:

(i) $G(j\omega)$ is the Fourier transform of a real function g satisfying

$$\int_{-\infty}^{\infty} |g(t)| \, dt < \infty, \tag{2}$$

(ii)
$$1 + \tfrac{1}{2}(a + b)G(j\omega) \neq 0 \tag{3}$$

for all $\omega \; \varepsilon \; (-\infty, \infty)$, and

(iii)
$$1 + K_{eq}G(j\omega) \neq 0 \tag{4}$$

for all $\omega \; \varepsilon \; (-\infty, \infty)$, where K_{eq} is the equivalent gain of the nonlinear characteristic $f(\cdot)$ obtained by statistical linearization; that is,

$$K_{eq} = \frac{E[mf(m)]}{E[m^2]}. \tag{5}$$

In equation (5), $E[\cdot]$ denotes expectation taken over the probability distribution of m. The input r to the feedback system is assumed to be a zero-mean, stationary gaussian random function with the power spectral density given by $\sigma_r^2 \phi_r(\omega)$.

We further assume that m can be represented by a zero-mean gaussian probability distribution. That m can be zero-mean follows
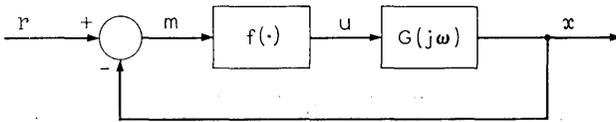
Fig. 1 — Basic feedback system.

from $f(\cdot)$ being odd. This assumption is consistent with the usual one made in connection with a statistical linearization analysis of non-linear feedback systems.[5]

If the nonlinearity $f(\cdot)$ is replaced by the equivalent gain K_{eq}, then the variance of m can easily be determined from

$$\sigma_m^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sigma_r^2 \phi_r(\omega)}{\mid 1 + K_{eq}G(j\omega) \mid^2} \, d\omega. \tag{6}$$

From equation (6), it is seen that

$$\frac{d}{d\sigma_m} (\sigma_m^2 \mid 1 + K_{eq}G(j\omega) \mid^2) > 0 \tag{7}$$

for all $\omega \ \varepsilon \ (-\infty, \ \infty)$ is sufficient to guarantee*

$$\frac{d\sigma_r}{d\sigma_m} > 0 \qquad \text{for all } \sigma_r \,. \tag{8}$$

Condition (8) implies that σ_m is a monotonically increasing function of σ_r, which in turn implies that there is a unique value of σ_m given by equation (6) for a given σ_r. This is the context in which the term "uniqueness" is used in this paper. Suppose that $(d\sigma_r/d\sigma_m) < 0$ in a certain interval of the values of σ_r. Then, the curve of σ_m versus σ_r will have the shape given by either Fig. 2a or b. Figure 2a indicates nonunique σ_m, and hence nonunique responses, or the presence of jump phenomena in the nonlinear feedback system of Fig. 1. Thus, the condition given by equation (7) is sufficient for the absence of jump phenomena in the system of Fig. 1.

Rewriting equation (7) with $H(j\omega) = G^{-1}(j\omega)$, Re $H(j\omega) = \xi_\omega$, and Im $H(j\omega) = \eta_\omega$, one obtains

$$\left(\xi_\omega + K_{eq} + \frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m}\right)^2 + \eta_\omega^2 > \left(\frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m}\right)^2. \tag{9}$$

Thus, inequality (7) is equivalent to the condition that the locus of $H(j\omega) = G^{-1}(j\omega)$, when plotted on the complex plane for $\omega \ \varepsilon \ (-\infty, \ \infty)$, remains outside of the circle centered at

$$\left(-\left[K_{eq} + \frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m}\right], 0\right) \tag{10}$$

and with radius

* Inequality (7) may be considered to be necessary as well as sufficient for condition (8), in the sense that if the inequality is reversed in inequality (7), then there is at least one $\phi_r(\omega)$; for example, $\phi_r(\omega) = \delta(\omega - \omega')$, such that condition (8) is violated
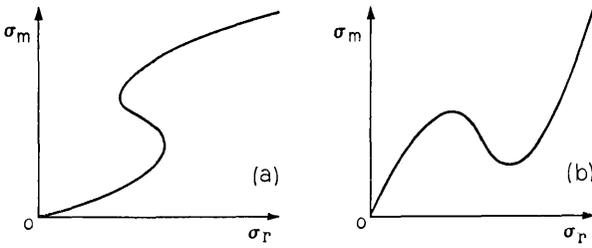
Fig. 2 — Curves of σ_m versus σ_r.

$$\rho = \left| \frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m} \right|. \tag{11}$$

The union of all such circles for all nonnegative real values of σ_m gives a region on the $H(j\omega)$-plane such that the sufficient condition (on the basis of statistical linearization) for unique response or for the absence of jump phenomena is that the locus of $H(j\omega) = G^{-1}(j\omega)$ remains outside of that region as ω is varied on $(-\infty, \infty)$.

As in Ref. 3, the circles defined by equations (10) and (11) will be called the iso-σ_m circles, and the union of all iso-σ_m circles for positive σ_m will be referred to as the jump phenomena region. Both the iso-σ_m circles and the jump phenomena region are determined by the characteristics of nonlinearity only.

III. JUMP PHENOMENA REGIONS FOR TYPICAL NONLINEARITIES

Centers and radii of iso-σ_m circles for several typical nonlinearities are tabulated in Table I along with their normalized characteristics. Figure 3 shows the jump phenomena regions of these nonlinearities.

IV. LIMIT JUMP PHENOMENA REGION

Fukuma and Matsubara have shown that, using the describing function method for the system of Fig. 1 subject to sinusoidal inputs, the jump resonance regions for idealized saturation and idealized deadzone include the jump resonance regions for all other nonlinearities satisfying

$$0 \leqq f'(m) \leqq 1, \tag{12}$$

in addition to being single-valued and odd.[3] The idealized saturation is given by

TABLE I—CHARACTERISTICS OF NONLINEARITES

	(a) RELAY	(b) RELAY WITH DEADZONE	(c) SATURATION
NONLINEAR CHARAC-TERISTIC			
K	$\left(\frac{2}{\pi}\right)^{1/2}\frac{\ell}{\sigma_m}$	$\left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sigma_m}\,\mathrm{EXP}\left(\frac{-a^2}{2\sigma_m^2}\right)$	$\left(\frac{2}{\pi}\right)^{1/2}\frac{\ell}{a}\int_0^{a/\sigma_m}\mathrm{EXP}\left(\frac{-t^2}{2}\right)dt$
ρ	$\frac{1}{(2\pi)^{1/2}}\frac{\ell}{\sigma_m}$	$-\frac{1}{(2\pi)^{1/2}}\frac{a^2\ell}{\sigma_m^3}\,\mathrm{EXP}\left(\frac{-a^2}{2\sigma_m^2}\right)$	$\frac{1}{(2\pi)^{1/2}}\frac{\ell}{\sigma_m}\,\mathrm{EXP}\left(\frac{-a^2}{2\sigma_m^2}\right)$
NORMAL-[†]IZATION	$\ell = 1$	$\ell = 1,\ a = 1$	$\ell = 1,\ a = 1$

	(d) SATURATION WITH DEADZONE		(e) DEADZONE
NONLINEAR CHARAC-TERISTIC			 $h = \mathrm{TAN}\ \gamma$
K	$\left(\frac{2}{\pi}\right)^{1/2}\frac{\ell}{(a-b)}\left[\int_0^{a/\sigma_m}\mathrm{EXP}\left(\frac{-t^2}{2}\right)dt-\int_0^{b/\sigma_m}\mathrm{EXP}\left(\frac{-t^2}{2}\right)dt\right]$		$h\left[1-\left(\frac{2}{\pi}\right)^{1/2}\int_0^{a/\sigma_m}\mathrm{EXP}\left(\frac{-t^2}{2}\right)dt\right]$
ρ	$\frac{1}{(2\pi)^{1/2}}\frac{\ell}{(a-b)\sigma_m}\left[a\,\mathrm{EXP}\left(\frac{-a^2}{2\sigma_m^2}\right)-b\,\mathrm{EXP}\left(\frac{-b^2}{2\sigma_m^2}\right)\right]$		$-\frac{1}{(2\pi)^{1/2}}\frac{ah}{\sigma_m}\,\mathrm{EXP}\left(\frac{-a^2}{2\sigma_m^2}\right)$
NORMAL-[†]IZATION	$\ell = 1,\ a = 2,\ b = 1$		$a = 1,\ \gamma = \frac{\pi}{4}$
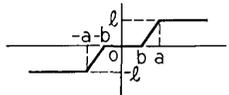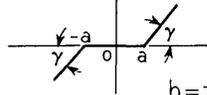
	(f) $f(m) = Nm^2\ \mathrm{sgn}\,(m)$	(g) $f(m) = Nm^3$
NONLINEAR CHARAC-TERISTIC		
K	$2\left(\frac{2}{\pi}\right)^{1/2}N\sigma_m$	$3N\,\sigma_m^2$
ρ	$-\left(\frac{2}{\pi}\right)^{1/2}N\sigma_m$	$-3N\,\sigma_m^2$
NORMAL-[†]IZATION	$N = 1$	$N = 1$

*IN ALL CASES CENTER IS GIVEN BY $-\lambda + j0$, RADIUS IS GIVEN BY $|\rho|$, WHERE $\lambda = K - \rho$.
† NORMALIZATION OF THE PARAMETER VALUES OF NONLINEAR CHARACTERISTIC USED IN FIGURE 3.
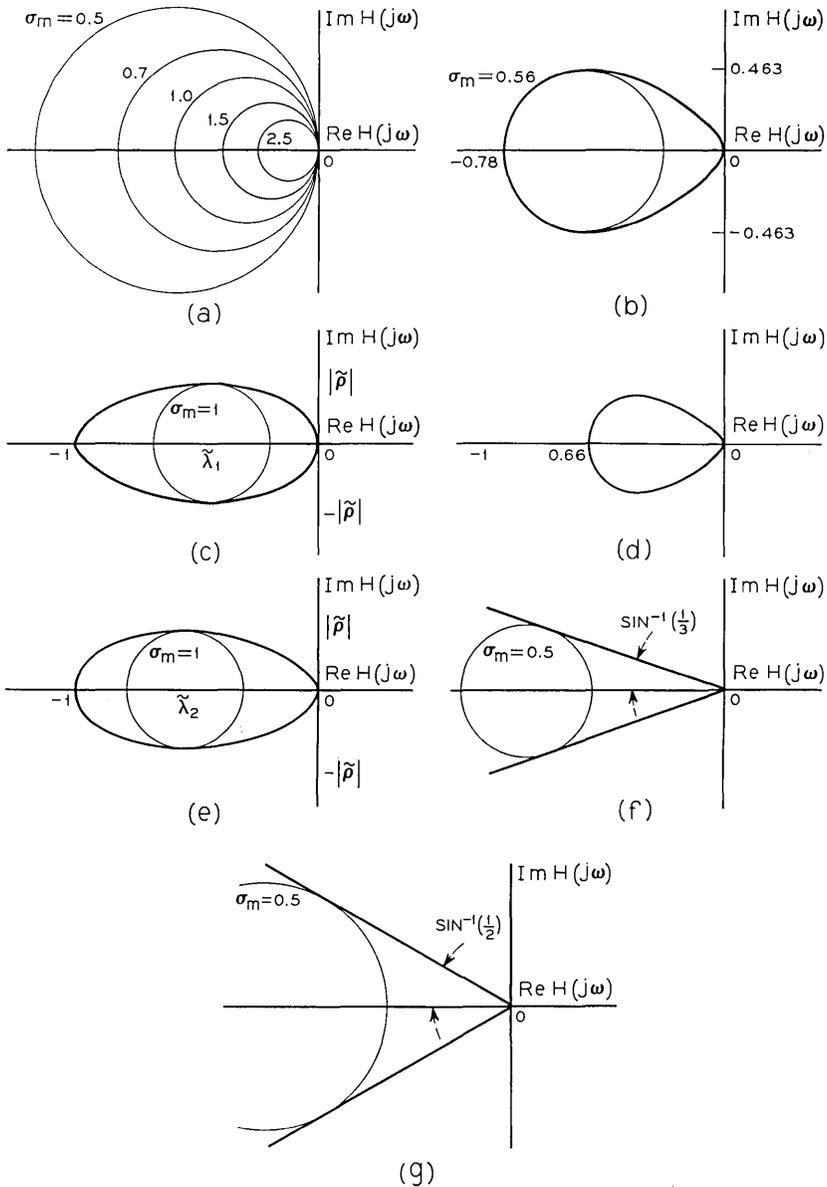
Fig. 3 — Jump phenomena regions: (a) relay, (b) relay with deadzone, (c) saturation, (d) saturation with deadzone, (e) deadzone, (f) $f(m) = m^2 \operatorname{sgn}(m)$, and (g) $f(m) = m^3$.

$$f(m) = \begin{cases} -\alpha; & m < -\alpha, \\ m; & -\alpha \leqq m \leqq +\alpha, \\ +\alpha; & +\alpha < m, \end{cases} \tag{13}$$

and the idealized deadzone is given by

$$f(m) = \begin{cases} m + \beta; & m < -\beta, \\ 0 ; & -\beta \leqq m \leqq +\beta, \\ m - \beta; & +\beta < m, \end{cases} \tag{14}$$

where α and β are positive real constants. Such limit jump resonance regions are determined by finding the nonlinearity satisfying inequality (12) which maximizes the radius of the circle given the coordinates of the center of the circle.

In this section we show that the idealized saturation and idealized deadzone give a limit jump phenomena region also in the case of random inputs, if $f(\cdot)$ is restricted to those satisfying inequality (12).

Notice that

$$\frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m} = \frac{d}{dv} E[mf(m)] - K_{eq},$$

where $v = \sigma_m^2$. From a theorem given in Ref. 7, $(d/dv)E[mf(m)] = E[f'(m)] + \frac{1}{2}E[mf''(m)]$, where prime denotes differentiation with respect to the argument. Integrating the first term on the right by parts,

$$E[f'(m)] = K_{eq}. \tag{15}$$

These relations reduce to

$$\frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m} = \frac{1}{2}E[mf''(m)].$$

If $f(\cdot)$ is such that $f''(m)$ is piecewise continuous, then the right side of the above equation may be integrated by parts to give

$$\frac{\sigma_m}{2} \frac{dK_{eq}}{d\sigma_m} = -\frac{1}{2} \int_{-\infty}^{\infty} f'(m)[p(m) + mp'(m)] \, dm. \tag{16}$$

For the gaussian probability density function for $p(m)$,

$$mp'(m) = -\frac{m^2}{\sigma_m^2} p(m).$$

Therefore, equation (16) may be rewritten as

$$\frac{\sigma_m}{2}\frac{dK_{eq}}{d\sigma_m} = \frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{m^2}{\sigma_m^2} - 1\right)f'(m)p(m)\,dm. \tag{17}$$

If $f'(m)$ is only piecewise continuous (as in the case of saturation and deadzone given by equations (13) and (14), respectively) then $f''(m)$ is not piecewise continuous, and the integration by parts used above to obtain equation (16) may not be valid in the ordinary sense. However, if the meaning of the integration

$$E[mf''(m)] = \int_{-\infty}^{\infty} mp(m)f''(m)\,dm$$

is extended, and is considered as an operation of a distribution $f''(m)$ on an infinitely smooth function $mp(m)$, then a generalized integration by parts can be used.[8] The use of integration by parts, in the generalized sense, does not change the result in the present case, and equation (17) remains valid.

Now, combining equation (17) with equation (15),

$$-\left(K_{eq} + \frac{\sigma_m}{2}\frac{dK_{eq}}{d\sigma_m}\right) = -\frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{m^2}{\sigma_m^2} + 1\right)f'(m)p(m)\,dm. \tag{18}$$

Suppose that the coordinate of the center of the circle is fixed, that is,

$$-\left(K_{eq} + \frac{\sigma_m}{2}\frac{dK_{eq}}{d\sigma_m}\right) = -\lambda, \tag{19}$$

where λ is a constant. Clearly from equation (18), $0 \leq \lambda \leq 1$ for $f'(m)$ satisfying inequality (12). From equations (18) and (19),

$$\int_{-\infty}^{\infty}(m^2 + \sigma_m^2)f'(m)p(m)\,dm = 2\lambda\sigma_m^2. \tag{20}$$

The nonlinearity that gives the limit jump phenomena region is found by determining $f'(m)$ such that it maximizes

$$\left|\frac{\sigma_m}{2}\frac{dK_{eq}}{d\sigma_m}\right| = \frac{1}{2\sigma_m^2}\left|\int_{-\infty}^{\infty}(m^2 - \sigma_m^2)f'(m)p(m)\,dm\right| \tag{21}$$

subject to the constraints given by equations (12) and (20).

By using Pontryagin's maximum principle the appendix shows that the solution of above optimization problem is given by an idealized saturation and an idealized deadzone, or the nonlinearities of the form of equations (13) and (14), respectively. In other words, the idealized saturation and idealized deadzone yield the limit jump phenomena regions among all nonlinearities which are single-valued and odd, and

satisfy $0 \leqq f'(m) \leqq 1$, in the case of gaussian random input, as well as in the case of sinusoidal input.

Suppose that the unit of the signals r, m, and so on, is normalized such that σ_m is taken as the unit. Then the appendix also shows that the jump phenomena circles giving the maximum radius are centered at $(-\tilde{\lambda}_1, 0)$ for the idealized saturation with $\alpha = 1$ in equation (13) and at $(-\tilde{\lambda}_2, 0)$ for the idealized deadzone with $\beta = 1$ in equation (14), where

$$\tilde{\lambda}_1 = \frac{1}{2(2\pi)^{\frac{1}{2}}} \left[\int_0^1 e^{-\gamma/2} \gamma^{1/2} \, d\gamma + \int_0^1 e^{-\gamma/2} \gamma^{-1/2} \, d\gamma \right], \qquad (22)$$

$$\tilde{\lambda}_2 = \frac{1}{2(2\pi)^{\frac{1}{2}}} \left[\int_1^\infty e^{-\gamma/2} \gamma^{1/2} \, d\gamma + \int_1^\infty e^{-\gamma/2} \gamma^{-1/2} \, d\gamma \right]. \qquad (23)$$

In both cases, the magnitude of the maximum radius is given by

$$\tilde{\rho} = \frac{1}{2(2\pi)^{\frac{1}{2}}} \left[\int_0^1 e^{-\gamma/2} \gamma^{-1/2} \, d\gamma - \int_0^1 e^{-\gamma/2} \gamma^{1/2} \, d\gamma \right]. \qquad (24)$$

The values of the integrals in equations (22) through (24) are tabulated in Ref. 9; it is found that

$$\tilde{\lambda}_1 = 0.44072, \qquad \tilde{\lambda}_2 = 0.55928, \qquad \tilde{\rho} = 0.24197.$$

V. COMPATIBILITY OF CONDITIONS

In this section, we compare inequality (7), which is an approximate condition for the uniqueness of response or the absence of jump phenomenon based on statistical linearization, with a related rigorous condition to obtain further evidence concerning the validity of statistical linearization approximation. Section II showed that inequality (7) implies the condition that the locus of $H(j\omega) = G^{-1}(j\omega)$ must remain outside of the circle defined by equation (10) and (11) on the complex plane as ω is varied on $(-\infty, \infty)$.

On the other hand, the rigorous sufficient condition for the convergence of a successive approximation is given in Ref. 4 as

$$\sup_{\omega \varepsilon (-\infty, \infty)} \left| \frac{G(j\omega)}{1 + \frac{1}{2}(a + b)G(j\omega)} \right| \frac{1}{2}(b - a) < 1. \qquad (25)$$

Inequality (25) implies that the locus of $H(j\omega) = G^{-1}(j\omega)$ on the $H(j\omega)$-plane, as ω is varied on $(-\infty, \infty)$, must not enter the circle centered at

$$(-\tfrac{1}{2}[a + b], 0) \qquad (26)$$

with radius

$$\rho = \tfrac{1}{2}(b - a). \tag{27}$$

The circle defined by equations (10) and (11) intersects the real axis of the complex plane at $-K_{eq}$ and $-[K_{eq} + \sigma_m(dK_{eq}/d\sigma_m)]$ with its center lying on the real axis. Similarly, the circle defined by equations (26) and (27) intersects the real axis at $-a$ and $-b$ with its center also lying on the real axis. Thus it suffices to show

$$a \leqq K_{eq} + \sigma_m \frac{dK_{eq}}{d\sigma_m} \leqq b, \tag{28}$$

and

$$a \leqq K_{eq} \leqq b, \tag{29}$$

for all positive σ_m.

But inequality (29) follows immediately from equations (15) and (1). Combining equations (17) and (18),

$$K_{eq} + \sigma_m \frac{dK_{eq}}{d\sigma_m} = \frac{1}{\sigma_m^2} \int_{-\infty}^{\infty} m^2 f'(m) p(m) \, dm. \tag{30}$$

From equations (1) and (30), $\alpha \leqq K_{eq} + \sigma_m(dK_{eq}/d\sigma_m) \leqq \beta$, which is the inequality (28).

Thus, inequalities (28) and (29) are satisfied for all $\sigma_m > 0$, which implies the two conditions are compatible; that is the satisfaction of condition (25) implies the satisfaction of condition (7) for all $\sigma_m > 0$, and conversely, the violation of condition (7) for some $\sigma_m > 0$ implies the violation of condition (25).

VI. CONCLUDING REMARKS

The conditions for the unique response in a randomly excited nonlinear control system were studied using a statistical linearization approximation. The jump phenomena regions of several common nonlinearities were given. It was shown that, among nonlinearities satisfying $0 \leqq f'(m) \leqq 1$, the idealized saturation and idealized deadzone yield the limit jump phenomena regions.

It was also shown that, concerning the uniqueness of the response in nonlinear feedback systems subject to random input, the criterion obtained by the statistical linearization approximation is not contradicted by a related, although not equivalent, rigorous result. However, as mentioned in the introduction, the interpretation of this

result is largely open to debate since (*i*) the comparison made is between an approximate and a rigorous sufficient condition, and (*ii*) the two sufficient conditions are not concerned with exactly the same requirement. More specifically, the approximate criterion obtained in Section II is for a globally unique response, while the rigorous result of Holtzman, with which the comparison is made, is for the convergence of a specific successive approximation.

It is shown in Ref. 10 that, in a system closely related to that with which the present paper is concerned, there is a unique response up to an equivalence to an input r satisfying

$$\limsup_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |r(t)|^2 \, dt < \infty,$$

if the condition identical to condition (25) is satisfied. This result strongly suggests that condition (25) may be sufficient not only for the convergence of a specific successive approximation as shown in Ref. 4, but also for a globally unique response (up to an equivalence). If this is true, then the meaning of the result of comparison made in the present paper is correspondingly strengthened.

It is interesting to compare the limit jump phenomena regions of the present approximate analysis (Fig. 3c and e) with the circle of rigorous analysis, and to notice that the limit jump phenomena regions occupy substantial portions of the interior of the circle of exact analysis. Also notice that in view of inequalities (25) and (29), the statistical linearization analysis of the system of Fig. 1 always has a solution under the conditions dicussed in Section II.

APPENDIX

Optimization Problem

The following optimization problem is stated in Section IV: Maximize $|\rho|$, where

$$\rho = \frac{1}{2\sigma_m^2} \int_{-\infty}^{\infty} (m^2 - \sigma_m^2) f'(m) p(m) \, dm, \tag{31}$$

by choosing $f'(m)$, $-\infty < m < \infty$, satisfying the condition

$$0 \leqq f'(m) \leqq 1, \tag{32}$$

subject to the constraint

$$\int_{-\infty}^{\infty} (m^2 + \sigma^2)f'(m)p(m) \, dm = 2\lambda\sigma_m^2 , \tag{33}$$

where λ is a given constant such that $0 \leq \lambda \leq 1$. This problem may be solved by making use of Pontryagin's maximum principle.

Since both $(m^2 - \sigma_m^2)p(m)$ and $(m^2 + \sigma^2)p(m)$ are even functions of m, it suffices to find $f'(m)$ for $m \geqq 0$, and to let $f'(-m) = f'(m)$. Thus, the problem may be reformulated in the following way. Let

$$\dot{x}_1(m) = (m^2 - \sigma_m^2)p(m)f'(m), \tag{34}$$

$$\dot{x}_2(m) = (m^2 + \sigma_m^2)p(m)f'(m), \tag{35}$$

where $x_1(0) = x_2(0) = 0$. We want to minimize or maximize $x_1(\infty)$ subject to $x_2(\infty) = \lambda\sigma_m^2$. Pontryagin's maximum principle may be used to the above reformulation. The Hamiltonian function is

$$H = g_1(m)(m^2 - \sigma_m^2)p(m)f'(m) + g_2(m)(m^2 + \sigma_m^2)p(m)f'(m), \tag{36}$$

where $g_1(m)$ and $g_2(m)$ are the adjoint variables. Clearly, $\dot{g}_1(m) = \dot{g}_2(m) = 0$.

Suppose first that $x_1(\infty)$ is to be minimized. Then g_1 may be set as $g_1 = -1$; and maximizing the resulting H with respect to $f'(m)$ satisfying inequality (32), one obtains,

$$f'(m) = \tfrac{1}{2} + \tfrac{1}{2} \operatorname{sgn} \left[-(m^2 - \sigma_m^2) + g_2(m^2 + \sigma_m^2) \right]. \tag{37}$$

It is easy to determine that

$$-1 \leqq g_2 \leqq +1 \tag{38}$$

to satisfy the constraint $x_2(\infty) = \lambda\sigma_m^2$. For the values of g_2 satisfying inequality (38), equation (37) and $f'(m) = f'(-m)$ give

$$f'(m) = \begin{cases} 1; & |m| \leq \left(\dfrac{1 + g_2}{1 - g_2}\right)^{\frac{1}{2}} \sigma_m , \\[2mm] 0; & |m| > \left(\dfrac{1 + g_2}{1 - g_2}\right)^{\frac{1}{2}} \sigma_m , \end{cases} \tag{39}$$

as the one that minimizes $x_1(\infty)$. The actual value of g_2 is determined from equation (33), or

$$\int_0^\alpha (m^2 + \sigma_m^2)p(m) \, dm = \lambda\sigma_m^2 , \tag{40}$$

where $\alpha = (1 + g_2/1 - g_2)^{\frac{1}{2}}\sigma_m$.

Proceeding similarly, the function $f'(m)$ which maximizes $x_1(\infty)$ is

$$f'(m) = \begin{cases} 0; & |m| \leq \left(\dfrac{1-g_2}{1+g_2}\right)^{\frac{1}{2}}\sigma_m , \\[3mm] 1; & |m| > \left(\dfrac{1-g_2}{1+g_2}\right)^{\frac{1}{2}}\sigma_m , \end{cases} \tag{41}$$

where $-1 \leq g_2 \leq 1$. The actual value of g_2 is found from equation (33), or

$$\int_{\beta}^{\infty} (m^2 + \sigma_m^2)p(m)\,dm = \lambda\sigma_m^2 , \tag{42}$$

where $\beta = (1 - g_2/1 + g_2)^{\frac{1}{2}}\sigma_m$.

The functions $f'(m)$ of equations (39) and (41) correspond to idealized saturation and idealized deadzone, respectively. Thus, among nonlinearities giving $\rho < 0$, $f'(m)$ of equation (39) yields the limit jump phenomena region, and among the ones giving $\rho > 0$, $f'(m)$ of equation (41) yields the same.

Having determined the functions $f'(m)$ that maximize $|\rho|$, it is also of interest to determine the actual values of maximum $|\rho|$ and the location of the center of the corresponding circles on complex plane. In case of idealized saturation, the maximum of $|\rho|$ corresponds to the minimum of ρ, and

$$\tilde{\rho} = \frac{1}{\sigma_m^2} \int_0^{\alpha} (m^2 - \sigma_m^2)p(m)\,dm, \tag{43}$$

where α is given following equation (40). We want to find the value of λ, $0 \leq \lambda \leq 1$, such that $\tilde{\rho}$ above is further minimized, and to find that minimum value of $\tilde{\rho}$. Differentiating equation (43) with respect to λ,

$$\frac{d\tilde{\rho}}{d\lambda} = \frac{1}{\sigma_m^2} (\alpha^2 - \sigma_m^2)p(\alpha) \frac{d\alpha}{d\lambda}. \tag{44}$$

But, from equation (40),

$$p(\alpha) \frac{d\alpha}{d\lambda} = \frac{\sigma_m^2}{\sigma_m^2 + \alpha^2}. \tag{45}$$

Thus, equation (44) becomes

$$\frac{d\tilde{\rho}}{d\lambda} = \frac{\alpha^2 - \sigma_m^2}{\alpha^2 + \sigma_m^2}. \tag{46}$$

For minimum $\tilde{\rho}$, $\alpha = \sigma_m$ or $g_1 = 0$. Thus,

$$\tilde{\rho}_{\min} = \frac{1}{\sigma_m^2} \int_0^{\sigma_m} (m^2 - \sigma_m^2)p(m)\,dm, \tag{47}$$

and the corresponding value of λ is given by

$$\bar{\lambda}_1 = \frac{1}{\sigma_m^2} \int_0^{\sigma_m} (m^2 + \sigma_m^2) p(m) \, dm. \tag{48}$$

In order to obtain the results which are independent of the particular signals used, suppose that the idealized saturation being considered is further normalized such that

$$f(m) = \begin{cases} -1; & m < -1, \\ m; & -1 \leqq m \leqq 1, \\ 1; & 1 < m. \end{cases} \tag{49}$$

The units of the signals are also normalized such that σ_m is taken as the unit. With these normalizations, the integrals of equations (47) and (48) may be evaluated using the tables in Ref. 9 to obtain $\bar{\lambda}_1 = 0.44072$, $\bar{\rho}_{\min} = 0.24197$.

In a similar manner, for the normalized idealized deadzone given by

$$f(m) = \begin{cases} m + 1; & m < -1, \\ 0 \;; & -1 \leqq m \leqq 1, \\ m - 1; & 1 < m, \end{cases} \tag{50}$$

it is found that $\bar{\lambda}_2 = 0.55928$, $\bar{\rho}_{\max} = 0.24197$.

REFERENCES

1. Sandberg, I. W., "On the Response of Nonlinear Control Systems to Periodic Input Signals," B.S.T.J., *43*, No. 3 (May 1964), pp. 911–926.
2. Hatanaka, H., "The Frequency Responses and Jump Resonance Phenomena of Nonlinear Feedback Control Systems," Trans. Amer. Soc. Mechanical Engineers, J. of Basic Eng., *85*, No. 2 (June 1963), pp. 236–242.
3. Fukuma, A., and Matsubara, M., "Jump Resonance Criteria of Nonlinear Control System," Trans. IEEE Automatic Control, *AC-11*, No. 4 (October 1966), pp. 699–706.
4. Holtzman, J. M., "Analysis of Statistical Linearization of Nonlinear Control Systems," SIAM J. Control, *6*, No. 2 (May 1968), pp. 235–243.
5. Booton, R. C., Jr., "The Analysis of Nonlinear Control Systems with Random Input," *Proc. Symp. on Nonlinear Circuit Analysis,* New York: Polytechnic Institute of Brooklyn, 1953, pp. 369–391.
6. Holtzman, J. M., *Nonlinear System Theory,* Prentice-Hall: Englewood Cliffs, N. J., (in press).
7. Papoulis, A., *Probability, Random Variables and Stochastic Processes,* New York: McGraw-Hill, 1965, pp. 160–161.
8. Zemanian, A. H., *Distribution Theory and Transform Analysis,* New York: McGraw-Hill, 1965, pp. 67–70.
9. Pearson, E. S., and Hartley, H. O., *Biometrika Tables for Statisticians,* Vol. 1, 3rd Ed., Cambridge, England: University Press, 1966, pp. 9, 128–135.
10. Beneš, V. E., "A Nonlinear Integral Equation in the Marcinkiewicz Space $\mathfrak{M}_2$," J. Math. Phys., *44*, No. 1 (January 1965), pp. 24–35.

A Video Encoding System With Conditional Picture-Element Replenishment

By F. W. MOUNTS

(Manuscript received January 29, 1969)

This paper describes an experimental method for encoding television signals which takes advantage of the frame-to-frame correlation to reduce transmission bit rate. The technique encodes only those elements that change between successive frames instead of encoding every element of every frame. We have demonstrated the method in real-time using the head-and-shoulder view of a person in animated conversation as the picture source, such as is likely to be encountered in a visual communication system. An average transmission rate of one bit per picture element gives quality comparable with standard eight-bit PCM transmission.

I. INTRODUCTION

Most known methods for efficiently transmitting pictures over communication circuits exploit point-to-point correlation along a scanning line. In particular, point-to-point predictive quantizing systems have been successful, but it is known that there is more correlation between television picture elements in the frame-to-frame time dimension than there is between adjacent elements in a single frame. This is especially true when using the head-and-shoulder view of a person as the picture source, such as is likely to be encountered in a visual communication system. Now that devices for storing large amounts of data and integrated circuits for complex digital processing are available, it is not only possible to take advantage of this fact in picture coding, but it is also economically attractive.

In this paper, we describe one method of encoding television signals which takes advantage of the frame-to-frame correlation to reduce the transmission bandwidth. We also describe the experimental facility and the results of initial experiments.

We want to emphasize that the picture source for these experiments is a head-and-shoulder view of a person carrying on an animated

conversation in real time. Thus, the results are not particularly relevant to stationary graphics or to commercial television where frames are switched, panned, and zoomed. Our experiments have been confined to noninterlaced, 60 frames per second television pictures.

II. GENERAL DESCRIPTION

The technique encodes for transmission only those elements that change between frames instead of encoding every element of every frame. This method has been described previously by R. D. Kell, A. J. Seyler, and T. C. Damen.[1,2,3] Seyler has published statistics of frame-to-frame differences for commercial television signals[4] and has proposed coding methods[5] that are based on this information. E. R. Kretzmer has investigated the correlation between successive frames of motion-picture films showing that there is redundancy that may be exploited.[6] When using video-telephone-like signals with moderate motion in the scene, we find, on the average, that less than one-tenth of the elements change between frames by an amount which exceeds 1 percent of the peak signal. We regard such 1 percent changes as being significant.

We shall describe a complete transmission system that makes use of frame-to-frame redundancy to gain encoding efficiency. The technique which we call "conditional replenishment" is found to be particularly useful for the pictures encountered in visual communication systems. The conditional replenishment system uses a memory to store a reference picture, and only those elements of the picture that have changed significantly between frames are updated (or replenished). Only the picture information necessary to update the reference picture need be transmitted. At the receiver, this information is used to update a similar stored reference picture which is intended to track the one stored at the transmitter.

In order for the receiver to correctly update the picture elements, two pieces of information must be conveyed to the receiver—the new value and the position of the element to be replenished. Because this information occurs at a random rate, buffers are used to redistribute and present the information to the transmission channel at a uniform bit rate. In order to regulate the average replenishment rate to match the channel capacity, the threshold (which determines whether or not a significant change in the picture information has occurred) is varied as a function of the amount of information stored in the buffer.

This method of encoding requires that only the information pertain-

ing to the active region of the picture format be transmitted. The receiver reinserts the horizontal and vertical blanking interval within the reconstructed video information.

III. TRANSMITTER

Figure 1 shows the operations performed by the transmitter. The video signal from the camera is band limited, sampled, and digitized into eight-bit PCM. A selector switch is provided which either conveys new information to the input of the reference frame memory whenever a significant difference is detected or alternatively recirculates the information presently stored in the frame memory. The frame memory consists of delay lines and has sufficient capacity to store one complete frame of video information—each sample encoded as eight-bit PCM.

New information from the camera is compared with the reference picture stored in the frame memory by a subtractor circuit which yields the absolute difference between the new sample of information and the reference value corresponding to the same picture element. During each sample period, the control logic makes a decision, depending upon the magnitude of the difference signal, as to whether a significant difference between the signal values exists. If the difference is significant, the output of the control logic operates the selector switch to strobe the new signal value into the frame memory.
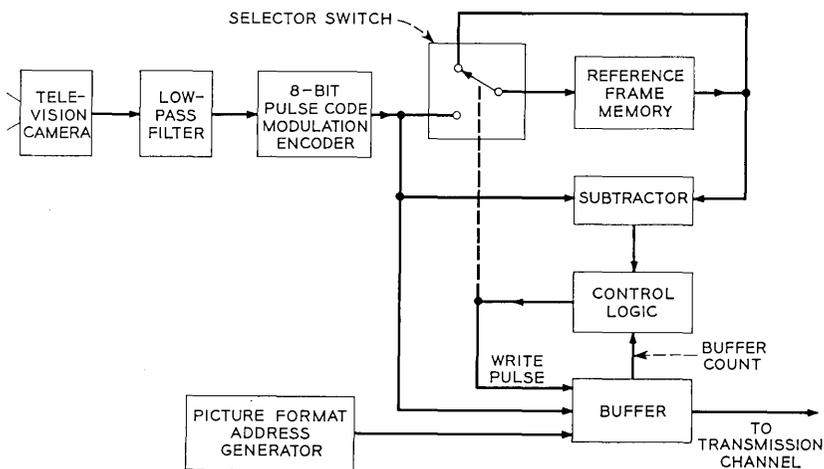
Fig. 1 — Conditional replenishment transmitter terminal.

If the difference is insignificant, the signal value stored in the frame memory is recirculated. In addition to replenishing new information in the frame memory, the control logic also causes the new signal value, accompanied by its address, to be stored in a buffer. The buffer store matches the varying data rate to the constant bit rate of the transmission channel. The information stored in the buffer is read at a constant rate, first-in, first-out.

In the implementation of the experimental system, the amplitude information is expressed as eight-bit PCM with an additional seven bits being used to identify the position information—a total of 15 bits comprising each word transmitted to the receiver. Seven bits in the address is sufficient only to give the horizontal position along the active region of a scanning line. Ambiguity in the vertical position is avoided by always sending the first active sample of each line whether it changes or not. A unique code word defines the first active element of the frame.

In order to force the average replenishment rate to match the channel capacity, the significant change threshold is varied as a function of the amount of information stored in the buffer. This may be accomplished by the control logic characteristic shown in Fig. 2. We express the absolute value of the frame-to-frame difference signal, derived by the subtractor circuit, along the ordinate with a range of 0 to 255 discrete levels. The number of replenished elements stored in the buffer is expressed along the abscissa and may range from 0 to M—the maximum capacity of the buffer. The staircase curve represents the threshold corresponding to each buffer state. The area above the curve represents a significant change in picture information where the control logic forces replenishment. The shaded area below and to the right of the curve represents an insignificant change where the control logic causes the information stored in the frame memory to be retained.

Three properties of this control function should be noted:

(*i*) As the subject becomes more active, causing an increased number of samples to be stored in the buffer, the value of the significant change threshold is increased to permit only the more significant changes in the picture to be replenished. As the subject becomes less active, causing fewer elements to be stored in the buffer, the threshold is decreased in value permitting the less significant changes to be corrected.

(*ii*) It is desirable to keep some minimum amount of data in the buffer at all times so that data is always available for transmission,
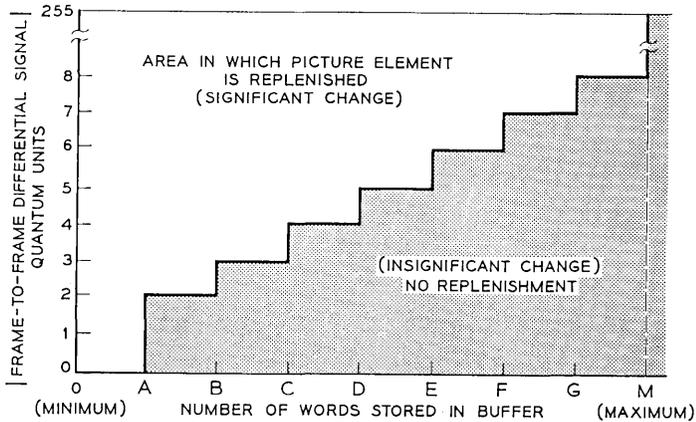
Fig. 2 — Control logic characteristic.

especially in readiness for the vertical blanking interval when data leaves the buffer but none enters. To ensure the buffer does not empty, the significant change threshold is lowered to zero whenever the buffer count falls below a chosen amount.

(*iii*) Whenever the number of samples stored in the buffer is equivalent to the capacity of the buffer, all replenishment is stopped—independent of the frame-to-frame difference. This causes picture breakup as shown in Fig. 5b.

IV. RECEIVER

Figure 3 shows a buffer placed at the receiver to store the received picture information until it can be strobed, in the proper time sequence, into the receiver's frame memory. A transfer of new information from the buffer to the frame memory occurs whenever the output of the picture-format address generator agrees with the address information of the picture element to be read from the buffer. This agreement is determined by the address comparison circuit which operates the selector switch to enable the new amplitude information to flow from the buffer into the frame memory. The buffer readout then advances to the next element. When the addresses do not coincide, the information stored in the frame memory recirculates and the readout cell of the buffer is held fixed. The information stored in the frame memory, when decoded, provides the video information for visual display.
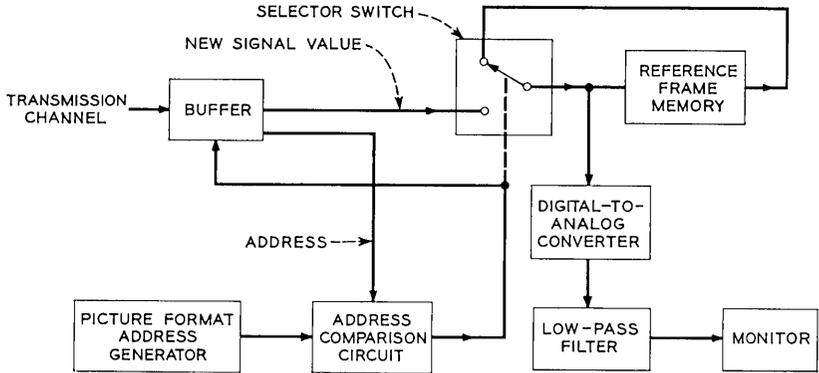
Fig. 3 — Conditional replenishment receiver terminal.

V. EXPERIMENTAL SYSTEM

In order to evaluate this method of encoding video information in real time, only the equipment for the transmitter terminal was assembled as shown in Fig. 4. The information stored in the transmitter's reference frame memory is decoded to recover the video information for visual display. The functional blocks are the same as described for the transmitter except for the buffer which is replaced
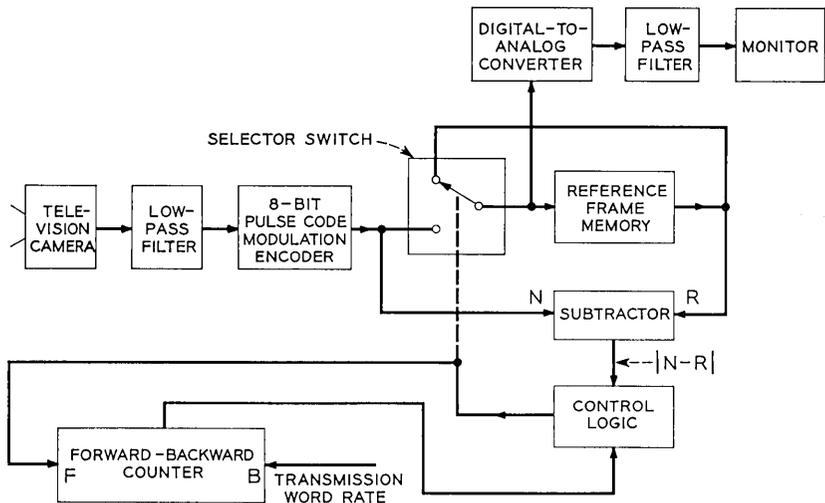
Fig. 4 — Conditional replenishment test terminal.

by a forward-backward counter in order to obtain a count of the data that would have been stored in a buffer had one been used. The count is increased by "one" whenever a picture element is replenished and decreased by "one" each time a word is transmitted. The state of the counter provides feedback to the control logic representing a measure of buffer fullness. In this way, we ignore transmission error and we have eliminated the actual buffer for experimental purposes.

In the initial experiment, the simulated buffer was assumed to have sufficient capacity to hold data relating to as many elements as there are in two complete frames, a total of 51,240 words, each word comprised of 15 bits. In practice, this would result in an inherent one-way signal delay of one-half second. The backward count rate of the counter was set to be one-eighth of the transmission rate required to send the picture directly as eight-bit PCM. The head-and-shoulder view of a person, as might be used in a visual communication system, was used as the picture source. The video information generated by the camera was band limited to 0.75 MHz. The picture format was composed of 140 picture elements per line with 183 lines per frame sequentially scanned at a rate of 60 frames per second. The active region of the picture format was composed of 120 picture elements by 171 lines. The highlight luminance of the picture was 70-80 fL (24-27.4 cd/m^2) and the ambient illumination was 125 fc (1350 lm/m^2).

VI. RESULTS

Photographs of a single frame of video information are used to illustrate the effects of conditional replenishment. These photographs are not very effective in portraying picture quality subjectively since impairments are produced only in the presence of motion.

The following results have been obtained for the experimental system outlined above:

(*i*) When motion is moderate, the picture quality is nearly the same as for eight-bit PCM coding as shown in Fig. 5a.

(*ii*) As the motion of the subject becomes more rapid, the number of picture elements stored in the buffer increases. This causes the significant change threshold to be increased so that small changes in the picture are not reproduced. The reproduction becomes somewhat poorer since all picture elements are not represented with the same
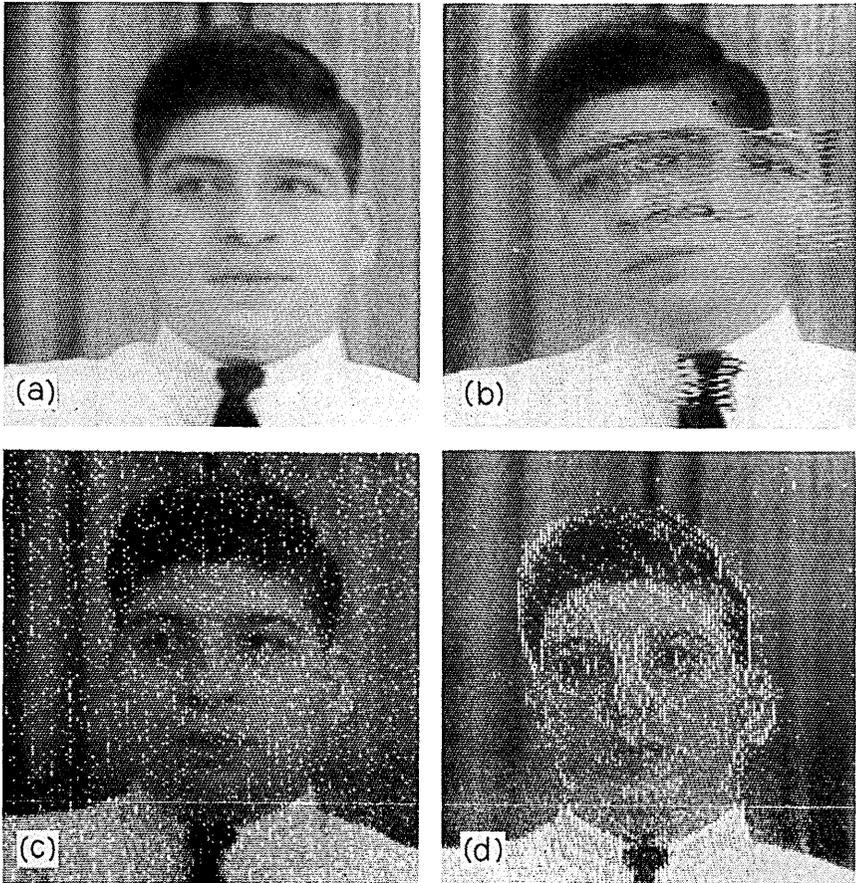
Fig. 5 — Single frames of video information processed in real time by the conditional replenishment system.

accuracy and the result resembles a scene viewed through a dirty window.

(*iii*) When the subject becomes very active, that is, when the picture contains sustained rapid motion covering a large part of the field of view, the buffer becomes saturated, allowing no more changes to be accepted. This condition is referred to as buffer overload and causes picture breakup as shown in Fig. 5b. Picture breakup is only momentary and a quick recovery takes place as soon as the subject slows down. This might be a serious defect except that there is still much

that can be done with buffer and threshold strategy to reduce this effect.

Figures 5c and d show pictures depicting the output of the system with markers or flags purposely superimposed upon the picture whenever a picture element is replenished (they are not a defect in the system). When the subject is not moving, as shown in Fig. 5c, the points are replenished more or less at random since the peaks of noise exceed the low threshold. As soon as the subject moves, as shown in Fig. 5d, the changing elements of the picture take precedence and one can see that replenishment concentrates on the subject as one might expect. The background noise causes very few picture elements to be replenished.

By viewing the accumulation of markers representing replenished points, we observe that the picture is replenished very quickly around the moving subject and that it takes a much longer time to randomly replenish the other elements. Left to chance, there are a few parts that are not replenished for a long time. We demonstrated that it is more efficient to gradually update all picture elements according to a predetermined pattern, rather than to lower the threshold to permit noise to cause replenishment.

VII. SUMMARY

We have presented a method of encoding television signals taking advantage of frame-to-frame redundancy. Only the address and amplitude of elements that have changed significantly between successive frames are transmitted. Varying the significant change threshold value helps to match the average rate of replenishment to the capacity of the transmission channel. A buffer then smooths the data flow for transmission.

Conditional replenishment lends itself to many ways of efficiently encoding pictures for transmission. We think that the buffer capacity and transmission requirement can be considerably reduced over that demonstrated here.

VIII. ACKNOWLEDGMENTS

I would also like to express my very sincere thanks to J. C. Candy for his help and advice, to J. B. Pestrichelli for assistance in setting up and testing the experimental equipment, to R. C. Brainard and E. S. Bednar for designing the frame memory, to R. L. Eilenberger for providing the television camera equipment, and to Sal Farina for his assistance.

REFERENCES

1. Kell, R. D., British Patent No. 341811, 1929.
2. Seyler, A. J., "Frame Run Coding of Television Signals: A New Method for Bandwidth Reduction." Research Laboratory Report No. 5064, Commonwealth of Australia, Postmaster General's Department, September 1959.
3. Damen, T. C., unpublished work.
4. Seyler, A. J., "Probability Distributions of Television Frame Differences," Proc. IREE (Australia), 26 (November 1965), p. 355.
5. Seyler, A. J., "The Coding of Visual Signals to Reduce Channel-Capacity Requirements," The Institution of Electrical Engineers Monograph No. 535E, July 1962.
6. Kretzmer, E. R., "Statistics of Television Signals," B.S.T.J., 31, No. 4 (July 1952), pp. 751–763.

Design of Dither Waveforms for Quantized Visual Signals

By J. O. LIMB

Dither signals may be added to coarsely quantized picture signals to mask undesirable contours. We show that a class of differential quantizers is equivalent to ordinary quantizers with respect to the design of dither signals. We give a design method for a number of deterministic and random dither waveforms and evaluate their visibility using a simple model of threshold vision.

I. INTRODUCTION

Television signals are invariably generated in an analog form. To obtain the advantages of digital transmission, it is necessary to quantize the signal in some way. In ordinary quantization the output levels of the quantizer are uniformly spaced throughout the range of the input signal; in the absence of any coding it would require six bits to send a signal quantized to 64 levels. In practice, at least 64 levels are required to produce a high quality picture.

A strong incentive to reduce the number of levels is that it would reduce the number of bits transmitted. For example, if the quantizer step size is doubled, the number of levels can be halved and the bit rate of the source can be reduced from six to five bits per sample. If this is done, the picture quality is degraded, but primarily for only one type of picture material, those areas in which the luminance changes slowly. These areas will be referred to as low-detail areas. The degradation takes the form of curved lines which look very much like contour lines on a map; thus this type of degradation is referred to as contouring.* The problem, then, is to eliminate the objectionable effect of contouring, which occurs only in the low-detail part of the picture, without using a larger number of levels.

---

* For example, see Fig. 3b of Ref. 1 for a differentially quantized picture showing contouring.

An effect similar to increasing the number of levels can be achieved by adding a dither signal to the true input signal. The dither signal produces rapid switching between the quantizer levels on either side of the true input signal. This switching is arranged so that the time one spends at a level depends on how close the true input signal is to that level. Thus in Fig. 1a when the signal lies midway between two levels, it oscillates between the two levels, spending equal time at each. In Fig. 1b the input lies a quarter of the distance up from the lower level; consequently, the required switched waveform should be down for three samples and up for one. The output waveform obtained when a dither signal is added to the input will be referred to as the chopped waveform or chopping pattern.

One could ask why such a strategy should be any good. While it is true that on the average the output signal will have the same amplitude as the input, it now has an additional error component which could degrade the signal further. Thus it is necessary to compare the visibility of the chopped waveform with the visibility of the contours that would otherwise be seen. Visibility is used here in the subjective sense of how easy is it to see an object. An objective visibility scale can be constructed using a fairly well defined subjective point on the visibility scale, that is, threshold, the point at which an object just becomes (or just ceases to become) visible. If the objective measure of the amplitude of a stimulus at threshold level is large, the stimulus has low visibility; conversely, the smaller the amplitude of the stimulus, the greater the visibility.

For signals near threshold, the visual system acts like a low-pass filter so that the chopped waveforms with the highest frequency components will be attenuated most and hence will be the least visible. Thus the pattern of Fig. 1a will be less visible since its repetition frequency is twice that of the pattern of Fig. 1b. In choosing suitable chopped waveforms we attempt to select those signals which have the least visibility.

The chopping patterns can be random or deterministic. Figure 1c shows a typical sample of a random pattern for an input halfway between the two quantizer levels (the same input amplitude as in Fig. 1a). The probability of each sample being high or low is one-half and is independent of previous samples. Since there is a finite chance that a given segment of the random sequence contains frequency components lower than those of the waveform of Fig. 1a, the random sequence of Fig. 1c is more visible than the deterministic pattern of Fig. 1a.

Let us now consider the problem of designing a dither signal which will produce an optimum chopping pattern at various levels. Goodall first observed that by adding a small amount of noise to the input, contouring was almost eliminated at the expense of a small increase in the granularity (or noisiness) in the picture.[2] Roberts examined the problem quantitatively and showed that in order to produce a random chopping pattern, which always averaged out to the same amplitude as the input, the probability density function (PDF) of the noise should be rectangular with a maximum amplitude of plus and minus half a quantizing interval.[3] He further showed that if one subtracted at the quantizer output the same noise that was added at the input, the root mean square error in the output signal is halved (if one forgets the correction term for the quantizing intervals at the end of the range). Limb considered the visibility of the granulation in the quantizer output.[4] Using a simple model of the visual process, it was shown that the visibility of granulation resulting from independent random samples with a rectangular probability density function of the correct amplitude is zero when the input equals a quantizer output level, and reaches a maximum when the input is midway between levels. Further, by introducing negative correlation between samples, the visibility can be reduced by about 50 percent.
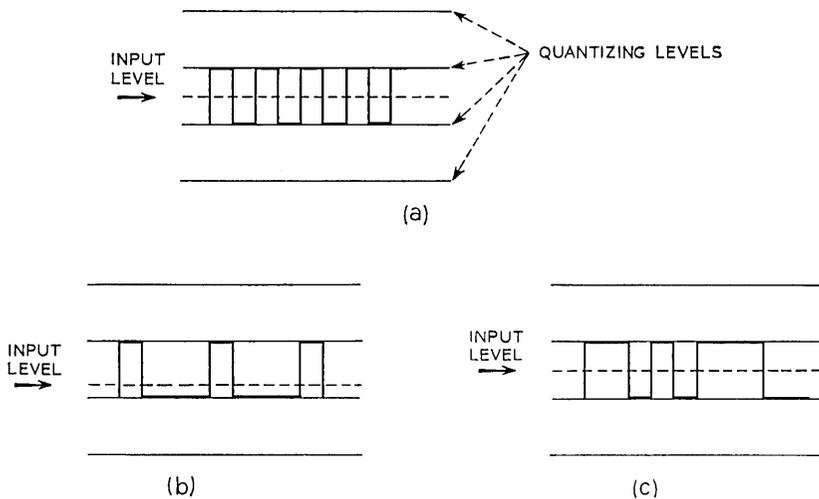
Fig. 1 — Chopping patterns for (a) input half way between levels, (b) input quarter way between levels, and (c) random pattern with input half way between levels.

In this paper we look at the problem of applying dither to differential quantization as opposed to ordinary quantization. The approach is the same as with ordinary quantization; design a dither waveform which, when added to the input, produces the required chopping pattern at the output (see Fig. 2a). All the components of the differential quantizer are assumed to be ideal. The chopped waveforms produced by the differential quantizer will depend on how the levels of the quantizer within the loop are positioned close to the zero level. Two commonly used configurations are (*i*) a decision, or input, level placed at zero (Fig. 2b), and (*ii*) a representative, or output, level placed at zero (Fig. 2c). We consider only the second configuration (however, see Section VII). We show that under fairly general conditions a differential quantizer, containing a quantizer stage with a representative level at zero, behaves the same as an ordinary quantizer (with equal level spacing) with respect to dither. We design a set of dither signals, both random and deterministic, which produces chopping patterns with a low visibility. The visibility of the chopping patterns is calcu-
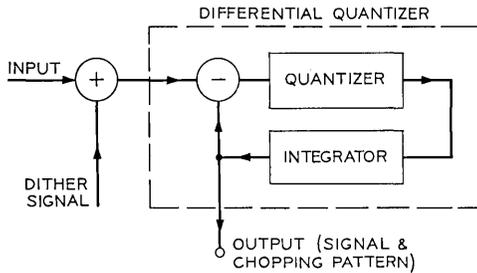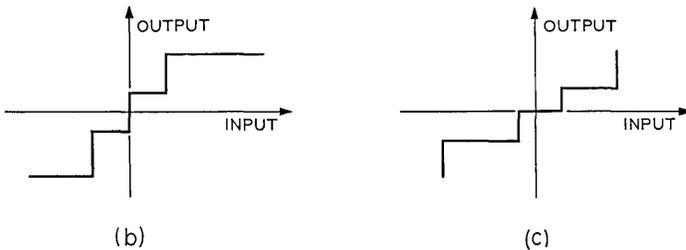
Fig. 2 — (a) Dither applied to differential quantization. (b) Transfer characteristic of quantizer with decision level at zero. (c) Transfer characteristic of quantizer with representative level at zero.

lated, enabling a comparison to be made between random and deterministic dither. It is anticipated that two-dimensional dither signals will reduce the visibility of contours by a factor of six.

If one decides to subtract the dither signal from the averaging pattern as Roberts did, the rules for generating the best dither waveforms have to be rederived. When this is done, it is found that subtracting dither signals is barely superior to not subtracting them.

II. DIFFERENTIAL QUANTIZER—QUANTIZER CHARACTERISTIC

A quantizer may be divided into two sections, the classifier which divides the signal into a number of ranges according to the position of its decision levels, D_i, and the weighter which assigns a value to each range according to the settings of the representative levels R_i. Figure 3a shows the quantizer characteristic as it is generally drawn. An alternative representation is given in Fig. 3b, where the vertical dashes represent the decision levels, and the crosses represent the representative levels. This representation is more convenient since we are concerned with the positions of the representative levels relative to the positions of the decision levels.

The input level to the classifier, in the absence of the dither signal, is denoted by Δ and expressed as a fraction of r, the distance from R_0 to R_1 (Fig. 4). Since we are considering slowly changing input signals, Δ will always lie in the range R_{-1} to R_1.

We assume that the quantizer has a representative level at zero as Fig. 4 shows. The only levels that affect the design of the dither signal are the two decision levels, D_{-1} and D_1, lying closest to zero and the adjacent representative levels R_{-1} and R_1. Furthermore, we assume that R_{-1}, D_{-1}, R_0, D_1, and R_1 are equally spaced. This is probably the most useful configuration since it satisfies Max's first condition for minimizing error, that is, the decision levels should lie midway between the corresponding representative levels.[5] In addition, $R_1 = 2D_1$ which is on the stability boundary and hence corresponds to the maximum setting of R_1 if limit cycles are not to occur.[6]

III. DESIGN OF DITHER SIGNAL

When rectangular random noise is used as a dither signal, the chopped waveform has a granular appearance and the visibility of this granulation depends on the amount of correlation in the waveform. For example, when the correlation is positive, the waveform is
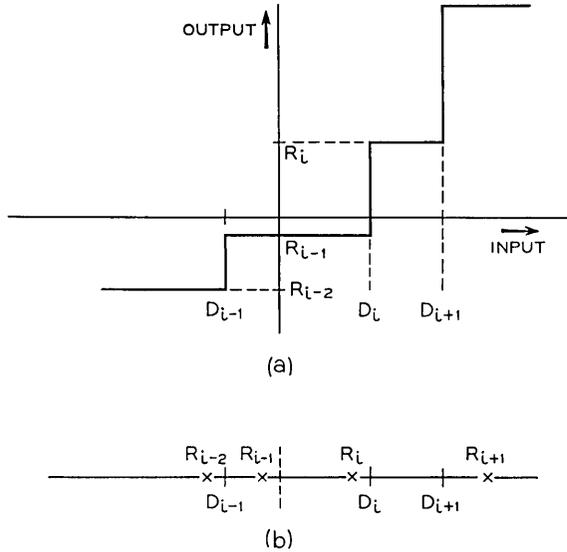
(a)

(b)

Fig. 3 — Quantizer characteristic (a) usual representation and (b) alternative representation.

more likely to contain large runs of 0's and 1's (if we use 0 and 1 to denote the two quantizer levels between which the output is chopping); if the waveform is negatively correlated, a 1 is more likely to follow a 0 and the waveform will switch back and forth more rapidly. Notice that the visibility of a perturbation is approximately proportional to the area when the area is small. Consequently, long runs of 1's or 0's are much more visible than the negatively correlated waveform containing a higher probability of short runs.

If we restrict the chopped waveform to be described completely by a second order probability density function, there are limits on the
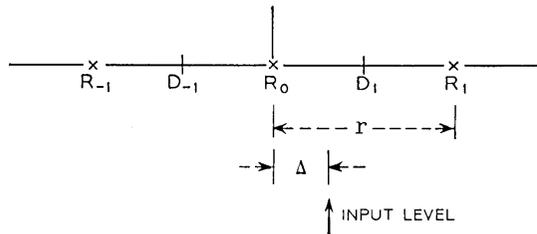
Fig. 4 — Quantizer characteristic—configuration with representative level at zero.

amount of negative correlation that can be achieved.[4] This restricted chopped waveform with maximum negative correlation can be generated with a dither signal having the second order probability density function shown in Fig. 5. Here x_1 and x_2 represent adjacent samples of the dither signal. The negative correlation produces a sharp minimum in the visibility of granulation in the waveform when the input to the classifier is close to D_1 (or D_{-1}), that is, when $\Delta = 0.5$. The dashed curve is for uncorrelated noise and is shown for comparison.

The dither noise can also be represented as shown in Fig. 6a, which better illustrates the time series nature of the process. For example, a sample occurring at random in the top half of the amplitude range will, for the next sample, occur in the lower half. The nature of the second order probability density function ensures that the random sample oscillates between the upper and lower half of the range. This type of dither will be referred to as two-step random dither.

Dither waveforms can be generated for any number of steps, although with an increase in the number of steps, the visibility of granulation will reach a minimum and then start to increase. Figure 6b shows an example of four-step dither. Notice that when the input level lies on the boundary between two steps, deterministic chopping patterns are produced. Furthermore, these patterns should have the least visibility of any chopping pattern with the same average level. In general, a low visibility pattern (LVP) has the minimum allowable cycle length (for example, cycle length of 3 at $\Delta = \frac{1}{3}$) and has the
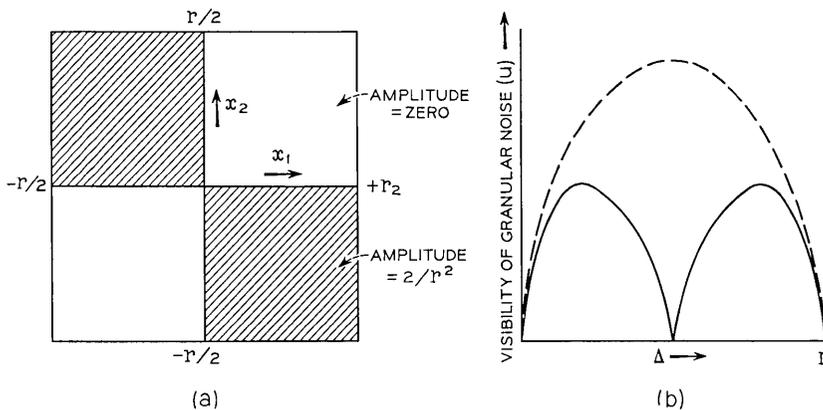
Fig. 5 — Two-step random dither signal: (a) probability density function of correlated noise and (b) visibility of noise.
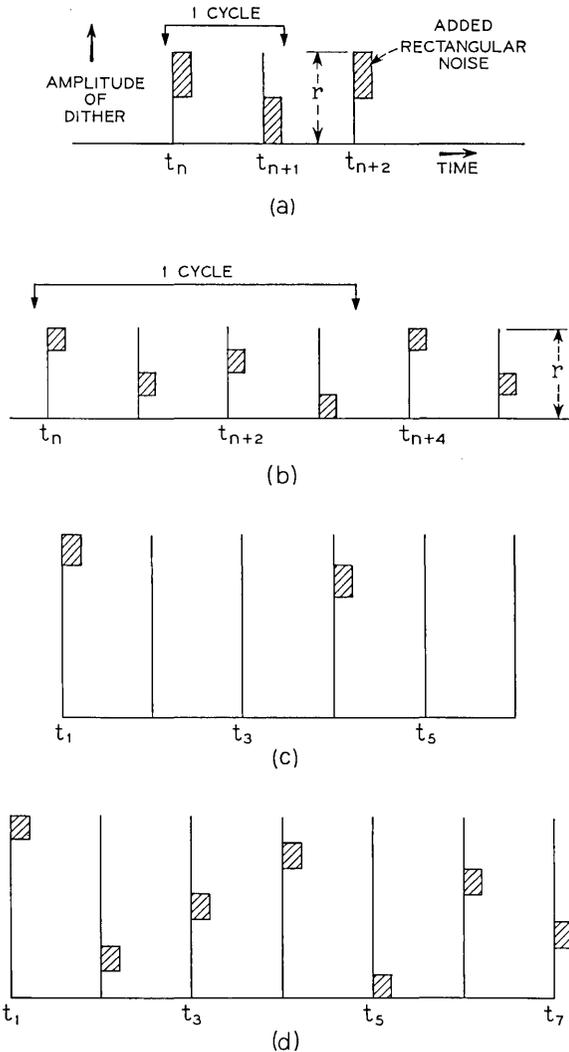
Fig. 6 — Dither signal represented in time series form: (a) two-step dither, (b) four-step dither, (c) attempt to construct six-step dither, and (d) seven-step dither.

shortest maximum run of either value (for example, 1010100 is a low visibility pattern for $\Delta = 3/7$ but 1100100 is not).

Can an n-step low-visibility chopping pattern be generated with second order noise for any value of n? The answer is no, as the following attempt to reconstruct patterns for six-step and seven-step

dither shows. In Fig. 6c the first step has been assigned arbitrarily to t_1, while the second step must then be assigned to t_4 so that when the input gives $\Delta = \frac{1}{3}$, every third sample exceeds threshold ($LVP = 1, 0, 0$). There is no sample to which the third step can be assigned to give a low visibility pattern of $(1, 0)$ as required for $\Delta = \frac{1}{2}$.

For the seven-step case (Fig. 6d), the first step is assigned to t_1, the second step may be assigned to either t_4 or t_5, both giving low visibility patterns (assume t_4). The third step, if assigned to t_6, will again give a low visibility pattern $(1, 0, 0, 1, 0, 1, 0)$. Similarly, all the other steps can be assigned to give low visibility patterns.

In the general case of n-step dither, tests for low visibility patterns can be made as follows:

Assign first step $\quad - t_1$
$\qquad$ second step $- t_{n/2} \qquad n$ even
and $\quad$ third step $\quad - t_{n/4} \qquad n$ divisible by 4.

To have the least visibility, the resulting pattern after assigning the third step must not have runs of 0's differing in length by more than one, otherwise the position of a 1 could be moved to shorten the longest run. However, this would affect steps 1 and 2 which have given low visibility patterns. Thus any multiple of 4 equal to or greater than 8 will not give low visibility patterns. Again:

Assign third step $-t_{(n\pm2)/4} \qquad n$ even, not divisible by four.

By the same argument as above for $n \geq 6$ and even, low visibility patterns are not obtained. Similarly, the odd numbers can be tested. In all, low visibility patterns can be obtained for $n = 2, 3, 4, 5$, and 7.

In the scheme considered so far, each step in the quantizing interval has been filled with rectangular noise of amplitude equal to the step height. Random patterns are produced whenever Δ lies within a step, while a deterministic pattern is generated when Δ lies exactly on the boundary between two steps. Consider changing the rectangular noise to a fixed level at the midpoint of the step. The chopping pattern will now switch from one deterministic pattern to another as Δ changes. We examine the visibility of granulation associated with both random and deterministic dither in Section IV.

Implementation of either random or deterministic dither schemes would be a simple matter requiring little additional hardware. Fig. 7 shows the output from a computer simulation of a differential quantizer with four-step deterministic dither in Fig. 7a and seven-step deterministic dither in Fig. 7b. The visibility of granularity in these two dither schemes will differ; in Section IV, calculations of visibility are made to enable the most promising schemes to be selected.
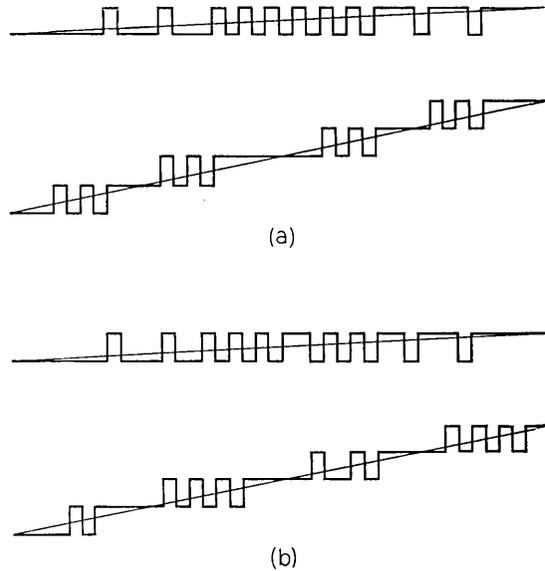
(a)

(b)

Fig. 7 — Chopping patterns resulting from deterministic dither signal for (a) four-step dither and (b) seven-step dither. The straight lines denote the inputs.

IV. VISIBILITY OF THE CHOPPED WAVEFORM

The measure of the visibility of the discrete waveform is based upon a simple model of threshold vision that has proved reasonably accurate.[4,7] Briefly, threshold vision is assumed to act like a spatial low-pass filter, and the difference in visibility between two displays (in this case the display resulting from the analog signal and the display resulting from the quantized signal) is measured by the difference between the filtered version of the two signals.* Two different measures of the difference are considered: one is the mean square, and the other is the mean modulus. The measure of visibility is denoted by $U(\Delta)$, which depends on Δ since the value of Δ determines the shape of the chopped waveform.

4.1 Deterministic Patterns

The solid-line curves in Figs. 8 and 9 show the calculated visibility of granulation for three- and four-step patterns at a viewing distance of 36 inches. The visibility is shown for only half the range of Δ,

---

* Appendix B gives more detail.
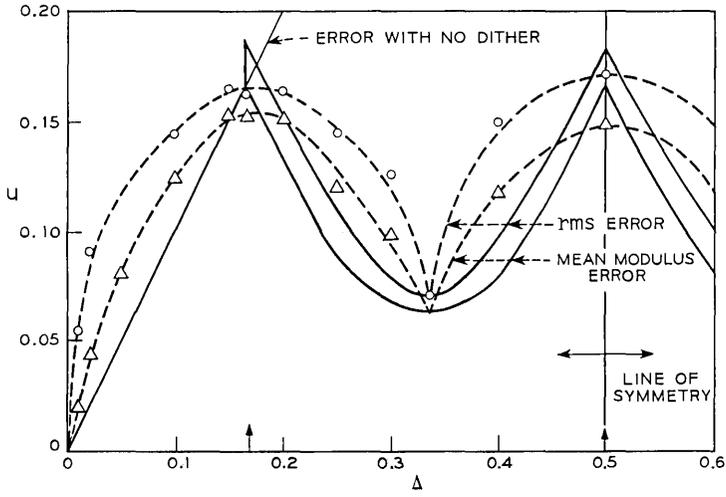
Fig. 8 — Visibility of granularity produced by three-step dither at 36 inches viewing distance (— — random; ——— deterministic).
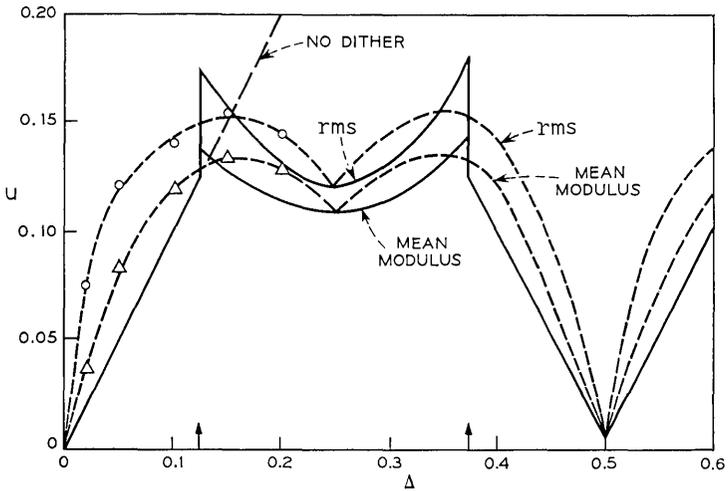
Fig. 9 — Visibility of granularity produced by four-step dither at 36 inches viewing distance (— — random; ——— deterministic).

since $U(\Delta)$ is symmetrical about $\Delta = 0.5$. The arrows on the abscissa indicate the amplitude levels of the dither pattern. Thus until the value of Δ exceeds 0.167 in Fig. 8, no switching occurs in the output. There are minima at $\Delta = \frac{1}{3}$ and $\frac{2}{3}$ as expected for three-step dither. The curves of $U(\Delta)$ for the two criterion functions are similar in shape, the rms curve lying slightly above the mean modulus curve. In Fig. 9 the minimum at $\Delta = 0.25$ is not very large, and one would expect granulation to be more visible for patterns having a greater number of steps. Figure 10 clearly shows this for a five-step pattern which has a higher minimum than Figs. 8 and 9. By comparing the average value of U for three-, four-, and five-step patterns, four-step is just better than three-step, and both are superior to five-step patterns.

Figure 10 also gives U for a five-step dither at a viewing distance of 72 inches. The spread of the visual impulse response is now much greater in relation to the size of a picture element. In fact, $U(\Delta)$ is not very different from what would be expected with infinite smoothing by the eye. With infinite smoothing all minima would be zero and joined to the maxima at 0.1 by straight lines; that is, five equal triangles of amplitude 0.1. The similarity of U to the result expected for infinite smoothing would suggest that a pattern with a larger number of steps would be superior. Going to the maximum of seven steps
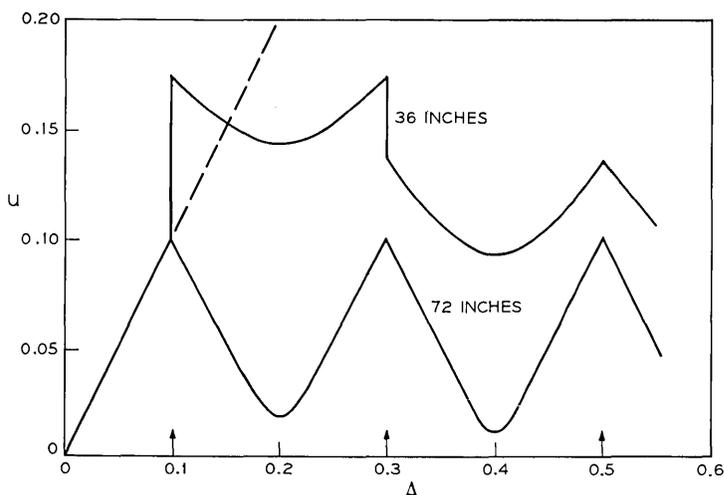
Fig. 10 — Visibility of granularity produced by a five-step dither at 36 and 72 inches viewing distance; mean square error criterion.

(Fig. 11) significantly reduces the mean value of U, and the curve is no longer similar to the curve for infinite smoothing. By comparing Fig. 11 with Fig. 9, one can see that increasing the viewing distance by a factor of two has reduced the calculated visibility of granulation by about one-half for the best pattern in each case—a not altogether surprising result.

4.2 *Random Patterns*

Random noise was added to the deterministic patterns in the manner shown in Figs. 6 and 7. Notice that the pattern generated after quantization is deterministic when the decision level lies at the junction of two steps and is the same as the pattern produced in the absence of noise. Figs. 8, 9, and 11 give the calculated visibility of random patterns for mean square and mean modulus error criteria. As required, the random curves touch the deterministic curves between steps, and in most other places the curves lie above them. Four-step dither still gives the smallest average U, $\langle U \rangle_{av}$; three-step dither is the next best.

4.3 *Deterministic versus Random Patterns*

With deterministic patterns, n-step dither results effectively in inserting n-1 levels in the original quantization interval. The brightness at these new levels is not constant, however, and has a variance about the true analog input value given by the minima, $U(\Delta_{min})$. As Δ changes from Δ_{min}, the variance remains unchanged but a constant error is introduced since the average value of the output no longer equals the average value of the true analog input.

With random patterns, the average value of the output always equals the average value of the input. Thus at the maxima of $U(\Delta)$, the variance of the perceived image with deterministic patterns is less than with random patterns, but there is an additional error resulting from differences in the perceived average values of the true analog input and the chopped waveform. By using a decision theory model of threshold vision, the visibility in the two situations could be compared. However, such models have not proved accurate enough to apply to this type of second order effect.

On the basis of the mean square and mean modulus criteria it appears that deterministic dither is slightly superior; but because each case has different distributions for the perceived brightness, such comparisons are risky and best wait experimental confirmation.
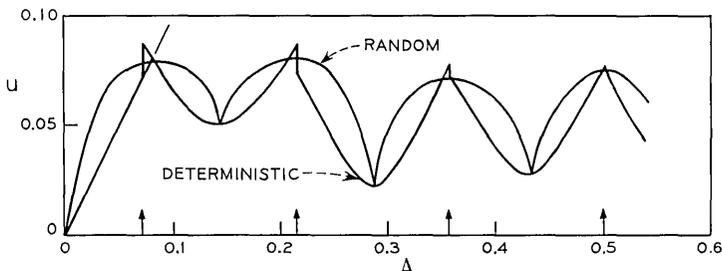
Fig. 11 — Visibility of granularity produced by seven-step dither at 72 inches viewing distance; random and deterministic, mean square error criterion.

V. DITHER APPLIED IN TWO AND THREE DIMENSIONS

Devising low visibility patterns in two dimensions is more difficult than in one dimension. In fact it appears that there are only two equivalent, trivial low visibility patterns. These patterns occur for two by two step interpolation; they are,

$$\begin{array}{c} -\ x\ \to \\ \Big\downarrow\ y \begin{array}{cc} 1 & 3 \\ 4 & 2 \end{array} \end{array} \quad \text{and} \quad \begin{array}{c} -\ x\ \to \\ \Big\downarrow\ y \begin{array}{cc} 1 & 4 \\ 3 & 2 \end{array} \end{array}.$$

For larger patterns it appears that we must settle for something less ideal. A four by four step pattern was generated by considering it to consist of four two by two patterns, which were themselves generated in the manner of a two by two pattern, as the partly completed pattern in Fig. 12a shows.

The computer program used previously for the one-dimensional case was extended to calculate the visibility of the four by four pattern. $U(\Delta)$ is shown in Fig. 13 for a viewing distance of 36 inches and a mean square error criterion. $\langle U \rangle_{av}$ has been reduced to about one-third in going from one dimension to two in this example. This pattern does not have minimum visibility. This can be seen for $\Delta = \frac{1}{4}$ where a lower visibility pattern could be obtained by the chopping pattern of Fig. 12b. This would make little difference to $\langle U \rangle_{av}$, however, since $U(\Delta)$ for $\Delta = \frac{1}{4}$ is already very small. Undoubtedly patterns approximating the ideal could be found for a larger number of steps.

In applying dither in the time dimension, care must be taken not to introduce "temporal granularity," that is, flicker. To study the visibility of flicker would require an entirely new model, accounting

for the variation in sensitivity over the retina to temporal changes in luminance. Flicker occurs when large picture areas differ in luminance periodically from frame to frame. By arranging for the average luminance of an area to change as little as possible from frame to frame, flicker can be minimized. Thus the two-dimensional pattern

$$\text{Time} \downarrow \quad
\begin{array}{cccc}
\multicolumn{4}{c}{-\ \text{Distance} \rightarrow} \\
1 & 15 & 7 & 10 \\
12 & 5 & 14 & 4 \\
8 & 9 & 2 & 16 \\
13 & 3 & 11 & 6
\end{array}
\quad
\begin{array}{c}
\text{Frame 1} \\
\text{Frame 2} \\
\text{Frame 3} \\
\text{Frame 4}
\end{array}$$

which was built up with the help of the two by two low visibility pattern, will have an average luminance which varies at most by $\frac{1}{16}$ of a quantizing interval from frame to frame. This is not true of the sequence,

$$\text{Time} \downarrow \quad
\begin{array}{cccc}
\multicolumn{4}{c}{-\ \text{Distance} \rightarrow} \\
1 & 13 & 4 & 16 \\
9 & 5 & 12 & 8 \\
3 & 15 & 2 & 14 \\
11 & 7 & 10 & 6
\end{array}
\quad
\begin{array}{c}
\text{Frame 1} \\
\text{Frame 2} \\
\text{Frame 3} \\
\text{Frame 4}
\end{array}$$

which is simply derived from the two by two pattern and nearly identical to the pattern of Fig. 12a. Notice that if the input signal has a uniform distribution over the quantizing interval, the average luminance of each frame will be the same. For example, for $0 < \Delta < 0.25$, frames 1 and 3 have greater average luminance, while for $0.5 < \Delta < 0.75$, frames 2 and 4 have greater average luminance. Since the probability of obtaining signals that do not vary (lie within one step) over "large" areas is small for high quality pictures (which
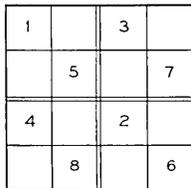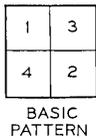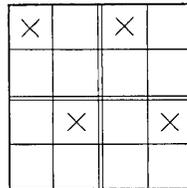
Fig. 12 — Generation of four by four step pattern: (a) partially completed pattern (b) two dimensional chopping pattern having lower visibility than the corresponding pattern resulting from (a).

consequently have a small step size), the probability of obtaining flicker should be small. There is advantage in using the second pattern since it provides better smoothing.

Dither applied in the time domain should be more successful than in one spatial dimension at a 36-inch viewing distance since the temporal impulse response, even at high ambient illumination, probably has a greater spread; the problem of flicker, however, should be kept in mind. Another advantage of the temporal dimension is that the amount of smoothing should be independent of the viewing distance.

In comparing deterministic patterns with random patterns, temporal smoothing has often been neglected; this leads to incorrect conclusions. For example, if deterministic two-dimensional spatial dither is compared with random dither, the random pattern would provide smoothing in three dimensions since added noise components in adjacent frames are uncorrelated, and, as just shown, the improvement in smoothing provided by an additional dimension is large. A valid comparison could be made by using "frozen" noise, that is, noise that repeats from frame to frame.

Section 3.43 of Ref. 1 describes results that were obtained when two types of dither waveform were added at the input of a differential quantizer. The first pattern was a one dimensional four step waveform added vertically. The second pattern was a four by four step pattern added horizontally and vertically. Figure 3d of Ref. 1 shows the effect of adding the two dimensional dither to a picture while Figs. 6c and 6d of Ref. 1 show the effect of adding one and two dimensional dither respectively, to a low amplitude ramp waveform.

VI. RECEIVER SUBTRACTION

Roberts added pseudorandom noise having a rectangular probability density function to the signal prior to quantization, and subtracted the same noise from the signal at the receiver.[3] Neglecting end effects from the smallest and largest quantization levels, a reduction of one-half in the variance of the output signal is obtained. Roberts states that adding noise to the input and subtracting it from the output is equivalent to adding a level of noise to the signal, but that this is not the same noise as was added to the input. Since we are concerned with the exact sequence in the output signal (this will critically affect the visibility of the added noise), the relation between the added input noise and the equivalent output noise will be derived.

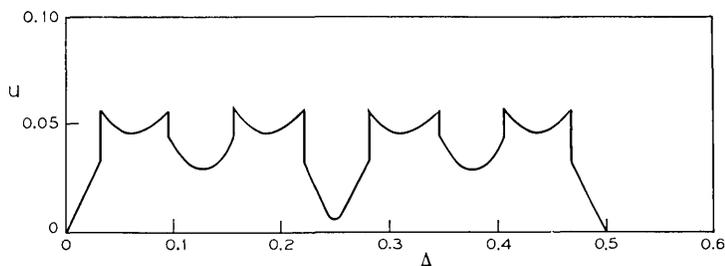In Fig. 14, rectangular noise is added to the input signal of value

Fig. 13 — Visibility of granularity produced by two dimensional four by four step pattern. Viewing distance is 36 inches with mean square error criterion.

$R_n + \Delta$. All noise components which cause the input signal to exceed D_n are represented by R_{n+1}, and all components producing a combined signal less than D_n are represented by R_n. Thus, in subtracting the input noise from the quantized signal, components lying between $r/2 - \Delta$ and $r/2$ are subtracted from R_{n+1}, while the other components are subtracted from R_n. When the noise is subtracted one sees that the whole process is equivalent to adding noise of the same amplitude to the unquantized signal. The noise to be added can be obtained from the input noise by inverting separately amplitudes greater and less than $r/2 - \Delta$ as Fig. 14 shows. For example, amplitudes above D_n [such as $(r/2 - \Delta) + \Gamma$] go to $r/2 - \Gamma$ where Γ is any increment between 0 and Δ. This relationship is very useful since now we can forget the quantization and consider just the distortion of the added noise component.

If an n-step dither sequence is quantized and the original sequence subtracted, inversion occurs at every step except where Δ is less than $r/2n$. Consequently, $n - 1$ new sequences will be produced and only in special cases will the new sequences be the same as the original. A technique will be developed for rapidly estimating the new output sequences from the input sequence.

A sequence can be written as a function of time,

$$\text{Time} \qquad 1, \quad 2, \quad 3, 4, \; \cdots \, , n$$
$$\text{Amplitude} \quad A_1 \, , A_2 \, , \qquad \cdots \, , A_n$$

where A_i is an integer between 1 and n denoting the step. A sequence can also be written

$$\text{Amplitude} \quad 1, \quad 2, \quad 3, \quad \cdots \, , n$$
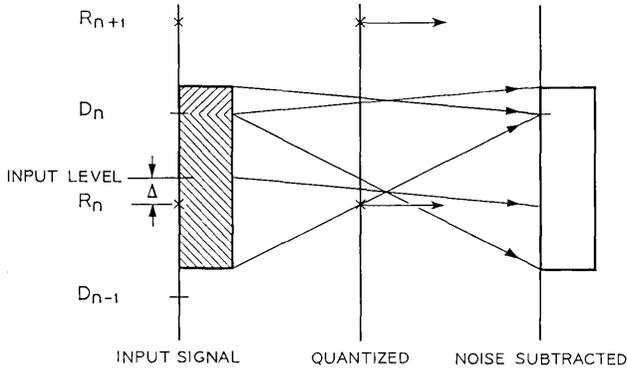$$\text{Time} \qquad T_1 \, , T_2 \, , T_3 \, , \; \cdots \, , T_n$$

Fig. 14 — Derivation of the properties of add-subtract noise patterns.

where T_i is an integer between 1 and n denoting the time slot in which the ith amplitude occurs. Conversion from the time representation to the amplitude representation can be simply accomplished. For example, at amplitude A_1 the corresponding time slot is 1; that is, T corresponding to A_1 is 1.

Consider the sequence $T_1 \ldots T_i, T_j \ldots T_n$. For Δ lying between the ith and jth steps, the new sequence is $T_i \ldots T_1, T_n \ldots T_j$. Notice that the cyclic order is reversed but otherwise unchanged. Thus if Δ changes by i steps, the amplitude sequence shifts by i steps but the order is unchanged unless Δ is less than $r/2n$, in which case the order reverses. However, since the visibility of a sequence does not change if the order is reversed, this may be neglected. A cyclic shift in the amplitude sequence must now be converted to the time sequence, since we use the time sequence to calculate visibility. A shift by one step in the amplitude representation corresponds to an addition or subtraction by one, modulo n (depending on the direction of the shift), in the time representation. Thus if the time sequence was $1, 2, \ldots, n$ (a bad sequence from the point of view of visibility), the sequence at the ith level would be $n - i + 1, \ldots 1, 2, \ldots, n - i$, which is in fact the same sequence. This particular case is one of a set of sequences that remain unchanged as Δ changes from step to step.

6.1 *Visibility of Sequences*

There are at most $(n - 1)!/2$ different sequences that can be generated for a particular value of n where sequences are regarded as different if they have different visibilities; that is, they are not shifted

in time or reversed versions of another sequence. There is only one unique sequence for $n = 3$ and three unique sequences for $n = 4$. For $n = 4$ the three possible sequences are: (i) 1, 2, 3, 4; (ii) 1, 3, 2, 4; (iii) 1, 2, 4, 3. Sequence i produces three output sequences which are the same as itself. Sequence ii produces the output sequences

$$
\begin{array}{lll}
\text{Input} & 1, 3, 2, 4 = 1, 3, 2, 4 & \text{Output} \quad 0 \\
& 4, 2, 1, 3 = 1, 3, 4, 2 & \qquad\quad\ 1 \\
& 3, 1, 4, 2 = 1, 3, 2, 4 & \qquad\quad\ 2 \\
& 2, 4, 3, 1 = 1, 3, 4, 2 & \qquad\quad\ 3 \\
& 1, 3, 2, 4
\end{array}
$$

and iii produces the output sequences

$$
\begin{array}{lll}
\text{Input} & 1, 2, 4, 3 = 1, 3, 4, 2 & \text{Output} \quad 0 \\
& 4, 1, 3, 2 = 1, 3, 2, 4 & \qquad\quad\ 1 \\
& 3, 4, 2, 1 = 1, 3, 4, 2 & \qquad\quad\ 2 \\
& 2, 3, 1, 4 = 1, 3, 2, 4 & \qquad\quad\ 3 \\
& 1, 2, 4, 3
\end{array}
$$

These outputs are just shifted versions of one another and they should yield the same overall value of U. The output sequence 1, 3, 2, 4 provides better smoothing than 1, 3, 4, 2 which contains lower and hence more visible frequency components. This can be seen in the curves of U for the two sequences which were calculated independently of the arguments of this section (Fig. 15).

For comparison, $U(\Delta)$ is shown for the case previously considered in which the dither waveform is not subtracted from the output. The subtraction method gives a slightly lower average value of U (< 2 percent lower). The value of U for uncorrelated rectangular noise with subtraction is also shown. U is now independent of Δ. However, the problems associated with comparing random and deterministic dither schemes should be borne in mind (see Section 4.3).

6.2 Constant Sequences

Here is a technique for finding sequences which do not change as Δ changes from step to step (constant sequences). A step change in Δ results in an increment, modulo n, of each number in the sequence; thus, the numbers must be arranged so that the order remains unchanged after a shift. Constant sequences can be constructed simply by using a geometric method. In Fig. 16, five points are spaced equally around a circle, each point corresponding to a number in the sequence. Starting from any point, a line is drawn to another point to cor-
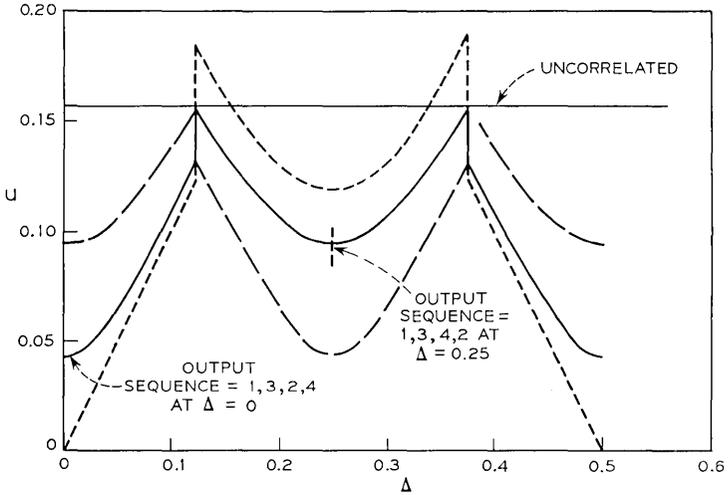
Fig. 15 — Visibility of granularity produced by four-step add-subtract dither pattern. Viewing distance is 36 inches (———— input sequence 1, 3, 2, 4; ———— ———— input sequence 1, 2, 4, 3; - - - - addition only 1, 3, 2, 4).

respond to a shift of the number of points cut off by the line: 1 in Fig. 16a and 2 in Fig. 16b. This second point is then shifted the same distance in the same direction. The shifting process is repeated until all points are covered, and we arrive back at the starting point if the number of points shifted is not a divisor of n (excluding 1). The number of unique constant sequences for an n-step pattern is equal to the number of nondivisor integers less than $n/2$ plus 1. Thus Figs. 16a and b represent the two constant sequences for $n = 5$. Unfortunately, for larger n, constant sequences do not have low visibility, as the five constant sequences,

$$1, \quad 2, 3, \quad 4, 5, \quad 6, \quad 7, \quad 8, \quad 9, 10, 11$$
$$1, \quad 7, 2, \quad 8, 3, \quad 9, \quad 4, 10, \quad 5, 11, \quad 6$$
$$1, \quad 5, 9, \quad 2, 6, 10, \quad 3, \quad 7, 11, \quad 4, \quad 8$$
$$1, \quad 4, 7, 10, 2, \quad 5, \quad 8, 11, \quad 3, \quad 6, \quad 9$$
$$1, 10, 8, \quad 6, 4, \quad 2, 11, \quad 9, \quad 7, \quad 5, \quad 3$$

show for $n = 11$. Probably the best sequence is the third, but this is significantly inferior to a sequence such as

$$1, 11, 2, 10, 3, \quad 9, \quad 4, \quad 8, \quad 5, \quad 6, \quad 7.$$

Figure 17 is a graph of $U(\Delta)$ for $n = 5$ for the constant sequence 1, 4, 2, 5, 3 and sequence 1, 2, 5, 4, 3. The constant sequence has an
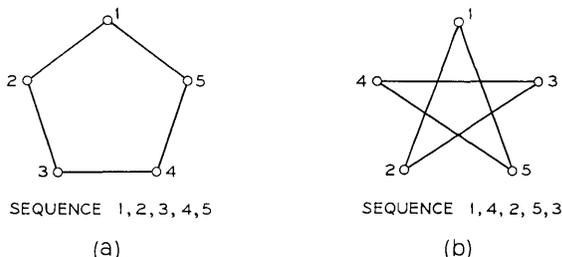
Fig. 16 — Generation of constant sequences.

average value $\langle U \rangle_{av}$ of 0.095, which is lower than the other sequence and the low visibility pattern derived in Section 4.1 (which is also shown for comparison). The subtracted sequences for $n = 4$ (Fig. 15) give a slightly greater value of $\langle U \rangle_{av}$ (0.099) compared with $n = 5$. Notice the very low minimum at $\Delta = 0.2$ for the nonconstant sequence. The sequence producing this minimum may be calculated by subtracting one from each digit of the input sequence and is thus 1, 4, 3, 2, 5.

VII. DISCUSSION

The quantizer configuration with a decision level at zero was referred to briefly in the introduction. This configuration results in an
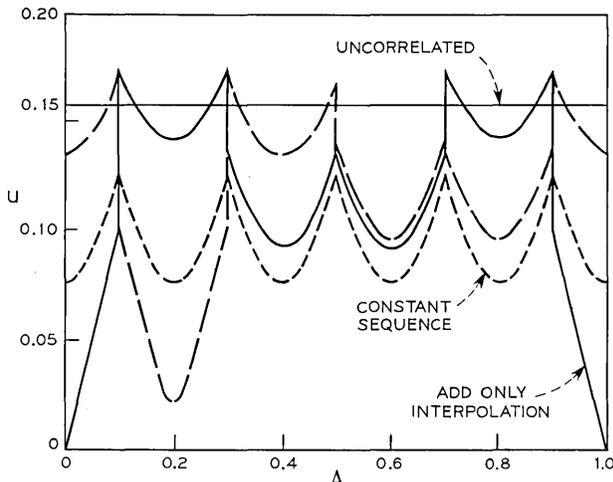
Fig. 17 — Visibility of granularity produced by five-step add-subtract dither pattern. Viewing distance is 36 inches.

even number of representative levels and has received more attention in the literature. In the absence of a dither signal the output will oscillate between R_1 and R_{-1}, but otherwise has no inherent dithering ability of its own. It will produce contours in low detail areas with much the same visibility as the quantizer configuration we have investigated. For an uncorrelated random dither signal, the switching waveform is not constrained to lie between the two adjacent quantizer levels as it is with a representative level at zero. For illustrative purposes, two switching waveforms have been generated for two different input levels assuming a random, uncorrelated dither signal (Fig. 18). Although it would be more complex to do so, one could calculate the visibility of these types of waveforms as done previously and compare the results with those just obtained. One problem is to decide upon the relative amplitudes of R_1 and R_{-1} for the two configurations.

VIII. SUMMARY AND CONCLUSIONS

The design of dither signals for ordinary quantizers is the same as the design for differential quantizers for quantizer characteristics of specific types. The requirements for equivalence are that the char-
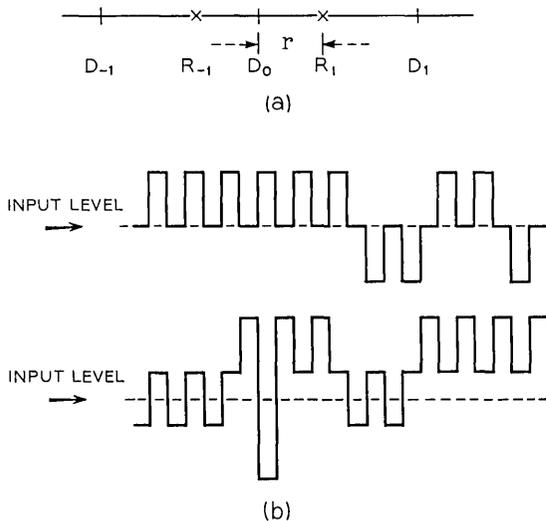
Fig. 18 — (a) Quantizer characteristic with decision level at zero. (b) Chopping pattern for input at $\Delta = 0$. (c) Chopping pattern for input at $\Delta = 1/2\ R_1$.

acteristic have a representative level at zero and uniform spacing of the adjacent pair of representative and decision levels.

Dither patterns may be three dimensional in design, varying horizontally, vertically, and from frame-to-frame. A pattern that varies in only one dimension can be generated having two, three, four, five, or seven amplitude levels (and no others), such that the visibility of the added pattern is a minimum for each level of the pattern.

We predict that a deterministic four-level pattern will give minimum visibility or granularity for *Picturephone*® visual telephone viewed at 36 inches. The use of this one dimensional dither signal should reduce the visibility of contours by a factor of about two when compared with a picture with no dither.

At 72 inches viewing distance (or say 36 inches with twice the sampling frequency) seven level dither should be used.

Four-level dither applied in two dimensions should reduce the visibility of contours by a factor of six compared with a picture having no dither. A further significant reduction should occur when dither is applied to the temporal dimension as well.

The dither signal may be subtracted from the received signal to further reduce the visibility of the added waveform. But the rules for determining the best patterns are different. For four-level dither the best addition-subtraction patterns give results that are only marginally better than the best patterns when they are not subtracted from the receiver.

APPENDIX A

Equivalence of Dither for Quantization

The method of proof is to show that for a Markov dither pattern (having an arbitrary conditional probability density function) the conditional probability of the switching pattern being at either level, given the previous value of the dither signal, is the same for both ordinary and differential quantizers.

Assume a Markov dither pattern described by the transition probability density function $P(x_i/x_{i-1})$ where $| x_i | \leq r/2$. The pattern could be either deterministic or random. For ordinary quantization, the probability of obtaining level R_n and level R_{n+1} for an input analog amplitude of $R_n + \Delta$ (see Fig. 19b) is

$$\Pr \{R_n/x_{i-1}\} = \int_{(-r/2)}^{(r/2)-\Delta} P(x_i/x_{i-1}) \, dx_i = I_1 \qquad (1)$$
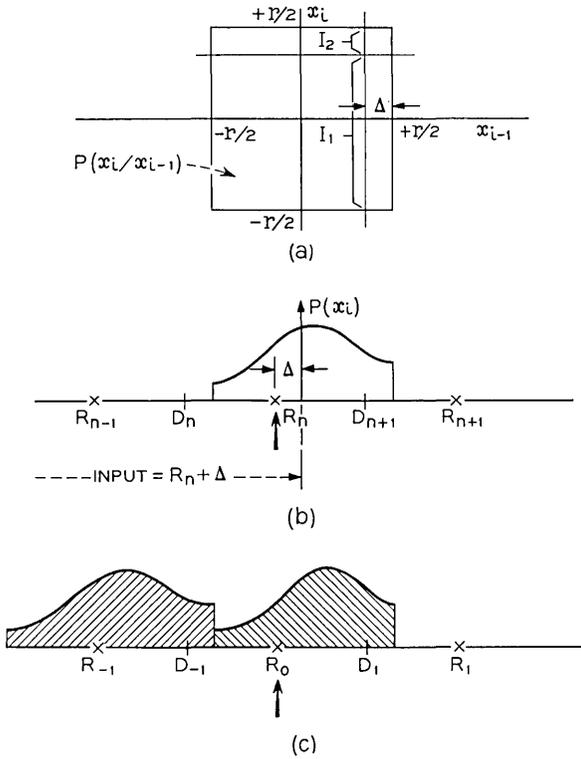
Fig. 19 — Equivalence of dither for ordinary and differential quantization— definition of terms—(a) conditional probability density function, (b) ordinary quantization, and (c) differential quantization.

and

$$\Pr\{R_{n+1}/x_{i-1}\} = \int_{(r/2)-\Delta}^{r/2} P(x_i/x_{i-1})\, dx_i = I_2\,, \qquad (2)$$

where $\Pr\{a/b\}$ is the probability of event a occurring given that event b has occurred. Thus the probability of obtaining levels R_n and R_{n+1} is conditional only on x_{i-1}.

For differential quantization, feedback occurs from the previous sample value, and it becomes necessary to distinguish between the output of the quantizer (primed) and the output of the complete differential quantizer (unprimed). Again, for an analog input of $R_n + \Delta$, assuming equal spacing of R_{-1}, D_{-1}, R_0, D_1, and R_1 we have

$$\Pr \left\{ R_n / x_{i-1} \; ; x_{i-1} < (r/2) - \Delta \right\}$$

$$= \int_{-(r/2)}^{(r/2)-\Delta} P(x_i / x_{i-1} \; ; x_{i-1} < (r/2) - \Delta) \, dx_i \qquad (3)$$

$$= \Pr \left\{ x_{i-1} < (r/2) - \Delta \right\} \int_{-(r/2)}^{(r/2)-\Delta} P(x_i / x_{i-1}) \, dx_i$$

$$= \Pr \left\{ x_{i-1} < (r/2) - \Delta \right\} \left\{ I_1 \right\}. \qquad (4)$$

Now

$$\Pr \left\{ R_n / x_{i-1} \; ; x_{i-1} > (r/2) - \Delta \right\}$$

$$= \Pr \left\{ R'_{-1} / x_{i-1} - r; x_{i-1} > (r/2) - \Delta \right\}$$

$$= \Pr \left\{ x_{i-1} > (r/2) - \Delta \right\} \Pr \left\{ R'_{-1} / x_{i-1} - r \right\}. \qquad (5)$$

But since

$$\Pr \left\{ R'_{-1} / x_{i-1} - r \right\} = \Pr \left\{ R'_0 / x_{i-1} \right\} = \Pr \left\{ R_n / x_{i-1} \right\}, \qquad (6)$$

$$\Pr \left\{ R_n / x_{i-1} \; ; x_{i-1} > (r/2) - \Delta \right\} = \Pr \left\{ x_{i-1} > (r/2) - \Delta \right\} \left\{ I_1 \right\}.$$

Thus from equations (3) and (6) one can see that $\Pr\{R_n/x_i\}$ is independent of the previous state of the differential quantizer and equal to the value obtained for the ordinary quantizer. By a similar argument, $\Pr\{R_{n+1}/x_{i-1}\}$ can be equated for the two quantizers.

APPENDIX B

Calculation of Visibility of Dither Signals

B.1 *Model of Vision*

Figure 20 shows a simple model used to describe the visibility of small amplitude signals.[7] $I(x, y, t)$ represents the spatial and temporal luminance pattern incident at the eye. The filter $\lambda(x, y, t)$ accounts for spread of the signal in space and time caused by the optics, the receptors, and subsequent neural processing. The amplitude of the hypothetical signal $E(x, y, t)$ is proportional to the observed visibility of the display. Thus the difference in visibility between two
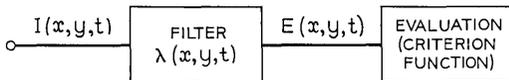

Fig. 20 — Model of threshold vision.

displays can be measured by evaluating the average of some function of the difference between the value E resulting from one display, and the value of E resulting from the other.

We wish to know how well the discrete waveform with added dither approximates the analog signal in flat areas where contours are most bothersome. Thus it is reasonably accurate to represent the analog signal by a constant amplitude E_a, and the measure of the visibility of the discrete waveform is

$$U(\Delta) = E\{f[E_a - E_\Delta(x)]\},$$

where $E\{\cdot\}$ denotes the expected value and, as before, Δ denotes the position of the input within the quantizing interval. $E(x)$ varies with Δ since the chopping pattern $I(x)$ changes as Δ is varied. In a number of cases, $U(\Delta)$ has been evaluated for two different f functions, the square and the modulus.

B.2 *Visibility of Waveform*

The method of evaluating the visibility differs from that used previously.[4] Earlier, $E(x)$ was calculated for every possible input combination occurring in a signal segment of the length of the significant part of the impulse response. The probability density function of $E(x)$ was then calculated by weighting each output by the probability that the corresponding input occurred. From the probability density function the error can be calculated for the required criterion function.

The method now used is to first calculate a combined impulse response for the reconstruction filter and visual filter: this is then convolved with the input signal to obtain an output from which a measure of the granularity is derived. This technique is fast and accurate for deterministic signals which repeat after a short length, but slower if accurate results are required for random inputs. Fortunately, most of the signals investigated were deterministic.

Denoting the impulse response of the low-pass filter by $h_1(x)$ and the visual spatial filter (for example, in the horizontal dimension) by $h_2(x)$, then the combined impulse response is given by[8]

$$\lambda(x) = \int_{-\infty}^{\infty} h_1(y)h_2(x - y) \, dy.$$

This integral was evaluated for

$$h_1(x) = \frac{1}{x'} \left(\sin \frac{\pi x}{x'} \right) \bigg/ \frac{\pi x}{x'}$$

and

$$h_2(x) = \pi^{-1/2} \exp\left[-0.0833\left(\frac{Ax}{x'}\right)^2\right],$$

where x' is the spatial Nyquist interval and A, which depends upon the viewing distance, is the width of a picture element in minutes of arc; $h_2(x)$ is the same impulse response as used previously.[4]

Figure 21 shows the combined impulse response of the normalizing low-pass filter and the visual system for Mod. II *Picturephone*® visual
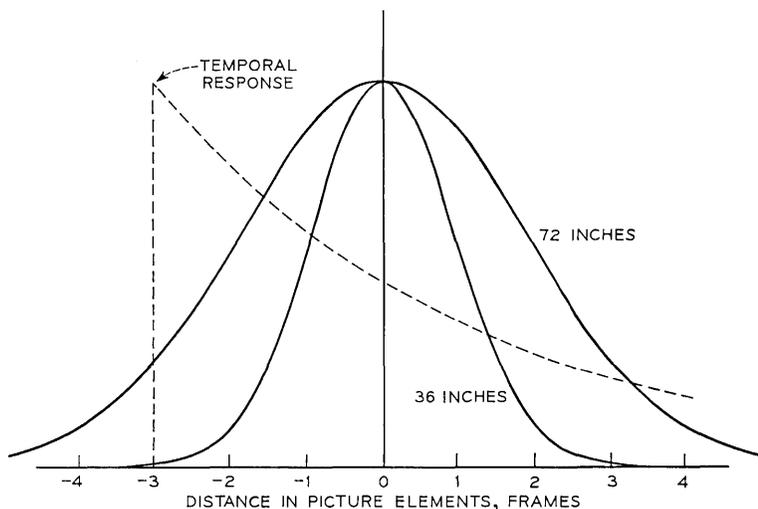

Fig. 21 — Combined spatial impulse response in one direction at viewing distances of 36 and 72 inches.

telephone viewed at 36 inches. The corresponding impulse response for a viewing distance of 72 inches (or alternatively, for 36 inches at twice the sampling frequency) is also shown and agrees to three decimal places with the impulse response of the visual system itself. In other words, the visibility of threshold detail is almost completely unaffected by the horizontal resolution of the display at 72-inch viewing distance (resolution limited by eye).

For a given input $I(x)$ the output is (using the convolution theorem),

$$E(x) = \int_{-\infty}^{\infty} I(y)\lambda(y - x)\, dy.$$

The limits of the integral can be reduced so as to integrate over only those values of $(y - x)$ for which λ is significantly greater than zero (in practice, greater than 0.1 percent). For random inputs, simulations were run for between 300 and 900 samples.

REFERENCES

1. Limb, J. O., and Mounts, F. W., "Digital Differential Quantization," B.S.T.J., this issue, pp. 2583–2599.
2. Goodall, W. W., "Television by Pulse Code Modulation," B.S.T.J., *30*, No. 1 (January 1951), pp. 33–49.
3. Roberts, L. G., "Picture Coding Using Pseudo-Random Noise," IRE Trans., *IT-8*, No. 2 (February 1962), pp. 145–154.
4. Limb, J. O., "Coarse Quantization of Visual Signals," Australian Telecommunication Res., *1*, Nos. 1 and 2 (November 1967), pp. 32–42.
5. Max, J., "Quantizing for Minimum Distortion," IEEE Trans. Inform. Theory, *IT-6*, No. 1 (March 1960), pp. 7–12.
6. Graham, R. E., "Predictive Quantizing of Television Signals," IRE Wescon Conv. Rec. *2*, part 4 (1958), pp. 147–157.
7. Budrikis, Z. L., "Visual Thresholds and Visibility of Random Noise in TV," Proc. Inst. Radio Eng. (Australian) *22*, No. 12 (December 1961), pp. 751–759.
8. Mason, S. J., and Zimmerman, N. J., *Electronic Circuits, Signals, and Systems*, New York: John Wiley, 1960, p. 327.

Digital Differential Quantizer for Television

By J. O. LIMB and F. W. MOUNTS

Correct tracking between the transmitter and receiver is difficult to main-
tain when a long integrator time constant is used. We describe a differential
quantizer which has a digital integrator; this integrator enables perfect
*tracking to be achieved at the output of the integrator without any adjust-*
ments.

The differential quantizer gives high quality pictures when seven, eight,
and nine quantizer output levels are used. We present a scheme for trans-
mitting the nine-level signal at the rate of three bits per picture element.

Picture quality is improved significantly by adding low-amplitude dither
patterns to the input signal to mask contours. The coder is more susceptible
to transmission errors than coders having an analogue integrator with a
short time constant; we discuss two methods for reducing the susceptibility.

I. INTRODUCTION

Differential quantization is well suited to visual signals for two
reasons.[1] First, the quantizer acts like a predictive encoder, taking
advantage of the large amount of correlation between adjacent ele-
ments of a picture to obtain a good prediction of the amplitude of the
point being quantized.[2] Thus, the differential quantizer makes use of
some of the statistical redundancy in the source. Second, the quantiza-
tion can be partially matched to the changing sensitivity of vision.[3]
To understand this, remember that the sensitivity of the visual system
to small differences in luminance decreases markedly at boundaries be-
tween light and dark areas. The signal that is applied to the quantiza-
tion stage of a differential quantizer is very nearly equal to the change
in amplitude between adjacent elements. Thus, by quantizing small
amplitude samples finely and large amplitude samples more coarsely
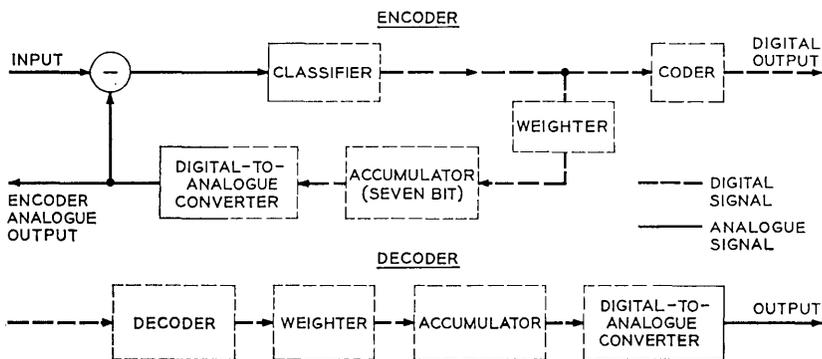a picture can be obtained which is partially matched to visual require-
ments.

Figure 1a is a block diagram of a differential quantizer; it differs in
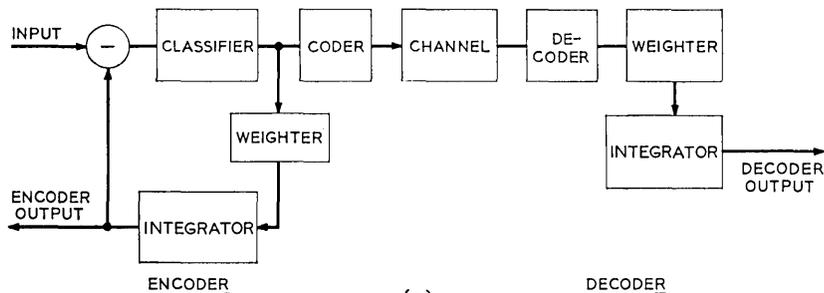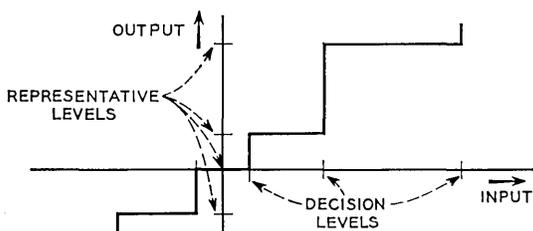
Fig. 1 — (a) Differential quantizer (DPCM coder-decoder). (b) Transfer characteristic of a quantizer. (c) Digital differential quantizer.

one small aspect from the usual representation. The quantizer section of the differential quantizer is considered as two separate parts. The first part, the classifier, contains the decision levels which divide the input signal range into a number of intervals (see Fig. 1b). Thus, the signal at the output of the classifier is digital and just denotes the interval in which the signal sample occurred. The signal at this stage

is encoded for transmission to the receiver. The second part, the weighter, assigns an amplitude or weight (representative value) to each section that can have either a digital or analogue value. The integrator is usually an analogue device in which case the weights are generated as analogue values.

In practice, it is difficult to resynthesize a high-quality picture at the decoder. The longer the time constant of the encoder and decoder integrators, the greater the precision required in implementation to prevent the received picture from differing from the sent picture (referred to as mistracking). A short integrator time constant, on the other hand, makes mistracking less of a problem but introduces effects similar to noise. These, generally, are not too serious if the inner pair of levels alone is used to make the correction for integrator leak. Mistracking can stem from three sources:

(i) The representative levels at the encoder and decoder can be mismatched. While the setting of the largest pair of levels is not quite as critical since it is seldom used, the smallest levels must be adjusted quite accurately, especially when the integrator time constant is long.

(ii) The frequency response of the integrators can differ.

(iii) An analogue component of the signal can bypass the classifier stage (hence analogue breakthrough) and feed through the analogue weighter into the integrator. The analogue breakthrough is generally prevented from reaching the integrator in the decoder by digital regeneration in the signal path, and so a mismatch between the encoder and decoder can occur. The problem can be overcome by carefully gating the digital output of the classifier stage to remove any vestige of analogue signal.

In an attempt to overcome decoder mistracking and still have a long integrator time constant, it was decided to perform the operations of weighting and integrating digitally. This should ensure exact tracking of the decoder under all conditions except for the obvious cases of either a digital circuit malfunction or an error occurring during transmission.

Figure 1c is the block diagram of the digital differential quantizer. The analogue parts are distinguished from the digital parts. An extra block is required since it is necessary to convert the output of the accumulator (digital integrator) to an analogue quantity prior to subtracting.

The digital encoder-decoder completely eliminates mistracking; thus, the picture at the receiver cannot be distinguished from the

quantized picture at the transmitter. High precision analogue weighters and integrators are now replaced by digital circuits and, further, the classifier design need not be precise. Indeed, analogue breakthrough, unless it is large enough to switch a digital circuit, is ineffectual and small variations in the position of the decision levels have negligible effect on picture quality. Requirements on the digital-to-analogue (D-A) converter are not very strict. Nonlinearity in the characteristics of the D-A converters has the effect of producing a change in the gamma of the signal, of which the eye is not very critical. In fact, changes in the D-A converter produce similar effects to changes in the D-A converter of an ordinary PCM system.

In quantizers used for ordinary PCM encoding the appearance of contour lines in low-detail areas sets the lower limit on the number of quantizing levels that can be used. For the differential quantizer a similar effect occurs. As the weight assigned to the smallest pair of representative levels is increased, contour lines become more visible. Addition of random and pseudorandom noise improves picture quality for ordinary quantizers having less than about 100 levels.[4,5] A theoretical study of the use of dither signals with differential quantizers suggests that even at a close viewing distance dither should prove effective in improving picture quality for a given number of levels.[6]

This paper first describes the digital differential quantizer and the results obtained with it. The reduction in the visibility of contours by adding specially designed dither signals in both the horizontal and vertical directions is then explored. The problem of overcoming the effects of channel errors is also briefly discussed.

II. DESCRIPTION OF SYSTEM

2.1 *Analogue Section*

See Fig. 1c. The subtractor is an emitter-coupled pair circuit that is ac coupled at both the input and feedback terminals. The classifier comprises a set of eight threshold circuits connected in parallel. Their threshold levels can be adjusted independently to give the desired partition of the input. For example, setting two decision levels to the same value reduces the number of intervals by one.

The classifier design is simplified by making the positive and negative stages identical. This is made possible by feeding them separately with signals of opposite polarity; such signals are generated by the emitter-coupled pair of the subtractor. Sampling is inherent in the operation of the decision circuits. Narrow sample pulses with a base

width of approximately 20 ns are amplitude modulated by the signal. The decision as to whether the sample exceeds the threshold is then made using a high speed flip-flop. When the flip-flop is set, the signal is considered to have exceeded threshold. A reset pulse is applied to all flip-flops prior to the occurrence of the next sample pulse. The classifier successfully uses only emitter-coupled integrated circuit logic elements (for essentially analogue operations) to obtain fast decisions (< 30ns) and good stability.

2.2 Digital Section

Because a parallel classifier is used, all threshold circuits with thresholds less than the input signal value are triggered. For example, if the signal to the classifier exceeds level number three, then an output occurs on level number three and also on levels number two and number one. The outputs of the threshold circuits are combined logically to give a sign bit and four other binary signals—one for each level. A particular signal takes the value "one" when the input to the classifier falls in the interval associated with that level. This code allows no more than one of the four outputs to be a "one" for any sample. A zero on all four bits denotes the zero interval of the classifier. This code change is not essential, but it allows the weights associated with each interval to be controlled independently, which is convenient in an experimental coder.

Each output is connected to a word generator which generates a binary number specifying the amplitude of the representative level. Notice that this method of weighting means that only symmetrical weighter configurations can be investigated because the magnitude of the levels is generated independently of the sign. The experimental arrangement allowed any set of digital weights to be wired on a small plugboard.

The accumulator uses a seven bit adder that sets the precision with which the weights can be assigned (Fig. 2). The contents of the adder for the previous sample are fed back into the adder together with the new difference signal. For a zero difference signal it can be seen that the same seven bit amplitude signal would circulate through the adder and delay stage without change. The operation of the accumulator can be expressed as $y_n = y_{n-1} + x_{n-1}$, where x_n and y_n would be the values of the input and output, respectively, of the accumulator for the nth sample.

Under certain conditions the adder could overflow or underflow, say for a large peak in the input video signal. This is prevented by the
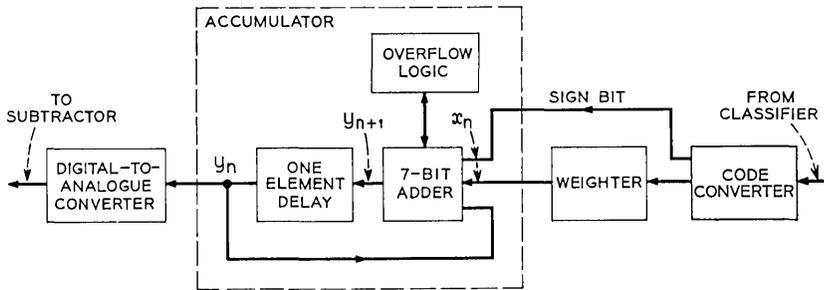
Fig. 2 — Digital section of differential quantizer.

addition of overflow logic. The circuit holds the adder output at level 127 if overflow occurs or at level zero if the adder underflows, that is,

$$y_n = 127 \quad \text{if} \quad y_{n-1} + x_{n-1} \geqq 127$$

and

$$y_n = 0 \qquad \text{if} \quad y_{n-1} + x_{n-1} \leqq 0.$$

The overflow logic effectively clamps the accumulator when it is under-driven; we make use of this action to fix the dc level in the coder. The delay period of one element is realized with a clocked flip-flop.

The digital-to-analogue converter uses the ladder method of conversion. The seven-bit converter was built using selected 1 per cent resistors and has a settling time of less than 50 ns. Resampling for display purposes was not considered necessary; within the loop the classifier resamples.

2.3 *Alignment*

Setting the input and output levels of a quantizer accurately is normally a tedious problem requiring precise equipment. In this section we describe a self-alignment technique using the digital-to-analogue converter of the differential quantizer; the level adjustment problem then becomes quite trivial.

Since the representative levels of the weighter are assigned digitally, they can be set exactly with the wired plugboard. The decision levels, however, are analogue levels, and they need to be set up accurately both in relation to themselves and to the representative levels. To do this, the feedback loop is broken between the classifier, and the weighter and the classifier levels are set up digitally in the weighter at twice the desired value. The sign bit to the accumulator is alternated each sample period so that the output of the digital-to-analogue con-

verter is a square wave of twice the desired amplitude. The input video signal is disconnected so that only the square wave is coupled to the classifier. Since the signal is ac coupled, the excursions from the mean are of the right value to set both the positive and negative decision levels. The control on the position of the threshold is adjusted until the decision stage just triggers. Nothing more complex than a voltmeter is required to make this adjustment.

III. INVESTIGATIONS

3.1 *Ideal Weighter-Integrator*

Probably the most significant difference between previous analogue implementations of the differential quantizer and the present digital implementation is the fact that the weights are known exactly and the integrator accumulates exactly until it is reset at the end of a line. Thus, the weights of the larger levels can be set at values which are exact multiples of the smallest (or inner) levels. For example, if the inner pair of levels was set at 3/128ths and the other three pairs of levels were set at 6/128ths, 12/128ths, and 21/128ths, this would be called a multiple setting of the weighter. A multiple setting is a necessary condition for producing clear contour patterns that are the same as one gets with ordinary quantization. Figure 3b, which shows a picture processed by the digital differential quantizer, illustrates these contour patterns; for comparison Fig. 3a shows the original signal. The picture was chosen because of its flat background which has the effect of emphasizing contours. A coarse quantizer setting with only seven levels is used to make them more visible.

By going to a nonmultiple setting, the coded amplitude in a flat area of the picture will vary from line to line depending upon what levels were used in the previous part of the line. It might be argued that this effect could be used to mask contouring. It can, but it is not very successful as Fig. 3c shows. There is still contouring on the left side of the picture; the right side is quite streaky compared with what can be done using other methods (Fig. 3d). These methods are discussed further in Section 3.4.

Thus, for a multiple setting of the weighter the low-detail areas of the picture are quite free from random noise. Quantizing errors show up as more or less visible contours (depending on the level setting and the type of picture material) which are quite sharp if the input signal-to-noise ratio is high. Quantization error at edges and in high detail areas, on the other hand, is more random in nature.

Fig. 3 — Pictures processed by digital differential quantizer: (a) original analogue signal, (b) processed picture—seven levels with multiple setting of levels, (c) processed picture—seven levels with nonmultiple setting, and (d) processed picture—same level setting as (b) with 4 × 4-step dither signal added. (The scan lines and printing screen cause moiré patterns that are not in the originals.) Glossy prints of this figure can be obtained by writing to the authors.

3.2 Quantizer Characteristic

We are limited to symmetrical configurations of the decision and representative levels because of the way the quantizer was designed. However, there is nothing to suggest that a nonsymmetrical setting would have any significant advantage. Many settings were tried using nine representative levels—four positive levels, four negative levels, and a zero level. The scale of Table I was found to give good results—no contouring is apparent but the skilled viewer can detect slight degradation at edges. The best setting changes only slightly with the subject matter. For contrasty pictures and graphics, edges can be improved by expanding the scale slightly. This is done by

reducing the input amplitude and increasing the output amplitude by a compensating amount.

Eight-level settings of the weighter were investigated using the same configuration as with nine levels, except that the zero representative level was removed by moving the first pair of decision levels at ±1 percent to zero. The smallest output step of this configuration, however, is unchanged. Edges should also be reproduced with the same fidelity.

The pictures obtained with eight levels are very similar to the nine-level pictures. Contours still occur with about the same visibility for coarse-level settings but now a constant amplitude consists of an oscillation between two levels separated by an amount equal to the smallest step size. This oscillation is largely removed by the filtering at the receiver and the filtering taking place in the eye.

Since there is little choice between the picture qualities of the eight-level setting and the nine-level setting, the eight-level setting would be preferred since fewer levels are required. However, the nine-level configuration of Table I can be altered to significantly reduce the visibility of contours by placing the first pair of decision levels at half the value given in Table I. One can show that this is equivalent to adding a deterministic, two amplitude dither signal (which is random in the vertical and temporal directions) in the horizontal direction (see Section 3.4 and Ref. 6). The quantizer levels can now be expanded (taking advantage of the reduction in visibility of contours) to reproduce high-detail areas more accurately, giving an overall improvement in picture quality.

In a number of analogue differential quantizers built previously by others, representative level settings of approximately 1, 3, 7, and 20 percent (for example, Ref. 2) have been found satisfactory. This setting compares with approximately 2, 6, 14, and 24 percent for the digital differential quantizer. The difference in the inner level settings

TABLE I—LEVEL POSITIONS OF QUANTIZER*

Level		±1		±2		±3		±4	
Decision Level		±1		±4		±10		±19	
Representative Level	0		±2		±6		±14		±24

* Expressed as percent of peak-to-peak signal.

is quite large and is almost surely the result of the different integrator characteristics.

The picture quality is rather insensitive to a change in input signal amplitude. For example, an increase of 4 dB in signal level causes a slight loss of sharpness at edges with a decrease in the visibility of contours, while a decrease of 4 dB improves the edges and makes the contours a little more visible.

3.3 Sign Predictor

If there is to be no further coding of the digital signal (apart from assigning a constant length word to each output sample), it is more efficient to have the number of quantizer output levels equal to a power of two. Thus, the improvement in quality obtained using nine output levels would be negated if four bits instead of three had to be assigned to each picture sample.

Now we describe a simply implemented scheme (which may also be used with differential quantizers having analogue integrators) for reducing the required number of levels by one. The scheme enables the nine-level setting to be used with a channel transmitting at the rate of three bits per picture element.

The probability of having the largest level (level number four) preceded by a level of the opposite sign is small. One factor tending to reduce this probability is the smoothing provided by the normalizing filter at the input to the differential quantizer.* Consequently, the sign of an outside level can be predicted fairly accurately by assuming that it is the same as the previous sign. If the prediction is wrong, a number three level, rather than a number four level, is used and will thus have the correct sign.

Thus, a level is effectively eliminated since instead of indicating that the fourth positive or the fourth negative level has occurred (one of two possible events), it is only necessary to indicate to the decoder that the fourth level has occurred (one event) and the decoder then assigns the sign of the previous sample. The signal is modified in this way immediately after the classifier stage; this modification is best regarded as an adjunct to classification. Thus, the encoder and decoder keep in track; in the event of an outside level being preceded by a level of the opposite sign, the slope capability of the encoder and decoder is reduced. For most pictures it is difficult to detect any

---

* The input filter is 3 dB down at 670 kHz and 14 dB down at half the sampling frequency (1 MHz). The output filter is 6 dB down at 1 MHz, giving an overall attenuation of 20 dB at 1 MHz.

change in picture quality when level elimination is used with nine levels. It is simple to implement—in fact it requires only an additional flip-flop and two gates.

3.4 *Dither*

3.4.1 *Straight Quantization with Dither*

In quantizers used for straight PCM encoding, contouring becomes obvious if less than about 100 levels are used. However, small amplitude, high frequency waveforms may be added to the coarsely quantized signal to reduce the visibility of the contours. The penalty associated with adding dither (as these added waveforms will be called) is that the background noise level in the picture is increased slightly. The task in designing dither waveforms is to select that waveform which has the minimum visibility and hence disturbs the picture the least. In the past, random and pseudorandom waveforms have been investigated,[5,7] but more recent calculations have been made with deterministic waveforms (or patterns) indicating their superiority.[6]

3.4.2 *Differential Quantization With Dither*

Dither can be used to advantage in the digital differential quantizer. The quantizer levels are generally set so that contouring is not detectable; but if dither is used, the level spacing can be expanded to either enable the number of levels to be reduced (say from nine to seven) or improve the picture quality by reproducing edges more sharply.

For a certain quantizer configuration, the design of the dither waveform becomes identical to the design for ordinary quantizers.[6] Figure 4 shows this configuration, which has a representative level at zero and the first pair of representative levels set at twice the value of the first pair of decision levels. On the other hand, dither with a decision level at zero produces complex multilevel output waveforms and is not considered here.

Dither may be applied in two ways; the design of the waveform depends on the way it is applied. In the first way it is just added at the input. In the second way, besides adding the waveform at the input it is subtracted from the output. For random uncorrelated waveforms Roberts has shown that for random dither the addition-subtraction technique is superior to addition alone[5]—the variance of the output waveform is reduced by one half.
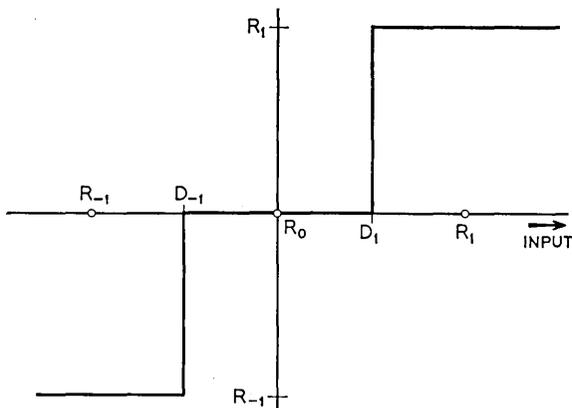
Fig. 4 — Configuration of quantizer with representative level at zero.

However, for a deterministic waveform it has been predicted that there is little difference in the visibility of waveforms designed for addition and waveforms designed for addition-subtraction.[6] This prediction has been tested using a four-amplitude (four-step) pattern with the sequence 1, 3, 2, 4 applied to one dimension, the vertical dimension (Fig. 5). Notice, that although the added waveform is written 1, 3, 2, 4, it has a mean of zero and the levels are positioned uniformly within the quantizing interval as shown in the figure. Theoretically, this is the best one-dimensional, four-step dither waveform for the addition method. The design technique for the addition-subtraction waveform is different from the design for the addition waveform and, in fact, there are two waveforms that have minimum visibility; they are sequence 1, 3, 2, 4 (as for the addition method) and sequence 1, 2, 4, 3.

3.4.3 *Result with Dither*

We now describe the results obtained by adding the waveforms to two particularly sensitive types of display. The first display is a ramp applied in the horizontal direction (Fig. 6); the second is a picture with large flat areas (Fig. 3). For the addition method, the waveform introduces a small amount of frame flicker as a result of interlace since the component added to the first and third lines (field one) is not equal to the component added to the second and fourth lines (field two). Although the add-subtract method does appear to give a smoother looking display, the difference is very slight.
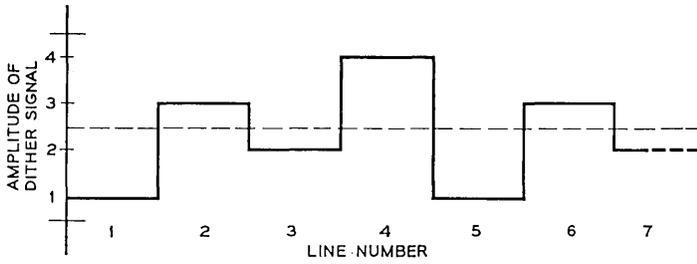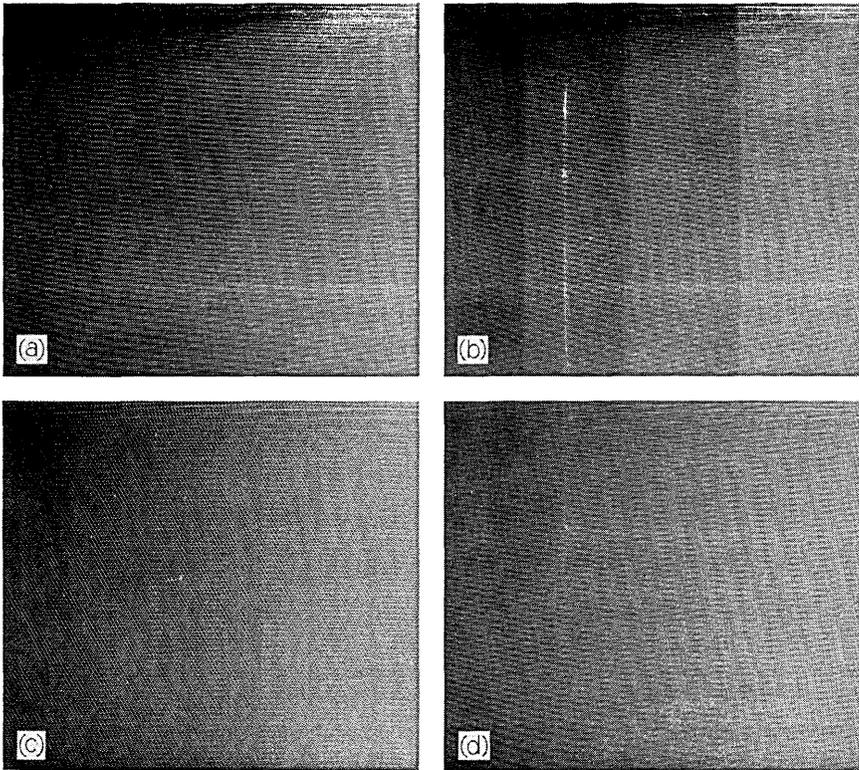
Fig. 5 — Four-step dither waveform.

Fig. 6 — Ramp signal processed by digital differential quantizer: (a) Original analogue signal, (b) processed signal—no dither, (c) processed signal—dither added in vertical direction with four-step pattern, and (d) processed signal— dither added in vertical and horizontal directions with 4 × 4-step pattern. (The scan lines and printing screen cause moiré patterns that are not in the originals.) Glossy prints of this figure can be obtained by writing to the authors.

The 1, 2, 4, 3 pattern, as expected, gave the same results as the 1, 3, 2, 4 pattern for the addition-subtraction method. For the addition method, frame flicker was eliminated with a 1, 2, 4, 3 pattern but the output waveform was more visible since the dither waveform is less suitable.

A two dimensional 4 × 4 step dither waveform was generated for the horizontal and vertical dimensions. Table II shows the pattern. With reference to Table II, the sequence 1, 9, 3, 11, 1, 9, 3, 11 . . . was added to the first line; 14, 6, 16, 8, 14, 6, 16, 8 . . . was added to the second line, and so on. Thus 1, 9, 3, 11 . . . would be added again to the fifth line.

Since we have interlace, line one is in a different field than line two. The average contribution to each field can be found by adding the numbers in each line. The first and third lines contribute to the first field and have a total of 24 + 28 = 52. The second and fourth lines contribute to field two and have a total of 44 + 40 = 84. Hence, the average contribution to field one is not equal to the average contribution to field two and a small amount of 30 Hz flicker results.

The pattern in Table III does average out over a frame and hence would produce no flicker but was not investigated at this time. Figure 3d shows the improvement obtained by using the dither signal of Table II, compared with not using dither (Figure 3b). The photographs indicate fairly accurately the improvement due to dither, since this particular dither waveform does not rely on time averaging.

An attempt was made to quantitatively assess the improvement resulting from dither. Theoretical predictions were that four-step interpolation would reduce the visibility of contours by 7.1 dB and 4 × 4-step dither by 16.8 db.[6] The method used to test these figures was to add dither to the quantized ramp display (Fig. 6b). A subject attenuated the displayed picture signal until the visibility of the display without dither was equal to the display with dither. The amount of attenuation was then recorded. The viewing distance was

TABLE II—TWO DIMENSIONAL 4 × 4-STEP DITHER WAVEFORM
WITH UNBALANCED FIELD CONTRIBUTIONS

	Horizontal →				Line Totals	
Vertical ↓	1	9	3	11	Field 1	24
	14	6	16	8	Field 2	44
	4	12	2	10	Field 1	28
	15	7	13	5	Field 2	40

36 inches. For the four-step pattern (Fig. 6c), four technical subjects gave an average value of 9 dB with a spread of 4 dB, probably reflecting the difficulty of making a match in the presence of flicker. For the 4 × 4-step pattern (Fig. 6d), three subjects gave an average attenuation of 14.3 dB with a spread of 1 dB. These measurements support the theoretical figures of 7.1 dB and 16.8 dB.

A quantitative comparison with picture material would be more difficult and has not been attempted. The method above is not suitable because attenuating the amplitude of the output signal for the picture without dither produces displays that are quite different in appearance. Qualitatively, the reduction in visibility of contours in low-detail areas of the picture is quite dramatic.

3.5 *Further Coding*

When subsequent coding is permitted, the whole approach to the design of the encoder changes. For example, applying dither in the horizontal direction would lead to a less efficient Huffman code (as would a short integrator time constant) while dither applied vertically or temporally would have little effect.

3.6 *Transmission Errors*

A chief disadvantage of the digital accumulator is that an error in transmission will affect subsequent picture elements until the accumulator is reset. Thus, each error will produce a horizontal streak which starts at the point where the error is made and persists to the right edge of the picture where the accumulator is reset. With an analogue integrator the length of an error streak is commensurate with the time constant of the integrator (time constants as short as six picture elements have been used); however, as mentioned previously, a short time constant has other disadvantages.

The length of an error streak in the digital implementation can be shortened by updating the accumulator during the line. For example,

TABLE III—TWO DIMENSIONAL 4 × 4-STEP DITHER WAVEFORM
WITH BALANCED FIELD CONTRIBUTIONS

	Horizontal →				Line Totals	
	1	14	3	16	Field 1	34
Vertical	10	5	12	7	Field 2	34
↓	4	15	2	13	Field 1	34
	11	8	9	6	Field 2	34

a full seven-bit PCM signal may be transmitted halfway through each line. The accumulator at the receiver could be updated to this value thus truncating any error which may have occurred. Of course, this process could be repeated more often (with a consequent reduction in information transmission efficiency) for a high transmission error rate. This method of reducing the effect of errors is analogous in a way with shortening the time constant of an analogue integrator without the same disadvantage.

The precision of digital integration leads to another method for reducing the effect of errors. Assume that overflow or underflow of the transmit accumulator is inhibited. Then, when the coder at the transmitter is reset to a pre-assigned value at the end of a block of data (say a line), the decoder at the receiver should recover to the same value. If it does not, transmission errors have occurred. Errors that would escape detection in this way are self-correcting errors, for example, level "a" is received as level "b" followed by level "b" received as level "a". In practice, the probability of an error being of this type would be small.

When a block of data is detected as being in error it can be replaced by an estimate of that block. Now since there is a large amount of unexploited redundancy in a television signal, a reasonable estimate of the line can be made. For example, the previous line could be used or the next and previous lines could be averaged. This technique would probably be satisfactory down to error rates where the probability of obtaining errors in adjacent blocks becomes significant. If the block length was one quarter of a line, error rates of one per line or approximately one in 10^3 might still give a reasonable picture. The degradation would appear as a slight loss in vertical resolution. This proposal has the disadvantage that at least two lines of storage would be required at the receiver.

IV. SUMMARY

We were able to construct a rugged, adjustment-free quantizer with low precision components by using digital techniques at certain points in the path of a differential quantizer. High quality pictures are obtained by using either seven, eight, or nine quantizer output levels. The quality appears somewhat different from the pictures obtained with analogue implementations. The pictures appear less noisy. This is attributed to two facts: (i) the values of the quantizer output levels are assigned digitally allowing the larger levels to be set at an exact multiple of the smallest level; (ii) integration is also per-

formed digitally, resulting in a virtually infinite time constant.

A simple technique was described whereby the number of quantizer output levels required to be transmitted can be reduced by one. Thus, a nine-level picture, which gives an improvement in picture quality over an eight-level picture can be transmitted at the rate of three bits per picture element.

By expanding the quantizing scale (spacing the levels further apart), edges were reproduced more sharply; however, contouring (exactly as encountered in straight quantization) becomes obvious in low-detail areas of the picture. By adding specially designed dither waveforms to the input signal, contours were "washed out" at the expense of a very small increase in background noise. Changing the input amplitude by ±4 dB produced little change in picture quality.

Because of the long effective time constant of the digital integrator, transmission errors are more visible than in differential quantizers employing a short integrator time constant. Two methods for reducing the visibility of such errors were discussed.

In a visual communication system, a coder must be reliable and have a long adjustment-free life under a wide variety of environmental conditions. Further, a decoder must be capable of working with every encoder in the system. The digital differential quantizer is ideally suited to such a situation.

V. ACKNOWLEDGMENT

REFERENCES

1. Cutler, C. C., "Differential Quantization of Communication Signals," Patent No. 2,605,361, applied for June 29, 1950, issued July 1952.
2. Graham, R. E., "Predictive Quantizing of Television Signals," IRE Wescon Conv. Rec., 2, part 4 (1958), pp. 147–157.
3. Limb, J. O., "Source-Receiver Encoding of Television Signals," Proc. IEEE 55, No. 3 (March 1967), pp. 364–379.
4. Goodall, W. W., "Television by Pulse Code Modulation," B.S.T.J., 30, No. 1 (January 1951), pp. 33–49.
5. Roberts, L. G., "Picture Coding using Pseudo Random Noise," IRE Trans. on Inform. Theory, IT-8, No. 2 (February 1962), pp. 145–154.
6. Limb, J. O., "Design of Dither Waveforms for Quantized Visual Signals," this issue, pp. 2555–2582.
7. Limb, J. O., "Coarse Quantization of Television Signals," Australian Telecommunications Research 1, Nos. 1 and 2 (November 1967), pp. 32–42.

Contributors to This Issue

C. P. BATES, B.E., 1958, Nova Scotia Technical College, Halifax, Nova Scotia; M.E., 1960, Nova Scotia Technical College; Ph.D., 1966, University of Illinois; Bell Telephone Laboratories, 1966—. Since joining Bell Telephone Laboratories, Mr. Bates has been engaged in research on the propagation of very low frequency (VLF) waves in the earth-ionosphere waveguide, the solution of certain Wiener-Hopf boundary value problems, pulse scattering, and propagation in waveguide bends. Member, IEEE, Sigma Xi.

PAUL T. BRADY, B.E.E., 1958, Rensselaer Polytechnic Institute; M.S.E.E., 1960, Massachusetts Institute of Technology; Ph.D., 1966, New York University; Bell Telephone Laboratories, 1961—. He has worked in modeling on-off speech patterns and speech level distributions, especially as they occur in two-way conversation over circuits containing voice-operated devices and transmission delay. Member, Acoustical Society of America, Sigma Xi.

MRS. H. J. CHEN, M.S. (mathematical statistics), 1959, University of Michigan; Bell Telephone Laboratories, 1960—. She has worked on the development of statistical methods and their applications in statistical testing procedures, sequential analysis, analysis of variance, time-slice work sampling, and economic forecasting. Member, American Statistical Association.

LOUIS H. ENLOE, B.S.E.E., 1955, M.S.E.E., 1956, Ph.D. (E.E.), 1959, University of Arizona; Bell Telephone Laboratories, 1959—. His early work was in modulation and noise theory in connection with space communications. Later work has been with lasers, coherent light, and holography with emphasis upon communication and display. He is head of the Opto-Electronics Research Department. Member, IEEE, Phi Kappa Phi, Sigma Xi, Tau Beta Pi, Pi Mu Epsilon, Sigma Pi Sigma.

CHARLES A. FRITSCH, B.M.E., 1958, University of Dayton; M.S.M.E., 1960, Ph.D., 1962, Purdue University; Bell Telephone Laboratories, 1961—. He has worked on problems in the thermal sciences associated

with hardening structures to withstand nuclear weapon effects, cooling electronic equipment, and developing gas lenses. He is Supervisor of the Fluid Mechanics and Heat Transfer Group of the Engineering Mechanics and Physics Department. Member, American Society of Mechanical Engineers, American Physical Society, Sigma Xi.

EDGAR N. GILBERT, B.S., 1943, Queens College; Ph.D., 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–1946; Bell Telephone Laboratories, 1948—. Mr. Gilbert has done research in several branches of applied mathematics and is interested in communication theory. Member, American Mathematical Society, IEEE.

J. E. GOELL, B.E.E., 1962, M.S., 1963, and Ph.D. (E.E.), 1965, Cornell University; Bell Telephone Laboratories, 1965—. While at Cornell Mr. Goell was a teaching assistant and held the Sloan Fellowship and the National Science Cooperative Fellowship. At Bell Telephone Laboratories, he has worked on solid-state repeaters for millimeter wave communication systems and optical integrated circuits. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi, IEEE.

J. E. IWERSEN, B.S., 1949, Wagner College; M.A., 1951, Ph.D., 1955, The Johns Hopkins University. Bell Telephone Laboratories, 1955—. Mr. Iwersen has been almost continually engaged in work on semiconductor devices. Until recently he headed the Exporatory Device Department, whose activities include the development of new structures and new applications for devices and integrated circuits. He now heads the Advanced Circuit Technology Department, responsible for the exploratory development of small central offices, optical stores, remote line concentrators, and electronic switching networks.

L. U. KIBLER, B.S., 1950, U. S. Coast Guard Academy; M.S.E.E., 1956, Massachusetts Institute of Technology; Ph.D., 1968, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1956—. Mr. Kibler has been concerned with experimental research in the fields of parametric amplifiers, tunnel diodes, lasers, microwave photo diodes, and Schottky-barrier diode converters. He participated in the design and operation of the receivers for the Echo and *Telstar*® communications satellite projects. Now he is engaged in millimeter wave antenna investigations. Member, IEEE, Eta Kappa Nu, Sigma Xi.

SANG H. KYONG, B.S., 1961, University of Rhode Island; Ph.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966——. Mr. Kyong has worked on radar signal processing and on the analyses of guidance and control systems. He has also been on the adjunct faculty of New York University. Member, American Association for the Advancement of Science, American Nuclear Society, IEEE, SIAM, Phi Kappa Phi, Sigma Xi, Tau Beta Pi.

ARTHUR B. LARSEN, B.S.E.E., 1959, M.S.E.E., 1961, Ph.D., 1966, Case Institute of Technology; Bell Telephone Laboratories, 1966——. Mr. Larsen has been investigating applications of holography and coherent light to visual communications systems. He is also involved with studies of camera systems for color *Picturephone*® visual telephone service. Member, IEEE, Optical Society of America, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

JOHN O. LIMB, B.E.E., 1963, Ph.D., 1967, University of Western Australia; Bell Telephone Laboratories 1967——. Mr. Limb has worked on the coding of television signals to reduce channel capacity requirements. He is currently working on methods of reducing frame-to-frame redundancy in moving pictures for *Picturephone*® visual telephone applications.

E. A. J. MARCATILI, Aeronautical Engineer, 1947, and E. E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–54; Bell Telephone Laboratories, 1954——. He has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated on optical transmission media. Fellow, IEEE.

STEWART E. MILLER, B.S. and M.S. in E.E., 1941, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1941——. He first worked on coaxial carrier repeaters and later worked on microwave radar systems development. At the close of World War II he returned to coaxial carrier repeater development until 1949, when he joined the radio research department. There his work has been in circular electric waveguide communication, microwave ferrite devices, and other components for microwave radio systems. As Director,

Guided Wave Research Laboratory, he heads a group engaged in research on communication techniques for the millimeter wave and optical regions. Fellow, IEEE; member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

F. W. MOUNTS, E.E., 1953, and M.S., 1956, University of Cincinnati; Bell Telephone Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, IEEE, Eta Kappa Nu.

VASANT K. PRABHU, B. E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu has been concerned with various theoretical problems in solid-state microwave devices, noise, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

DAVID J. PRAGER, B.A.E., 1961, New York University; M.S., 1965, and Ph.D. (Aeronautics and Astronautics), 1967, Stanford University; Bell Telephone Laboratories, 1967—. He has been engaged in analysis of thermal convection and of underwater acoustical transmission. More recently, Mr. Prager has been concerned with the fluid mechanics of nuclear blasts. Member, American Physical Society.

G. H. ROBERTSON, B.Sc., 1943, and Post Graduate Certificate (natural philosophy) 1948, University of Glasgow; Bell Telephone Laboratories, 1948—. Until 1958 Mr. Robertson was engaged in electronics research and a variety of electron tube development projects. Since 1958 he has been working on signal propagation and processing studies in the Underwater Research and Systems Departments. Associate member, IEEE; member, AAAS.

M. V. SCHNEIDER, M.S., 1956, and Ph.D., 1959, Swiss Federal Institute of Technology, Zurich, Switzerland; Bell Telephone Laboratories, 1962—. Mr. Schneider has been engaged in experimental work on thin-film solid-state devices, optical detectors, and microwave integrated circuits. Mr. Schneider is now working on hybrid integrated circuits at microwave frequencies and in the millimeter-wave fre-

quency range for solid-state radio systems. Member, IEEE, American Vacuum Society.

DAVID A. SPAULDING, A. B., 1959, M.S., 1960, Dartmouth College; M.S., 1961, Ph.D., 1965, Stanford University; Bell Telephone Laboratories, 1967—. Mr. Spaulding has been concerned with network studies for data transmission systems. Member, IEEE, Phi Beta Kappa, Sigma Xi.

W. H. WILLIAMS, B.A. (mathematics), McMaster University, 1954; M.S., 1956, and Ph.D., 1958 (statistics), Iowa State University. Before joining Bell Laboratories in 1960 he held faculty appointments at both universities. At Bell Laboratories he has worked on developing statistical methodology for survey sampling, time and cost analysis, and economic forecasting. He is a consultant to the U. S. Census Bureau and to the Executive Office of the U. S. President. Member, American Statistical Association, American Economic Association, Econometric Society, American Finance Association; Fellow, Royal Statistical Society.

JACK K. WOLF, B.S. in electrical engineering, 1956, University of Pennsylvania; M.S.E., 1957, M.A., 1958, and Ph.D., 1960, Princeton University; Bell Telephone Laboratories, 1968–1969. Mr. Wolf is an Associate Professor of Electrical Engineering at the Polytechnic Institute of Brooklyn; for the academic year 1968–1969 he was on a leave of absence to the Communications Theory Department at Bell Laboratories, Murray Hill, New Jersey. His main interests are in information theory, algebraic coding theory, and detection theory. Member, Tau Beta Pi, Sigma Xi, Sigma Tau, Eta Kappa Nu, Pi Mu Epsilon, IEEE, American Association for the Advancement of Science, American Association of University Professors.

C. P. WU, B.S., 1956, The National Taiwan University; M.S., 1959, and Ph.D., 1962, Ohio State University; Bell Telephone Laboratories, 1962—. Mr. Wu was an assistant instructor at the National Taiwan University during the 1956–57 academic year. He has done research in electromagnetic radiation in anisotropic media. His present work includes phased array antennas and development of numerical techniques for application in electromagnetic scattering and waveguiding problems. Member, IEEE, Sigma Xi.

CONTENTS

(Continued from front cover)