THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIII MAY 1954 NUMBER 3

P-N-I-P and N-P-I-N Junction Transistor Triodes

By J. M. EARLY

(Manuscript received March 18, 1954)

Theory indicates that the useful frequency range of junction transistor riodes may be extended by a factor of ten by a new structure, the p-n-i-p, which uses a thick collector depletion layer of intrinsic (i-type) semiconductor to reduce greatly the collector capacitance and to increase the collector breakdown voltage. This structure will permit simultaneous achievement of high alpha cutoff frequency, low ohmic base resistance, low collector capacitance, and high collector breakdown voltage. Because of the high breakdown voltages and larger areas per unit capacitance, permissible power dissipations appear much larger than for other high frequency junction types. Theoretical calculations indicate that oscillations at frequencies as high as 3,000 mcps may be possible.

Early exploratory models have verified the basic theory. Progress toward initial design objectives has been encouraging. In general, the observed performance has been consistent with the materials used and the structure achieved. The highest frequency of oscillation obtained to date is 95 mcps. Better performance is expected as technical control of materials and structures is improved.

In the five years since the announcement of the junction transistor by Shockley, great steps have been made in extending its useful frequency range and its power-handling capacity. Recent developments, particularly those which have increased the frequency range,[2, 12, 13] have brought the performance of practical devices close to ultimate limits

prescribed by structure and material. Further extension of frequency range and, to a lesser degree, of power capability must be sought in new materials or in improved structures. The p-n-i-p* transistor employs a new structure which in theory promises to increase the useful frequency range of junction triodes by a factor of at least ten. In the p-n-i-p, the n region of the base and the p region of the collector are separated by a relatively thick region of i-type (i.e., intrinsic or near-intrinsic, almost free of donor and acceptor centers) semi-conductor. This permits establishment of a thick collector depletion layer at relatively low voltages, thus producing low collector capacitance and several other desirable features.

The advantages of the new structure may be seen by study of the limitations of previous triode structures. In general, high frequency performance of conventional units, such as p-n-p alloy[1] transistors, is improved by making the base region thinner to increase the alpha cutoff frequency (f_α), by using lower resistivity base material to reduce the ohmic base resistance (r_b'), and by decreasing the area of emitter and collector junctions to reduce the collector capacitance (C_c). These equivalent circuit parameters are of nearly equal importance as may be seen from the gain-bandwidth expression discussed below.

The design changes required to improve the parameters involve conflicts, and compromises are necessary. For example, the decrease of base thickness which increases alpha cutoff frequency also increases (less rapidly) the ohmic base resistance.† The decrease in base resistivity which reduces base resistance also increases (again, less rapidly) the collector capacitance and decreases the collector breakdown voltage, thus decreasing power capacity. The reduction of junction area which decreases collector capacitance reduces the current rating and thereby the possible power rating. For transistors having circular electrodes, it may also increase the ohmic base resistance.

For these reasons, conventional junction triodes designed for high frequency application tend to be very small and to have very low voltage, current, and power ratings. Ultimately, the decrease of collector reverse breakdown voltage sets a lower limit to usable base resistivity and thereby to the thickness of the collector depletion region. This sets a lower limit on base region thickness, since average base layer thickness should be two or more times depletion layer thickness. For base layers thinner than this, irregularities in thickness or in impurity distribution may permit the depletion layer to contact the emitter, producing the

* And its homologue, the n-p-i-n.

† In the junction tetrode, this increase of base resistance is overcome by crowding the minority carrier emission close to one of the base contacts, thus producing low ohmic base resistance. See Reference 2.

ac collector-to-emitter short circuit effect called "electrical punch-through." Lower limits of junction areas are set by desired operating currents and by mechanical reasons. Diminishing returns are reached for structures a few mils in diameter and a fraction of a mil thick.

To facilitate comparisons, the limitations described qualitatively above have been interpreted quantitatively in terms of a gain-band figure of merit,*

$$\Gamma_0 \cdot B^2 = \frac{f_\alpha}{25 r_b' C_c} ; \tag{1.1}$$

in which Γ_0 is low frequency available power gain in the common emitter connection† and B is the frequency at which the gain is 3 db down from its low frequency value. A reasonable upper limit on this (power gain) $\times$ (bandwidth-squared) product is 4×10^{16}, which indicates that a 0–10 mcps video gain of 26 db may be obtained by improvement of conventional triode structures.

The same figure of merit, for a p-n-i-p of equal junction area, is approximately 10^{19}. Calculation shows that units may be designed to produce 10 db or more gain at 1,000 mcps. Although many of its operating principles are similar to those of the p-n-p and the n-p-n, the p-n-i-p differs from the earlier triodes in that low collector capacitance is obtained by means of a thick collector depletion (space-charge) layer of intrinsic semi-conductor. The section view of a p-n-i-p in Fig. 1 illustrates its major features. The wide depletion layer (electric field region) produces small collector capacitance (C_c) and gives a high reverse breakdown voltage, while the very thin base region of low resistivity gives simultaneously a low ohmic base resistance (r_b') and a very high alpha cutoff frequency (f_α). The design with four regions, emitter, base, depletion layer, and collector, increases the (power gain) $\times$ (band-squared) figure of merit ($f_\alpha/25 r_b' C_c$) about two decades, thus increasing the useful frequency range about one decade.

The thick collector depletion layer of intrinsic or near-intrinsic semiconductor provides advantages in addition to the reduction of the collector capacitance. Because base layer resistivity does not limit collector breakdown voltage as it does in previous structures, much lower base resistivities may be used, thus producing lower ohmic base resistances. Furthermore, the thick depletion region makes the structure much more rugged for very high alpha cutoff units since the very thin base layer is

* This figure of merit is essentially identical with one described by R. L. Prichard at the A.I.E.E. Winter Meeting in New York City, Jan. 22, 1954.

† It is assumed that the input terminals of the transistor are shunted by an external resistance which determines the input impedance and therefore the bandwidth. Power gain decreases approximately 6 db per octave at frequencies greater than B.

a surface layer on an 0.5–2.0 mil intrinsic layer, rather than a thin and unsupported web.

When operating biases are applied to a p-n-i-p transistor, holes injected at the forward biased emitter diode diffuse across the n region of the base then drift at high velocities through the field region to the reverse biased collector p region just as in a PNP transistor. However, in the p-n-i-p, the drift transit time through the collector field is comparable to the diffusion transit time through the base and contributes to phase shift of the short-circuit current-transfer ratio, alpha. In addition, the emitter depletion layer capacitance, C_{Te}, which is unimportant in previous triodes, is relatively large in the p-n-i-p and degrades performance at very high and microwave frequencies by providing a low impedance shunt around the emitter junction.

The details of structure and operation, design theory, a comparison of p-n-p and p-n-i-p units and some experimental results are discussed in the following sections. The concluding summary reviews the theoretical and experimental work.

STRUCTURE AND OPERATION

Impurity Distribution

In general, device characteristics depend on structure and on operating conditions. However, structure is more basic than operating conditions. The spatial distribution of fixed charge centers (donors and ac-

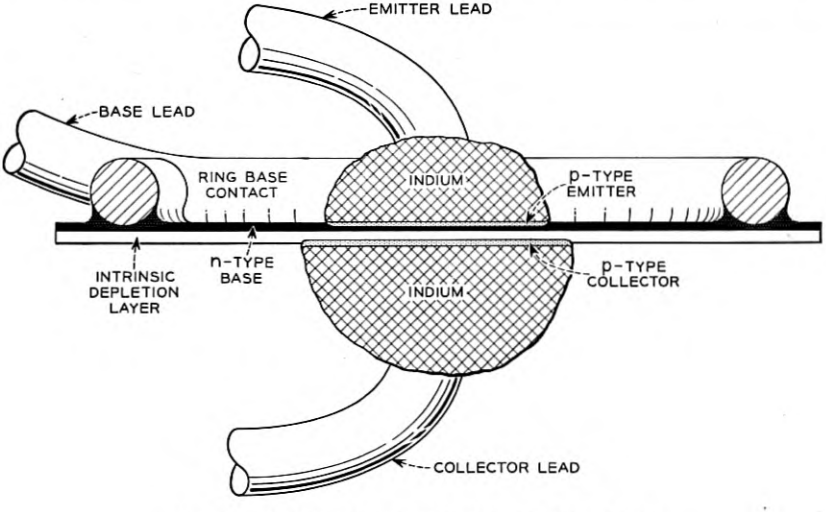
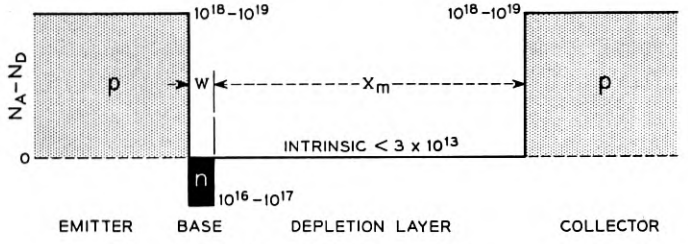


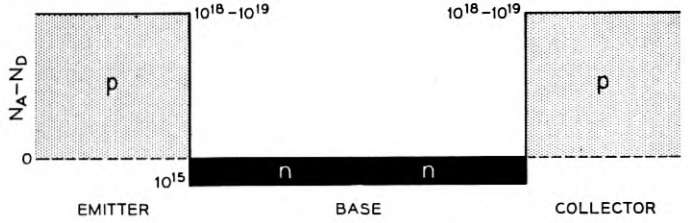
Fig. 1 — Sectional view of a p-n-i-p transistor.

ceptors) is the fundamental structural characteristic of the junction transistor. Fig. 2(a) shows an impurity density profile for a p-n-i-p along an axial line running through emitter, base, collector space-charge layer, and collector. Similar profiles for step junction (alloy) and graded junction (grown crystal) p-n-p's are shown in Figs. 2(b) and (c).

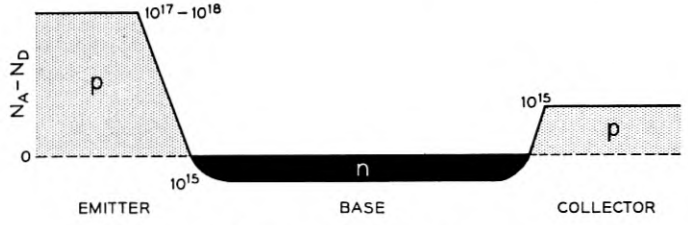
The emitter and collector regions of the p-n-i-p have very high impurity concentrations (low resistivities), while the impurity density in the base is moderately high and the depletion layer is almost free of impurities. The high acceptor density in the emitter forces most of the emitter current to flow as holes, giving an injection ratio (γ) close to unity. The high density in the collector gives a low collector body resistance and fixes the position of one face of the collector depletion layer.



(a) STEP-BASE p-n-i-p



(b) STEP (ALLOY) p-n-p



(c) GROWN (GRADED) p-n-p

Fig. 2 — Impurity density profiles.

The high donor concentration in the base region leads to low ohmic base resistance (r_b') and fixes the position of the base face of the depletion layer. In the depletion layer, the concentration of impurities is so low that the field region (space-charge layer) extends from the n-type base to the p-type collector at low voltages.

Depletion Layer

The properties of the depletion layer which are important at high frequencies are the capacitance across it (C_c) and the carrier transit time through it (τ_c). These are determined primarily by the impurity density, the thickness of the region, and the base-to-collector voltage. Potential and field distributions in the depletion layer for both small and

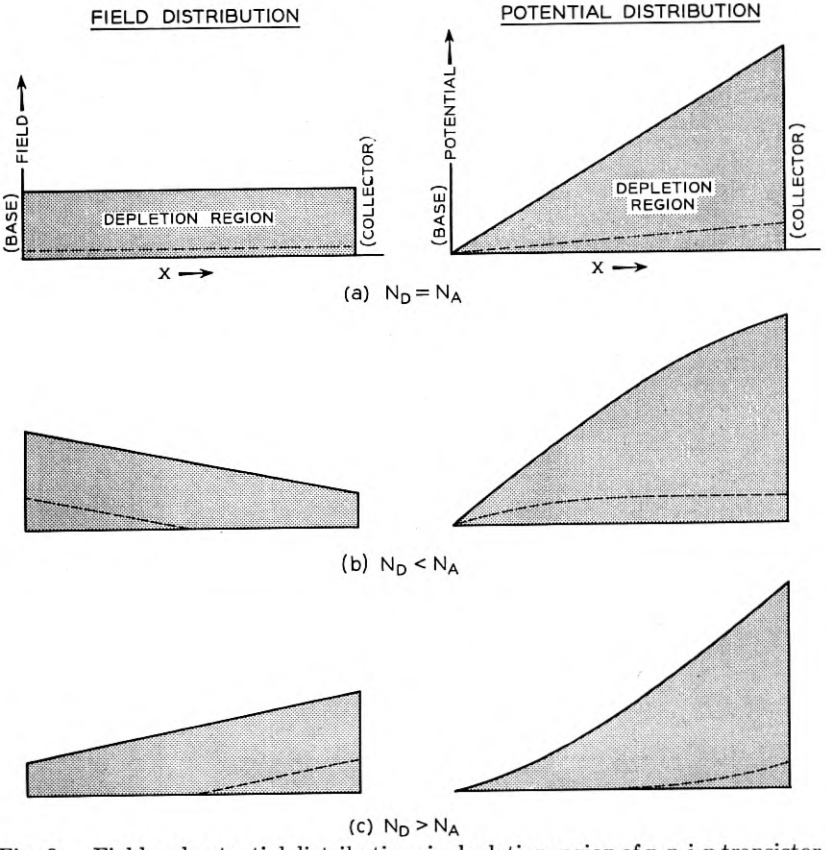


Fig. 3 — Field and potential distributions in depletion region of p-n-i-p transistor.

typical applied voltages are shown in Fig. 3 for p-n-i-p structures in which the depletion layer contains no net impurities (a), a small acceptor dominance (b), and a small donor dominance (c). When collector voltage is increased from zero, the space charge layer thickens until it extends from base to collector. Further increase of voltage simply increases the field strength in the region, without significant further increase in its thickness.

The capacitance initially changes inversely as the square root of collector potential, but becomes constant when depletion region thickness becomes constant. The time required for holes to drift from base to collector decreases with increase of depletion region field until scattering-limited carrier velocities are reached (about 5×10^6 cm/sec for holes, at 10,000 volts/cm).[10] It should be noted that normal operation does not occur until the depletion layer extends from base to collector (particularly if the depletion region is slightly n-type so that effective base thickness is large at low collector voltages, see Fig. 3(c)). The breakdown voltage of the collector is very high,* since the field strength in the depletion region is relatively uniform by comparison with that in older types of units, the region is wide, and strong fields are required to produce carrier multiplication.

Base Region

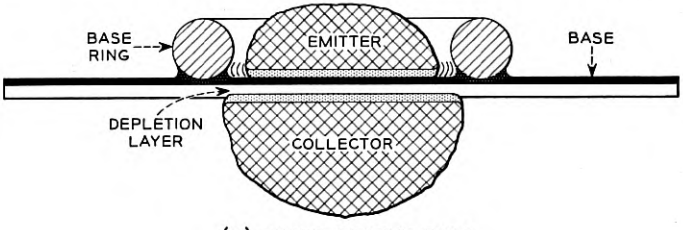
Base region design seeks the conflicting objectives of short diffusion transit time, requiring a thin region, and low ohmic base resistance, requiring a thick region. In practice, the region is made as thin as feasible, but of low resistivity material, and base contact geometry is chosen to minimize the ohmic resistance. In the p-n-i-p, very low base resistivity is practical, because the collector breakdown potential is fixed by the thickness of the intrinsic depletion layer rather than by the base resistivity as in fused junction p-n-p's.

The large donor density in the base region together with the very high frequencies of operation make the emitter depletion layer capacitance (C_{Te}) both larger and more important than in previous transistors. In order to reduce this capacitance, the emitter junction area is made small, thus leading to emitter current densities of 1 to 100 amperes/cm^2. In general, as the dc current density is increased, the minimum dc collector voltage must also be increased in order to preserve

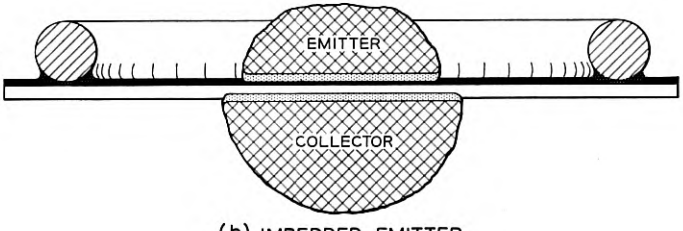
* An avalanche mechanism similar to a Townsend discharge in gases is now believed responsible for reverse voltage breakdown in junction structures. See Reference 3.

emission-limited current flow. Insufficient voltage may result in space-charge limited operation.[4, 5]

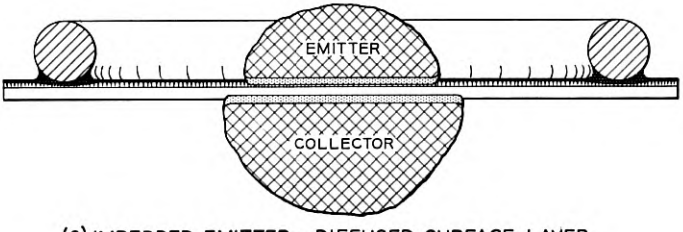
Three structures which may be used to obtain low ohmic base resistance are shown in Fig. 4. Obviously, the base contact ring may be placed arbitrarily close to the emitter, as in Fig. 4(a), so that the base resistance is that of the region beneath the emitter. Since this is somewhat difficult, the ring may be placed at a distance from the emitter, and the emitter imbedded in the base n-region as in Fig. 4(b), reducing the resistance between the emitter periphery and the base ring at only a small cost in alpha cutoff frequency. In addition, as shown in Fig. 4(c), the n-region used may be of graded resistivity such as results from impurity diffusion from the surface. The large impurity concentration at the surface minimizes both edge emission and radial base resistance.



(a) CLOSE-SPACED RING



(b) IMBEDDED EMITTER



(c) IMBEDDED EMITTER - DIFFUSED SURFACE LAYER

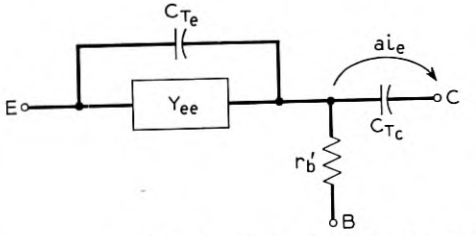
Fig. 4 — Low-base resistance structures.

These advantages are, however, balanced in part by an increase in the emitter depletion region capacitance associated with the low resistivity base material.

DESIGN THEORY

General

The principal objectives in the initial p-n-i-p design have been high alpha cutoff frequency, low collector capacitance, and low ohmic base resistance. The equivalent circuit employed is shown in Fig. 5. The output and feedback admittances which are important in earlier junc-



$$a \simeq \frac{a_0}{1 + i\,\frac{f}{f_a}}$$

$$Y_{ee} \simeq \frac{q I_e}{kT}\;\frac{\left[2\left(1-a_0+\frac{i\omega}{\omega_a}\right)\right]^{\frac{1}{2}}}{\tanh\left[2\left(1-a_0+\frac{i\omega}{\omega_a}\right)\right]^{\frac{1}{2}}}$$

Fig. 5 — Equivalent circuit of the p-n-i-p transistor.

tion triodes are omitted, since the space charge layer widening factor (H_{12} or μ_{ec}, $\frac{kT}{qw}\frac{\partial w}{\partial V_c}$) is very small.[6, 7] The transfer admittance is shown as a current generator (ai_e) with cutoff frequency ($|a^2| \sim 3$ db down) of f_α because this gives explicit recognition to base region diffusion transit time τ_b and allows it to be combined with space charge layer transit time τ_c.

Emitter Region Design

Emitter region acceptor concentration should be very large ($10^{18} - 10^{19}$ atoms/cc) in order to keep the injection ratio γ close to unity at both low and high frequencies.[8] At low frequencies, γ is determined by emitter resistivity and carrier life path or diffusion length, base resistivity and width, as

$$\gamma = \gamma_0 = \frac{1}{1 + \frac{\sigma_b w}{\sigma_e L_{ne}}}$$

At high frequencies, γ is determined by the ratio of acceptor density in the emitter to donor density in the base as

$$\gamma_{hf} = \frac{1}{1 + \frac{N_{Db}}{N_{Ae}} \sqrt{\frac{D_n}{D_p}}}$$

Obviously, since the effective donor density in the base must be large to give low ohmic base resistance, the effective acceptor density in the emitter must be even larger if high frequency γ is to be close to unity.

Base Region Design

Base region thickness, w, and the diffusion constant, D_p, determine the diffusion transit time for holes from injection by the emitter to collection by the field of the depletion layer.

$$\tau_b = \frac{w^2}{2D_p}. \tag{1}$$

For circular electrodes, which are useful, easily made, and easily analyzed, the ohmic base resistance for the active region of the base between emitter and collector depends on base resistivity, ρ_b, and base thickness as follows:

$$r_b' = \frac{\rho_b}{8\pi w} = \frac{1}{q\mu_n N_D 8\pi w}. \tag{2}$$

If w is made small, r_b' can be reduced only by making N_D large. Although large reductions in r_b' can be made, increasing N_D is ultimately a self-defeating procedure for several reasons: as N_D is increased both D_p and the electron mobility, μ_n, decrease, thus increasing hole transit time and also partially off-setting the reduction in r_b' by N_D. In addition, the capacitance of the emitter depletion region varies approximately as $N_D^{1/2}$, thus diverting more ac emitter current from hole injection. This capacitance is

$$C_{Te} = \kappa\epsilon_0 \left(\frac{qN_d}{2\kappa\epsilon_0 V_e'}\right)^{1/2} A_e, \tag{3}$$

where V_e' is the average electrostatic potential across the emitter depletion layer. Equations (1) to (3) show the conflicts which necessarily arise in base region design for very high frequencies. The limiting design combines very small w, large N_D, small emitter area A_e, and relatively

large dc emitter current I_e so that the minority carrier emitter admittance y_{ee} is at least of the order of magnitude of $j\omega C_{Te}$. Total emitter admittance is

$$y_{ee} + j\omega C_{Te} = \frac{\frac{qI_e}{kT}(1 + j\omega\tau)^{1/2} \coth \left[\frac{(1 + j\omega\tau)w^2}{(D\tau)}\right]^{1/2}}{\coth \left(\frac{w^2}{D\tau}\right)^{1/2}} + j\omega C_{Te} \quad (4)$$

Depletion Layer Design

As mentioned previously, the most important characteristics of the depletion layer in the p-n-i-p are the transit time for holes, τ_c, and the capacitance, C_{Tc}. The minimum voltage for normal operation, V_{min}, and maximum or breakdown voltage, V_{max}, are also significant.

The minimum voltage for "normal" operation is reached when the electric field between the n-type base and p-type collector is strong enough so that the holes drift at their limiting velocity of 5×10^6 cm/sec.* The collector to base voltage required for normal operation is the product of the minimum field strength for the limiting velocity and the thickness of the depletion layer and is given by

$$V_{min} = 10{,}000\, x_m \quad (5)$$

in which x_m is depletion layer thickness in cm. The maximum field obtainable before reverse voltage breakdown is not known exactly, but is in practice near 100,000 volts/cm, so that

$$V_{max} \simeq 100{,}000\, x_m . \quad (6)$$

Depletion layer capacitance is nearly independent of collector voltage in normal operation and is inversely proportional to layer thickness.

$$C_{Tc} = \frac{\kappa\epsilon_0 A_c}{x_m} \quad (7)$$

Transit time for holes increases directly with layer thickness, however, being

$$\tau_c = \frac{x_m}{5 \times 10^6} \quad (8)$$

Since increase of τ_c decreases the alpha cutoff frequency f_α, the choice

* At lower field strengths, the transit time for holes is longer, giving a lower alpha cutoff frequency. The "normal" is the best, rather than the only possible, operating condition.

of x_m is a design balance between C_c and f_α, with any desire for low voltage operation weighting the scales toward smaller x_m.

As collector voltage and therefore field strength is reduced below that required for normal operation, transit time is increased because of the reduced drift velocity. In addition, the holes in transit interact with the ac field of the layer, thus increasing the output conductance g_{cc}. Further, a larger density of holes in the layer is required to carry the same current, disturbing the field distribution. If the voltage is reduced greatly, space-charge limited emission may occur,[4, 5] producing much longer effective transit times.

If output voltage is reduced sufficiently, the collector field will not extend all the way from base to collector. If the layer is somewhat n-type, the field region collapses toward the p-region of collector. If it is somewhat p-type, the field collapses toward the n-region of the base. The latter arrangement has the advantage that f_α is less drastically reduced. Further, in normal operation, the negatively charged acceptor atoms of a slightly p-type layer will neutralize the charge of the holes in transit, thus making the field more nearly constant from collector to base. The effects of low voltage on the collector field distribution are indicated approximately by the dashed lines of Fig. 4.*

Collector Region Design

Acceptor concentration in the collector should be large for several reasons. This gives a low collector body resistance, which virtually eliminates internal series loading of the collector, and it aids operation by fixing the position of the collector edge of the depletion layer. The advantages obtained may be seen by considering a unit in which the collector body is made somewhat p-type and a collector contact is attached at some distance from the depletion layer. If 10 ohm-cm p-material is used for the collector body and a collector contact fastened 2.5 mils from the collector resistance of 250-500 ohms will result. In addition, because of the weak drift field at the collector edge of the depletion layer, the hole transit time is about twice that for a true p-n-i-p.

Alpha Cutoff Frequency

A current transmission cutoff frequency f_α for the p-n-i-p is given approximately by†

* The field distributions occurring in an intrinsic depletion layer at low field strengths have been discussed in Reference 11.

† It is assumed that alpha is given by $\alpha = \alpha_0(1 + jf/f_\alpha)$. Equation (9) represents the phase of this expression quite well, but the amplitude rather poorly.

$$f_\alpha = \frac{1}{2\pi(\tau_b + \tau_c/2)} \tag{9}$$

Equation (9) implies (correctly) that the delay time for total current passing through the depletion layer is about one-half the transit time for the carriers. This results from the induction of charge on the base and collector electrodes by the carriers in transit. If $\varphi = \omega\tau_c$ is carrier transit angle and $J_c = e^{j\omega t}$ is the *conduction* current of holes entering the depletion layer from the base, the *total* current entering the depletion layer from the base can be shown to be

$$J = e^{j\omega t}\left(\frac{1 - e^{-j\phi}}{j\phi}\right) \tag{10}$$

which reduces for small φ to

$$J \simeq e^{j\omega t - j\phi/2}$$

It may be noted that the total current J of equation (3.6-2), when written in the form $J_{\max} \angle\theta$ in which θ is the phase shift of the total current with respect to the conduction current entering from the base, is approximately $0.973 \angle -22.5°$ for $\varphi = 45°$, $0.901 \angle -45°$ for $\varphi = 90°$, and $0.636 \angle -90°$ for $\varphi = 180°$.

DESIGN COMPARISON

General

Comparison of figures of merit is the best, albeit unsatisfactory, means for comparative evaluation of devices. For junction transistors, one non-controversial figure of merit is established — the noise figure. Two transmission figures of merit for junction transistors are suggested at the bottom of Table I. It should be pointed out that the p-n-i-p figures are for theoretical design possibilities, some features of which have already been realized experimentally.

The Units

Table I gives parameters of interest for several types of transistors. Structural, material, and electrical parameters for the Bell Telephone Laboratories' developmental M1778 p-n-p unit are averages for large numbers of units. The electrical parameters of the plated-contact transistor recently announced by Philco were taken from a talk by W. H. Forster before the Philadelphia I.R.E., Dec. 3, 1953.[12] The structural and material parameters have been estimated. The p-n-i-p structures

TABLE I — TRANSISTOR DESIGNS

	M1778	Philco	P-N-I-P(Calculated Values)		
			No. 1	No. 2	No. 3
w_b — mils	1.0	0.2	0.13	0.8	0.04
ρ_b — ohm cm	1.5	0.5	0.14	0.05	0.02
dia_e — mils	15	4	10	6	5
dia_c — mils	30	6	15	8	5
x_m — mils	0.1	0.05	0.63	0.36	0.7
N_b — atoms/cc	10^{15}	3.5×10^{15}	1.4×10^{16}	4.2×10^{16}	1.2×10^{17}
f_α — mcps	2.0	55	100	200	360 (600)*
r_b' — ohms	50	65	34	20	16
C_c — mmf	25	2.5	1.0	0.5	0.1
$\omega_a C_c$ — mhos	—	—	0.023	0.038	0.102
C_e — mmf	—	—	36	22	27
$\omega_a r_b' C_c$	0.0157	0.056	0.0214	0.0126	0.0060
$(f_\alpha/25\ r_b' C_c)^{1/2}$ — mcps	8	115	340	900	3000

* First value calculated by Equation (9); second value is for diffusion through base n-region only (i.e., $\tau_c = 0$).

and materials were assumed and electrical parameters were calculated from them by the Equations (1) to (11). Mobilities measured for low resistivities by M. B. Prince[9] were used in the calculations.

Figures of Merit

The last row of Table I gives $(f_\alpha/25 r_b' C_c)^{1/2}$, which was discussed previously as a gain-bandwidth figure of merit for a broad band common emitter amplifier. It is also related to the maximum frequency at which reliable oscillations may be obtained. The figure of merit $\omega_\alpha r_b' C_c$ is the open circuit voltage feedback ratio at the alpha cutoff frequency and gives some indication of the balance between the two time constants, $1/\omega_\alpha$ and $r_b' C_c$. It is also approximately the ratio of input impedance to output impedance in a common emitter broadband amplifier at high frequencies.

Comments

It should be noted that the emitter depletion layer capacitance is significant in all the p-n-i-p designs and that barrier transit time reduces alpha cutoff frequency some forty per cent in the highest frequency design. Despite this, it is probable that p-n-i-p or n-p-i-n germanium junction triodes will serve as oscillators and perhaps amplifiers at frequencies as high as 3,000 mcps.

EXPLORATORY MODELS

Objectives

While the p-n-i-p transistor will be useful for high voltage and high power operation, our exploratory development work has been directed toward good performance at very high frequencies. The initial electrical objectives set were those of p-n-i-p No. 1 of Table I: $f_\alpha = 100$ mcps, $C_c < 1.0$ mmf, and $r_b' = 34$ ohms. The base thickness of 0.13 mil and base resistivity of 0.14 ohm-cm are the critical structural parameters.

Fabrication

Although p-n-i-p's might conceivably be built in a single operation, one procedure used has two major parts. The first is the production and evaluation of 2-mil thick wafers of intrinsic germanium with a skin or surface layer of 0.1–1.0 ohm-cm n-germanium 0.3-0.5 mils thick. The second step is the alloying of collector, emitter, and base electrodes to these wafers.

Wafers with n-type skins have been made by three methods. Intrinsic crystals growing from a melt by the Teal-Little technique have been doped with arsenic, grown for a few seconds longer (another 0.5–1.0 mils), and snatched mechanically from the melt. The resulting crystal surface has a mirror finish and is relatively flat. N-type skin layers have also been produced by alloying the wafer surfaces with lead-arsenic and lead-antimony mixtures and by the diffusion of arsenic into wafer surfaces.

Collector and emitter electrodes are alloyed by the indium germanium process with times, temperatures, and quantities of indium selected to give desired alloying depths. Ring-base connections of antimony and gold plated kovar have been used.

Measurements

Progress toward the initial design objectives mentioned previously has been encouraging. The predicted behavior has been verified semi-quantitatively. The capacitance of a 15-mil diameter collector is usually less than 1.0 mmf at $V_c = -25$ volts as predicted in design No. 1 of Table 1. Ohmic base resistances generally less than 50 and as low as 5 ohms have been measured. However, the highest alpha cutoff frequency obtained as yet is 25 mcps. This has been limited primarily by the thickness of the base layer. At present this is of the order of 0.30 mils so that an alpha cutoff frequency of 25 mcps is about what would be predicted. Further development of the technology of fabrication seems reasonably

straight-forward at least to the design objectives of No. 1 and No. 2 of Table I.

The best unit measured to date showed $\alpha_0 > 0.96$, $f_\alpha \simeq 25$ mcps, $r_b' \simeq 60$ ohms, and $C_c \simeq 1.8$ mmf. These values agree quite well with those expected from the resistivities and layer thicknesses employed. The unit oscillated at 95 mcps with $V_c = -30$, $I_e = 1.0$ ma. Connected in a common emitter video amplifier working from a 75-ohm generator impedance into a load resistance of 2,150 ohms shunted by 5 mmf of capacitance, this unit produced a power gain of 23 db at 500 kc, falling to 20 db at 3 mcps and 15 db at 10 mcps.* In an uncompensated common emitter tuned circuit, this unit gave 20.5 db at 10 mcps with 3 mcps bandwidth between the three db points.* It has been operated with a collector voltage of -90 volts.

SUMMARY

The designed elimination of donors and acceptors from a thick collector depletion layer introduces a new design variable in junction transistor triodes. The new structure (p-n-i-p or n-p-i-n) is believed capable of development into the microwave frequency range. Several factors which were of second order importance in p-n-p and n-p-n units such as emitter depletion layer capacitance and collector transit times become significant in limiting ultimate performance. The thick depletion layer permits operation at higher voltages than were previously possible in any but low frequency units.

Moderately good results have been obtained already. Units having 10 mil emitter diameter, 15 mil collector diameter have produced stable gains without compensation of 20.5 db at 10 mcps and have oscillated at 95 mcps.

The junction transistor now promises to be a serious competitor to high vacuum triodes over a much larger range of frequencies and power levels than before.

ACKNOWLEDGMENTS

J. A. Morton and R. M. Ryder have strongly supported and encouraged this work. J. W. Peterson and W. C. Hittinger have collaborated in and contributed to the experimental studies. The models constructed and tested are the products of the persistent efforts and many useful suggestions of J. A. Wenger, J. McGlasson, and L. P. Meola. Many others, particularly those engaged in semiconductor materials research

* These measurements were made by L. G. Schimpf.

and development, have also assisted us. Discussions with colleagues have been most helpful in preparation of this report.

REFERENCES

1. J. S. Saby, Fused Impurity p-n-p Junction Transistors, I.R.E. Proc., **40,** pp. 1358–1360, Nov., 1952.
2. R. L. Wallace, Jr., L. G. Schimpf and E. Dickten, A Junction Transistor Tetrode for High-Frequency Use, I.R.E. Proc., **40,** pp. 1395–1400, Nov., 1952.
3. K. G. McKay, and K. B. McAfee, Electron Multiplication in Silicon and Germanium, Phys. Rev., **91,** pp. 1079–1084, Sept. 1, 1953.
4. W. Shockley, and R. C. Prim, Space-Charge Limited Emission in Semiconductors, Phys. Rev., **90,** pp. 753–758, June 1, 1953.
5. G. C. Dacey, Space-Charge Limited Hole Current in Germanium, Phys. Rev. **90,** pp. 759–763, June 1, 1953.
6. J. M. Early, Effects of Space-Charge Layer Widening in Junction Transistors, I.R.E. Proc. **40,** pp. 1401–1406, Nov., 1953.
7. J. M. Early, Design Theory of Junction Transistors, B. S. T. J., **32,** pp. 1271–1312, Nov., 1953.
8. W. Shockley, M. Sparks and G. K. Teal, The p-n Junction Transistors, Phys. Rev., **83,** p. 151, July, 1951. See also Reference 7. Nov. 1953, op cit.
9. M. B. Prince, Drift Mobilities in Semiconductors. I. Germanium. Phys. Rev., **92,** pp. 681–687, Nov. 1, 1953.
10. E. J. Ryder, Mobilities of Holes and Electrons in High Electric Fields, Phys. Rev., **90,** p. 766, June, 1953.
11. R. C. Prim, D. C. Field in a Swept Intrinsic Semiconductor, B. S. T. J., **32,** pp. 665–694, May, 1953.
12. W. E. Bradley, et al, The Surface Barrier Transistor, I.R.E. Proc., **41,** pp. 1702–1720, Dec., 1953.
13. C. W. Mueller and J. I. Pankove, A p-n-p Alloy Triode Transistor for Radio Frequency Amplification, RCA Review, **14,** pp. 586–598, December, 1953.

Arcing of Electrical Contacts in Telephone Switching Circuits

Part III—Discharge Phenomena on Break of Inductive Circuits

By M. M. ATALLA

(Manuscript received November 16, 1953)

This is a presentation of a study of the discharge phenomena occurring between contacts on break of an inductive load. The main objectives are: (1) to forward some detailed explanations of the main components of a break transient in terms of basic conduction and emission processes, and (2) to establish the conditions that determine the nature of the transients. The study covered the following: (1) occurrence of interrupted and steady arcs, (2) initiation of reversed arcs in one breakdown, (3) arc initiation under dynamic conditions, (4) initiation and maintenance of glow discharge, and (5) glow-arc transitions.

INTRODUCTION

An important phase in the study of discharge phenomena between contacts is that involving the break of an inductive circuit. A typical switching circuit in its simplest form consists of a battery in series with a coil (electro-magnet), a cable or lead and a pair of contacts. Coils now in use may have inductances of the order of tens of henries and may store as much energy as 10^6 ergs. On break of the circuit an appreciable portion of this energy may be dissipated between the contacts through a steady arc, a series of interrupted arcs, a glow discharge or any of their combinations. In most cases, the energies involved are too high to provide satisfactory contact life from the standpoint of electrical erosion.

The discharge transients obtained are usually complex in nature.[1] A close examination of these transients reveals a great deal of rather curious effects that have not been previously considered in detail. This is a presentation of a recent study of the break transient with the primary objective of furnishing some explanation of the more pertinent phenomena involved in terms of the basic concepts of surface emission and gas conduction.

NOTATION

a Arc radius or equivalent characteristic length of cross section
c Local capacitance at the contacts
e Electron charge
i_a Current density in the arc
i_{th} Thermionic emission current density
i_{ng} Normal glow current density
i_{ag} Abnormal glow current density
$(i_{g\text{Limit}})$ Limiting glow current density preceding glow-arc Limit transition
k Boltzman constant
l Local inductance at the contacts
m Mass of contact metal atom
n Number of consecutive arcs in *one* breakdown
r Resistance of the local contact circuitry
s Separation between the contacts
t Time
t_{ch} Charging time between breakdowns
t_{dei} Deionization time following an arc
t_g Glow duration
u_s Velocity of contact separation
u_{ch} Charging velociy defined as s/t_{ch}
u_{at} Velocity of the metal atoms
v Arc voltage
$\bar{v}_n$ Residual voltage at the contacts following a breakdown of n-consecutive arcs
z Impedance $(l/c)^{1/2}$
A Constant in the thermionic equation
A_a Area of arc spot
C Circuit capacitance
E Battery voltage
F Field strength
I Current
I_g Current in a glow discharge
I_m Minimum arcing current
I_o Initial closed circuit current
L Circuit inductance
R Circuit resistance
T Absolute temperature
T_b Absolute boiling temperature

T_o Absolute initial temperature
V Voltage
V_{ai} Arc initiation voltage
V_{gi} Glow initiation voltage
V_g Voltage drop across the contacts wtih normal glow
α Thermal diffusivity
φ Work function
ω Angular frequency $(lc)^{-1/2}$

GENERAL

A typical circuit consisting of a battery, a coil of an electro-magnet, a cable or lead and a pair of contacts is shown in Fig. 1(a). Due to the usual magnetic core of the coil, this circuit presents some unnecessary complications in making interpretations of the observed contact phenomena. Since our main objective is an understanding of the basic phenomena occurring between the contacts, it appeared justifiable to restrict our work to circuits and circuit elements that lend themselves to simple treatment. Figure 1(b) shows the circuit used in most of this work. All coils used have air cores.

When the contacts are closed, a steady state current $I_o = E/R$ is established in the circuit. At the first physical separation between the contacts, the circuit current will charge the capacitance C causing a voltage rise at the contacts at an initial rate of I_o/C. In the meantime, the separation between the contacts will increase. The first breakdown will occur when the voltage across the contacts first reaches or exceeds the arc initiation voltage corresponding to the separation attained, the atmosphere involved and the *contact surface condition*. Fig. 2 represents diagrammatically the occurrence of the first discharge. abc is the arc initiation voltage versus separation line for a "normal" contact.[2] The

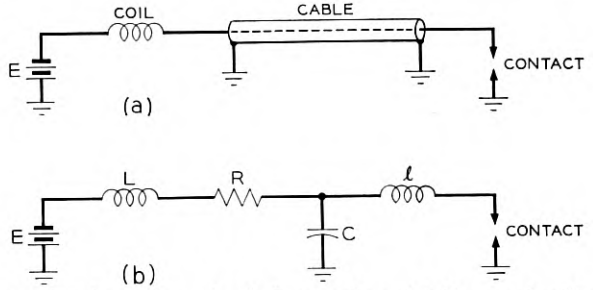


Fig. 1 — (a) Typical relay circuit in practice. (b) Linear circuit used in this study.

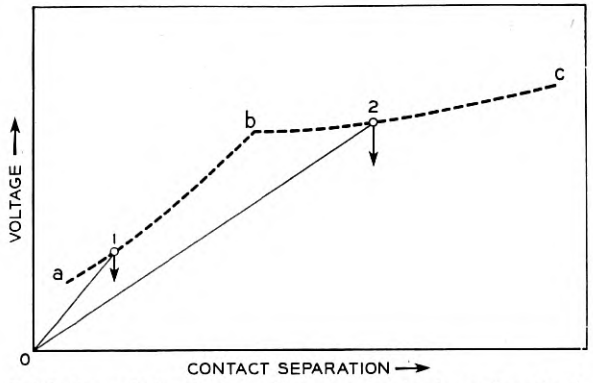


Fig. 2 — Initiation of the first arc between contacts on break of an inductive circuit.

portion bc corresponds to the sparking potentials in the atmosphere. ab corresponds to the range of small separations, of the order of or less than the mean free path of an electron in the atmosphere, where the arc is initiated by field emission through the influence of surface contamintions or films. As was shown in Reference 2, when the cathode surface was carefully cleaned, the constant field line was not obtained and the arc was initiated at the minimum sparking potential of the atmosphere. It occurred on the sides of the contacts along a path much longer than the minimum separation between the contacts.*

Lines 0–1 and 0–2 represent the voltage rise at the contact with small and large shunt capacities. Points 1 and 2 are the respective first discharge points. In the first case, the arc is initiated at a smaller separation and higher field strength without direct influence of the atmosphere. In the second case the arc is initiated at a lower field strength at the spark potential of the atmosphere.†

The first arc established may or may not be maintained depending on conditions that are discussed in the next Section. When an arc is inter-

* With Pd contacts a gross field of 20×10^6 volts/cm was reached between clean contacts without initiating an arc along the shortest gap. According to the Fowler-Nordheim equation a field of about 50×10^6 volts/cm is required to give the necessary initiatory electrons. It is possible, however, that before such a high field is attained a metal bridge is pulled electrostatically[3] to short the gap. The electrostatic stress is roughly given by $0.5 \times 10^{-12} F^2$ Kg/cm^2 where F is the field strength in volts/cm. At $F = 50 \times 10^6$ volts/cm, the stress is 1250 Kg/cm^2 which may exceed the yield stress for the contact metal.

† The first arc may be initiated at an appreciably lower voltage than predicted by the above static consideration. The first break at the contacts usually follows the explosion of molten bridge drawn between the contacts. Thermionic emission can then furnish the initiatory electrons of the arc. This is only possible, however, if the voltage across the contacts exceeds the ionization potential of the metal atoms before excessive cooling of the cathode has occurred.

rupted, it is followed by a recharging process to a new arc initiation voltage when a second arc is initiated. Under certain conditions, the second arc may be initiated at a lower voltage than the first arc due to residual effects of the first arc which may alter the conditions in the gap. This effect is discussed later.

A transient on break with a series of interrupted arcs is shown in Fig. 3. The first arc was initiated at 230 volts and a gross field of 2.5×10^6 volts/cm. All the following arcs were initiated at the spark breakdown potentials in air corresponding to the separations involved. Fig. 4 shows a transient where the arc was sustained with occasional interruptions.

In addition to arcing, one may obtain glow discharge. Fig. 5 shows a transient where glow discharge predominates. Glow initiation and glow-arc transitions are discussed in a later Section.

Fig. 6 shows the methods used for current and voltage measurements. As indicated, direct voltage measurements at the contacts were avoided to eliminate the unnecessary complications of the measuring circuit.

INTERRUPTED ARCS

Conditions for Obtaining Interrupted Arcs

A breakdown from a voltage V_{ai} into an arc corresponds to a rapid voltage drop at the contacts from V_{ai} to the arc voltage v. For most prac-

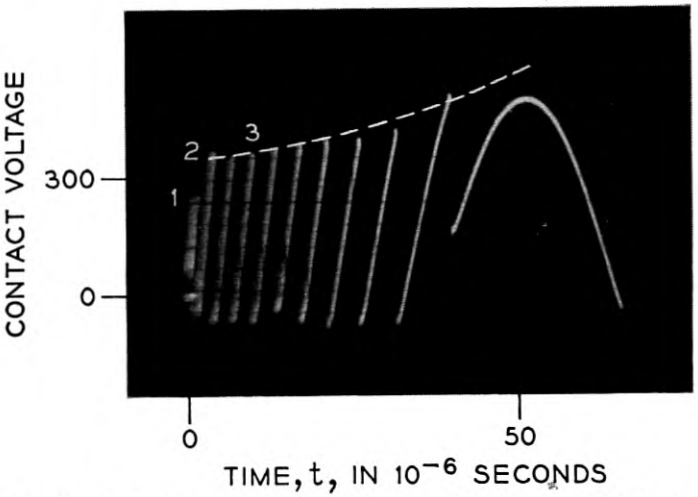


Fig. 3 — Typical contact voltage transient on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.2$ henry, $R = 950$ ohms and $C = 510 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

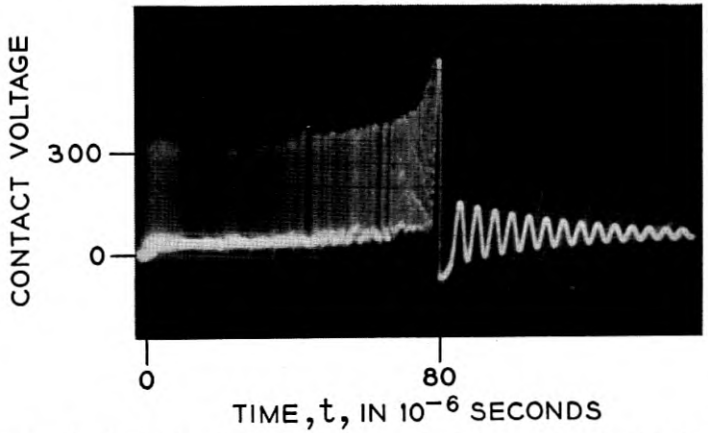


Fig. 4 — Contact voltage transient with sustained arc on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.025$ henry, $R = 115$ ohms, $C = 20 \times 10^{-12}$ farad. Velocity of contact separation 40 cms/sec.

tical purposes one may neglect the voltage drop time which is the initiative period of the arc. For the circuit in Fig. 1b, the current through the arc is the summation of the main circuit current and the transient current from the l-c circuit. The transient current is $(V_{ai} - v)(\frac{c}{l})^{1/2} \sin {}^{t}/(lc)^{1/2}$. Fig. 7, (a) and (b), represent diagrammatically the voltage and current transients for lumped and distributed circuits. In both cases the arc is

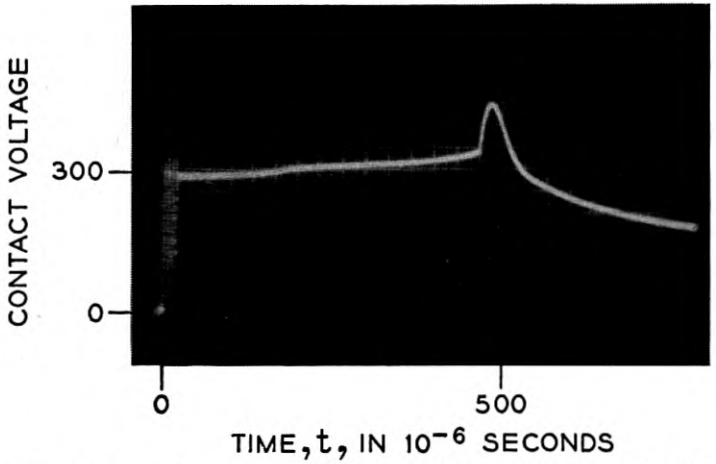


Fig. 5 — Contact voltage transient with glow discharge on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, 700 ohms relay coil and $C = 200 \times 100^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

terminated when the current drops to the minimum arcing current I_m. It is evident that the condition for obtaining an interrupted arc is:

$$I_0 - (V_{ai} - v)\left(\frac{c}{l}\right)^{1/2} < I_m \qquad (1)$$

It may be pointed out that surface contamination, such as organic activation, tends to decrease both I_m and V_{ai}[2, 4]. According to equation 1, one may conclude that contact surface contaminations usually tend to cause a transition from an interrupted arc transient to a steady arc transient. The latter is usually associated with appreciably higher energy dissipation between the contacts and much lower contact life due to erosion.

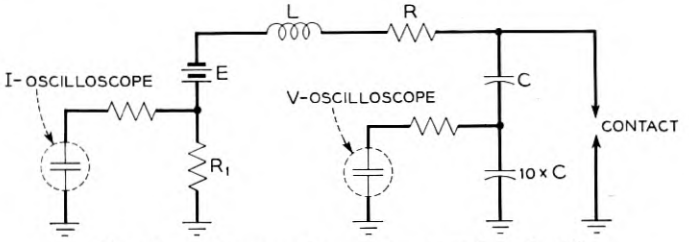


Fig. 6 — Voltage and current measuring circuit.

Residual Voltage Following an Interrupted Arc*

At the interruption of the first arc the voltage at the contact is v, the arc voltage, and the voltage at the capacitor C, Figure 1b, is $\bar{v}_1$ which is usually negative. If the local contact circuit is non-dissipative, the residual voltage is $\bar{v}_1 = 2v - V_{ai}$. For a dissipative circuit with a resistance r corresponding to the frequencies involved:

$$\bar{v}_1 = v - (V_{ai} - v)e^{-(\pi/2)\cdot(r/z)} \qquad (2)\dagger$$

for an oscillating circuit, as is usually the case, where $z = (l/C)^{1/2}$. The capacitor C at $\bar{v}_1$ will then recharge the local contact capacity c, $c \ll C$, through the inductance l. If the voltage attained at the contacts is sufficient and the conditions in the gap and at the contact surface are favorable, a reversed arc may be re-initiated, as previously discussed. This process may repeat several times and the residual voltage $\bar{v}_n$ will change sign and decrease progressively. At the end of n arcs, it can be shown that the residual voltage $\bar{v}_n$ is given by:

* The term "recovery has also been used in the literature.

† Equation 2 and 3 are valid only for small values of r/z. These are approximations of the more general expression given by Germer.[14]

$$(-1)^n \bar{v}_n = v + (V_{ai} - v)e^{-(\pi/2)\cdot(r/z)\cdot n} - 2v \sum_{n=0}^{n=n-1} e^{-(\pi/2)\cdot(r/z)\cdot n} \quad (3)$$

This equation indicates that $\bar{v}_n$ is negative for odd numbers of arcs and positive for even numbers of arcs.* If r/z is neglected, Equation 3 is reduced to

$$(-1)^n \bar{v}_n = V_{ai} - 2vn \quad (3a)$$

For $V_{ai} = 300$ volts and $v = 14$ volts, the residual voltages following the first four arcs are respectively -272, $+244$, -216 and $+188$. These

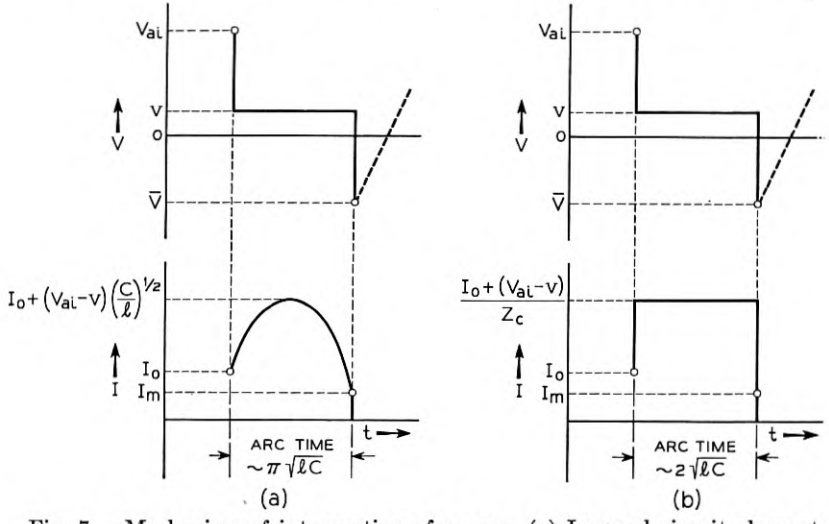


Fig. 7 — Mechanism of interruption of an arc. (a) Lumped circuit elements (b) Distributed elements.

values are numerically higher than measurements due to neglecting the term r/z. For the circuits used in our experiments r/z ranged between 0.1 and 0.5 and as many as 4 or 5 consecutive arcs have been obtained in one breakdown. Figure 8 shows a transient with both positive and negative residual voltages corresponding to even and odd number of arcs respectively.†

* Except when $\bar{v}_n$ is not too much higher than the arc voltage v.

† The following alternative explanation for the occurrence of high positive residual voltage was considered: the first arc may be extinguished by the formation of a metal bridge due to the arc[2]. This may occur before the capacitor C has attained a negative voltage. This possibility, however, was eliminated. From the measured residual voltages the energies in the arcs were calculated. The heights of the bridges produced were computed (reference 2) and were found to be too small compared with the contact separations.

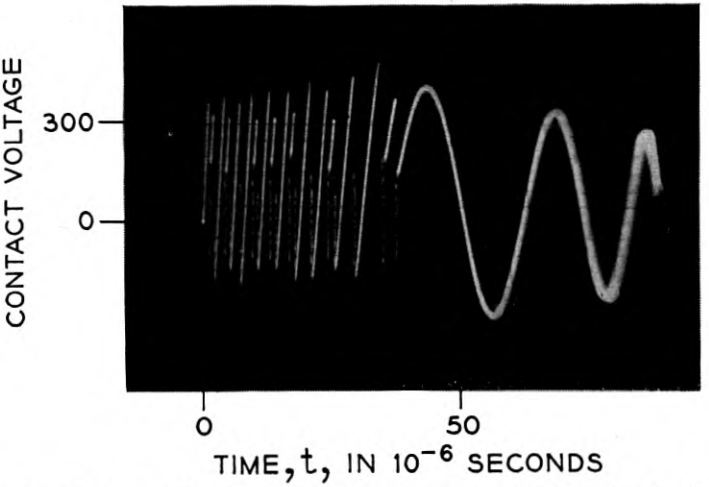


Fig. 8 — Contact voltage transient with interrupted arcs on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.010$ henry, $R = 40$ ohms and $C = 900 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

Initiation of Reversed Arcs in One Discharge

In one breakdown from a voltage V_{ai} it is commonly observed that a succession of reversed arcs may be obtained. It was shown in equation 3 that the residual condenser voltage $\bar{v}_n$ progressively decreases, numerically, with the number of arcs n. Following the interruption of the first arc, the condenser voltage is $-|\bar{v}_1|$ and the contact voltage is $+v$, the arc voltage. The capacity C will then recharge the local capacitance at the contact through a small lead inductance l. If the circuit resistance is neglected, the maximum voltage the contact will acquire is $-(2|\bar{v}_1| + v)$. If this equals or exceeds the original arc initiation voltage V_{ai}, a second arc is obtained. For illustration, consider a breakdown initiated at $V_{ai} = 300$ volts and $v = 14$ volts. From Equation 3, $\bar{v}_n$ was calculated for the first four arcs at $r/z = 0.0$ and 0.2 The corresponding maximum contact voltages acquired after each arc were also calculated and the results are given in Table I. For $r/z = 0$, column 3, one may obtain, according to this simple circuit consideration, more than 4 arcs, actually 5. For $r/z = 0.2$, which is a reasonable practical value, only 2 arcs may be obtained, column 5, since following the second arc the maximum voltage attained at the contacts is only 256 volts which is less than the initial arc initiation voltage.

It is possible in some cases, however, to obtain a few more arcs than

TABLE I — INITIATION OF REVERSED ARCS BY OVERCHARGING OF CONTACT CAPACITANCE

(Calculated)

$\frac{r}{z} = 0$			$\frac{r}{z} = 0.2$	
(1) Arc No.	(2) $\bar{v}_n$	(3) Max. Cont. Voltage	(4) $\bar{v}_n$	(5) Max. Cont. Voltage
1	−272	−558	−195	−404
2	+244	+502	+121	+256
3	−216	−446	−60	−134
4	+188	+390	+22	+58

V_{ai} = 300 volts, v = 14 volts.

predicted above. These additional arcs have appeared to be initiated at lower voltages than the first arc. This is undoubtedly due to the residual surface and gap effects of the previous arc.* These are discussed in the following section.

Arc Initiation Under Dynamic Conditions — Introduction

In Reference 2 measurements have been presented of the arc initiation voltage between contacts at different separations and surface conditions. These tests are "static" in the sense of allowing enough time to elapse between two arcs to obtain a complete reconditioning of the contact surfaces and gap. With successive arcing, as obtained on break of an inductive circuit or during one breakdown, it was observed that the arc may be initiated at appreciably lower voltages compared with static test results.

One arc may enhance the initiation of a shortly following arc possibly through the effects of: residual ions in the gap or on a cathode surface film, residual metal atoms in the gap and residual thermionic emission. Exactly how each of these effects can enhance the initiation of the arc can be determined only after an understanding of the mechanisms of initiation of the first arc, its maintenance and its termination. It is in order at this point to present a sketchy outline of some plausible mechanisms which are largely of speculative nature. This discussion is also limited to short arcs initiated and maintained with no direct influence of the surrounding atmosphere.

* The additional arcs observed may be partially accounted for by a consideration of the actual value of the arc terminating current which was taken as zero in the above calculations.

a. Arc Initiation

(1) The first initiatory electrons are produced by field emission. The necessary field strength is largely dependent on cathode surface conditions. It is highest for perfectly clean cathode surfaces and appreciably lower in the presence of cathode surface films.[2, 4, 6] This is probably due to lower work functions or due to the presence of positive ions on a cathode film causing local field intensification.[7] (2) The field emission electrons will travel to the anode where, to qualify for setting the second step in arc initiation, should be able to produce, through evaporation, some anode metal atoms* or possibly atoms of an adsorbed gas or a surface film. (3) The potential drop across the contacts should exceed the ionizing potential of the evaporated atoms to allow ionization by electron collision. (4) Ions produced, on approaching the cathode, will cause local fields high enough to produce electron avalanches. (5) the above processes will rapidly multiply leading to the establishment of an arc.

b. The Established Arc

One main characteristic of the short arc is its very high cathode current density.† This high emission rate indicates that the short arc is not only initiated *but also maintained by field emission.*‡§ Since the total voltage drop across the arc is only of the order of 10 volts, the cathode drop thickness should be very small compared to the total arc length. The cathode drop is followed by the arc column or plasma which is a high conduction medium with equal electron and ion densities, a small potential drop and a relatively high neutral atom density. To maintain the arc: (1) enough metal atoms should be produced to maintain the necessary ionization medium, (2) ions lost by collection at the cathode, by recombination and by lateral diffusion should be replaced by an

* The arc may also be initiated without the assistance of the anode atoms or ions[8]. The field emission current density at the cathode in this case, was found to reach a critical value before the arc is initiated. It is thought[9] that at this current density the emission spot can attain its melting point through resistive heating. The cathode in this case will furnish the necessary metal atoms for the subsequent steps of arc initiation.

† Recent measurements by the author obtained from arc tracks on Pd contacts produced by short duration *constant* current arcs indicated current densities as high as 50×10^6 amp/cm^2.

‡ Paper by P. Kislink to be published in the Journal of Applied Physics.

§ Recent analytic considerations, to be published by the author, indicate that in such arcs the current density should be dependent on the work function of the cathode material as well as on the product "pressure × separation" in the arc. For instance, for work functions of 2 and 5 volts, our calculations show that the minimum current densities are, respectively, 5×10^5 and 1.4×10^7 amp/cm^2.

equal number of ions obtained by electron-atom collision in the arc column.

c. Arc Termination

In general, the arc may be terminated by disturbing one or more of the steady state conditions discussed above. For instance, if the potential across the contacts is decreased to or below the ionization potential of the metal atoms, the necessary ionization process will stop and a deficiency of ions in the arc will result. The negative space charge will immediately upset the arc potential distribution interrupting the high electron emission, etc. The arc is also interrupted when the current drops to the minimum arcing current value. This is a well established experimental characteristic of the arc which has yet to be explained in terms of the more basic concepts. It is thought, however, that a decreasing arc current decreases the pressure and the atom density in the arc column. It is possible that when a limiting current is reached the ionization rate becomes too small to maintain the condition of equal space charges in the arc column. One should expect, accordingly, that providing the contact surfaces* with a film of low evaporation energy should furnish a more adequate supply of atoms to the arc which may then be maintained at lower currents. This is in accordance with observations obtained for active contacts.[4]

Arc Initiation Under Dynamic Conditions; Observations on Break

It appeared of interest to examine the relations between arc initiation voltage and contact separation during the break transient and compare them with measurements made under static coditions.[2] In Fig. 3, the increase in arc initiation voltage with separation is in accordance with the static relation shown as a broken line. During the period 2–3, the breakdowns occurred along longer paths than the minimum contact separation and at the minimum value of the sparking potential. By measurement $t_3 = 20 \times 10^{-6}$ sec, $s_3 = 8 \times 10^{-4}$ cm and $ps = 0.61$ mm Hg $\times$ cm. This is roughly the ps value at the minimum sparking potential in air.[10].

By gradually decreasing the charging times of the transient, by adjusting circuit parameters, it was observed that a point was generally reached when a portion of the breakdowns was initiated at voltages well below the coresponding static initiation voltages. Fig. 9 illustrates this

* The necessary atoms may be obtained from either electrodes or both. Arc transfer observations generally indicate signs of evaporation from both electrodes.

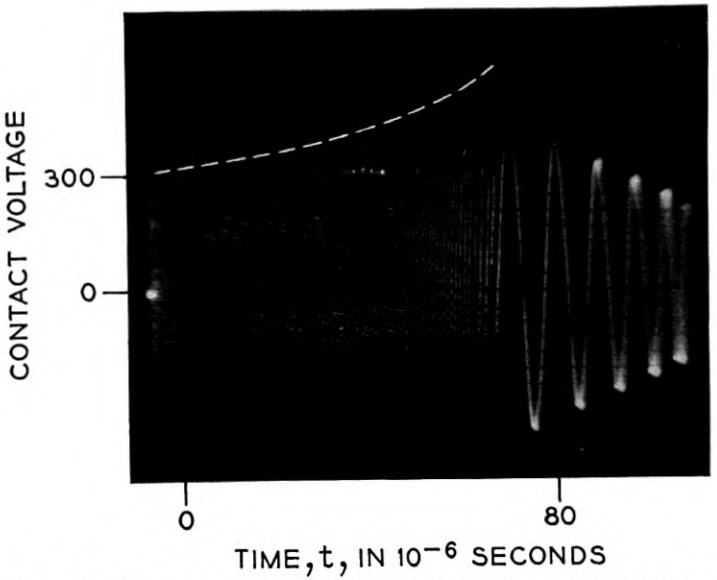


Fig. 9 — Lowering of arc initiation voltage under dynamic conditions. Transient on break of Pd contacts in atmospheric air. E = 50 volts, L = 0.010 henry, R = 40 ohms and C = 270 × 10^{-12} farad. Velocity of contact separation = 40 cms/sec.

effect. In contrast to the static line, shown as a broken line, the breakdown potential shows little change with separation for a major part of the transient. Towards the end, it shows a gradual increase which in this particular case fails to reach the static line. Figure 10(b) is a plot of the ratio $(V_{ai})_{dyn}/(V_{ai})_{stat}$ versus time along the transient.

This phenomenon is attributed to residual effects in the contact gap or on the contact surfaces. In this section, are discussed the possibilities of the presence of residual ions, residuad atoms and residual thermionic emission.

a. Deionization Time

This is determined by calculating the transit time of an ion across the contact gap under the applied field corresponding to the charging of the contact capacitance. For simplification, the initial motion of the ions and the initial field are neglected, the voltage rise is approximated by $V/V_{ai} = t/t_{ch}$ and the field is taken as V/s.

$$t_{deio} = \left(\frac{6m}{e} \cdot \frac{s^2 t_{ch}}{V_{ai}}\right)^{1/3} \tag{4}$$

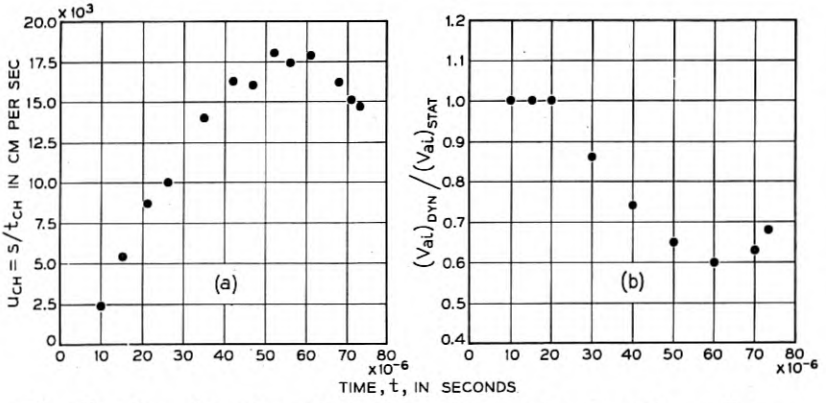


Fig. 10 — Lowering of arc initiation voltage under dynamic conditions.

Defining a charging velocity $s/t_{ch} = u_{ch}$ and a deionization velocity $s/t_{deio} = u_{deio}$ and substituting in equation 4 gives

$$u_{deio} = \left(\frac{eV_{ai}}{6m} \cdot u_{ch}\right)^{1/3} \tag{4a}$$

Following an arc, the contact voltage increases until a new breakdown occurs at V_{ai}. At this instant residual ions from the previous arc could be present in the gap only if $u_{ch} > u_{deio}$, or if

$$u_{ch} > \left(\frac{eV_{ai}}{6m}\right)^{1/2} \tag{5}*$$

This is a convenient expression to apply to our measurements, Fig. 9. For any breakdown point on the transient V_{ai} is measured and u_{ch} is calculated from the corresponding circuit current, capacity C and contact separation. For illustration, for Pd contacts and $V_{ai} = 300$ volts, equation 5 shows that for the presence of residual ions, *the charging velocity u_{ch} must be greater than 10^6 cms/sec.* For $I = 0.3$ amp. and $C = 10^9$ farad, $t_{ch} = V_{ai}C/I = 10^{-6}$ sec and for the presence of residual ions the separation between the contacts must be greater than 1.0 cm. This separation is much greater than most separations involved in our field of study. In Fig. 10(b) are plotted the values of u_{ch} during the transient. u_{ch} reaches a maximum of about 1.8×10^4 cms/sec. This maximum occurs because u_{ch} is proportional to sI which is a product of two monotonic functions one increasing and the other decreasing. It is of interest to note that the decrease in u_{ch} caused an increase in the ratio $(V_{ai})_{dyn}/(V_{ai})_{stat}$.

* Deionization by recombination and lateral diffusion were neglected.

From a group of transients similar to Fig. 9, obtained at different conditions, the plot in Fig. 11 was made. It indicates that in general, the ratio $(V_{ai})_{dyn}/(V_{ai})_{stat}$ starts decreasing at about $u_{ch} = 2 \times 10^3$ cms/sec and at 2×10^4 the arc initiation voltage is only 50 per cent of the corresponding static value. As shown in the figure a deionizing velocity of 10^6 cms/sec is just about two orders of magnitude too high to account for this phenomenon. It should be added, however, that while all the ions have cleared the gap, it has been proposed[7] that the life time of an ion on a surface film can be long enough to enhance the initiation of the next arc. If this mechanism is accepted, our data would indicate that the life time of the ions was only of the order of 10^{-7} second.

b. Residual Atoms

After an arc, the contact gap contains some metal atoms evaporated from the electrodes by the arc. These atoms will clear the gap by travelling to and condensing on the electrodes and by lateral diffusion. A crude approximation is given here of the time of recollection of the atoms on the electrodes based on their initial momentum.

One may visualize the arc spot on an electrode to have a temperature distribution extending from submelting temperatures to a range of boiling temperatures, corresponding to the arc pressures. The lowest temperature is probably the normal boiling temperature of the contact metal. At the termination of the arc, the metal atoms produced at the lowest boiling temperature are the slowest and last to recondense on the

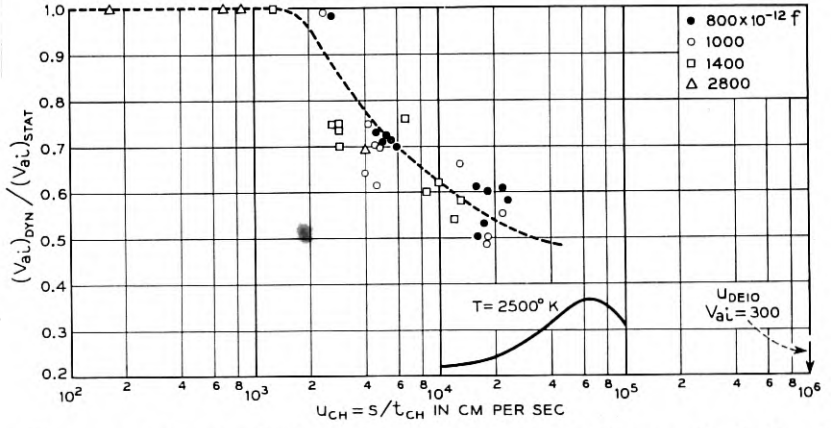


Fig. 11 — Apparent relation between arc initiation voltage and velocity of charging. $E = 50$ volts, $L = 0.010$ henry, $R = 40$ ohms and C as indicated for Pd contacts in atmospheric air.

opposite electrode. An estimate of their velocity may be obtained by assuming thermal equilibrium to have preceded the arc extinction and by using the Maxwellian velocity distribution. The most probable velocity of the metal atoms at the boiling temperature T_b is:

$$u_{at} = \left(\frac{2kT_b}{m}\right)^{1/2} \tag{6}$$

Due to subsequent collisions of the atoms, the velocities thus obtained are probably too high. For Pd at $T_b = 2500°\text{K}$, $u_{at} = 6.4 \times 10^4$ cms/sec. In Fig. 11 is plotted a portion of the velocity distribution at the above conditions. It appears that residual atoms can still be present in the gap at the initiation of the next arc. If it is assumed that the presence of Pd atoms in the gap is alone responsible for the lowering of the arc initiation voltage, one may conclude that the sparking potential in Pd vapor is lower than in air. No evidence, however, is available to support this. On the other hand, at least for contacts with gaps short enough to exclude the surrounding atmosphere, or for vacuum contacts in general, it is quite probable that the presence of metal atoms in the gap could enhance arc initiation. This, as pointed out previously, is because the arc cannot be initiated until atoms from the electrode surfaces are evaporated, by electron bombardment or otherwise, to be subsequently ionized.

c. Cooling Time of The Arc Spot, Maintenance of Thermionic Emission

At the interruption of the first arc, the arc spot initially at the boiling temperature of the metal, will start cooling mainly by conduction to the bulk of the surrounding metal. For a certain period, however, it will remain at temperatures high enough to furnish enough thermionically emitted electrons that may enhance the initiation of the following arc. Assuming the arc spot to be a hemisphere of radius "a" initially at a temperature T_b while the rest of the metal is at T_o, the temperature T at the center of the hemisphere is given by [11]:

$$(T - T_0)/(T_b - T_0) = \frac{4}{\pi^{1/2}} \int_0^{a/2(\alpha t)^{1/2}} z^2 e^{-z^2}\, dz \tag{7}$$

Numerically, for $T_b = 2500°\text{K}$ and $T_o = 300°\text{K}$, T drops to 2400°K and to 1600°K at $a/2\ (\alpha t)^{1/2} = 2.0$ and 1.2, respectively. It is evident that the cooling time is proportional to the area of the arc spot. If the current at which the arc is terminated is I_m and the arc current density is i_a, the area of the arc spot is $A_a = I_m/i_a$ and $a = (I_m/\pi i_a)^{1/2}$. For

$i_a = 10^7$ amp/cm^2[27], and $I_m = 0.5$ amp one gets: $T = 2400°$K at $t = 4 \times 10^{-9}$ sec and $T = 1600°$K at $t = 1.1 \times 10^{-8}$ sec for Pd.

The corresponding thermionic emission is obtained from

$$i_{th} = AT^2 e^{-\frac{e\phi}{kT}}$$

with $A = 60$ amp cm^{-2} deg.$^{-2}$ and $\varphi = 4.99$ volts for Pd[12]. At the termination of the arc, $t = o$, $i_{th} = 0.048$ amp/sec^2, at $t = 4 \times 10^{-9}$ sec, $i_{th} = 0.032$ amp/cm^2 and at $t = 1.1 \times 10^{-8}$ sec, $i_{th} = 6 \times 10^{-8}$ amp/cm.^{2} The respective rates of electron emission from the arc spot are 1.5×10^{10}, 1.0×10^{10} and 1.9×10^4 electrons/sec. *This indicates that the initiating electrons may be furnished by thermionic emission if the charging time following the first arc is of the order of or less than about* 5×10^{-9} *sec.* This time is more than an order of magnitude too small compared to the charging times involved in the data of this section. One may, therefore, exclude the thermionic emission as an explanation for the low arc initiation voltages obtained.

The initiation of reversed arcs, however, may be enhanced by thermionic emission from the previous arc spots since the recharging times involved, $\pi(lc)^{1/2}$, are usually very small. l and c are usually of the orders of 10^{-7} henry and 10^{-11} farad and the charging time is of the order 10^{-9} sec.

ESTABLISHMENT OF GLOW DISCHARGE AND TRANSITION INTO AN ARC

For the circuit in Fig. 1(b), it was observed that on break of the contact, glow discharge was observed under certain circuit and contact surface conditions. An obvious requirement was that the voltage across the contacts should exceed the glow discharge voltage of the contact in the surrounding atmosphere. This requirement alone, however, was not sufficient as in some cases no glow could be detected, in others glow was established and maintained and in other instances glow was followed by a transition into an arc. In this section is presented an experimental study of the conditions that determine the nature of the discharge.

Cathode Current Density in Static Normal Glow

First, measurements were made of the cathode current density in a static normal glow. This was done for palladium and gold contacts in dry atmospheric air at 25°C. In each case the cathode was the flat end of a cylinder and the anode was a larger parallel flat surface of the same material as the cathode. The circuit in Fig. 12 was used. The contacts were

cleaned by filing then washing with methyl alcohol and distilled water. The contacts were slowly brought together until glow discharge was established. Before measurements were made the circuit current was increased to allow the glow to cover the entire cathode flat surface as well as a portion of its cylindrical surface. By allowing the contacts to glow for about 20 minutes, the occasional arcing first observed was eliminated and a steady glow was established. The cathode was observed under a microscope and the current was adjusted to obtain a glow just covering the flat cathode area. From the measured current and the cathode area, the cathode current density was determined. The results are given in Table II.

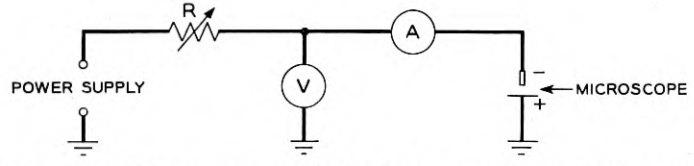


Fig. 12 — Circuit for measurement of cathode current density in normal glow discharge at static conditions.

Observations on Glow Maintenance and Glow-arc Transitions

The simplified circuit in Fig. 13 was used. The contact cathode was the flat end of a cylinder. The cylindrical portion was tightly fitted into a block of an insulating material allowing an exposure of the flat end and a cylindrical area less than 10 per cent of the flat area. The anode was a parallel plain surface of the same material.

To avoid the unnecessary complications of a measuring circuit connected to the contacts, the plates of a cathode ray oscilloscope were, instead, connected across a capacitor, 10 times C, in series with the circuit capacitor C. From the transients obtained, it was possible to identify glow discharge, steady arcs and interrupted arcs. Four typical transients are shown in Fig. 14. Transient A shows a case where glow discharge was established and maintained for the entire half period of the circuit. In transient B glow was not detected and, instead, interrupted arcs occupied the entire half period. In transient C, glow discharge was maintained for a short duration 1–2 followed by interrupted arcing, 2–3. At point 3 the circuit current was high enough to maintain an arc and a steady arc was obtained, 3–4. Transient D is similar to B where glow discharge was undetectable. The multiple discharge in D, however, lead to the steady arc 2–3.

Before presenting our measurements and discussion, a review is given

TABLE II — CATHODE CURRENT DENSITY IN STEADY NORMAL GLOW IN DRY ATMOSPHERIC AIR AT 25°C

Electrodes	Cathode Diameter, cm.	Glow Current, amp.	Cathode Current Density, amp./cm²
Pd	0.05	0.010	5.1
	0.10	0.033	4.2
Au	0.05	0.017	8.6
Ag*	—	—	9

* Measurement by F. E. Haworth.[13]

here of the process of the initiation of the steady arc which was explained in detail in reference.[5] For the inductive circuit in Fig. 13, when the proper contact separation is reached, a first breakdown will occur discharging the local capacitance at the contacts. This is followed by recharging from C through L and a second breakdown. This will repeat while the circuit current will increase in a discontinuous fashion. If it reaches the minimum arcing current of the contact, a steady arc is established, otherwise, the transient will be made up entirely of local multiple discharges. Figures 14D and B are the main condenser voltage transients corresponding to the above two cases respectively.

The interrupted arcs, or multiple discharges, and the steady arc constitute the two processes of conduction that are commonly obtained when the voltages involved are below the spark breakdown potential of the surrounding atmosphere. In such cases, the arc initiation is dependent on the contact material and its surface condition and is independent of the atmosphere.[2] If the voltages involved are equal to or greater than the minimum sparking potential of the atmosphere, the initiation of a breakdown is primarily dependent on the atmosphere. This breakdown, however, may in addition lead to a glow discharge as discussed above. This immediately raises the question as to whether breakdowns leading to an arc and breakdowns leading to a glow discharge are initiated at the same potentials. For this purpose the following experiment was performed.

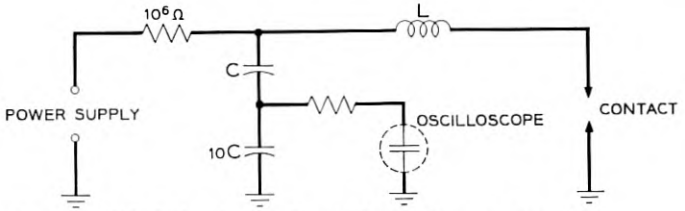


Fig. 13 — Simplified circuit for the study of glow-arc transitions.

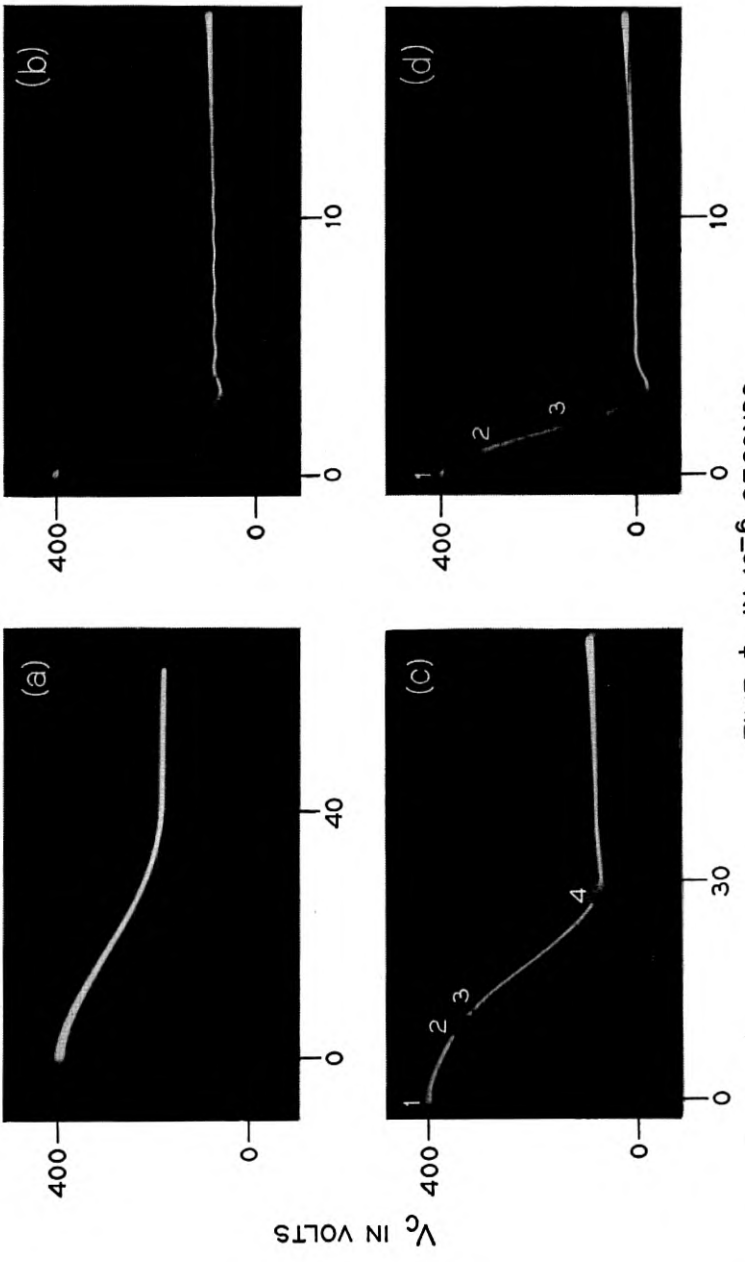


Fig. 14 — Voltage transients at circuit condenser. A: all glow. B: no glow, only a series of interrupted arcs. C: glow-interrupted arcs — steady arc. D: no glow, interrupted arcs — steady arc — interrupted arcs.

Initiation Voltage of Glow Discharge

The cantilever bar setup previously used for similar measurements of arc initiation voltage as function of separation[2] was used here. By varying the separation the corresponding glow initiation voltage was measured. For each separation a measurement was also made of the *arc* initiation voltage. The results are given in Table III. The results indicate that both arc and glow are initiated at the same voltage for the same separation. One may, therefore, conclude that *at least the first few steps involved in the process of the breakdown are the same whether they lead to a glow discharge or to an arc.* In many cases, it was observed that the arc was preceded by a period of glow discharge. This was not found, however, to be the general case as discussed in the following section.

TABLE III — GLOW AND ARC INITIATION VOLTAGES AS FUNCTIONS OF CONTACT SEPARATION FOR PD CONTACTS IN DRY ATMOSPHERIC AIR AT 25°C.

S: 10^{-4} cm.	1.5	3.0	4.5	6.0	7.5	9.0	10.5	12.0	13.5	15.0	16.5	18.0	19.5
V_{gi} : volts	320	320	340	380	400	420	450	480	500	520	540	560	590
V_{ai} : volts	310	320	340	370	400	420	450	470	490	510	540	570	590

Glow-arc Transition

The experimental setup used is shown in Fig. 13. By systematic variation of the circuit parameters *Vo, L and C*, a variety of transients was obtained and recorded. Samples of typical cases are shown in Fig. 14. For transient stability and reproducibility, it was found necessary to exercise extreme care in securing good contact surface cleanliness and in maintaining it during the experiment. The presence of organic vapors, humidity, films of grease or oil, fingerprints, etc., usually led to erratic results. *The general effect was an inclination towards more arcing and less glow discharge.* Only by proper cleaning of the contact surfaces and allowing the contact to arc heavily for about 20 minutes was it possible to obtain fairly reproducible results. Table IV shows a summary of results obtained from one of several sets of experiments performed.

Before stabilization of the transient, it was generally observed that the glow period was first short then gradually increased until it reached a limiting value which it did not exceed. These limiting values are given in column 5 as fractions of the half period $\pi(LC)^{1/2}$. They range from zero, actually glow was not detected with a time resolution of 1 per cent of the

TABLE IV — GLOW-ARC TRANSITION DATA FOR PD CONTACTS IN ATMOSPHERIC AIR-CATHODE DIAMETER 0.1 CM.

(1) V_0 volts	(2) C $10^{-12}f$	(3) L $10^{-3}h$	(4) $z = (L/C)^{1/2}$ ohms	(5) $t_g/\pi(LC)^{1/2}$	(6) $(V_0 - V_g)/L$ 10^4 amp./sec.	(7) (I_g) max. amp.	(8)‡ (i_g) max. amp./cm.^{2}	(9) $(i_g)/i_g$ max.
600	18,000	5	52	0*	6.0	<.018	<2.3	<0.5
500	—	—	—	0	4.0	<.012	<1.6	<0.4
450	—	—	—	0	3.0	<.009	<1.2	<0.3
400	—	—	—	1.0†	2.0	>.19	>23	>4.8
350	—	—	—	1.0	1.0	>.10	>13	>2.6
320	—	—	—	1.0	0.4	>.04	>5	>1.0
600	18,000	8	660	0	3.8	<.015	<2	<0.4
500	—	—	—	0.22	2.5	.19	24	4.8
450	—	—	—	0.44	1.9	.23	29	5.8
400	—	—	—	1.0	1.2	>.15	>19	>3.8
350	—	—	—	1.0	0.6	>.076	>10	>2.0
600	18,000	15	906	0.25	2.0	.23	29	5.8
550	—	—	—	0.30	1.7	.23	29	5.8
500	—	—	—	0.35	1.3	.20	26	5.2
450	—	—	—	1.0	1.0	>.17	>22	>4.4
400	—	—	—	1.0	0.7	>.11	>14	>2.8
600	18,000	20	1050	0.40	1.5	.28	36	7.2
550	—	—	—	0.40	1.2	.23	29	5.8
500	—	—	—	1.0	1.0	>.19	>24	>4.8
450	—	—	—	1.0	0.8	>.14	>18	>3.8

* No glow was detected with a time resolution of 1 per cent of a half period $\pi(LC)^{1/2}$.

† Uninterrupted glow occupied the entire half period.

‡ Obtained by dividing $(I_g)_{max}$ by the total cathode area.

transient time, to a full transient time. By calculation, the corresponding limiting currents and limiting current densities were obtained, columns 7 and 8 respectively. The ratios of the limiting current densities to the normal glow current density are also given in column 9. They show that at the interruption of the glow discharge the current density was 5 to 7 times the normal glow current density. This indicates a transition from normal glow to abnormal glow before the final transition into an arc. One may, therefore, conclude that if glow discharge is obtained it starts as normal glow which may occupy only a small fraction of the cathode area. By increasing the current the cathode glow area expands at constant current density until it covers the entire cathode area. Further current increase leads to a transition into abnormal glow with higher current densities. Transition of the abnormal glow into an arc occurs when the current density reaches a limiting value. This limiting current density is extremely sensitive to surface contamination and generally

increases with surface cleaning.* For clean Pd contacts in atmospheric air an average limiting current density of 30 amps/cm^2, or about 6 times the normal glow current density, was obtained. This sudden transition from the low current density glow to the very high current density arc represents a high rate of change in the emission process. With contaminated contacts, this is probably due to the presence of low work function high emission spots on the cathode. These spots may be eliminated by proper cleaning thus allowing glow discharge to be maintained at higher current densities. The observed glow-arc transitions for clean contacts, consistently occurring at about 30 amps/cm^2 for Pd, *may still be attributed to the formation of a surface film on the cathode through a cathode-atmosphere reaction.*†

Measurements have also indicated that under certain conditions, glow discharge cannot be obtained even at currents much below the limiting currents discussed above. It appears that there is a limiting rate of rise of current with time above which glow discharge cannot be maintained. In Table IV, column 6, the initial rates of current rise are given. In all cases where the rate of current rise was greater than about 3×10^4 amps/sec, lines 1, 2, 3 and 7, no glow was obtained. The experiment was repeated with two other cathode diameters of 0.2 and 0.05 cm. The limiting rates of rise obtained were approximately the same as given above, indicating that the limiting rate of current rise is independent of the cathode area. This seems reasonable since at the beginning of the transient the currents are very small and the emission area is only a very small fraction of the cathode area. No detailed explanation, however, can be furnished at this time as to why such a limit of the rate of current rise does exist. It is obvious, nevertheless, that while the rate of current rise can be increased without limit by manipulating the circuit parameters, the conduction mechanism in the contact gap, will, in general, have its own limitations as determined by the emission processes involved.

ACKNOWLEDGMENT

I am indebted to Miss R. E. Cox for assistance with many of the experiments and calculations reported here.

* With a contaminated cathode surface a transition into an arc may occur during the *normal* glow period well before the current is high enough to allow normal glow to cover the entire cathode surface. This is particularly true with larger cathode areas which are usually hard to clean satisfactorily by the above procedure.

† A recent unpublished study by F. E. Haworth has shown that in the absence of the usual surface contaminants, glow discharge is capable of activating palladium and silver contacts through the formation of surface films. These surface reactions appear to be strongly dependent on the atmosphere.

BIBLIOGRAPHY

1. A. M. Curtis, Contact Phenomena in Telephone Switching Circuits, B. S. T. J., **19,** p. 40, 1940.
2. M. M. Atalla, Arcing of Electrical Contacts in Telephone Switching Circuits — Part II, B. S. T. J., **32,** pp. 1493–1506, Nov., 1953.
3. G. H. Pearson, Phys., **32,** pp. 1493–1506, Rev., **56,** p. 471, 1939.
4. L. H. Germer, Arcing of Electrical Contacts on Closure — Part I, J. Appl. Phys., **22,** p. 955, 1951.
5. M. M. Atalla, Arcing of Electrical Contacts in Telephone Switching Circuits — Part I, B. S. T. J., **32,** p. 1231, 1953.
6. F. E. Haworth, Experiments on the Initiation of Electric Arcs, Phys. Rev., **80,** p. 223, 1950.
7. F. L. Jones, Initiation of Discharges at Electrical Contacts, Proc. Inst. Electrical Engineering I 124, 169, 1953.
8. W. P. Dyke, J. K. Trolan, E. E. Martin, and J. P. Barbour, The Field Emission Initiated Vacuum Arc — I, Phys. Rev., **91,** p. 1043, 1953.
9. W. W. Dolan, W. P. Dyke, and J. K. Trolan, The Field Emission Initiated Vacuum Arc — II, Phys. Rev., **91,** p. 1054, 1953.
10. J. J. Thomson and G. P. Thomson, *Conduction of Electricity Through Gases*, Vol. 2, p. 487.
11. H. S. Carslow, *Introduction to the Mathematical Theory of the Conduction of Heat in Solids*, 2nd Edition, p. 150, 1921.
12. S. Dushman, Rev. Mod. Phys. **12,** p. 381, 1930.
13. F. E. Haworth, Electrode Reactions in Glow Discharge, Jl. Appl. Phys., **22,** p. 606, 1951.
14. L. H. Germer, Erosion of Electrical Contacts on Make, J. Appl. Phys., **20,** pp.1085–1109, 1949.

Thickness Measurement and Control in the Manufacture of Polyethylene Cable Sheath

By W. T. EPPLER

(Manuscript received October 22, 1953)

The manufacture of multiple sheath for Alpeth and Stalpeth cables requires the application of a sheath of polyethylene over a sheath of corrugated metal which is flooded with a rubber asphaltic compound. For high quality and minimum cost, this outer sheath must be of uniform thickness throughout its length. One of the problems in cable sheath manufacture is to maintain the concentricity and average thickness of the extruded polyethylene sheath to close limits during manufacture. This article reports on: (1) The application of a capacitance sensitive bridge to the measurement of the eccentricity and average thickness of the sheath on cables moving at speeds of 20 to 100 feet per minute; (2) The method of thickness calibration; and (3) The use of the thickness measurements in maintaining the sheath concentricity and average thickness within close limits during the sheathing operation.

HISTORY

In the manufacture of multiple sheath for Alpeth and Stalpeth cables, an outer sheath of polyethylene is applied. It is desirable for high quality and low cost to make this outer sheath of a uniform thickness throughout. The construction of these cables is shown in Fig. 1. In both designs, the outer sheath is polyethylene extruded onto a corrugated metal undersheath which has been flooded with a rubber asphaltic compound.

The extrusion art had been unable to obtain a high degree of control, primarily because measurements of the thickness could not be obtained until after the sheath was applied to the cable core. Eccentric sheath must have a greater average thickness than concentric sheath, if the thickness of the thin side is not to fall below a required minimum thickness.

The symmetrical design of a typical core tube and die for sheathing is shown in Fig. 2. Concentric set-up of these extrusion tools around the

cable core will not produce concentric extruded sheath. This is caused by an unbalance in the plastic flow in the extruder. The flow makes a ninety degree turn from the extruder cylinder into the die head, and to reach the far side of the die, must flow around the core tube. The flow resistance also varies with changes in the temperature of the plastic and of the extruder screw speed.

The core tube is fixed in position in the extruder head. The die is located around the core tube and can be moved in any direction eccentric to it. Fig. 3 shows a core tube and die mounted in the extruder head and indicates the location of the four die adjusting screws by which movement of the die in relation to the fixed core tube is accomplished. The die must be located at some one eccentric position in relation to the core tube to compensate for the differences in flow resistances in the head.

To set the die for concentric sheath and to adjust for specified thickness the prevailing practice of the cable art of measuring the wall thickness of a sample taken from the lead or finish ends of the sheathed cable was of necessity resorted to because it was the best technique available. The cutting of a ring of sheath and the micrometer gage are shown in Fig. 4. These end samples only approximate sheath conditions because

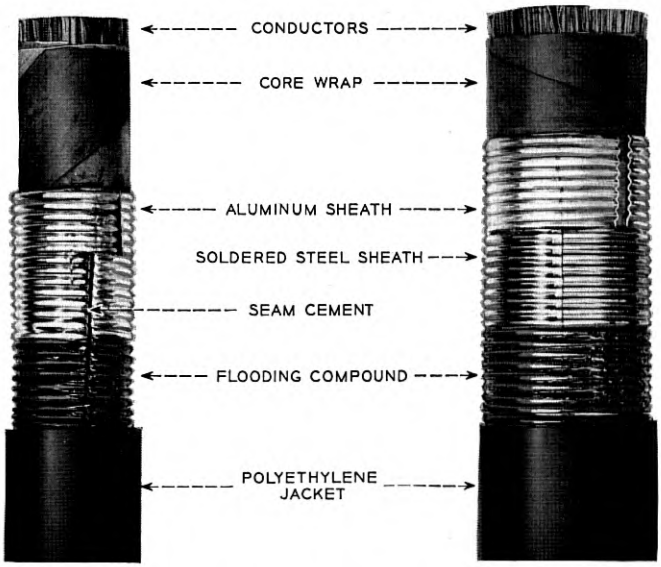


Fig. 1 — (Left) Telephone exchange cable of Alpeth design; (right) Stalpeth design.

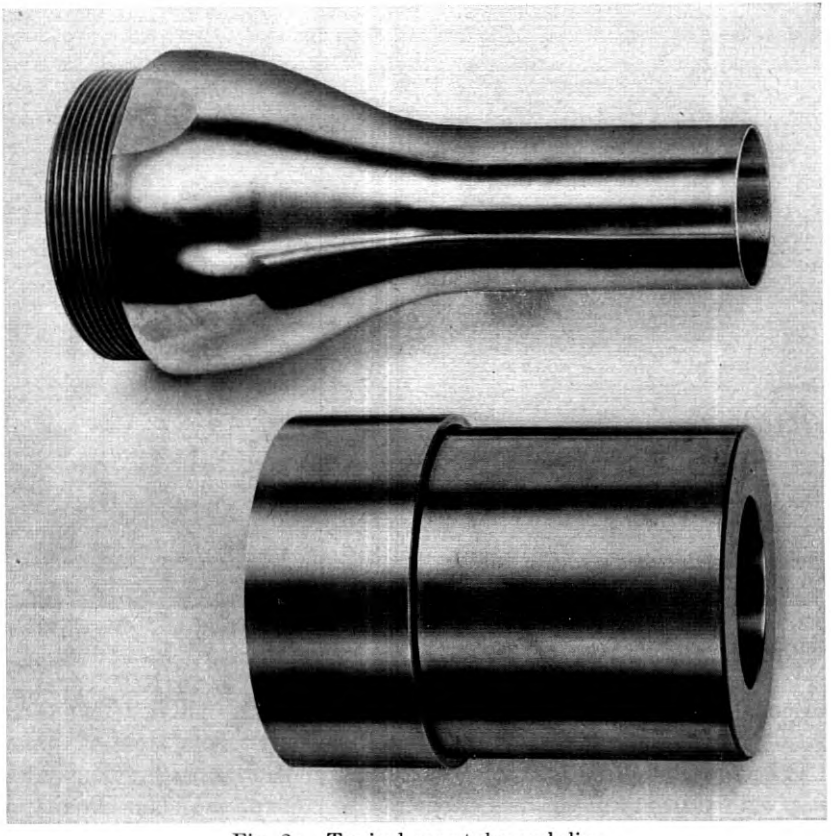


Fig. 2 — Typical core tube and die.

they are only short pieces to represent cables up to a few thousands of feet in length.

Sheath eccentricity is expressed as a percentage and is the difference between the thicknesses, of the thickest and the thinnest sides of a cross section, in relation to the specified wall thickness expressed in mils. Control from end sampling resulted in most cables having eccentricities of 30 per cent to 60 per cent. Also, it was difficult to keep the average thickness to within ± 0.010 inch of the specified average thickness.

The need for a better gaging method than end sampling, led to an investigation of determining the wall thickness in terms of the capacitance that would be formed by the metal undersheath and a probe sliding on the sheath surface.

A test set as shown in Fig. 5 was developed which responds to changes

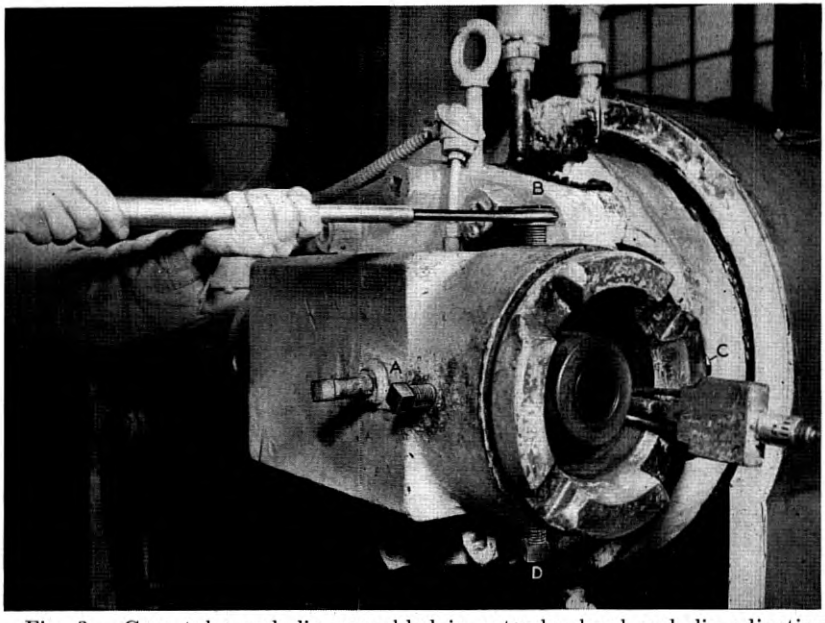


Fig. 3 — Core tube and die assembled in extruder head and die adjusting screws.

in capacitance. The capacitance response is in turn calibrated in thousandths of an inch of sheath thickness. The electronic system of the set has been described in the Bell System Technical Journal previously.* It is practical from test set measurements to control the concentricity of Alpeth cable to within 35 per cent and Stalpeth to within 20 per cent. Average thicknesses within ± 0.005 inch are maintained.

Formerly, the safe practice was to use an excess of approximately 10 per cent over specified average in order to keep the thin side of eccentric sheath within the minimum spot limit. Control from test set measurements eliminated the necessity of using an excess of polyethylene because sheath of improved concentricity maintained close to the specified average thickness does not vary below the specified minimum spot thickness. The quality of the sheath is improved because it is of consistently high dimensional uniformity not previously obtainable. Also, concentric sheath has better flexing characteristics since eccentric sheath concentrates the stresses of flexing in the thin side.

* Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath, B. M. Wojciechowski, B.S.T.J., **33**, pp. 353–368, Mar., 1954.

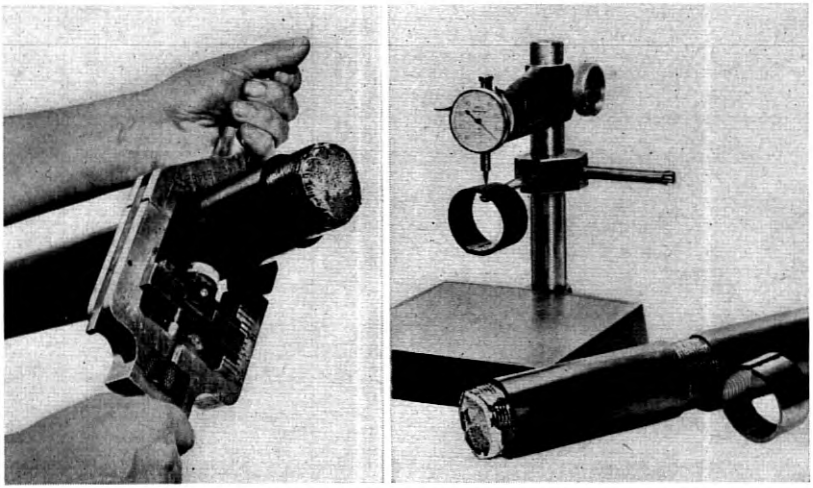


Fig. 4 — (Left) Removing test strip from end of cable; (right) performing micrometer measurements on test strip.

CALIBRATION OF THE TEST SET FOR SHEATH THICKNESS MEASUREMENTS

Calibration of capacitance into thickness was difficult because the capacitance is not a simple function of polyethylene thickness. It depends also on the curvature of the sheath surface, the size and shape of the probe, the amount of flooding and the height and shape of the corrugated metal. For a given probe, it depends chiefly on the thickness, the flooding and the sheath curvature. The flooding sometimes varies from a thin film to an excess that overfills the corrugations. The surface curvature is not uniform because the soldering of the metal overlap of Stalpeth cable generally produces a flattened sector and the capstan at the soldering operation results in an elliptical shape. Changes in the surface curvature and in the amount of flooding can be compensating or cumulative in varying the capacitance.

To determine whether a correlation between jacket thickness and capacitance existed, extensive spot checks for three sizes of cable were made. Marked points on cable were measured for capacitance and then with a micrometer. A slight error can exist because the micrometer measurement is only one spot in the center of an area which is effective to capacitance. This condition is shown by Fig. 6. Also, it is difficult to determine accurately the surface curvature associated with the capacitance measurement.

The relation of thickness to capacitance conditions in the samples is

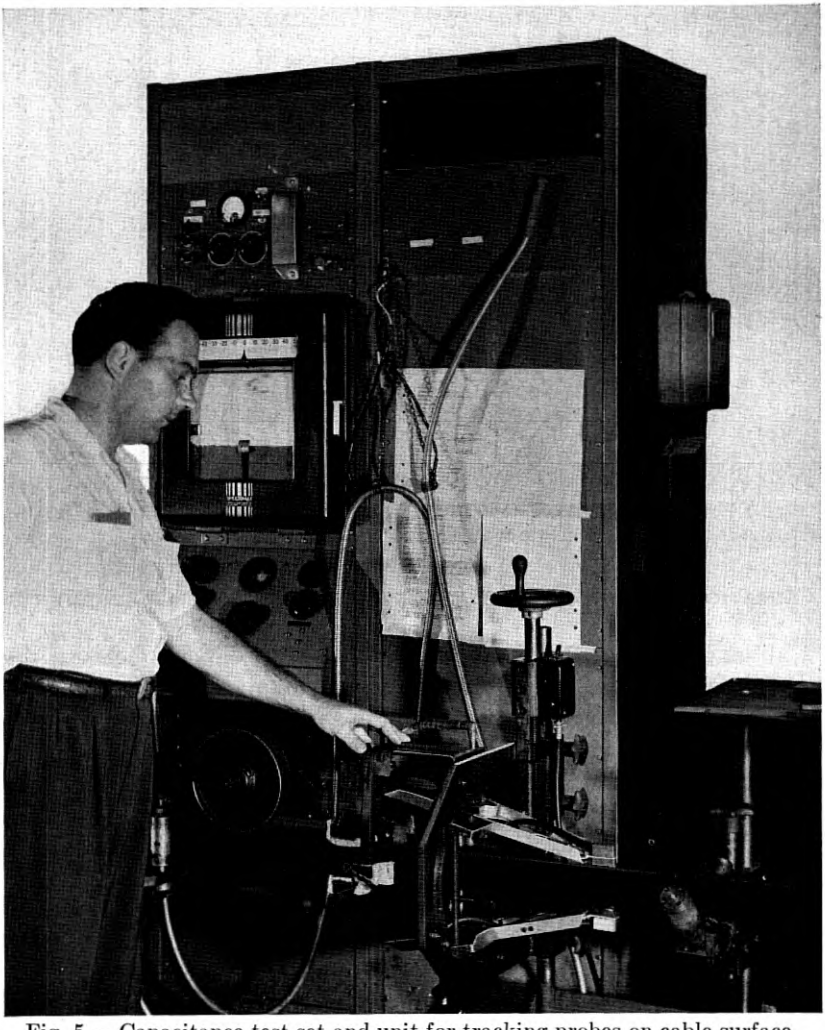


Fig. 5 — Capacitance test set and unit for tracking probes on cable surface.

shown by Fig. 7. The sheath thickness is specified as the distance between the outside surface of the sheath to the bottom of the corrugations formed into the polyethylene by the crests of the corregated metal sheath, as indicated by dimension T. The top sketch shows the normal amount of flood. The capacitance will be different in each of the three conditions of equal thickness shown. With excess flood, center sketch, the distance between plates is increased and the capacitance is decreased.

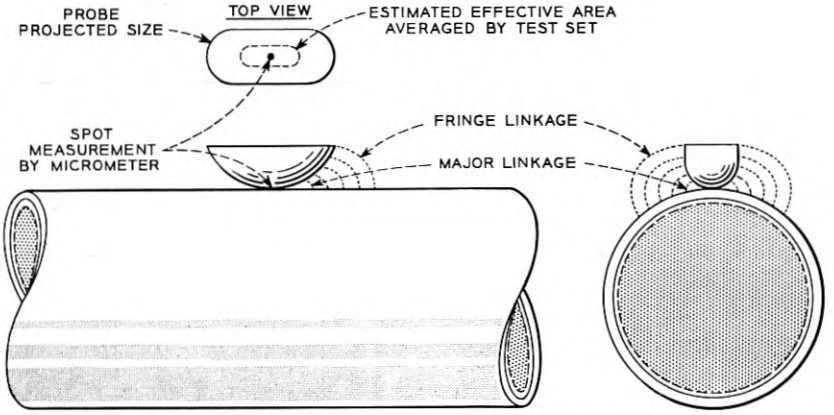


Fig. 6 — Thickness measured by direct calibration; spot by micrometer; area by capacitance.

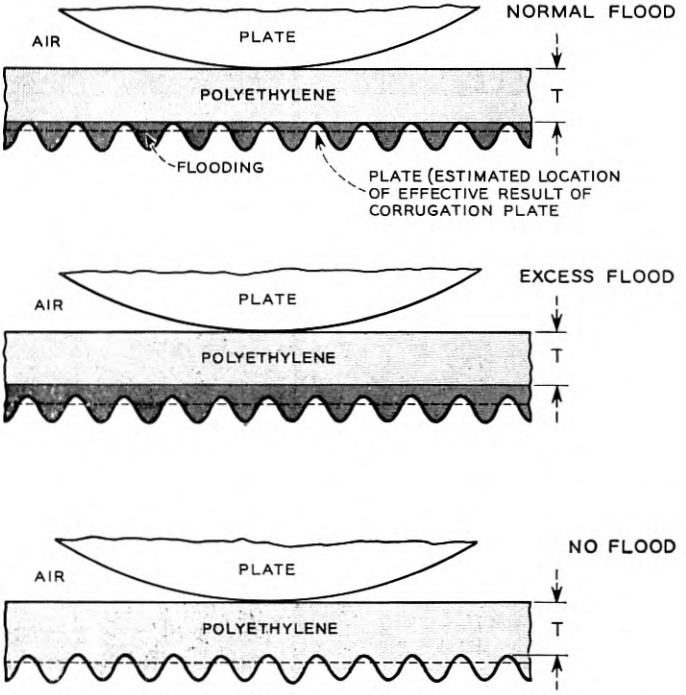


Fig. 7 — Equal thicknesses, different capacitances.

Insufficient flood, bottom sketch, alters the dielectric from polyethylene plus some flood, to all polyethylene. The capacitance is decreased.

A typical plot of points and a calibration curve are shown in Fig. 8. Each of the three cable sizes measured revealed a wide band of plot points. In each curve the points were more dense toward the left side of

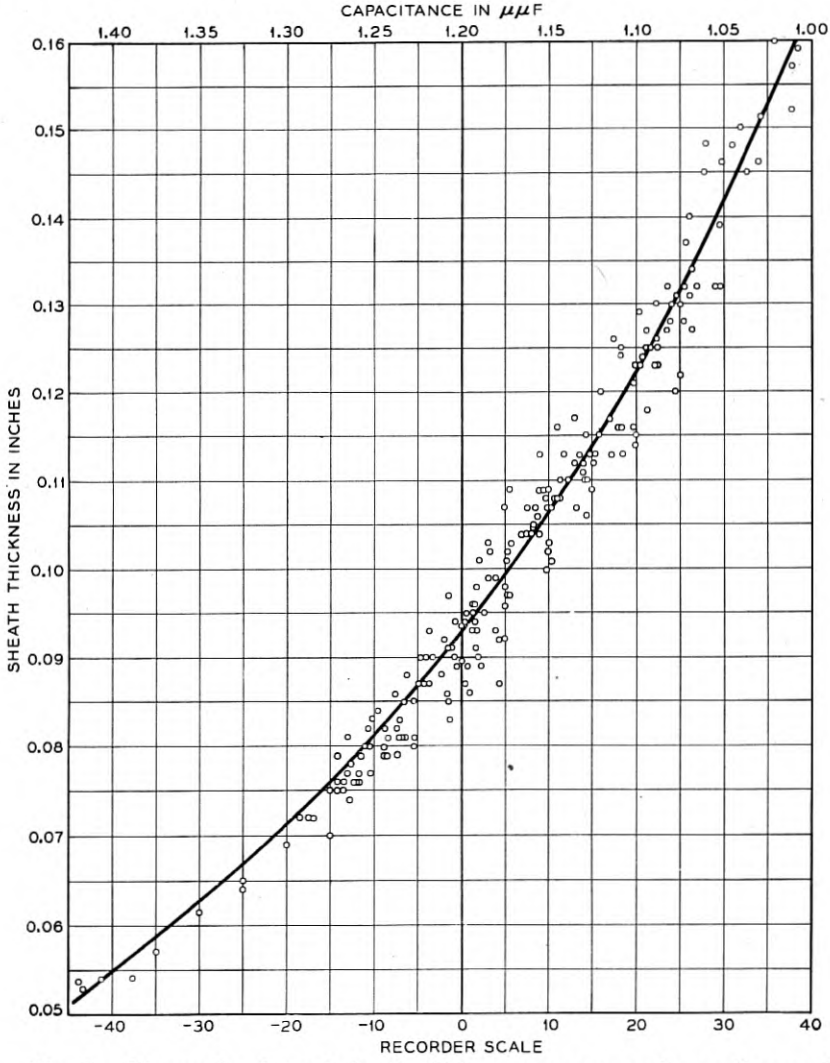


Fig. 8 — Measured points of sheath thickness versus recorder readings and developed calibration curve.

the band, becoming progressively less to the right across the band. The majority of points to the extreme right were found to be cases of excess flood. Many of the points, near the extreme right had insufficient flooding. Points close to the curve had the flood just filling the corrugation valleys. Other points consist of various other amounts of flood and/or are the result of deviation from correct surface curvature.

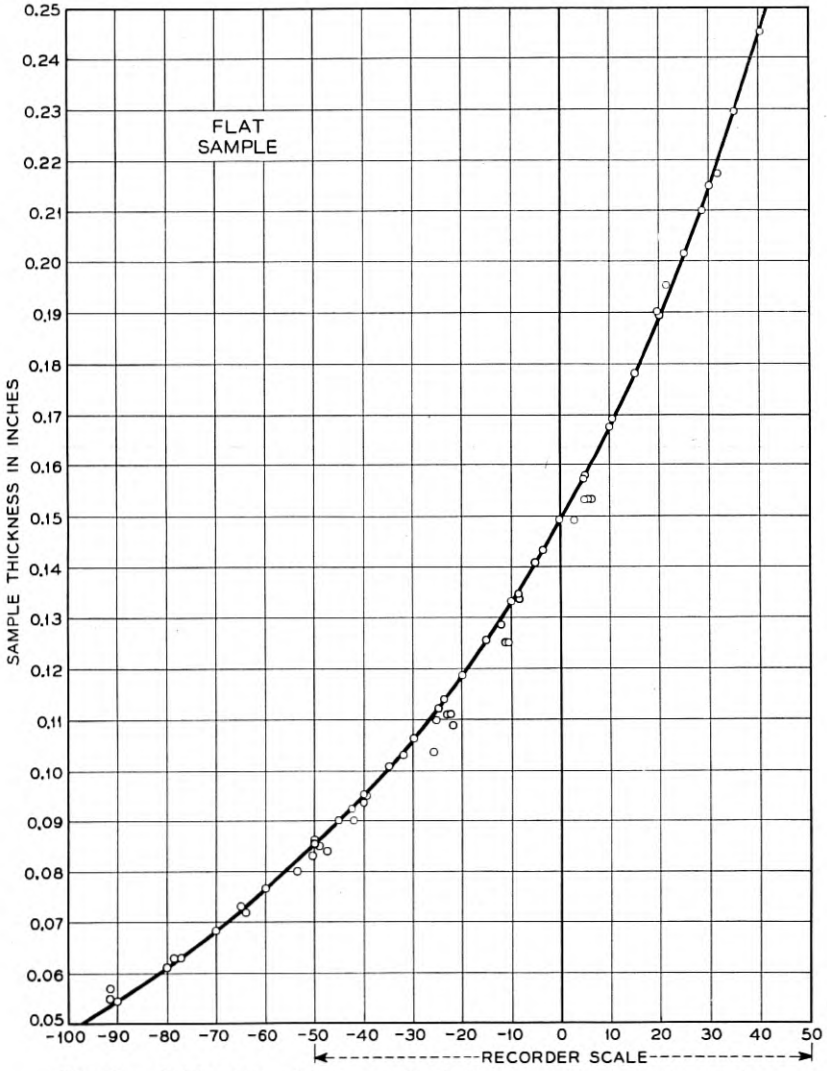


Fig. 9 — Calibration of flat sample thickness versus recorder scale.

Fig. 8 also shows that the greatest percentage of points are within a thickness range of approximately 0.010 inch. In moving downward from maximum thickness the concentration of measured thicknesses increases rapidly over approximately 0.003 inch and then becomes progressively less covering an additional 0.010 inch. The calibration curve was placed at about the location of maximum point concentration. By averaging the thickness indications along a short length of the cable, a measurement adjusted for the occasional extremes in flooding and surface variations is obtained. The accuracy for practical use is therefore within limits of ± 0.005 inch from the mean.

Investigation was also made of flat samples of Polyethylene placed upon a flat metal plate. Flat samples eliminate the variables introduced by the cable surface curvature, the corrugated metal undersheath and the flooding material. A plot of capacitance against thickness for flat samples is shown in Fig. 9. Each point represents an individual molded flat sample. The majority of points are within ± 0.003 inch of the curve.

The measurement of sufficient points to obtain curves for the many cable diameters would involve an impractical amount of work.

The calibration curves for the three cable sizes and the curve for flat samples drawn to the same capacitance versus thickness scale have similar form, but are displaced one from the other. The displacement of the calibration curves for cables of core diameters of 1.39 to 2.38 inches is shown by Fig. 10. The displacement is approximately 1 meter division for a diameter change of 0.1 inch.

Calibration curves for other cable diameters than the three measured were obtained by an approximation formula based on measuring a few points from each sheath diameter to determine the displacements and slopes and multiplying the flat sample curve values by the displacement and slope correction factors.

The curve for flat samples and the curve for 2.38 inch diameter cable plotted to the same scales is shown in Fig. 11. The two curves are sufficiently alike so that by multiplying the flat sample curve thickness values by a constant (K_1) obtained from the ratio of the cable sheath thickness to the flat sample thickness at zero recorder scale, the amount of curvature of the resultant curve and the measured sheath curve are essentially the same, and they have the same thickness and capacitance values at zero recorder reading. A multiplier (K_2) can then be added to adjust the slope of the percentage curve to make it practically coincide with the sheath thickness curve. Actually, there is a slight difference between the curvature of the flat sample curve and those of cable sheath. The amount of curvature increases as the cable diameter decreases.

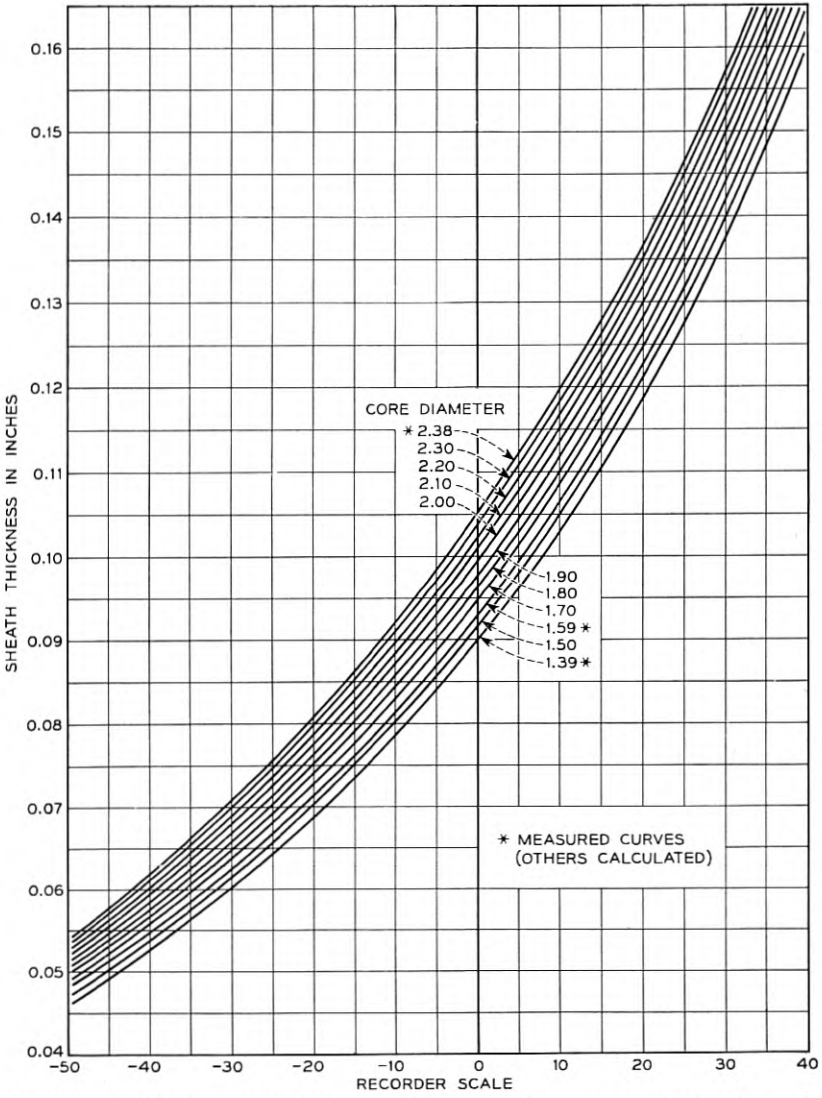


Fig. 10 — Calibration curves by core diameters, thickness versus recorder scale.

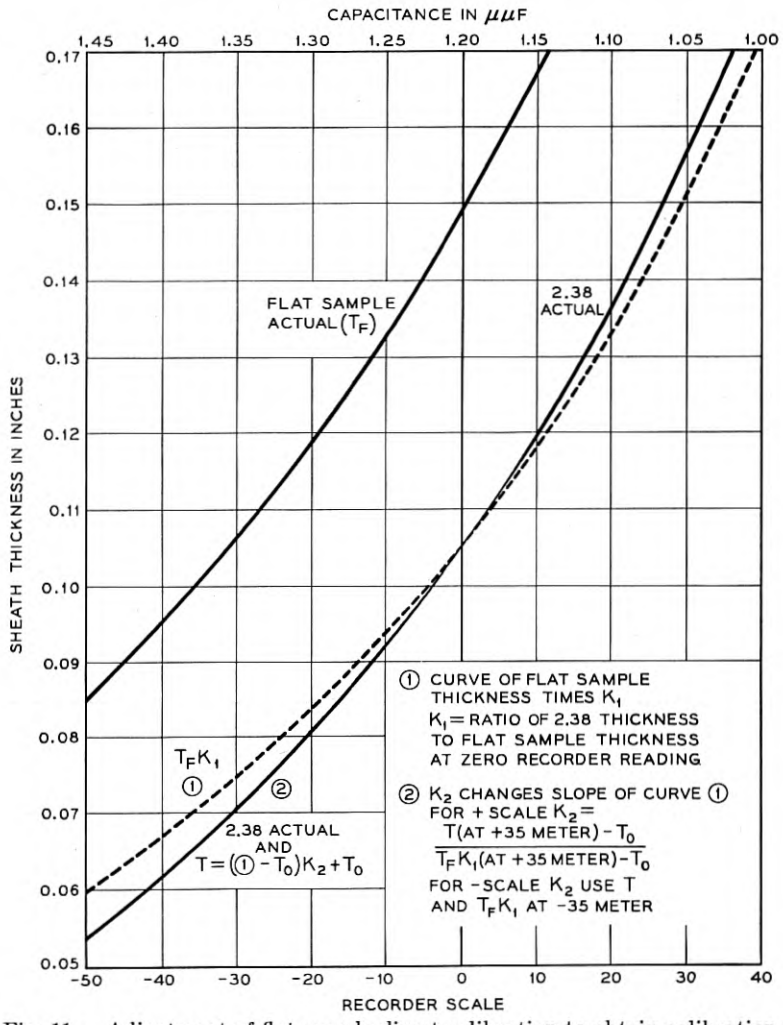


Fig. 11 — Adjustment of flat sample direct calibration to obtain calibration of 2.38-inch diameter cable sheath.

The result is the following approximation formula, from which the thickness calibration can be calculated within 0.001 inch with the error negligible over most of the working range.

$$T = T_F K_1 K_2$$

where T = Thickness in thousandth's of an inch of polyethylene cable sheath.

T_F = Thickness fo flat polyethylene sample at same recorder meter reading as for T.

K_1 = Ratio of actual cable sheath thickness to flat sample thickness at zero meter reading.

K_2 = Constant to change slope of $T_F K_1$ curve.

$$\text{For + meter readings } K_2 = \frac{T_{(\text{at} + 35\,\text{meter})} - T_0}{T_F K_{1(\text{at} + 35\,\text{meter})} - T_0}$$

$$\text{For - meter readings } K_2 = \frac{T_{(\text{at} - 35\,\text{meter})} - T_0}{T_F K_{1(\text{at} - 35\,\text{meter})} - T_0}$$

T_0 = Thickness in thousandth's of an inch of cable sheath at zero meter reading.

The K_1 factor accounts for the dimensional differences between the capacitor formed by a flat thickness of polyethylene on a flat plate compared to the actual capacitor construction of cable at zero meter. Both have the same capacitance of 1.20 uuF at zero meter reading. K_2 accounts for changes resulting from the curved surfaces of cable. K_1 and K_2 are different for each cable diameter.

Since zero meter is used as a reference point, the formula becomes:

$$T = (T_F K_1 - T_0) K_2 + T_0$$

ACCURACY CHECK UNDER OPERATING CONDITIONS

A check* was made of the accuracy of calibration and of the response under operating conditions of applying the sheath to the cable. The upper graph in Fig. 12 was obtained with the test set probe tracking at a cable sheathing speed of 50 feet per minute. The probe was shifted to different octant locations on the circumference for lengths of the cable as indicated on the graphs. The track of the probe was marked on the sheath surface and the sheath then removed, cleaned of flooding compound and the micrometer measurements of the thickness taken at six-inch intervals along the length. The lower graph is a plot of the thickness obtained by micrometer. The ability of the test equipment to track and respond to the thickness variations is apparent from comparison of the two graphs.

APPLICATION OF TEST EQUIPMENT FOR EXTRUSION CONTROL

The test set is placed at some distance after the extruder to prevent the probe from marking the plastic polyethylene. The machinery of the

* Test and measurements by courtesy of J. L. O'Toole, Bell Telephone Laboratories.

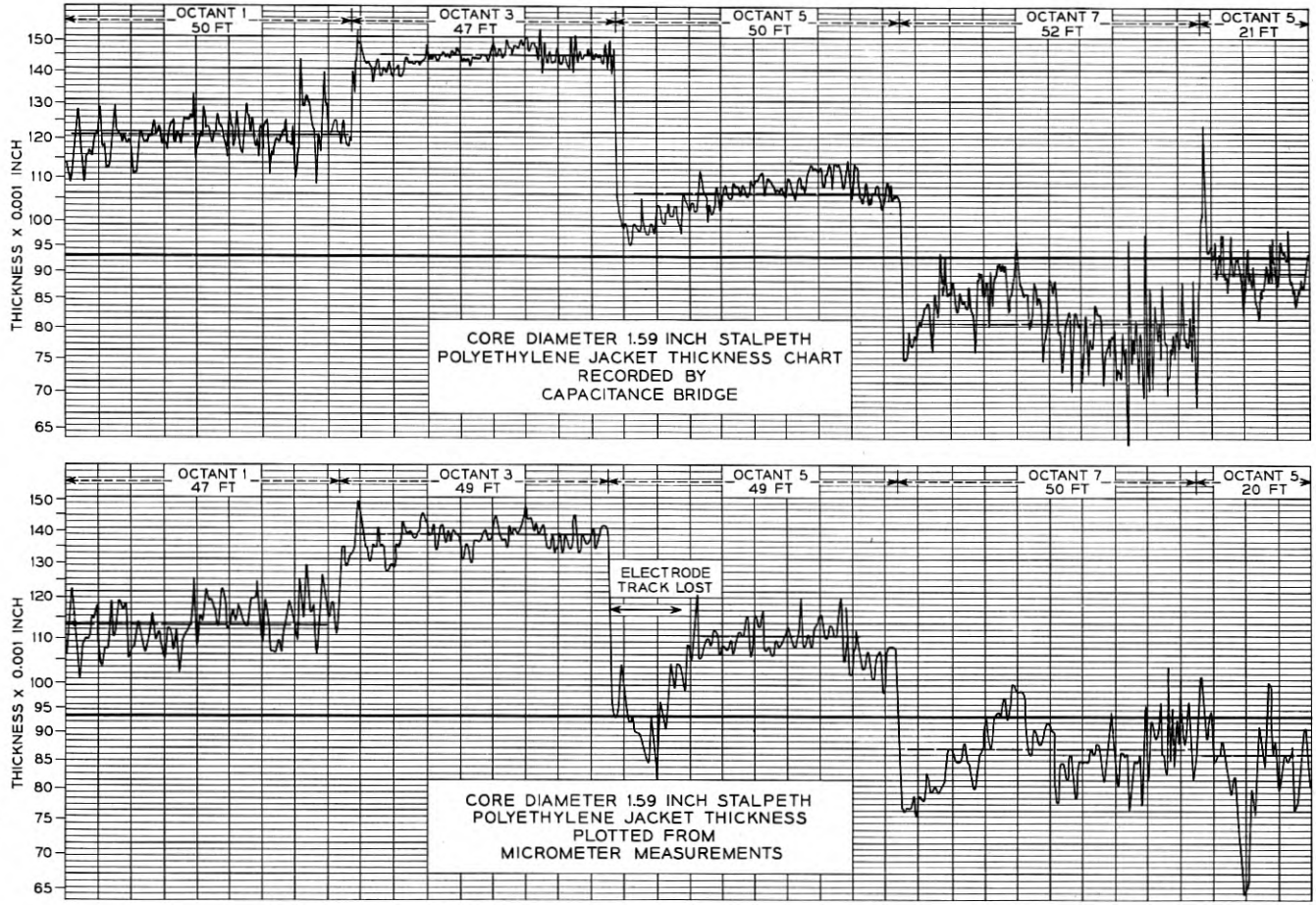


Fig. 12 — Accuracy check under operating conditions.

sheathing line is diagrammed in Fig. 13. At the top left is the supply reel of metal jacketed cable. The cable is pulled through the flood tank where the hot rubber asphaltic compound is flowed over the corrugated metal sheath. It then progresses through the die head of the extruder where the polyethylene sheath is extruded over the flooded metal sheath. The cable with plastic polyethylene then enters the cooling trough where it is cooled and solidified. At the exit of the cooling trough is an air blower for drying the water from the sheath surface. The test set is located after the dryer. The next unit is the capstan which pulls the cable. At the final unit to the right, the sheathed cable is taken up on the shipping reel.

A typical recorder graph taken along 360 feet of cable length with the sensing probe held at one location on the sheath circumference is shown in Fig. 14. With apparently stable conditions of extrusion the spot thickness indications will vary as much as plus or minus 0.010 inch while the lengthwise average remains stable as shown in Fig. 14. These fluctuations are sheath thickness variations which result from the complex interaction of the many sheathing line variables, but they may be increased or decreased by response to uneven flooding distribution and/or variations in surface curvature. However, it is practical to visually average this graph to within ± 0.001 inch.

For die adjustment, thickness measurements are obtained visually by estimating the average of the fluctuations of the recorder's visual indicator. Measurements are taken at quadrant locations corresponding to the locations of the four die adjusting screws. Opposite thicknesses give the amount of eccentricity. Die adjustments can be made accurately because the amount of eccentricity is known and the amount of die movement is governed by the adjusting screw pitch.

Adjustment to specified average sheath thickness is made by averaging measurements at eight positions equally spaced around the sheath. Increasing the speed of the cable in relation to the speed of extrusion increases the stretch of the polyethylene and decreases the average thickness. Decreasing the cable speed increases the average thickness.

APPLICATION OF TEST EQUIPMENT FOR SHEATH INSPECTION

The thickness test provides an accurate gage for the inspection organization to measure compliance of the sheath to specified requirements. Inspection possibilities with the thickness test set are many and the problem becomes one of an economic procedure that will assure the required quality. Continuous recording of the entire cable length is practical but is unnecessary from a manufacturing viewpoint. Recorder chart speed is one half inch per minute and cable speeds are from 20 to 100

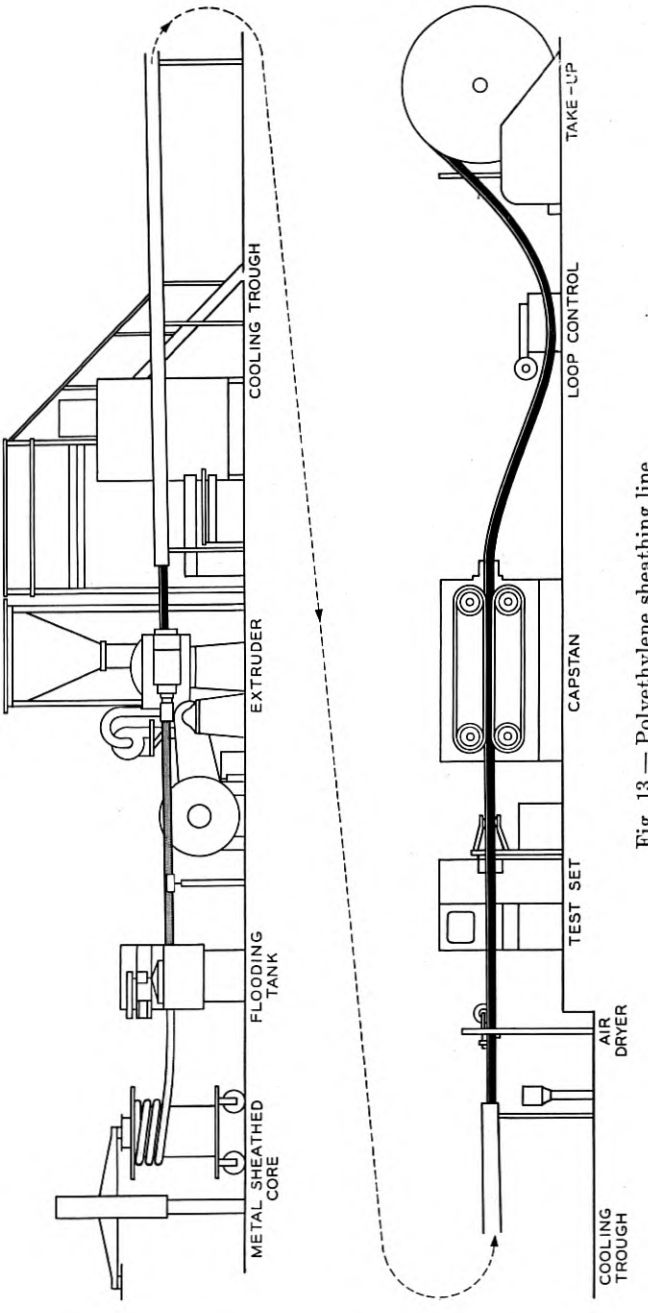


Fig. 13 — Polyethylene sheathing line.

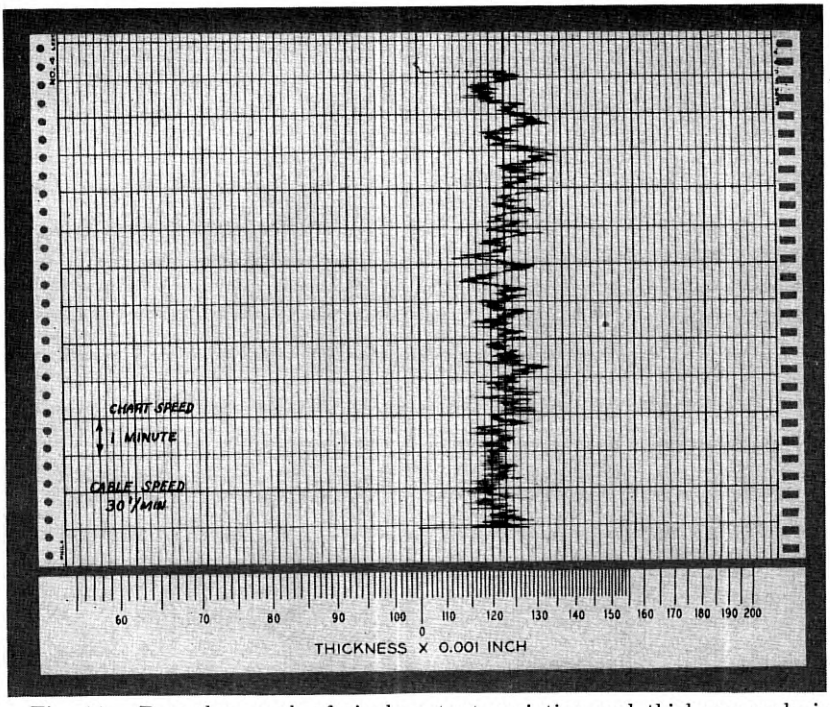


Fig. 14 — Recorder graph of single octant variation and thickness scale in thousandths of an inch.

feet per minute. It was found that the fluctuations or variation peaks of one line along the cable length could be averaged from chart lengths of ¼ inch. Also that by taking measurements consecutively by octants around the circumference a practical measure of the entire circumference is obtained and is sufficient coverage to locate the minimum wall thickness. The graphs of Fig. 15 show typical inspection recordings of two cable lengths.

Four thicknesses are specified for inspecting sheath, all of which are obtained from a graph of the consecutively recorded octants. These checks are:

1. The minimum spot thickness.
2. The average thickness lengthwise along the thinnest side. (Average of minimum octant.)
3. The average cross sectional thickness. (Average of octant averages).
4. The maximum difference between the lengthwise average of the thickest side (average of maximum octant) and the lengthwise average of the thinnest side (average of minimum octant).

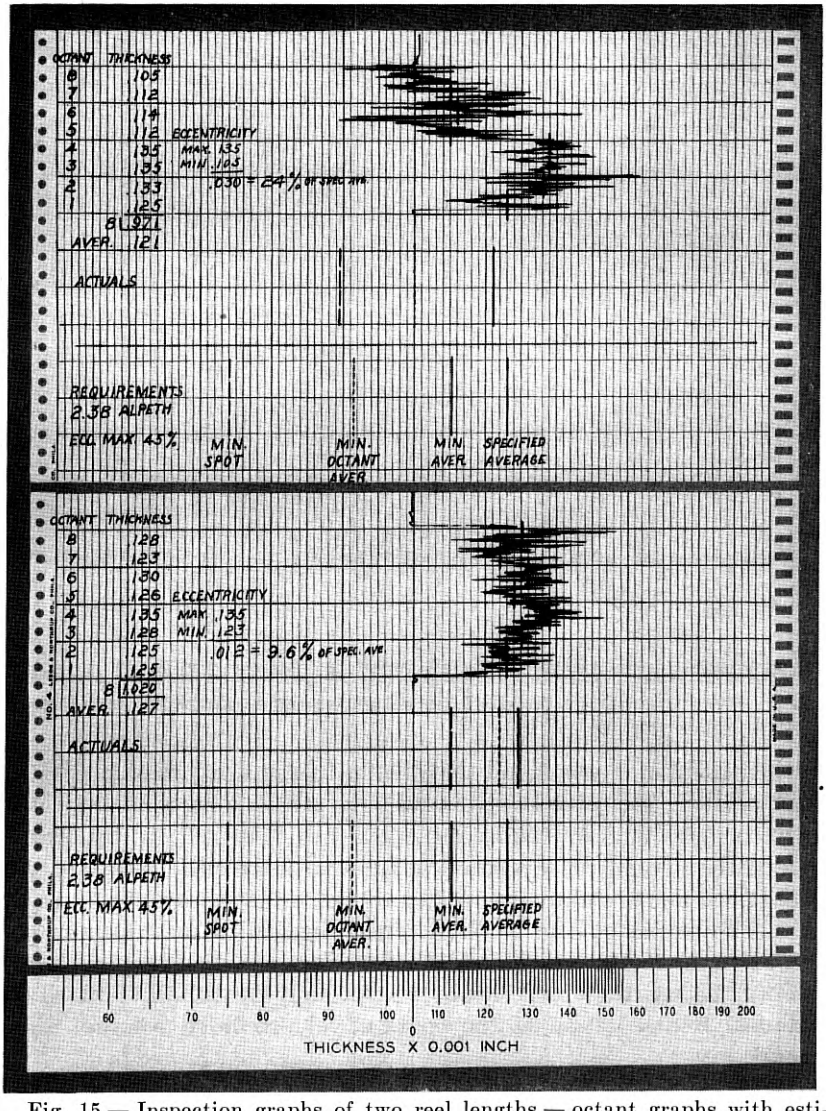


Fig. 15 — Inspection graphs of two reel lengths — octant graphs with estimated octant averages — calculation of average thickness and eccentricity; location of specified thickness and actual thickness, in thousandths of an inch.

The location of the four major thickness limits have been indicated below the test graphs.

CONCLUSIONS

This test equipment has proved to be a practical means for the control of the concentricity and the average thickness of the polyethylene sheath on Alpeth and Stalpeth cables. It is accurate, reliable and of rigid construction suitable for continuous shop use. It measures the sheath wall thickness directly in thousandths of an inch both visually and as a recorded graph and does so non-destructively as the sheath is applied.

Concentricity is maintained within 35 per cent on Alpeth and within 20 per cent on Stalpeth cable. Average thickness is controlled to within ± 0.005 inch of specified average thickness by the practice of visually averaging graphs of about twenty-five feet of cable length.

Polyethylene is conserved in two ways which reduce manufacturing costs. First, improved control permits operating at specified average thickness without varying below minimum spot limit. Previously, an excess over specified average thickness was necessary to prevent the wider range of variation from going below the specified minimum spot thickness. Second, the sheath is of consistently uniform dimensional quality not previously obtainable which made it practical to reduce the average wall thickness 11 per cent below previously specified thickness.

ACKNOWLEDGMENT

The writer wishes to express his appreciation of the co-operation of B. M. Wojciechowski of the Western Electric Company, who designed the capacitance test set and of Bell Telephone Laboratories cable engineers, in establishing the sheath requirements and for their encouragement in this project.

Topics in Guided-Wave Propagation Through Gyromagnetic Media

Part I — The Completely Filled Cylindrical Guide

By H. SUHL and L. R. WALKER

(Manuscript received January 26, 1954)

The characteristic equation for the propagation constants of waves in a filled circular guide of arbitrary radius is written in terms of magnetizing field and a carrier density, which are shown essentially to determine the dielectric and permeability tensors for a gas discharge plasma and for a ferrite. The complex structure of the spectrum of propagation constants and its dependence upon radius and the two parameters are analyzed by a semi-graphical method, supplemented by exact formulae in special regions. Thus the course of individual modes may be charted with fair accuracy.

1. INTRODUCTION

Any material medium which propagates electromagnetic disturbances possesses a local electric or magnetic structure and it is just the motion of the electric or magnetic carriers under the fields of the disturbance that determines how the propagation takes place. If a dc magnetic field be applied to the medium one may expect the local response to be altered and, consequently, to find changes in the character of the propagation. Gyromagnetic media are those for which such changes are sufficiently large to be experimentally significant. For plane waves and for optical frequencies the experimental effects and their explanation have been familiar for a great many years. The non-reciprocal rotation of the plane of polarization of light travelling parallel or antiparallel to an applied dc magnetic field, which is known as the Faraday effect, is such a phenomenon. So also is the fact that the medium becomes doubly refracting for arbitrary directions of propagation.

Interest in gyromagnetic media at longer wavelengths first arose in connection with radio propagation in the ionosphere. The ionosphere is essentially an ionic cloud and the earth supplies a magnetic field, which, for the charge densities involved, is sufficient to produce a large effect

upon propagation. Here, as in the earlier optical cases, the disturbances considered are essentially plane waves. In recent years, with the extensive development of microwave techniques, two gyromagnetic media have been investigated using guided waves. One of these is the gas discharge plasma, an ionic medium like the ionosphere, in which, however, the charge density may be varied over wide ranges in a controllable manner. The magnitude of the effects observed in such ionic media are governed by the relation of the applied frequency to the cyclotron frequency of the ions in the dc magnetic field. Goldstein and his associates[1] have studied the propagation of waves in a cylindrical waveguide within which a discharge is supported and to which a longitudinal magnetic field is applied. Among many effects which they have observed is a large Faraday rotation.

The other medium being actively investigated is the low-loss ferromagnetic medium, as exemplified by the ferrites. In this case the peculiarities of the medium have their origin in the precession of the magnetization of the ferrite about the applied field. This precession takes place with a frequency dependent upon the applied field strength and large changes in the nature of the propagation occur when the frequency of the r.f. applied field approaches this. Polder[2] worked out the effective properties of such a medium for plane waves and Hogan[3] has made various experimental studies of the propagation in cylindrical guides containing ferrite. Here, again, Faraday rotation and other non-reciprocal effects have been observed.

In this paper a variety of topics associated with the theory of guided waves in gyromagnetic media is considered, with the main emphasis laid on the ferrites. The exposition does not attempt to be systematic. Very few problems in this field admit of a thorough analytic treatment and, frequently, the more closely allied they are to the practical uses of ferrites in microwave devices the more fragmentary is the analysis. On the other hand since the problems can always be formulated it is always possible in specific cases to resort to a purely numerical solution. The problems considered here all arise in the effort to analyze the operation of various devices and different idealizations are utilized in particular cases.

In Part I the general properties of gyromagnetic media are discussed and the connection between the phenomenological constants of the medium and the underlying molecular model is derived for the ferrite and for the plasma. The assumptions necessary to render the ferrite problem tractable are discussed at some length. Maxwell's equations are written down for a general gyromagnetic medium and some of the salient features of their solution are noted. The propagation of circularly po-

larized waves in circularly cylindrical guide filled with ferrite or plasma is then considered. The characteristic equation connecting frequency and propagation constant is first derived. For the purpose of obtaining results which can be compared with experiment, a specific molecular model is chosen for the ferrite. In this way the ferrite itself is specified by a single parameter, its saturation magnetization, and its state by another, namely the applied field. The object of the calculation, then, is to find, for a given ferrite and a given guide radius, the mode spectrum of the wave guide and the variation of propagation constant with magnetic field. This is done by a semi-graphical method supplemented by exact analytic formulae in the neighborhood of certain critical points, series expansions in certain regions and some numerical computations in others. A sketch of a similar procedure applicable to the plasma is given.

It should be pointed out that the filled cylindrical waveguide is not a topic of the highest importance from the technical standpoint. It is for this reason that no effort is made here to obtain a comprehensive body of exact numerical information about the modes. One wishes, on the other hand, to exploit the simplifying features of the problem (as contrasted with the more useful case of a cylinder of ferrite not filling the guide) so that the discussion may be exhaustive, in the sense that the complete mode spectrum is exhibited.

In Part II we deal with cases of transverse magnetization. By that term we mean the following: the microwave fields propagate in a direction normal to the dc magnetization and they do not vary along the magnetization direction. They may then be separated into two independent sets of field components, of which only one explicitly depends on the dc magnetizing field. For these two fields wave impedances are defined which can be used for matching purposes. A few simple examples are then given. One special case, that of the "non-reciprocal helix" utilizing ferrite, is of importance in traveling-wave tube work and is discussed at length.[7] The slow-wave propagation along both a cylindrical and a "plane" helix are treated; magnetic loss is analyzed in some detail for the plane case, and general rules are given for its approximate determination in the cylindrical case.

In Part III perturbation theory and some miscellaneous topics are taken up. Suitable perturbation methods are developed for cases in which the wave guide fields are drastically modified over small volumes (as occurs if thin pencils or thin discs are inserted) and also for situations in which the local properties of the medium are but slightly disturbed over finite volumes. Among the miscellaneous topics dicusssed is the

propagation between infinite parallel planes filled with ferrite in a longitudinal magnetic field. The effect upon Faraday rotation of multiple reflections is considered.

2. THE PHYSICAL PROPERTIES

The propagation of electromagnetic waves in a medium is governed by Maxwell's equations which connect the space variations of $\underline{E}$ and $\underline{H}$, the electric and magnetic intensities with the time variations of $\underline{D}$ and $\underline{B}$, the electric displacement and magnetic induction. To characterize the particular medium relations may be given of the form $\underline{D} = \| \epsilon \| \underline{E}$ and $\underline{B} = \| \mu \| \underline{H}$ where $\| \epsilon \|$ and $\| \mu \|$ are the dielectric and permeability tensors. For disturbances whose amplitude is in some appropriate sense small, the elements of these tensors will be independent of rf amplitude, but will depend upon the dc state of the medium, upon the frequency of the signal and in unfavorable cases upon the wavelength of the latter. With the assumptions made in this paper the dependence upon wavelength will not arise.

The form of $\| \epsilon \|$ and $\| \mu \|$ may be known experimentally or it may be deduced from some molecular model of the medium. If the equations of motion of the parts of the medium are known under applied electric and magnetic fields, the displacement and magnetic induction resulting from this motion may be found explicitly. In isotropic media and in the absence of applied dc fields, each component of the displacement or of induction depends in the same way upon the associated component of $\underline{E}$ or $\underline{H}$. The tensors then become diagonal with equal elements. The application of a dc magnetic field, say in the z-direction, causes ions to circle about this field or magnetic dipoles to precess about it. It follows that a rf electric field in the ionic case or magnetic field in the ferrite, normal to the dc magnetic field, will produce a component of motion at right angles to itself and in time quadrature with it. From symmetry and from the equations of motion in a magnetic field the tensors may be expected to be now of the form

$$\begin{Vmatrix} a & -jb & 0 \\ jb & a & 0 \\ 0 & 0 & c \end{Vmatrix} \tag{1}$$

where a is an even function of magnetic field and b an odd function. c, in general, will be independent of the magnetic field.

That a and b at a given frequency and for a given sample of the medium are not independent but are related through the magnetizing dc field, H_0, is a fact of which we need not take cognizance when solving Maxwell's

equations subject to the appropriate boundary conditions. Their solution will determine the propagation constant β of a wave as a function of a and b, no matter what their interrelation. On the other hand, in a given experiment β is generally determined as a function of one parameter only: the magnetizing field H_0. Comparison of the family of calculated results $\beta = \beta(a, b)$, with the results $\beta = \beta(H_0)$, found experimentally will, of course, determine a and b as functions of H_0.

If, however, we have a prior knowledge of a and b in terms of H_0, either through postulating the correct dynamical model for the medium, or through independent experiments, we can utilize the functional form of a and b in our analysis of β, and thus arrive directly at β as a function of H_0. The distinction between the two methods is by no means academic; early introduction of such a functional form of a and b into the waveguide problem actually simplifies the analysis. Aside from this pragmatic consideration the latter method seems to us more appropriate for another reason: it is hardly the task of analysis of technical devices to check on the physical theories that give a and b as functions of H_0; such checks are made by experiments specifically designed to avoid the analytic complexities attending the solutions for most of the technically important structures.

Accordingly we adopt the more direct approach of expressing a and b in terms of H_0 (and, of course, in terms of the magnetic or electric carrier density of a given sample) throughout these papers, even in those few cases in which β can be expressed analytically as a function of a and b.

2.1 *Ferrites*

Most ferrites used in microwave applications are fully saturated in dc magnetic fields that are small compared with the dc field with which they are biased in operation. We shall therefore always postulate a fully saturated sample. Accordingly the magnetization vector $\underline{M}$ at a point in the sample will always be of constant magnitude, although its orientation will change in the ac field.

One equation of motion for M that takes this into account is

$$\frac{d\underline{M}}{dt} = \gamma[\underline{M} \times \underline{H}_T] - \frac{\gamma\alpha}{|\underline{M}|}[\underline{M} \times [\underline{M} \times \underline{H}_T]] \tag{2}$$

where $\underline{H}_T$ is a total effective magnetic field seen by the spins that make up $\underline{M}$, t is the time and γ is the gyromagnetic ratio appropriate to electron spins, whose g-factor is close to 2. The expression on the right hand side of (2) is in the nature of a torque; the force on $\underline{M}$ is always at right angles to $\underline{M}$, thus leaving its magnitude unchanged. The first term on

the right of (2) is quite well substantiated by quantum mechanical considerations. It is a vector normal to $\underline{M}$ and to the force $\underline{H}_T$ and is responsible for the precession. The second term is also a vector normal to $\underline{M}$, but is in the plane of $\underline{M}$ and $\underline{H}$ in a sense such as to reduce the angle of the precession. It thus represents a damping. Not much is known about the precise mechanism of the damping, so that its phenomenological representation by the second term of (2) is still in doubt.

$\underline{H}_T$, the total field acting on the electron spins, is made up of terms not all of which are of electromagnetic origin. It consists of the dc field $\underline{H}_0$ within the sample, the ac field $\underline{H}$, the anisotropy field, and the field ascribed to the quantum mechanical exchange forces between spins.

$\underline{H}_0$ in the sample must be calculated from the *applied* dc field $\underline{H}_{ext}$ by a purely magnetostatic calculation, which, in the case of sufficiently simple shapes, can be carried out with the help of the appropriate demagnetizing factors. Throughout this paper it is assumed that this problem has been solved, so that $\underline{H}_0$ is given. Furthermore it is assumed that $\underline{H}_{ext}$ and $\underline{H}_0$ are uniform. Boundary effects due to non-uniformities of $\underline{H}_0$ are neglected.

The microwave field $\underline{H}$ in the sample is one of the unknowns of the problem of propagation, and will appear in the solution of Maxwell's equations subject to the appropriate boundary conditions.

The anisotropy field, a property of a single crystal of ferrite, arises from the fact that through the medium of spin-orbit interaction, the electron spins can "see" the orbital wave-functions. Since these have the symmetry properties of the crystal, it is to be expected that the anisotropy field will be a vector function of $\underline{M}$, with the symmetry properties of the crystal. The samples of ferrite used in practice contain a great many small crystals randomly oriented, so that the net effect of the anisotropy field on microwave propagation must be obtained by means of an averaging procedure. The integrations involved are laborious and have not been carried out so far. We shall therefore neglect anisotropy altogether. Since anisotropy fields are usually of the order of a few hundred gauss, this will put our results in error below frequencies of about 3,000 mc/sec. (Corresponding to a precession frequency of $\gamma H_0 =$ 3,000 mc/sec., H_0 is about 1,100 gauss.)

The field between two spins ascribable to exchange forces will be zero when the two are parallel, and thus arises out of differences of spin orientation (that is, differences of $\underline{M}$) from place to place. In fact, analysis shows that this magnetic field is proportional to $\nabla^2\underline{M}$ for cubic crystals. Thus equation (2) really involves position coordinates as well as time. Hence the ac part $\underline{m}$ of $\underline{M}$ at a point will depend not only on the ac field

$\underline{H}$ at that point, but on values of $\underline{H}$ throughout the volume of the sample. Therefore $\underline{B}$, which is $\mu_0\underline{H} + \underline{m}$, will likewise be a functional of $\underline{H}$ over the whole sample. Fortunately it turns out that the spatial variation of $\underline{H}$ in a microwave structure is so much slower than that characteristic of the "spin waves" to which $\nabla^2\underline{M}$ gives rise that this effect is quite negligible at microwave frequencies. Only in the most immediate vicinity of gyromagnetic resonance could such effects become significant.

Thus, we shall regard $\underline{H}_T$ simply as the sum of the dc and ac magnetic fields, $\underline{H}_0 + \underline{H}$, and correspondingly $\underline{M}$ as the sum of the dc magnetization (directed along $\underline{H}_0$ in a saturated sample when anisotropy is neglected) plus an ac part $\underline{m}$. Equation (2) must now be solved for $\underline{m}$ in terms of $\underline{H}$. It is a non-linear equation, whose solution $\underline{m}$ will depend on $\underline{H}$ non-linearly, as will $\underline{B}$. Even if $\underline{m}$ could be determined in this way, Maxwell's equations would become non-linear, and hope of their solution remote. It is therefore necessary, and in the great majority of applications also quite sufficient, to assume that the ac quantities in (2) are so small that their products can be neglected and only linear terms taken into account. The terms $\underline{m}$ and $\underline{H}$ may now be assumed to vary as $\exp j\omega t$.

Under these circumstances, (2) becomes

$$\frac{d\underline{m}}{dt} = \gamma([\underline{m} \times \underline{H}_0] + [\underline{M}_0 \times \underline{H}]) - \frac{\alpha\gamma}{|M_0|}([\underline{M}_0 \times [\underline{m} \times \underline{H}_0]] + [\underline{M}_0 \times [\underline{M}_0 \times \underline{H}]]),$$

and is easily solved for $\underline{m}$ in terms of $\underline{H}$, and of the dc quantities H_0, M_0 which we shall assume to point in the z-direction. Each of the components m_x, m_y is a linear function of both H_x and H_y and when they are substituted in the components of the equation $\underline{B} = \mu_0\underline{H} + \underline{m}$, lead to expressions of the form (1) for $\underline{B}$ in terms of $\underline{H}$:

$$\begin{aligned} B_x &= \mu H_x - j\kappa H_y\,, \\ B_y &= j\kappa H_x + \mu H_y\,, \qquad \text{and} \\ B_z &= \mu_0 H_z\,. \end{aligned} \tag{3}$$

It is convenient to introduce two auxiliary quantities

$$\sigma = \frac{|\gamma|\,H_0}{\omega}; \qquad p = \frac{|\gamma|\,M_0}{\mu_0\omega},$$

and in terms of these one obtains the relations first derived by Polder:

$$\frac{\mu}{\mu_0} = 1 + p\,\frac{\sigma(1+\alpha^2) + j\alpha \operatorname{sgn} p}{\sigma^2(1+\alpha^2) - 1 + 2j\alpha\sigma \operatorname{sgn} p}, \qquad \text{and}$$
$$\frac{\kappa}{\mu_0} = \frac{-p}{\sigma^2(1+\alpha^2) - 1 + 2j\alpha\sigma \operatorname{sgn} p}, \tag{4}$$

where the function

$$\operatorname{sgn} p = +1 \qquad p > 0$$
$$= -1 \qquad p < 0$$

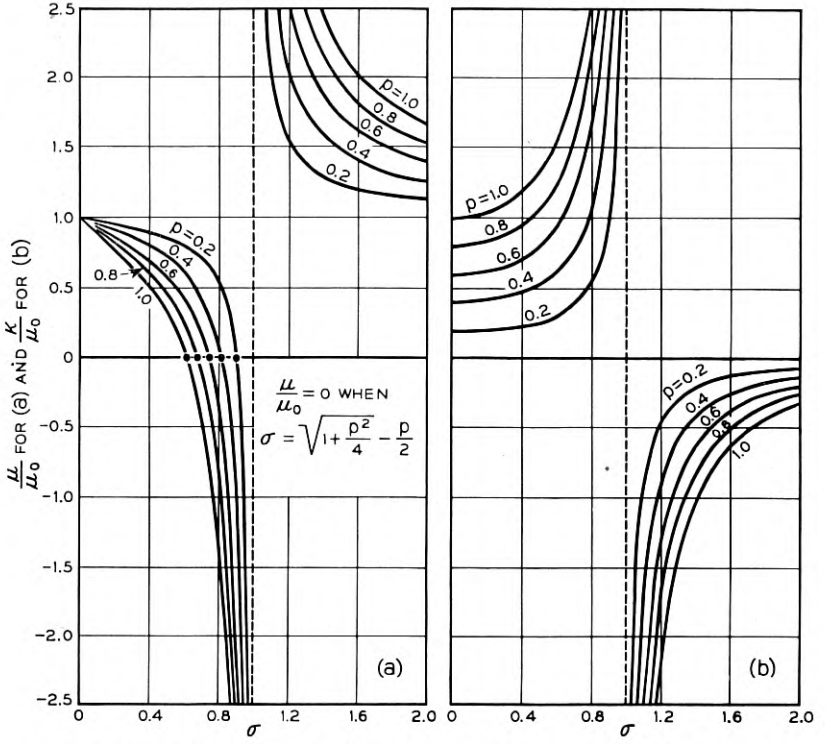
σ is the ratio of the natural precession frequency $\frac{1}{2\pi} \mid \gamma \mid H_0$ to the signal frequency. p is the ratio of a frequency $\frac{1}{2\pi} \mid \gamma \mid M_0/\mu_0$, associated with the saturation magnetization M_0, to the signal frequency. Note that σ and p always have similar signs: if H_0 is reversed, so is the saturation magnetization. Equations (4) are true only for a fully saturated sample. Therefore they hold good only for values of σ greater than the very small value corresponding to the amount of H_0 required to saturate the sample. In practice that value of H_0 is generally so small that this restriction is trivial. In the text a number of formulae will appear which apply "near $\sigma = 0$". These are to be understood as applying near the very small value of σ that corresponds to saturation.

Equation (4) has an interesting implication with regard to the loss parameter α. If α were zero, we would have

$$\frac{\mu}{\mu_0} = 1 - \frac{p\sigma}{1-\sigma^2}, \qquad \text{and}$$
$$\frac{\kappa}{\mu_0} = \frac{p}{1-\sigma^2}, \tag{5}$$

and these equations describe the loss-free case. If in equations (5), σ is replaced by $(\sigma + j\alpha \operatorname{sgn} p)$, the resulting expressions check (4) to order α. For small α, it follows that any propagation problem need be considered for the loss-free case (5) only.* The first order change due to loss in any formula so obtained can be deduced by differentiation of the formula

* A form of the damping term in Equation (2), no less justified experimentally than the one used above, is $-\frac{\alpha}{\mid \underline{M} \mid}\left(\underline{M} \times \frac{d\underline{M}}{dt}\right)$. When this expression is used the permeabilities are exactly functions of the variable, $\sigma + j\alpha \operatorname{sgn} p$.



Figs. 1(a) and 1(b) — The relative permeabilities μ/μ_0 and κ/μ_0 versus σ.

with respect to σ, and multiplication by $j\alpha$ sgn p. Of course this procedure is invalid close to resonance ($\sigma = 1$), when terms in α^2 play an important part. Equations (5), which will hereafter be called the Polder equations, are plotted in Fig. 1. The quantity

$$\rho_H = \frac{\kappa}{\mu} = \frac{p}{1 - p\sigma - \sigma^2}$$

is shown in Fig. 1(c). It occurs in the waveguide theory, and also in the theory of other microwave circuits considered later on. The ratio $\mu/\mu_0 = \mu/\mu_z$ will be denoted by ν_H. At a fixed p, μ/μ_0 decreases from unity at $\sigma = 0$, through zero at $\sigma = -p/2 + \sqrt{p^2/4 + 1}$ to $-\infty$ at $\sigma = 1 - 0$, and then from $+\infty$ at $\sigma = 1 + 0$ steadily down to unity at $\sigma = \infty$. κ/μ_0 increases from p at $\sigma = 0$ to $+\infty$ at $\sigma = 1 - 0$, and then again from $-\infty$ at $\sigma = 1 + 0$ to zero at $\sigma = \infty$.

It has already been mentioned that the anisotropy fields are of the order of a few hundred gauss. For most ferrites the saturation magnetiza-

tion is about 1,000 gauss. It will therefore be consistent with the neglect of anisotropy to assume that the applied frequency is such that p is less than unity and this will be done hereafter.

2.2 *Ion clouds or plasmas*

Since these are considered in much less detail in these papers, their physical properties are stated only briefly here.

Instead of a tensor relationship between $\underline{B}$ and, $\underline{H}$ we now have one between the displacement vector $\underline{D}$ and the electric field E. If the magnetizing field is along the z axis, we have

$$\begin{aligned} D_x &= \epsilon E_x - j\eta E_y , \\ D_y &= j\eta E_x + \epsilon E_y , \text{ and} \\ D_z &= \epsilon_z E_z . \end{aligned} \tag{6}$$

If the medium consists of equal densities R of positive ions and electrons, and if collisions and thermal velocities are neglected, ϵ and η can be calculated for weak ac disturbances $\underline{E}e^{j\omega t}$ from the equation of motion

$$\dot{\underline{v}} = \frac{e}{m}\underline{E}e^{j\omega t} + \gamma[\underline{v} \times \underline{H}],$$

where $\underline{v}$ is the velocity vector of the electron and $\gamma = e\mu_0/m$, in the usual notation. When this equation is solved and the abbreviations

$$\omega_0 = |\gamma| H_0; \qquad \sigma = \frac{\omega}{\omega_0}; \qquad q = \frac{\omega_p}{\omega}; \qquad \omega_p = \frac{Re^2}{m\epsilon_0}$$

are introduced, one obtains, from the fact that the total current is $j\omega\epsilon_0\underline{E} + R\underline{v}$, the heavy ions being assumed stationary,

$$\begin{aligned} \epsilon &= \epsilon_0\left(1 + \frac{q^2}{\sigma^2 - 1}\right), \\ \eta &= \epsilon_0 \frac{q^2\sigma}{\sigma^2 - 1}, \text{ and} \\ \epsilon_z &= \epsilon_0(1 - q^2), \end{aligned} \tag{7}$$

where ϵ_0 is the dielectric constant of vacuum. The waveguide theory will involve the parameters

$$\begin{aligned} \nu_E &= \frac{\epsilon}{\epsilon_z} = 1 - \frac{\sigma^2 q^2}{(1 - \sigma^2)(1 - q^2)}, \text{ and} \\ \rho_E &= \frac{\eta}{\epsilon} = \frac{q^2\sigma}{\sigma^2 + q^2 - 1}. \end{aligned} \tag{8}$$

These results apply to stationary plasmas only. If the plasma were an electron stream moving along the wave-propagation direction, for example, the dielectric constants would depend on wavelength also, and the propagation problem would be much more involved.

The variations of ϵ and η with σ are shown in Fig. 2. ϵ/ϵ_0 for a given q^2 starts at $\sigma = 0$, $\epsilon = \epsilon_0 (1 - q^2)$, decreases through zero at $\sigma = \sqrt{1 - q^2}$ to $-\infty$ at $\sigma = 1 - 0$, starts again from $+\infty$ at $\sigma = 1 + 0$, and decreases to ϵ_0 at $\sigma = \infty$. $\eta = 0$ when $\sigma = 0$, decreases to $-\infty$ at $\sigma = 1 - 0$ and then decreases from $+\infty$ at $1 + 0$ to zero at $\sigma = \infty$.

We note that similar formulae apply to the electron-gas in semiconductors at temperatures sufficiently low and frequencies sufficiently high so that damping is not important. However, the formulae have to be

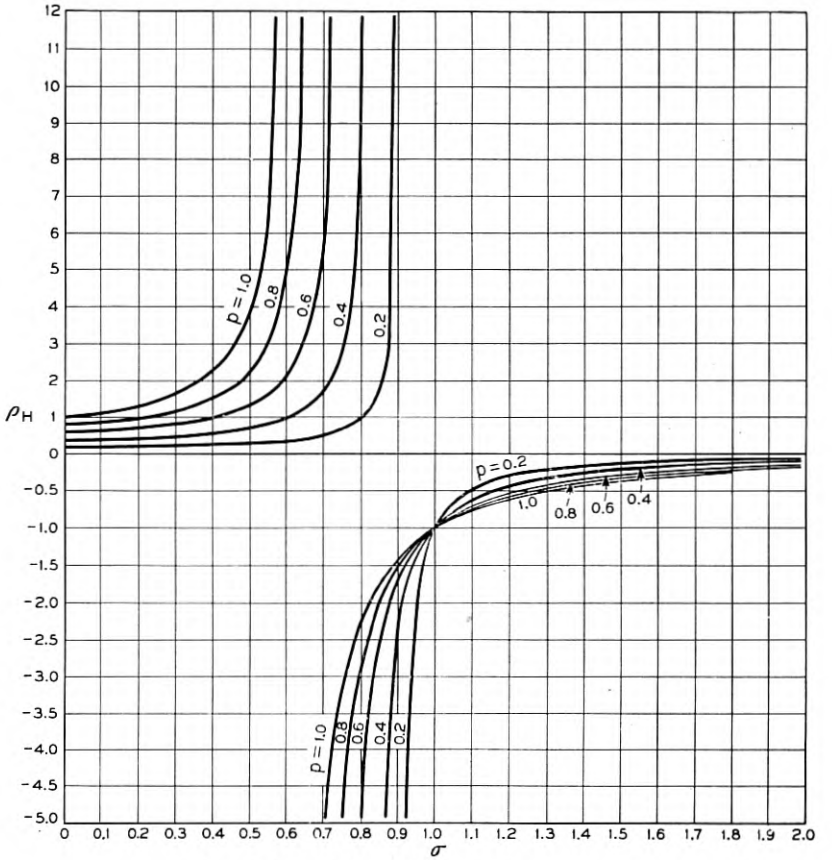


Fig. 1(c) — The ratio $\rho_H = \kappa/\mu$ versus σ.

generalized in some of those cases to take into account the existence of groups of electrons with different "effective masses" m.

3. THE SOLUTION OF MAXWELL'S EQUATIONS

Maxwell's equations will now be solved in a cylindrical waveguide filled with a hypothetical medium which contains the ferrite and the plasma as a special case. It will be supposed therefore that both its permeability and dielectric constant are tensors of the form previously considered.

3.1 *Field components*

The following notation will be found convenient. The projection of a vector $\underline{A}$ upon the plane normal to the z-axis will be written $\underline{A}_t$. If the components of $\underline{A}_t$ are α, β then an associated vector having components $(\beta, -\alpha)$ is devoted by $\underline{A}_t^*$. A similar notation is used for differential operators. Thus, if ∇ denotes $(\partial/\partial x, \partial/\partial y)$, ∇^* denotes $(\partial/\partial y, -\partial/\partial x)$†. Denoting scalar products by a dot, the following identities are evident

$$\underline{A}_t^* \cdot \underline{A}_t^* = \underline{A}_t \cdot \underline{A}_t ; \qquad (\underline{A}_t^*)^* = -\underline{A}_t ; \qquad \underline{A}_t \cdot \underline{A}_t^* = 0;$$
$$\underline{A}_t \cdot \underline{B}_t^* = -\underline{A}_t^* \cdot \underline{B}_t ;$$

and

$$\underline{A}_t \cdot \underline{B}_t^* = z\text{-component of } [\underline{A} \times \underline{B}].$$

Also if $\underline{k}$ is a unit vector along the positive z-axis, $\underline{k} \times \underline{A} = -\underline{A}_t^*$. Similar relations hold for differential operators. If one denotes the starring operation by the symbol P then clearly

$$P^2 = -1; \qquad P^{-1} = -P; \qquad \frac{1}{P + a} = \frac{1}{1 + a^2}(a - P),$$

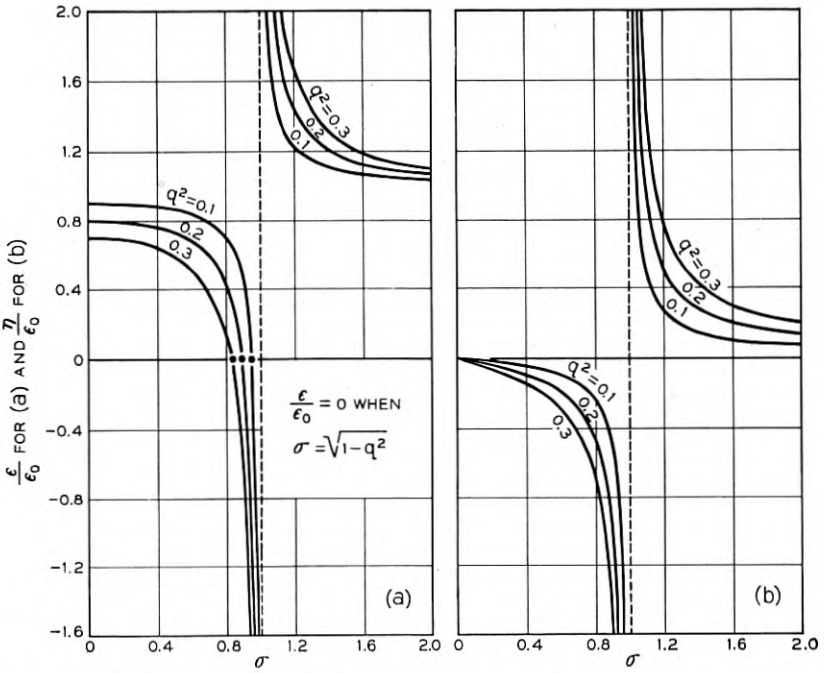
where a is a number.

Maxwell's equations may now be written, for that case in which the dependence of any component upon t and z is of the form $e^{j(\omega t - \beta z)}$, in the form:

$$\begin{aligned} \nabla^* H_z + j\beta \underline{H}_t^* &= j\omega\epsilon \underline{E}_t + \omega\eta \underline{E}_t^*, \\ \nabla \cdot \underline{H}_t^* &= j\omega\epsilon_z E_z , \\ \nabla^* E_z + j\beta \underline{E}_t^* &= -j\omega\mu \underline{H}_t - \omega\kappa \underline{H}_t^*, \text{ and} \\ \nabla \cdot \underline{E}_t^* &= -j\omega\mu_z H_z , \end{aligned} \tag{9}$$

where use is made of equations (3) and (6).

† The operator ∇^* is called "flux" by Schelkunoff. Strictly, one should write ∇_t and ∇_t^*, rather than ∇ and ∇^*, but this is needlessly cumbersome.



Figs. 2(a) and 2(b) — The relative dielectric constants ϵ/ϵ_0 and η/ϵ_0 versus σ.

It is desirable to remove scale factors as far as possible. A unit of length given by

$$\frac{1}{\beta_0} = \frac{1}{\omega\sqrt{\mu_z\epsilon_z}}$$

will be used to measure lengths. This unit is $\lambda_0/2\pi$, where λ_0 is the wavelength in an unbounded, unmagnetized medium. It will be assumed that β is in future measured in units of β_0. Finally all magnetic fields will be multiplied by $\sqrt{\mu_z/\epsilon_z}$ to give them the dimensions of electric fields. Using the definitions of the ν's and ρ's given in Section 2, Maxwell's equations may be put into the form:

$$\nabla^* H_z + j\beta \underline{H}_t^* = \nu_E(j\underline{E}_t + \rho_E \underline{E}_t^*), \tag{10a}$$

$$\nabla \cdot \underline{H}_t^* = jE_z, \tag{10b}$$

$$\nabla^* E_z + j\beta \underline{E}_t^* = -\nu_H(j\underline{H}_t + \rho_H \underline{H}_t^*), \tag{10c}$$

and

$$\nabla \cdot \underline{E}_t^* = -iH_z. \tag{10d}$$

$\underline{E}_t$ and $\underline{H}_t$ may now be eliminated yielding two simultaneous second order equations for E_z and H_z. These, in turn, may be combined to produce two independent second order equations each of which is satisfied by an appropriate linear combination of E_z and H_z. These equations may be solved and E_z and H_z expressed as linear combinations of the solutions. The transverse fields are then written in terms of E_z and H_z and, finally, the boundary conditions are applied leaving a transcendental equation in β^2.

Operating on (10a) and (10c) with $\nabla\cdot$ and taking account of (10b), (10d), one finds that

$$\begin{aligned} j\beta\nabla\cdot\underline{H}_t^* &= \nu_E(j\nabla\cdot\underline{E}_t + j\rho_E H_z) \quad = -\beta E_z\,, \text{ and} \\ j\beta\nabla\cdot\underline{E}_t^* &= -\nu_H(j\nabla\cdot\underline{H}_t + j\rho_H E_z) = \beta H_z \end{aligned} \tag{11}$$

Operating on (10a) and (10c) with $\nabla^*\cdot$, using $\nabla^*\cdot\nabla^* = \nabla^2$ and so on, one obtains, using (10b) and (10d),

$$\begin{aligned} \nabla^2\underline{H}_z + j\beta\nabla\cdot H_t &= \nu_E(-H_z + \rho_E\nabla\cdot\underline{E}_t), \text{ and} \\ \nabla^2\underline{E}_z + j\beta\nabla\cdot E_t &= -\nu_H(E_z + \rho_H\nabla\cdot\underline{H}_t). \end{aligned} \tag{12}$$

Now, elimination of $\nabla\cdot E_t$ and $\nabla\cdot H_t$ between (11) and (12) yields

$$\begin{aligned} \nabla^2 H_z + \nu_E\left(1 - \rho_E{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right)H_z &= j\beta(\rho_E + \rho_H)E_z\,, \text{ and} \\ \nabla^2 E_z + \nu_H\left(1 - \rho_H{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right)E_z &= -j\beta(\rho_E + \rho_H)H_z\,, \end{aligned} \tag{13}$$

equations which demonstrate that pure TE or TM fields no longer exist, as the result of the presence of ρ's. H_z or E_z might now be eliminated between these equations giving a single equation in ∇^2 and $(\nabla^2)^2$, but it is more convenient to find those linear combinations of E_z and H_z which satisfy a first order equation in ∇^2. Writing such a linear combination as

$$\psi = E_z + j\Lambda H_z\,, \tag{14}$$

and adding $j\Lambda$ times the first of equations (13) to the second, it is found that this is an equation in ψ alone of the form

$$\nabla^2\psi + \chi^2\psi = 0, \tag{15}$$

provided that Λ is a root of the quadratic

$$\Lambda^2 - \frac{\nu_E\left(1 - \rho_E{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right) - \nu_H\left(1 - \rho_H{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right)}{\beta(\rho_E + \rho_H)}\Lambda - 1 = 0. \tag{16}$$

The value of χ^2 is then given by

$$\chi_{1,2}^{\ 2} = \nu_E \left(1 - \rho_E^{\ 2} - \frac{\beta^2}{\nu_E \nu_H}\right) - \beta(\rho_E + \rho_H)\Lambda_{2,1}\,, \tag{17a}$$

or

$$\chi_{1,2}^{\ 2} = \nu_H \left(1 - \rho_H^{\ 2} - \frac{\beta^2}{\nu_E \nu_H}\right) + \beta(\rho_E + \rho_H)\Lambda_{1,2}\,, \tag{17b}$$

where Λ_1 and Λ_2 are the roots of (16) and $\chi_1^{\ 2}$, $\chi_2^{\ 2}$ are the corresponding χ^2. The labelling of the roots is not important, but consistency must be maintained. From (14) E_z and H_z must satisfy

$$E_z + j\Lambda_1 H_z = \psi_1\,,$$

and

$$E_z + j\Lambda_2 H_z = \psi_2$$

so that

$$E_z = \frac{\Lambda_2\psi_1 - \Lambda_1\psi_2}{\Lambda_2 - \Lambda_1}\,, \tag{18a}$$

and

$$H_z = j\,\frac{\psi_1 - \psi_2}{\Lambda_2 - \Lambda_1}\,. \tag{18b}$$

Solutions of (15) may now be sought in cylindrical coordinates. To satisfy the boundary conditions in circular guide it will be necessary to assume the solutions to vary as $e^{jn\theta}$, where θ is the polar angle and n is any integer, positive, negative or zero. Equation (15) then becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\,\frac{\partial\psi_{1,2}(r)}{\partial r}\right) + \left(\chi_{1,2}^{\ 2} - \frac{n^2}{r^2}\right)\psi_{1,2}(r) = 0,$$

if r is the radius. Solutions which are regular within the guide will have the form of constant multiples of $J_n(\chi_{1,2}r)$, where J_n is the n^{th} order Bessel function. The solutions of (15) are, then,

$$\psi_{1,2} = A_{1,2}J_n(\chi_{1,2}r)e^{jn\theta}, \tag{19}$$

where the A's are constants. E_z and H_z can be found now from (18), but further equations must be found to express $\underline{E}_t$ and $\underline{H}_t$. Using P to denote the starring operation, (10a) and (10c) may be re-written as

$$(j\nu_F P - \rho_E\nu_E)\underline{E}_t + j\beta\underline{H}_t = -\nabla\underline{H}_z\,,$$

and

$$-j\beta \underline{E}_t + (j\nu_H P - \rho_H \nu_H)\underline{H}_t = \nabla E_z ,$$

which yield

$$\{[\nu_E\nu_H(1 + \rho_E\rho_H) - \beta^2] - j\nu_H\nu_E(\rho_H + \rho_E)P\}\underline{E}_t = -j\beta\nabla E_z - \nu_H(jP - \rho_H)\nabla H_z ,$$

and

$$\{[\nu_E\nu_H(1 + \rho_E\rho_H) - \beta^2] - j\nu_H\nu_E(\rho_H + \rho_E)P\}\underline{H}_t = \nu_E(jP - \rho_E)\nabla E_z - j\beta\nabla H_z$$

The term in parentheses may be removed by using the rule for inverting such expressions in P which was given earlier. This process gives

$$\Omega\underline{E}_t = \left[\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right) + j(\rho_H + \rho_E)P\right] [- j\beta\nabla E_z - \nu_H(jP - \rho_H)\nabla H_z], \tag{20a}$$

and

$$\Omega\underline{H}_t = \left[\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right) + j(\rho_H + \rho_E)P\right] [\nu_E(jP - \rho_E)\nabla E_z - j\beta\nabla H_z], \tag{20b}$$

where

$$\Omega = \nu_E\nu_H\left[\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)^2 - (\rho_E + \rho_H)^2\right],$$

$$= \nu_E\nu_H\left[\frac{\beta^2}{\nu_E\nu_H} - (1 + \rho_E)(1 + \rho_H)\right]\left[\frac{\beta^2}{\nu_E\nu_H} - (1 - \rho_E)(1 - \rho_H)\right].$$

It may be noted that for plane waves in the unbounded medium along the z axis, which have $E_z = H_z = 0$, Ω must vanish and that the propagation constants for such plane waves are evidently given by

$$\beta^2 = \nu_E\nu_H(1 \pm \rho_E)(1 \pm \rho_H). \tag{21}$$

The values of E_z and H_z given by (18) may now be substituted in (20) and the operator P removed. This gives, finally,

$$(\Lambda_1 - \Lambda_2)\Omega\underline{E}_t = j\left[\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)(\beta\Lambda_2 - \rho_H\nu_H) + \nu_H(\rho_H + \rho_E)\right]\nabla\psi_1 - \left[(\beta\Lambda_2 - \rho_H\nu_H)(\rho_E + \rho_H) + \nu_H\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)\right]\nabla^*\psi_1 \tag{22a}$$

minus the same expression with suffixes 1 and 2 interchanged.

$$(\Lambda_1 - \Lambda_2)\Omega \underline{H}_t = \left[\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)(\Lambda_2\nu_E\rho_E - \beta) - \nu_E\Lambda_2(\rho_E + \rho_H)\right]\nabla\psi_1$$
$$- j\left[\nu_E\Lambda_2\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right) - (\rho_E + \rho_H)(\Lambda_2\nu_E\rho_E - \beta)\right]\nabla^*\psi_1 \qquad (22b)$$

minus the same expression with suffixes 1 and 2 interchanged.

Equations (22a) and (22b) may be written in a variety of equivalent forms by making use of the relations between Λ_1 and Λ_2. The manipulations which have been used in deriving (22a) and (22b) assume the use of rectangular coordinates, but the results are valid in polar coordinates if $\underline{E}_t$ means (E_r, E_θ) and ∇ means $\left(\frac{\partial}{\partial r}, \frac{1}{r}\frac{\partial}{\partial\theta}\right)$. That this is the case may be seen from the consideration that the rotation, $-\theta$, which carries the vector (E_x, E_y) into the vector (E_r, E_θ) also transforms $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$ into $\left(\frac{\partial}{\partial r}, \frac{1}{r}\frac{\partial}{\partial\theta}\right)$.*

3.2 *The characteristic equation*

The boundary conditions of the problem are that $E_z = 0$ and $E_\theta = 0$ at $r = r_0$, the radius of the guide. E_z is given by [see (18)].

$$(\Lambda_2 - \Lambda_1)E_z = [\Lambda_2 A_1 J_n(\chi_1 r) - \Lambda_1 A_2 J_n(\chi_2 r)]e^{jn\theta}, \qquad (23)$$

and vanishes at $r = r_0$ if

$$A_1 = \frac{J_n(\chi_2 r_0)}{\Lambda_2}; \qquad A_2 = \frac{J_n(\chi_1 r_0)}{\Lambda_1}.$$

Hence the relations hold:

$$\psi_{1,2} = \frac{1}{\Lambda_{2,1}} J_n(\chi_{2,1} r_0) J_n(\chi_{1,2} r) e^{jn\theta}.$$

From (22a) it follows that

$$(\Lambda_1 - \Lambda_2)\Omega E_\theta = \frac{J_n(\chi_2 r_0)e^{jn\theta}}{\Lambda_2}\left[-\frac{n}{r}\left\{\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)(\beta\Lambda_2 - \rho_H\nu_H)\right.\right.$$
$$\left.+ \nu_H(\rho_H + \rho_E)\right\} J_n(\chi_1 r) + \left\{(\beta\Lambda_2 - \rho_H\nu_H)(\rho_E + \rho_H)\right.$$
$$\left.\left. + \nu_H\left(1 + \rho_E\rho_H - \frac{\beta^2}{\nu_E\nu_H}\right)\right\}\chi_1 J_n'(\chi_1 r)\right]$$

* In Appendix III the field components in polar coordinates are written out fully for the ferrite and plasma cases with some changes in notation which are introduced in Sections 4.11 and 4.2.

minus the same expression with the suffixes interchanged. Hence

$$
\begin{aligned}
&(\Lambda_1 - \Lambda_2)\Omega E_\theta(r_0) \\
&= \frac{J_n(\chi_1 r_0)J_n(\chi_2 r_0)e^{jn\theta}}{r_0}\left[(\Lambda_1 - \Lambda_2)n\nu_H\left(\frac{\rho_H\beta^2}{\nu_E\nu_H} + \rho_E(1 - \rho_H{}^2)\right)\right. \\
&\quad + \left\{\beta(\rho_E + \rho_H) - \Lambda_1\nu_H\left(1 - \rho_H{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right)\right\}\frac{\chi_1 r_0 J_n'(\chi_1 r_0)}{J_n(\chi_1 r_0)} \\
&\quad \left. - \left\{\beta(\rho_E + \rho_H) - \Lambda_2\nu_H\left(1 - \rho_E{}^2 - \frac{\beta^2}{\nu_E\nu_H}\right)\right\}\frac{\chi_2 r_0 J_n'(\chi_2 r_0)}{J_n(\chi_2 r_0)}\right],
\end{aligned}
\tag{24}
$$

where use has been made of the relation $\Lambda_1\Lambda_2 = -1$. Therefore the characteristic equation for β^2 is obtained by equating the term in square brackets to zero. Because of the quadratic relation satisfied by Λ and the relation between Λ and χ, it is possible to write the characteristic equation in a great variety of ways. It will be convenient to introduce a function

$$
F_n(x) = F_{-n}(x) = \frac{xJ_n'(x)}{J_n(x)}. \tag{25}
$$

Using the F-function and replacing the Λ's by χ's the characteristic equation may be written:*

$$
\begin{aligned}
n\nu_H(\chi_2{}^2 - \chi_1{}^2)\left[\frac{\rho_H\beta^2}{\nu_E\nu_H} + \rho_E(1 - \rho_H{}^2)\right]\frac{1}{\beta(\rho_E + \rho_H)} &= \frac{\chi_2{}^2}{\Lambda_2}F_n(\chi_1 r_0) \\
&\quad - \frac{\chi_1{}^2}{\Lambda_1}F_n(\chi_2 r_0).
\end{aligned}
\tag{26}
$$

The asymmetry of this equation between ρ_H, ν_H and ρ_E, ν_E arises from the fact that the boundary conditions involve electric field components alone.

It may be noted that if the basic solution had been taken to vary as $\cos n\theta$ or $\sin n\theta$, the expression for E_θ would have been a linear combination of $\sin n\theta$ and $\cos n\theta$ that could not have vanished at the walls for all θ.

In passing we remark that for a guide of arbitrary cross-section, the

* The characteristic equations given in Reference 4 were specializations to the ferrite and plasma cases of the form in square brackets. They have also been derived by Kales[5] and Gamo[6]. These authors have given expressions for some, though not all, of the varieties of cut-off point derived in this paper and classified them as TE or TM according to the field configuration at cut-off. By contrast, they are classified here by their association with quasi-TE or quasi-TM limit modes which reduce to the usual TE and TM modes in the unmagnetized medium.

boundary value problem may be put into the form, of which (26) is a special case,

$$\nabla^2 f_1 + \chi_1^2 f_1 = 0,$$

and

$$\nabla^2 f_2 + \chi_2^2 f_2 = 0,$$

$$j\left(\frac{\chi_2^2}{\Lambda_2}\frac{\partial f_1}{\partial N} - \frac{\chi_1^2}{\Lambda_1}\frac{\partial f_2}{\partial N}\right) = \frac{\nu_H(\chi_2^2 - \chi_1^2)}{\beta(\rho_E + \rho_H)}\left[\frac{\rho_H\beta^2}{\nu_E\nu_H} + \rho_E(1 - \rho_H^2)\right]\frac{\partial f_1}{\partial S},$$

where $\partial/\partial N$ and $\partial/\partial S$ are normal and tangential derivatives at the guide surface, where, in addition, $f_1 = f_2$.

4. DISCUSSION OF THE PROPAGATION CONSTANTS

At this point we specialize the characteristic equation (26) to one or other of the two media.

4.1. *The ferrite* ($\rho_E = 0$, $\nu_E = 1$)

4.11. After some rearrangement the characteristic equation becomes

$$\frac{1}{\chi_1^2}\left[\frac{F_n(\chi_1 r_0)}{\lambda_1} - n\right] = \frac{1}{\chi_2^2}\left[\frac{F_n(\chi_2 r_0)}{\lambda_2} - n\right], \tag{27}$$

where $\lambda_{2,1} = \beta\,\Lambda_{1,2}$ and the λ satisfy

$$\lambda_{1,2}^2 - \frac{(1 - \nu_H)\left(1 - \dfrac{\beta^2}{\nu_H}\right) + \nu_H\rho_H^2}{\rho_H}\lambda_{1,2} - \beta^2 = 0. \tag{28}$$

The χ's are given by

$$\chi_{1,2}^2 = \left(1 - \frac{\beta^2}{\nu_H}\right) - \rho_H\lambda_{1,2}. \tag{29}$$

From Polder's equations for ρ_H and ν_H, (28) may be written

$$\lambda_{1,2}^2 - [p + \sigma(1 - \beta^2)]\lambda_{1,2} - \beta^2 = 0, \tag{30}$$

or

$$\lambda_1\lambda_2 = -\beta^2, \tag{31a}$$

$$\lambda_1 + \lambda_2 = p + \sigma(1 - \beta^2),$$

$$= p + \sigma + \sigma\lambda_1\lambda_2. \tag{31b}$$

If β^2 be eliminated between equations (28) and (29), $\chi_{1,2}^2$ may be ex-

pressed solely in terms of $\lambda_{1,2}$, ρ_H and ν_H in the form

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{\dfrac{1}{1 - \dfrac{\nu_H}{\rho_H}\lambda_{1,2}} - 1}.$$

Again using Polder's formulae, this becomes

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \sigma\lambda_{1,2}}. \tag{32}$$

With these expressions for the χ, the characteristic equation takes the form

$$G(\lambda_1, \sigma, r_0) = G(\lambda_2, \sigma, r_0), \tag{33}$$

where

$$G(\lambda, \sigma, r_0) = \frac{1 - \sigma\lambda}{1 - \lambda^2}\left[\frac{1}{\lambda} F_n\left(r_0\sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}\right) - n\right]. \tag{34}$$

Equations (31b) and (33) may now be considered for a fixed σ and p as determining associated pairs of values for λ_1 and λ_2. Such a pair in turn determines $\beta^2 = -\lambda_1\lambda_2$. Since β^2 must be positive for propagation λ_1, λ_2 must have opposite signs. The convention will be adopted that λ_1 is positive and λ_2 is negative. Equation (31b) will hereafter be called the Polder relation and Equation (33) the G-equation.

An important fact of which frequent use will be made is that the transformation

$$\lambda_1 \rightarrow -\lambda_2, \qquad \lambda_2 \rightarrow -\lambda_1, \qquad \sigma \rightarrow -\sigma, \qquad p \rightarrow -p$$

leaves the Polder relation and β^2 unchanged and converts n to $-n$ in the G-equation. It follows that it is necessary to consider positive n only, provided we allow the pair σ, p to take on negative as well as positive values. This corresponds to the physical fact that a right-circular wave in a backward-directed magnetizing field behaves like a left-circular wave in a forward field.

The discussion in this paper is confined to the first azimuthal mode number $n = \pm 1$. Accordingly the symbol F will replace F_1 in what follows.

Before commencing the graphical analysis of the G function it is advantageous to consider briefly the function $F(x) = xJ_1'(x)/J_1(x)$, which we require for real and for purely imaginary x. By logarithmic differentia-

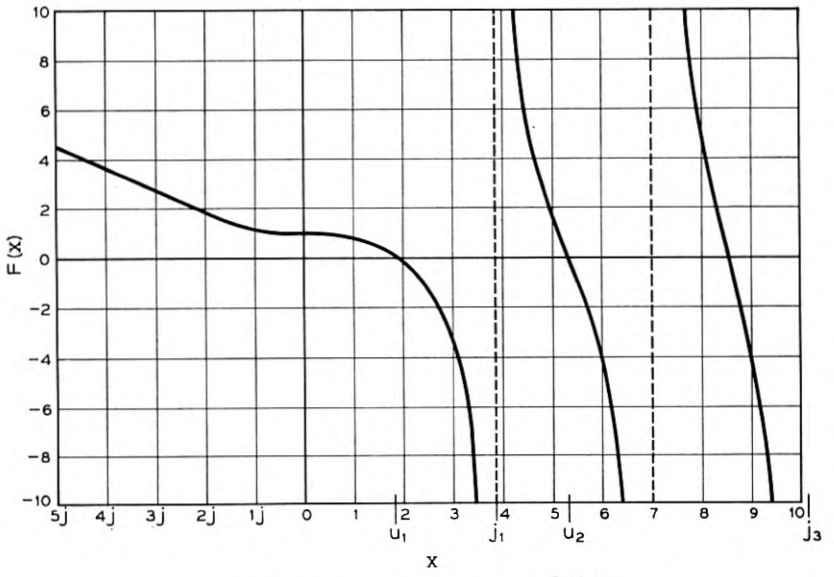


Fig. 3 — The function $F(x) = \dfrac{xJ_1'(x)}{J_1(x)}$.

tion of the infinite product for $J_1(x)$, $F(x)$ is found to be given by

$$F(x) = 1 - 2\sum_{n=1}^{\infty} \frac{x^2}{j_n^2 - x^2},$$

where the j_n's are the zeros of $J_1(x)$.

Thus, $F(x)$ is real if x^2 is, which is always the case here. For positive x, $F(x)$ is an always decreasing function of x, which has an infinite number of first order zeros and poles. The zeros are those of $J_1'(x)$ and will be denoted by u_n . The poles are the zeros of $J_1(x)$. It may be recalled from the properties of Bessel functions that for large n these zeros and poles are essentially equally spaced with a separation $\pi/2$. When x is a pure imaginary, equal to jy, $F(x)$ becomes $yI_1'(y)/I_1(y)$. This is a steadily increasing function of y, always positive, and behaving like $y - \frac{1}{2}$ for large y. The function F is shown in Fig. 3. Further formulae pertaining to F are given in Appendix I. The inverse function $F^{-1}(x)$, which is also of some importance, is a multivalued function of x, whose behavior is readily understood from the figure for $F(x)$. We are now ready to proceed with the graphical analysis of the G-equation.

In a rectangular coordinate system with λ as abscissa and σ as ordinate, a contour map is sketched of the function

$$G = \frac{1 - \sigma\lambda}{1 - \lambda^2}\left[\frac{1}{\lambda} F\left(r_0 \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}\right) - 1\right]$$

for all values of λ, σ from $-\infty$ to $+\infty$, r_0 being kept fixed. This can be done as accurately as desired by first drawing the contours $\chi^2 = \frac{1 - \lambda^2}{1 - \sigma\lambda}$ = constant (Fig. 4), along each of which G simply behaves like $A/\lambda + B$ and is easily evaluated with the help of a table of F. However,

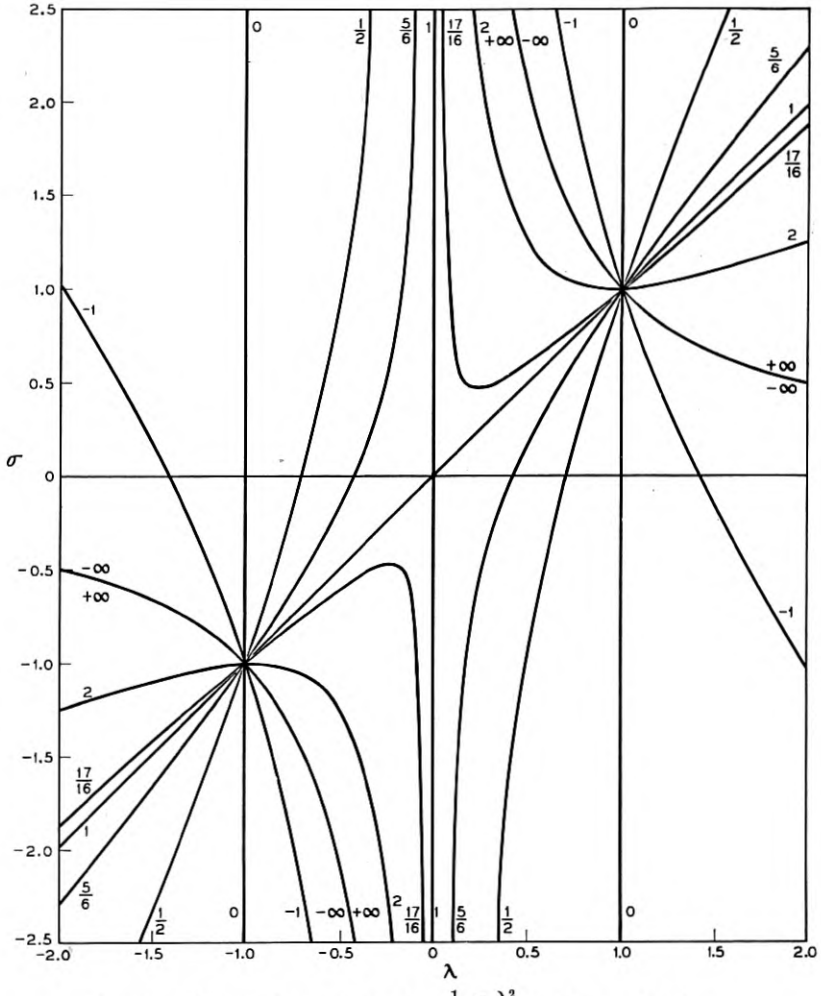


Fig. 4 — Curves of constant $\chi^2 = \frac{1 - \lambda^2}{1 - \sigma\lambda}$ in the $\sigma - \lambda$ plane.

many features of the G-contours are already determined by the position of the contours $G = \infty$, and $G = 0$, across which G changes sign (from $\pm\infty$ to $\mp\infty$ or from ± 0 to ∓ 0). Because of their special role in the subsequent analysis it is desirable to introduce a scheme for their enumeration. The infinity and zero curves in the right-hand half-plane will be denoted by I and 0, respectively, those in the left-hand half-plane by I' and $0'$. All but two of the I-curves arise from the poles j_n of F. Their equations are

$$\frac{1 - \lambda^2}{1 - \sigma\lambda} = j_n^2/r_0^2 \qquad n = 1, 2, \cdots$$

Each of these curves has two branches, one in the half-plane $\lambda > 0$, one in $\lambda < 0$ and these are called I_n, I_n' respectively. All I_n curves pass through $\lambda = 1$, $\sigma = 1$, all I_n' curves pass through $\lambda = -1$, $\sigma = -1$. The lines $\lambda = 0$, $\lambda = -1$ are also infinity curves to be denoted by I_A, I_B' respectively (As $\lambda \to +1$, G tends to a finite value).

Zero curves of G are given by

$$F\left(r_0\sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}\right) = \lambda,$$

or in a more readily computable form by

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda)]^2}. \tag{35}$$

The branches of $F^{-1}(\lambda)$ may be labelled according to the scheme: "0" for $-\infty < [F^{-1}(\lambda)]^2 < j_1^2$; "1" for $j_1^2 < [F^{-1}(\lambda)]^2 < j_2^2$ and so on. The nth branch of $F^{-1}(\lambda)$ gives rise to an 0_n curve for $\lambda > 0$ and to an $0_n'$ curve for negative λ. All $0_n'$ curves pass through $\lambda = -1$, $\sigma = -1$; all save one of the 0_n curves pass through $\lambda = 1$, $\sigma = 1$. The exceptional one, seen to be 0_0, is associated with the "0" branch of $F^{-1}(\lambda)$ on which $F^{-1}(1) = 0$. For fixed σ, G tends to zero as $\lambda \to \infty$, hence the vertical lines $\lambda = \pm\infty$ are also zero curves, to be denoted by 0_∞ and $0_\infty'$ respectively.

In a sense the two branches of $\sigma\lambda = 1$ are also zero curves, to be called 0_c and $0_c'$. 0_c and $0_c'$ are zero curves only when viewed from "one side." In the right half-plane, for $\lambda < 1$ as $\sigma\lambda \to 1 - 0$ and for $\lambda > 1$ as $\sigma\lambda \to 1 + 0$, the argument of F tends to infinity and remains real. Therefore G passes through all values an indefinite number of times and $\sigma\lambda = 1$ is a limit line of all contours, $G =$ constant. For $\lambda < 1$ as $\sigma\lambda \to 1 + 0$ and for $\lambda > 1$ as $\sigma\lambda \to 1 - 0$, the argument of F is

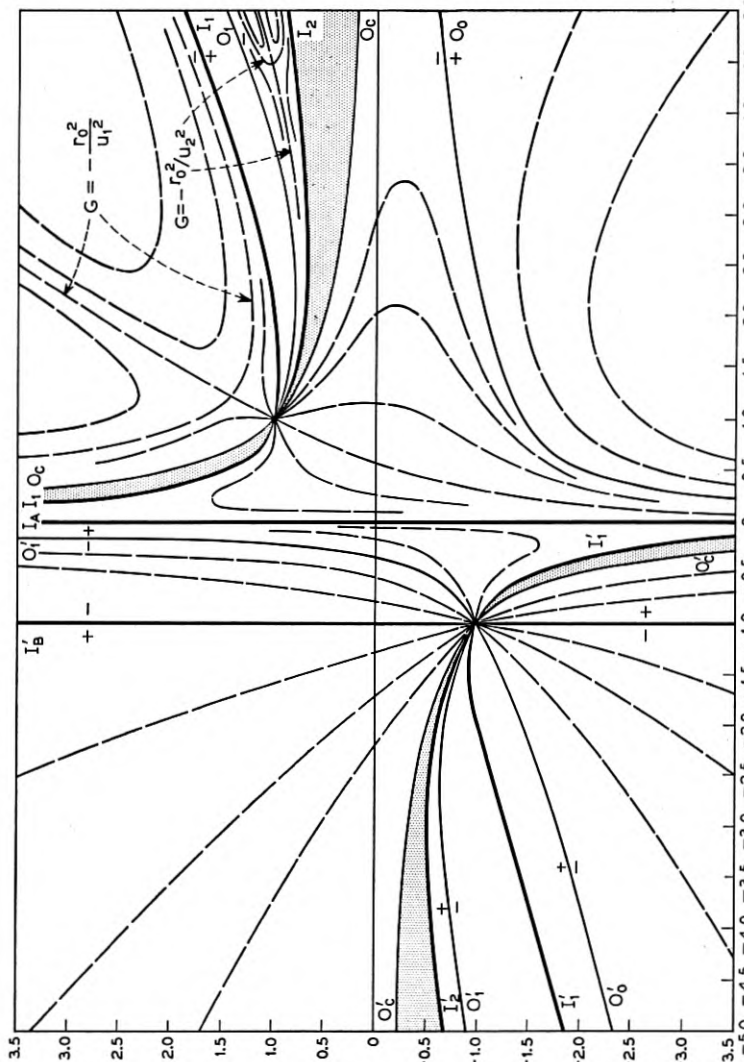


Fig. 5 — Contours of constant $G(\lambda, \sigma)$ for $r_0 = 2.2$. Scales distorted to show saddle points. G assumes any given value infinitely often in the shaded regions.

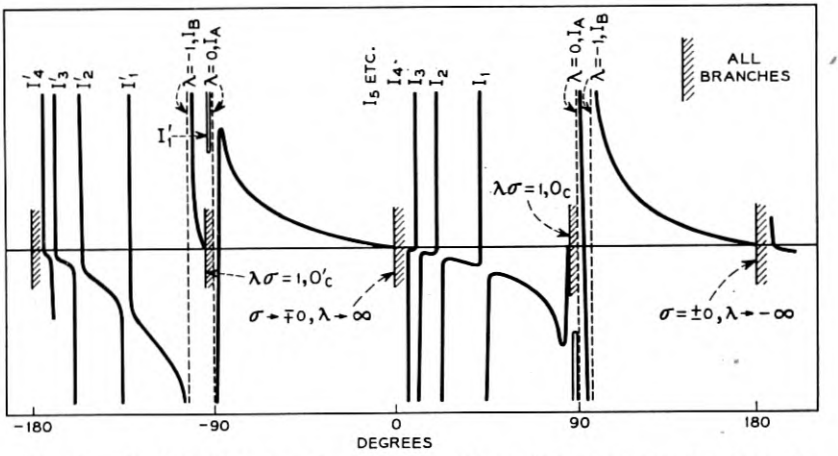


Fig. 6 — Qualitative behavior of $G(\lambda, \sigma)$ at large distances from the origin as a function of arc tan σ/λ. r_0 is about 2.

imaginary and F tends to infinity. However $F \Big/ \dfrac{1-\lambda^2}{1-\sigma\lambda}$ tends to zero so that G tends to zero.

To complete the picture of the G-function given by the form and position of the 0 and I curves it is necessary to see how it behaves at large distances from the origin. This is indicated in Fig. 5 and also by Fig. 6. The latter shows the value of G at large distances as a function of direction. In general, along the line $\sigma = c\lambda + d$ (c finite), G will tend to $-c$ for all d. For $c = r_0^2/j_n^2$ (which is the slope of the asymptotes to the I_n curves), G again tends to a constant. Now, however, the constant depends upon d and assumes all values from $-\infty$ to $+\infty$ as a function of d. In the first quadrant the sign of variation of the limiting value of G with direction c is opposite to that of its variation with d near $c = r_0^2/j_n^2$. Consequently local maxima and minima arise as a function of direction between successive I_n-curves. This suggests the existence of saddle points, which may be verified directly. In the third quadrant, the dependence of G upon c and d does not give such maxima and minima, and indeed no saddle points are found there. Finally it is necessary to consider the behavior of G as σ tends to infinity, while λ remains finite, corresponding to $(1/c) \to 0$. If λ remains fixed, then for $\lambda > -1$, $G \to \mp\infty$ as $\sigma \to \pm\infty$; and for $\lambda < -1$, $G \to \pm\infty$ as $\sigma \to \pm\infty$. As $\lambda \to 0$, the curves of constant G are asymptotic to $\lambda\sigma = \left(1 - \dfrac{r_0^2}{u_n^2}\right) - B\lambda$, where B goes from $-\infty$ to $+\infty$ with G. Interleaved with these families of curves are the curves

$G = \pm\infty$, which are $\lambda\sigma = \left(1 - \frac{r_0^2}{j_n^2}\right) + 0(\lambda^2)$. More detailed information on these matters will be found in Appendix II.

From the G-diagram it would be possible to determine pairs of λ-values with opposite signs, which, for a definite σ-value satisfy the characteristic equation, but, for a given p such pairs would not necessarily satisfy the Polder relation (31b). It is necessary to have a procedure which takes account of the latter systematically. Such a method may be based upon the fact that if, for σ and p positive, the Polder relation is solved for λ_1 in terms of λ_2 it can be thought of as a rather simple mapping of the whole λ_2-quadrant upon a part of the λ_1-quadrant ($\lambda_1 > 0$). Similarly for σ and p negative there is an analogous mapping of the λ_1-quadrant onto the λ_2-quadrant.

Considering first the case $\sigma, p > 0$, the Polder relation may be written in the forms

$$\lambda_1 = \frac{\sigma + p - \lambda_2}{1 - \sigma\lambda_2} = \frac{1}{\sigma} + \frac{\sigma + p - 1/\sigma}{1 - \sigma\lambda_2} = T(\lambda_2). \tag{36}$$

From (36) it may be seen that the curves $\lambda_2 = $ const. transform into a bundle of hyperbolae passing through the intersection of $\sigma = 1/\lambda_1$ and $\sigma = \lambda_1 - p$; that is, through λ_{10}, σ_0, where

$$\sigma_0 = -p/2 + \sqrt{\frac{p^2}{4} + 1}, \qquad \lambda_{10} = p/2 + \sqrt{\frac{p^2}{4} + 1}.$$

These hyperbolae have the vertical asymptotes $\lambda_1 = -1/\lambda_2$, and intersect $\sigma = 0$ at $\lambda_1 = p - \lambda_2$. For a fixed positive σ less than σ_0, λ_1 decreases from $1/\sigma$ to $\sigma + p$ as λ_2 increases from $-\infty$ to 0, but when σ is greater than σ_0, λ_1 increases from $1/\sigma$ to $\sigma + p$ under the same circumstances. Thus the whole λ_2-quadrant is transformed upon that part of the λ_1-quadrant which lies between the hyperbola $\lambda_1 = 1/\sigma$ and the straight line $\lambda_1 = \sigma + p$. It follows that points in the λ_1-quadrant which are, for a given p, excluded from this region, cannot be the site of acceptable solutions of the G-equation.

Since as has already been stated, the Polder relation is unchanged by the substitution $\lambda_1 \to -\lambda_2$, $\lambda_2 \to -\lambda_1$, $\sigma \to -\sigma$, and p to $-p$, it follows that for σ and p negative a similar mapping of the λ_1-quadrant upon part of the λ_2-quadrant takes place. The transforms of the lines $\lambda_1 = $ const. and so forth may easily be found by using these substitutions in the formulae already given.

Reference to Fig. 1(a) and (b) will show that $\pm\sigma_0$ are the values of σ at which μ reverses sign. Therefore we may expect σ_0 to play a special role

in the propagation theory, as also does $\sigma = 1$. The following scheme exists: for $0 < \sigma < \sigma_0$, κ and μ are both positive; for $\sigma_0 < \sigma < 1$, $\kappa < 0$ and $\mu < 0$, for $\sigma > 1$, κ is negative and μ positive. If σ is changed to $-\sigma$, μ goes into μ, and κ into $-\kappa$.

The procedure which will now be used to discuss the solution of the characteristic equation, observing the Polder relations, begins by writing the equation, for σ, p positive, in the form

$$G(\lambda_1, \sigma, r_0) = G(T(\lambda_1), \sigma, r_0)$$

We are already in possession of a contour map of the left hand side of this equation in the quadrant $\sigma > 0$, $\lambda > 0$, and of the function $G(\lambda_2, \sigma, r_0)$ in the quadrant $\lambda < 0$, $\sigma > 0$. The latter surface has now to be transformed into one in the λ_1-quadrant by the relation

$$\lambda_2 = T(\lambda_1) = (\sigma + p - \lambda_1)/(1 - \sigma\lambda_1)$$

(or equally well, $\lambda_1 = T(\lambda_2)$. This may be effected by considering the transformation of curves $G(\lambda_2, \sigma, r_0) =$ constant, onto the λ_1-quadrant. For the I' curves whose analytical expression in terms of σ and λ_2 is very simple, the corresponding explicit expression of the transformed curve in λ_1 and σ is simple. Contours other than I' are most easily transformed by replotting $G(\lambda_2, \sigma, r_0) =$ const. in the hyperbola-mesh formed by the lines $T(\lambda_2)$. However, information about particular points and about asymptotic behavior of these transformed curves is available in analytic form and is stated in Appendix II. The two surfaces so obtained will intersect in various curves, along whose projections on the $\lambda - \sigma$ plane both Polder relation and G-equation are satisfied. For each such projection λ_1 is a function of σ, λ_2 is then known in terms of σ and p, and finally $\beta^2 = -\lambda_1\lambda_2$ is known. In most cases the general course of these curves can be found without resort to much numerical analysis. Each of the curves is associated with a definite mode and it follows that the classification of the modes can be carried out fairly easily. The approximate location of the solution curves relies upon the fact that if the position of the infinity curves of both surfaces is known, continuity considerations will frequently assure the existence of an intersection within certain regions. Moreover, the neighborhood of certain special points on these solution curves can be investigated analytically. These are points at which one or both of the G-functions may be approximated by a simpler expression; included among these is the point at infinity.

It is clear that for σ and p negative the whole procedure outlined above may be carried out in a similar way, with the $\sigma > 0, \lambda > 0$ quadrant now being transformed on to the $\sigma < 0, \lambda < 0$ quadrant.

It is possible to translate such solution curves into $\beta^2 - \sigma$ curves in a direct graphical manner if a mesh of constant β^2 lines is drawn in the first quadrant. From (30) these are given by

$$\lambda_1^2 - [p + \sigma(1 - \beta^2)]\lambda_1 - \beta^2 = 0,$$

or

$$\frac{1 - \lambda_1^2}{1 - \left[\sigma + \dfrac{p}{1 - \beta^2}\right]\lambda_1} = 1 - \beta^2.$$

The contour, $\beta = b$, is just the contour $\chi^2 = 1 - b^2$ displaced along the σ-axis by an amount, $-p/(1 - b^2)$. The contours of constant β all pass through the point $\sigma = \sigma_0$, $\lambda = \sigma_0 + p$ and are shown in Fig. 7. Their

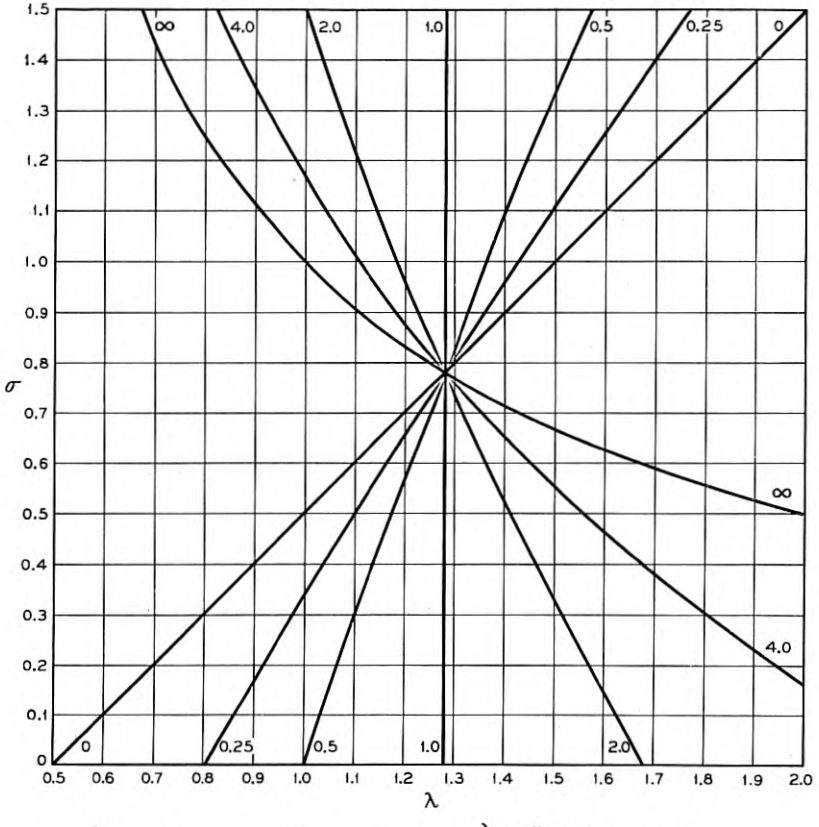


Fig. 7 — Contours of constant $\beta^2 = \lambda \dfrac{\lambda - p - \sigma}{1 - \sigma\lambda}$ for $p = 0.5$.

$[\beta^2(\lambda, \sigma, p) = \beta^2(-\lambda, -\sigma, -p)]$

course in the third quadrant is immediately found by reflection in the origin.

When $p = 0$ the magnetization of the ferrite vanishes and it is clear that we should then obtain just the modes of a guide filled with isotropic material ($\mu = \mu_z$; $\kappa = 0$). Superficially it might appear that, since the equations (31b) and (33) depend upon σ, even for $p = 0$, this result might not be attained. We now show that β^2 is indeed independent of σ for $p = 0$. It may first be noted that in this case if $\sigma \neq 1$, the Polder relation (31b) transforms

$$\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1} \text{ into } \frac{1 - \lambda_2^2}{1 - \sigma\lambda_2} \text{ and } \lambda_1 \text{ into } \frac{\sigma - \lambda_2}{1 - \sigma\lambda_2}.$$

The G-equation reads

$$\frac{1 - \lambda_2\sigma}{1 - \lambda_2^2}\left[\frac{1}{\lambda_2} F\left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}}\right) - 1\right]$$
$$= \frac{1 - \lambda_2\sigma}{1 - \lambda_2^2}\left[\frac{1 - \sigma\lambda_2}{\sigma - \lambda_2} F\left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}}\right) - 1\right].$$

Since $\lambda_1 \neq \lambda_2$ we must have

$$F\left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}}\right) = 0 \quad \text{or} \quad \infty,$$

or

$$\frac{1 - \lambda_{1,2}^2}{1 - \sigma\lambda_{1,2}} = \frac{u_n^2}{r_0^2} \quad \text{or} \quad \frac{j_n^2}{r_0^2}.$$

Thus $\lambda_{1,2}$ are roots of

$$\lambda^2 - \sigma\frac{u_n^2}{r_0^2}\lambda + \left(\frac{u_n^2}{r_0^2} - 1\right) = 0,$$

or else of

$$\lambda^2 - \sigma\frac{j_n^2}{r_0^2}\lambda + \left(\frac{j_n^2}{r_0^2} - 1\right) = 0,$$

In the first case

$$-\lambda_1\lambda_2 = \beta^2 = 1 - \frac{u_n^2}{r_0^2},$$

and in the second

$$-\lambda_1\lambda_2 = \beta^2 = 1 - \frac{j_n^2}{r_0^2}.$$

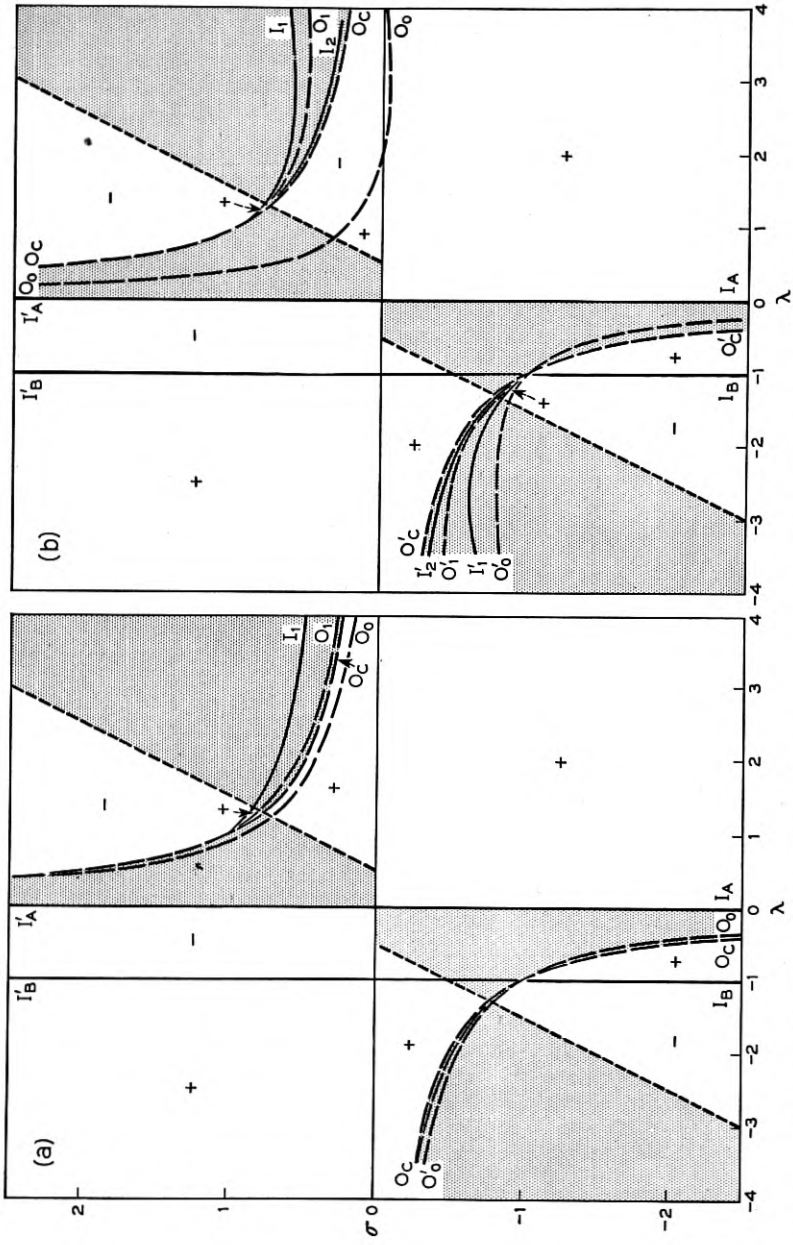


Fig. 8 — The division of the parts of the $\lambda - \sigma$ plane allowed by the Polder relation, Equation (31b), into regions of positive and negative G by the first few 0 and I curves. Excluded regions are shaded. $|p|$ is about 0.5. Fig. 8(a), $0 < r_0 < 1$; Fig. 8(b), $1 < r_0 < u_1$; Fig. 8(c), $r_0 = u_1$; Fig. 8(d), $u_1 < r_0 < j_1$; Fig. 8(e), $j_1 < r_0 < u_2$.

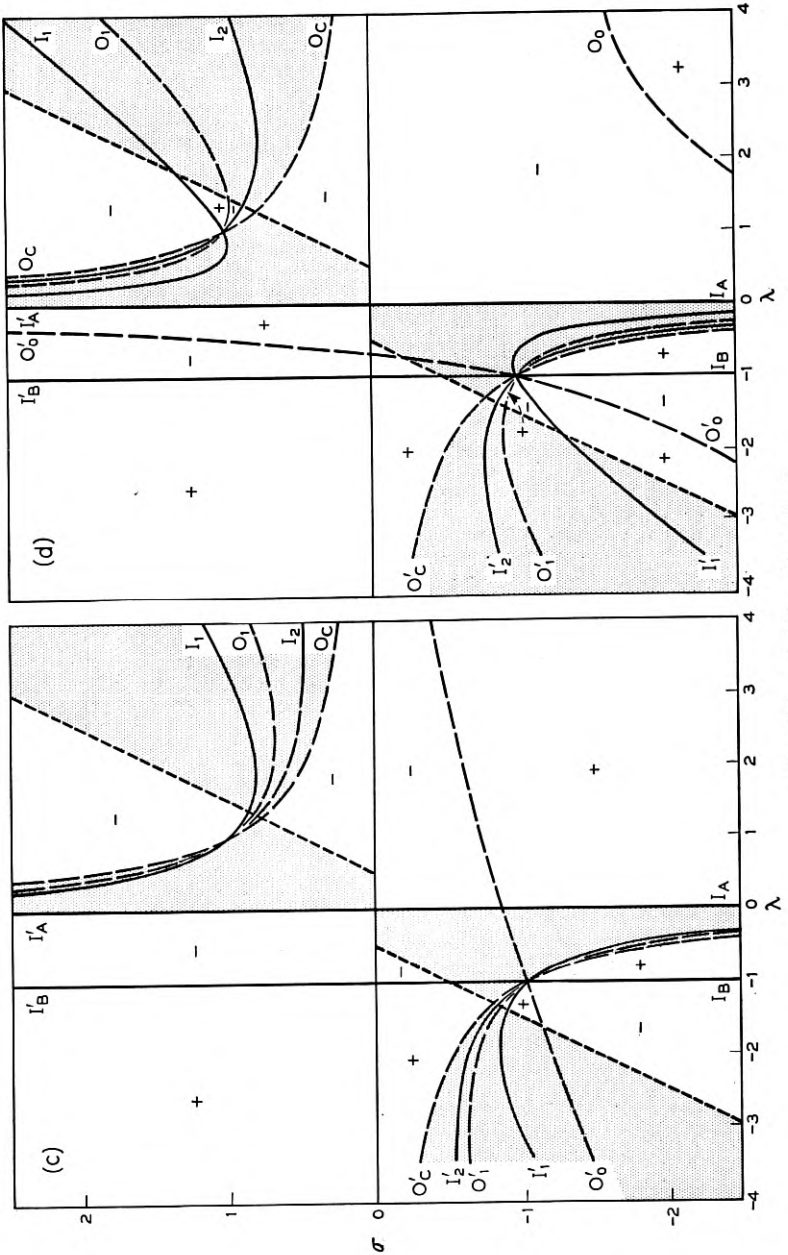


Fig. 8(c) and (d) — See Fig. 8.

Thus, when $\sigma \neq 1$, the β^2 values, for $p = 0$, are evidently independent of σ and are just those of an isotropic medium. When $\sigma = 1(\mu = \kappa = \infty)$, $p = 0$, β^2 is indeterminate and for p small, there is a small region near $\sigma = 1$, of width $\sim p$, in which β^2 differs appreciably from the isotropic value. The convergence of an expansion of β^2 in powers of p[(61) and (62)] shows a marked dependence on σ.

4.12. The scheme of analysis described above will now be illustrated in detail by a discussion for a radius r_0 between u_1 and j_1, which, if the ferrite were unmagnetized, would propagate the TE_{11}-mode alone.

Figs. 8(c) and 8(d) show the division of the $\lambda - \sigma$ plane into regions of positive and negative $G(\lambda, \sigma)$ by the various I and 0 curves. A few contours of constant G are plotted to indicate the behavior of the function in more detail. That part of the $\lambda - \sigma$ plane which is excluded by the

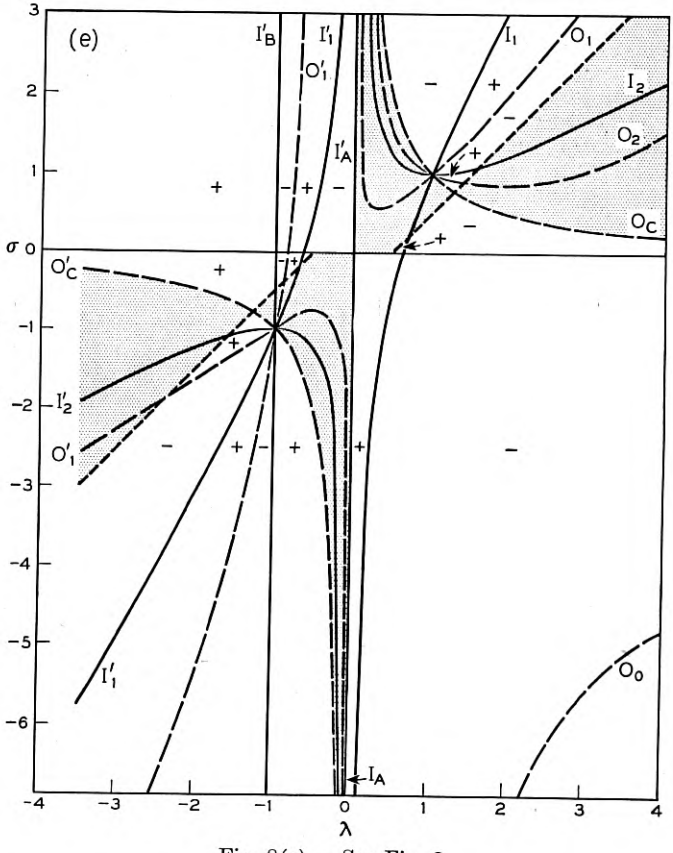


Fig. 8(e) — See Fig. 8.

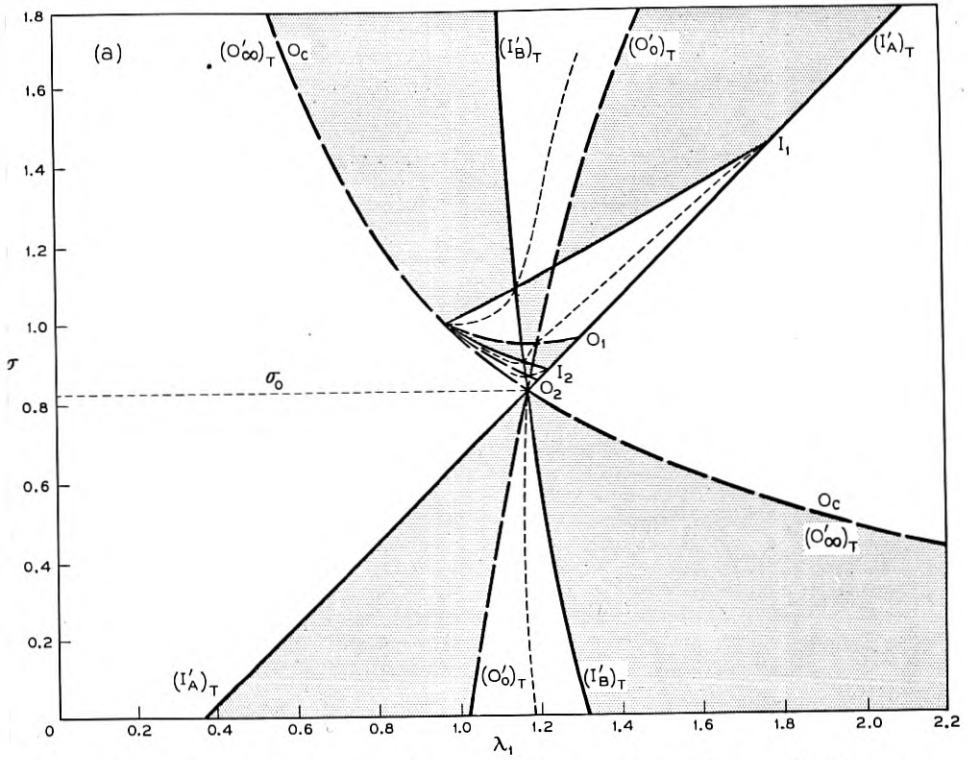


Fig. 9 — Geometrical exploration of the solution curves. The permitted areas of the $\lambda - \sigma$ plane are divided by the O, I, $(O)_T$, $(I)_T$ curves into regions in which $G(\lambda, \sigma)$ and $G(T(\lambda), \sigma)$ have like or unlike signs. Shaded regions are those of unlike signs. Solution curves (shown schematically by dotted lines) must lie in regions of like signs. Only the first few O and I curves are shown. Fig. 9(a) and (b), $r_0 \sim 3.0$; Fig. 9(c) and (d), $r_0 \sim 5.0$; Fig. 9(e) and (f), $u_2 < r_0 < j_2$; Fig. 9(g), $r_0 < u_1$. $|p| < 1$, throughout. The horizontal dashed line marks $|\sigma| = |\sigma_0|$.

Polder relation is indicated, for $p = \frac{1}{2}$, by shading. For other p-values the straight portion of the boundary of the excluded region is simply translated along the λ-axis.

In Fig. 9(a) the allowed region of the first quadrant is shown again, together with the transforms $(I_A')_T$, $(0_1')_T$, and $(I_B')_T$ of the only critical curves I_A', $0_1'$ and I_B' occurring in the second quadrant for the present radius. Regions in which $G(\lambda_1, \sigma)$ and $G(T(\lambda_1), \sigma)$ have opposite sign are shaded; the common signs in the remaining parts of the quadrant are as indicated. (In this diagram p is taken to be $\frac{3}{8}$).

From the disposition of the surfaces $G(\lambda_1, \sigma)$ in the region between 0_1

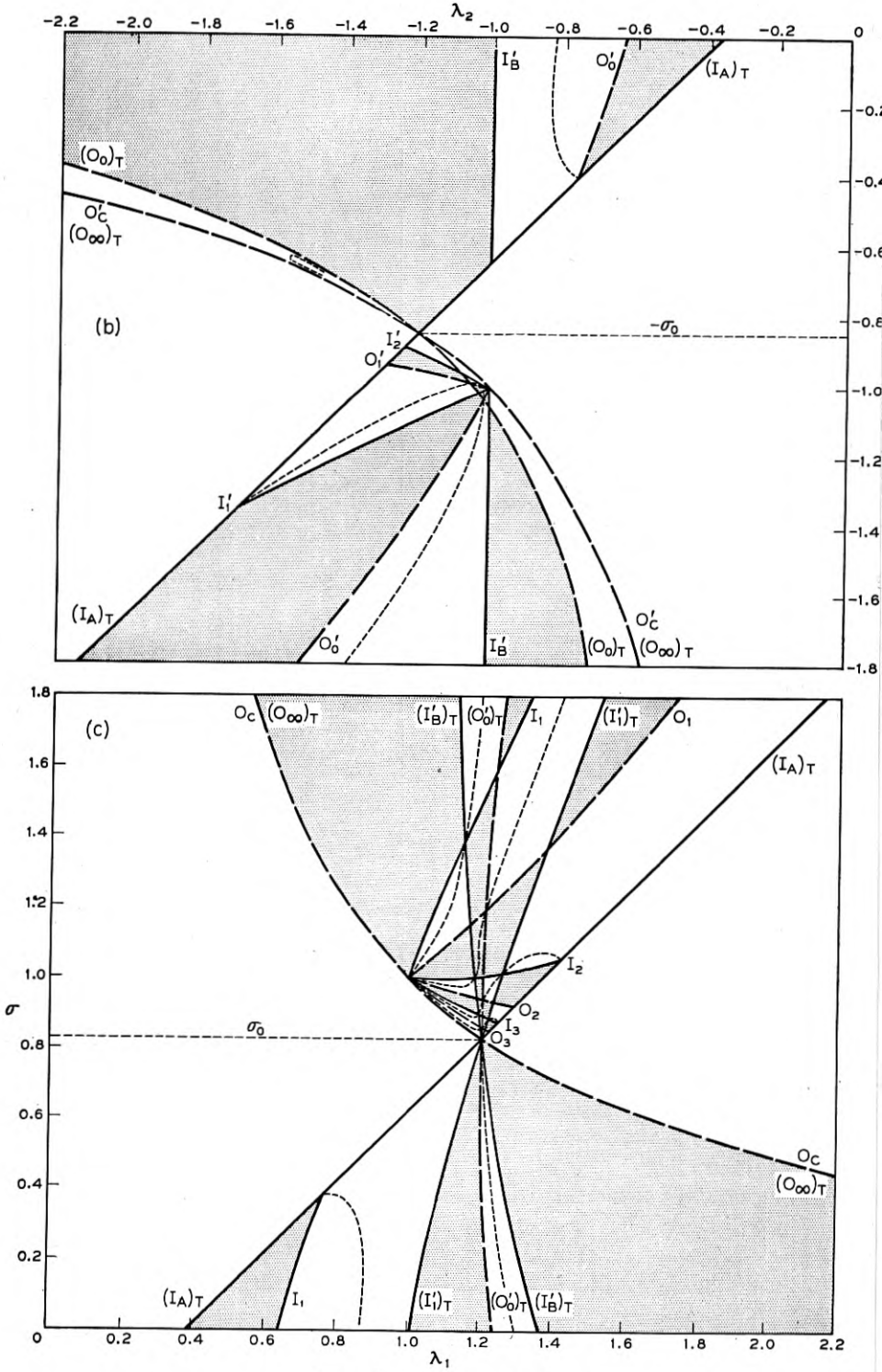


Fig. 9(b) and (c) — See Fig. 9.

and I_1 and of the surface $G(T(\lambda_1), \sigma)$ between $(0_\infty')_T$ and $(I_B')_T$ it is evident that along any contour such as $G(\lambda_1, \sigma) = K$, $G(T(\lambda_1), \sigma)$ will take on all positive values from 0 to ∞ and, in particular, K. Since this is true for any K, it follows that the region between 0_1, I_1, $(I_B')_T$ contains a solution curve. Two points on this curve are immediately obvious: the intersections of $(I_B')_T$, I_1 and the point (1, 1) on $(0_\infty')_T$, 0_c. The first is the intersection of the curves

$$\lambda = 1 + \frac{p}{1 + \sigma}, \qquad \frac{j_1^2}{r_0^2} = \frac{1 - \lambda^2}{1 - \sigma\lambda}.$$

At the point (1, 1), λ_2 is $-\infty$, λ_1 is unity and β^2 is therefore infinite. Armed with this knowledge we now investigate analytically the behavior of β^2 near $\sigma = 0$ directly from the original G-equation and Polder relations. Writing $\lambda_1 = 1 + c\epsilon$ and $\sigma = 1 + \epsilon$, $(1 - \lambda^2)/(1 - \sigma\lambda)$ is to zero

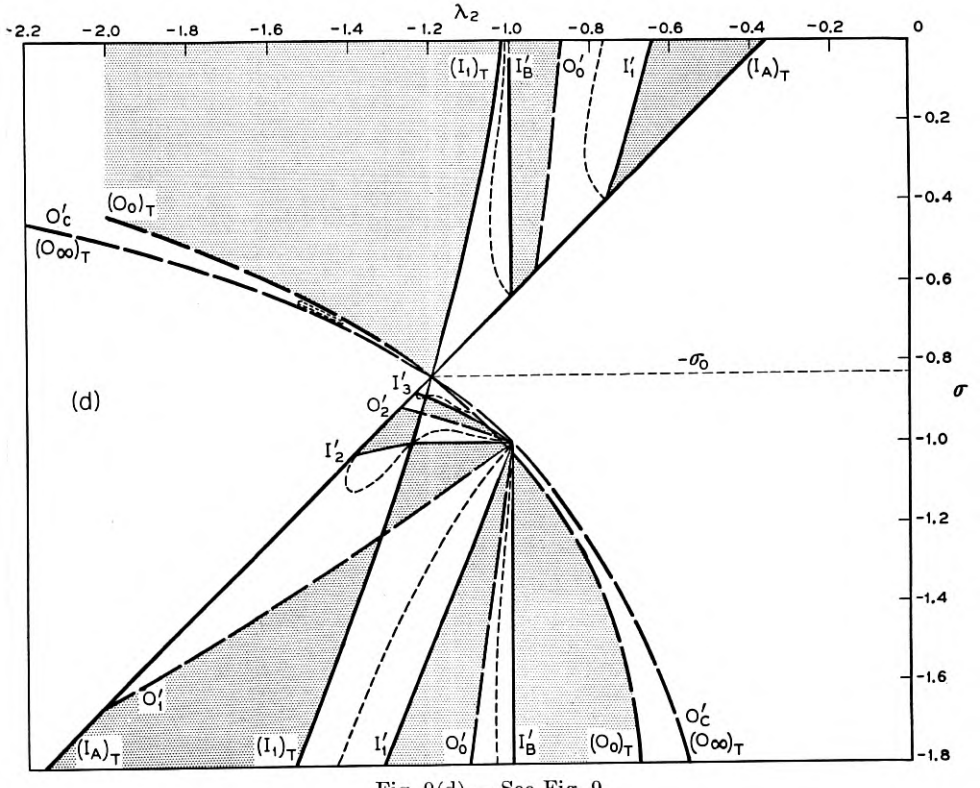


Fig. 9(d) — See Fig. 9.

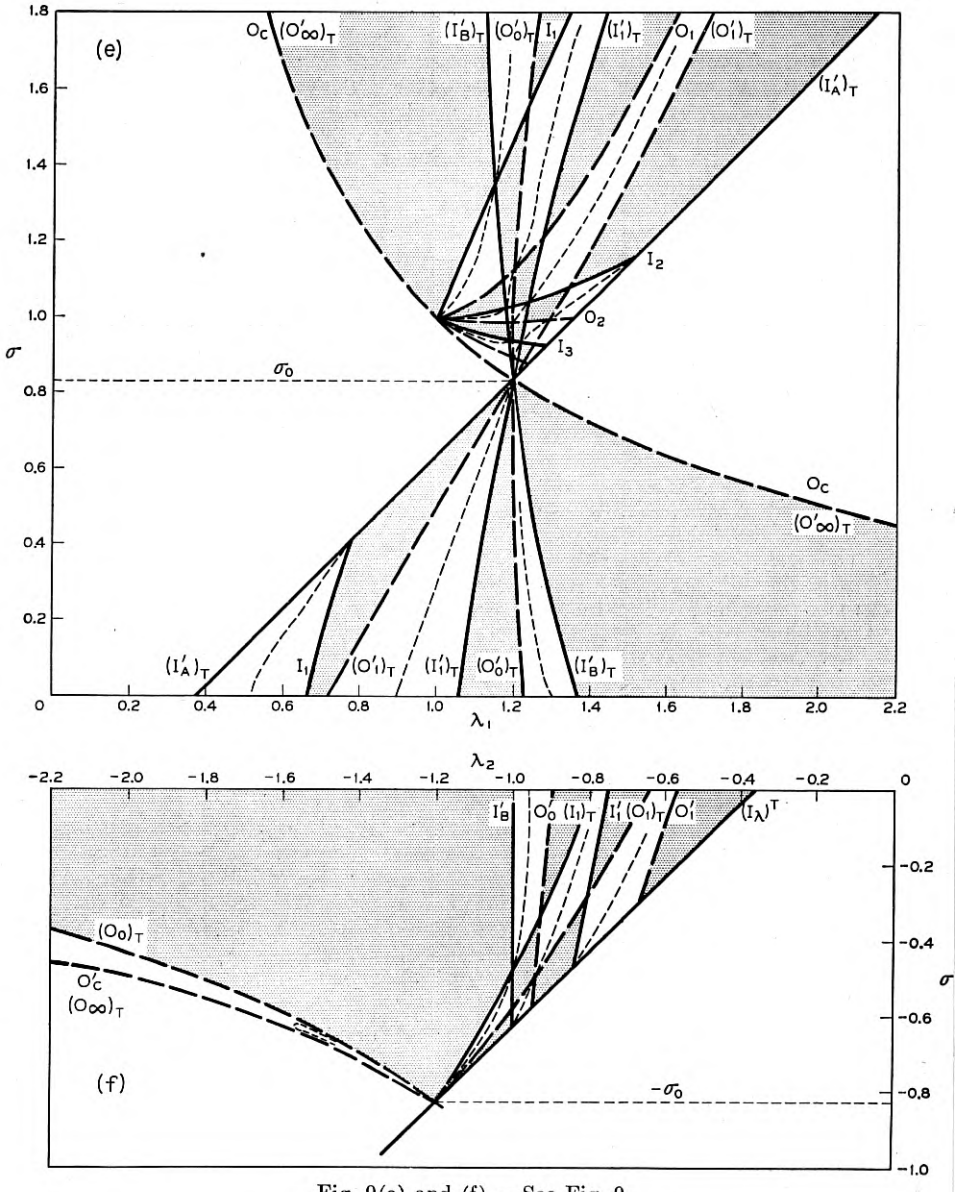


Fig. 9(e) and (f) — See Fig. 9.

order just $2c/1 + c$. Thus, $G(\lambda_1, \sigma)$ to zero order is

$$\frac{1+c}{2c}\left[F\left(r_0\sqrt{\frac{2c}{1+c}}\right) - 1\right].$$

But we have, in this case, $G(\lambda_2) = 0$, so that

$$\frac{2c}{1+c} = \frac{z_1^2}{r_0^2},$$

where z_1 is the smallest non-zero root of $F(z) = 1$. From the Polder relation, the leading term of λ_2 is $-p/(1+c)\epsilon$, and consequently the leading term of β^2 is

$$\frac{p}{(1+c)\epsilon} = \frac{p\left(1 - \frac{z_1^2}{2r_0^2}\right)}{\sigma - 1}.$$

This analysis is readily extended to the next order term, which is stated in Section (4.17).

From analogous considerations concerning the variation of one G-

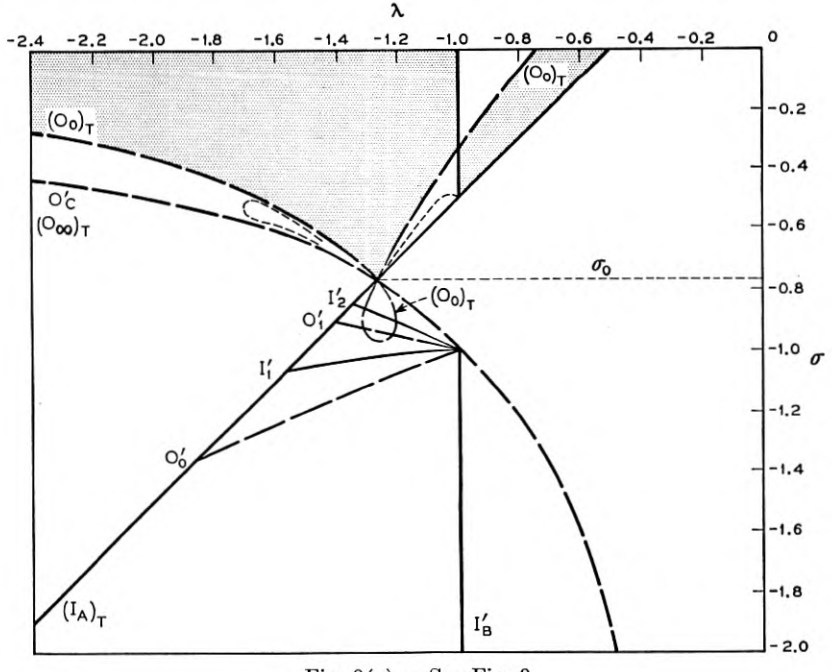


Fig. 9(g) — See Fig. 9.

function through all possible values of the other in the region bounded by I_1, $(I_B')_T$, $(0_1')_T$ we deduce that the solution curve just discussed continues into that region and persists as $\sigma \to \infty$. For, the asymptote of $(0_1')_T$ is $\sigma = \frac{r_0^2}{u_1^2}\lambda$, and between it and $\lambda = 1$, which is the asymptote of $(I_B')_T$, $G(T(\lambda_1), \sigma)$ takes on all values between 0 and $-\infty$; in particular, the limited range of values assumed by $G(\lambda_1, \sigma)$ in this region. The behavior of the solution curve for large σ may be deduced by using the asymptotic formulae for curves $G(\lambda_1, \sigma) = g$ and $G(T(\lambda_1)\sigma) = g$ which are given in the appendix. These are

$$\sigma = -g\lambda_1 - gF\left(\frac{r_0}{\sqrt{-g}}\right),$$

and

$$\sigma = \frac{r_0^2}{u_1^2}\lambda_1 - \left[p + 2\left(1 + \frac{gu_1^2}{r_0^2}\right)\frac{1 - \frac{r_0^2}{u_1^2}}{1 - u_1^2}\right].$$

It is clear that g at a point of intersection is given by $-r_0^2/u_1^2$ plus terms of order $1/\lambda_1$; substituting this value in the second equation gives the solution curve correctly to order $1/\lambda_1$ in the form.

$$\sigma = \frac{r_0^2}{u_1^2}\lambda_1 - p.$$

When the solution curve has such a linear asymptote it is convenient to calculate β^2 from the formula

$$\beta^2 = 1 + \frac{p}{\sigma} - \frac{\lambda}{\sigma} + \text{terms of order higher than } 1/\sigma$$

which is readily obtained from (30). In the present case

$$\beta^2 = \left(1 - \frac{u_1^2}{r_0^2}\right)\left(1 + \frac{p}{\sigma}\right) + \text{higher terms in } 1/\sigma. \tag{38}$$

As $\sigma \to \infty$, β^2 tends to the value appropriate to the TE_{11}-mode in an isotropic medium ($\mu \to \mu_z = \mu_0$, $\kappa \to 0$ as $\sigma \to \infty$). Thereby the whole solution curve is classified as specifying part of a TE_{11}-limit mode.

The remaining section of the TE_{11}-limit mode in the upper half-plane is again found in the region between $(0_1')_T$ and $(I_B')_T$ for $\sigma < \sigma_0$. Any line $\sigma =$ constant $< \sigma_0$ cuts these two curves at two values of λ_1. As λ_1 varies between these values, $G(T(\lambda_1), \sigma)$ varies from 0 to $-\infty$; it is, thus,

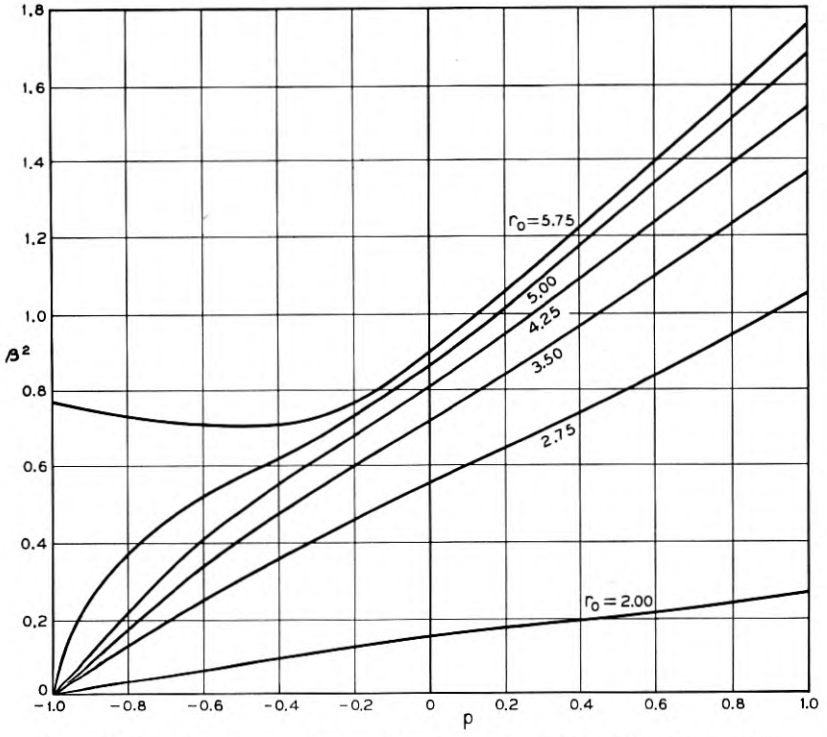


Fig. 10(a) — β^2 versus p for small values of σ — the TE_{11}-limit mode.

clearly equal to the finite (negative) $G(\lambda_1, \sigma)$ somewhere between. This situation persists up to $\sigma = \sigma_0 - 0$ and a solution curve therefore exists between $\sigma = 0$ and $\sigma = \sigma_0$. It meets $\sigma = 0$ for λ_1 satisfying

$$\frac{1}{1-\lambda_1^2}\left[\frac{1}{\lambda_1}F(r_0\sqrt{1-\lambda_1^2}) - 1\right] = \frac{1}{1-\lambda_2^2}\left[\frac{1}{\lambda_2}F(r_0\sqrt{1-\lambda_2^2}) - 1\right],$$

and

$$\lambda_1 + \lambda_2 = p. \tag{39}$$

These equations have been solved numerically; the corresponding $\beta^2 = -\lambda_1\lambda_2$ is shown in Fig. 10(a). For r_0 between u_1 and j_1 a value derived for β^2 from the first three terms of an expansion of β^2 in powers of p, equation (61), turns out to be in very good agreement with the numerical calculation up to $p = 1$, for $\sigma = 0$ and presumably is good for small σ.

At σ_0 (the point at which μ becomes negative), the solution curve is "cutoff". However, the corresponding β^2 is not zero. As σ_0 is approached from below $G(\lambda_1, \sigma) \to 0$ and so $G(\lambda_2, \sigma)$ tends to zero. Thus, λ_2 tends to

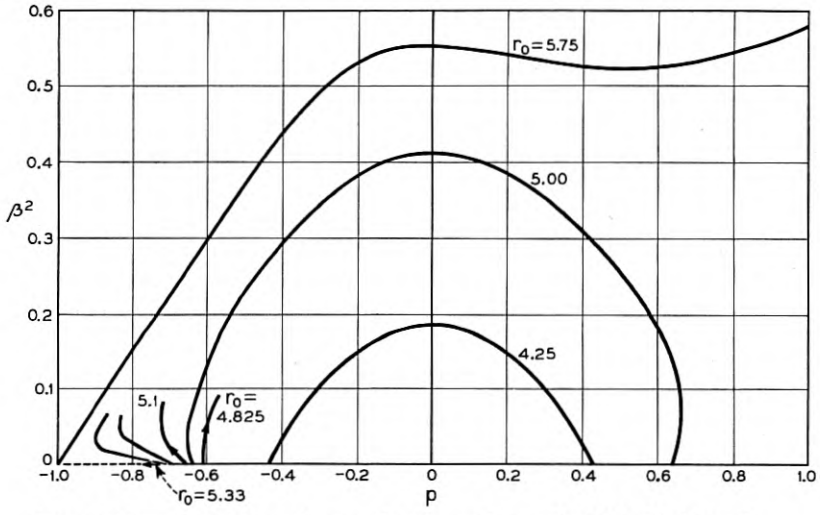


Fig. 10(b) — β^2 versus p for small values of σ — the TM_{11}-limit mode.

the negative root λ_{20} (unique for the present radius) of

$$F\left(r_0\sqrt{\frac{1-\lambda_2^2}{1-\sigma_0\lambda_2}}\right) = \lambda_2 .$$

The associated β^2 is $-\lambda_{20}\lambda_{10} = -\dfrac{\lambda_{20}}{\sigma_0}$ and is shown in Fig. 11(a). The way in which β^2 approaches this value as $\sigma \to \sigma_0$ can be found and is one of the more subtle examples of behavior of a mode near a special point. Writing $\sigma = \sigma_0 - \delta\sigma$, $\lambda_1 = \lambda_{10} - \delta\lambda_1$, we observe that, since $\sigma_0 + p - \dfrac{1}{\sigma_0} = 0$, the Polder relation in the form

$$\lambda_1 = \frac{1}{\sigma} + \frac{\sigma + p - 1/\sigma}{1 - \sigma\lambda_2}$$

fully determines $\dfrac{d\lambda_1}{d\sigma}$; any variation due to $\delta\lambda_2$ vanishes at $\sigma = \sigma_0$. $\delta\lambda_2$ can be determined from the G-equation. Near $\sigma\lambda = 1 - 0$ $(\lambda > 1)$, $G(\lambda_1, \sigma)$ is given by

$$-\frac{r_0}{\lambda_1}\sqrt{\frac{1-\sigma\lambda_1}{\lambda_1^2 - 1}},$$

which near σ_0, λ_{10} may be written

$$-\sqrt{\delta\sigma}\,\frac{r_0}{\lambda_{10}}\sqrt{\frac{\lambda_{10}\dfrac{d\sigma}{d\lambda_1}+\sigma_0}{\lambda_{10}^2-1}}\,.$$

The perturbed $G(\lambda_2,\sigma)$ which (since $G(\lambda_{20},\sigma_0)=0$) is $\dfrac{\partial G}{\partial\lambda_{20}}\,\delta\lambda_2+0(\delta\sigma)$ equals the preceeding expression and gives

$$\delta\lambda_2=-\sqrt{\delta\sigma}\,\frac{r_0}{\lambda_{10}}\sqrt{\frac{\lambda_{10}\dfrac{\delta\sigma}{\delta\lambda_1}+\sigma_0}{\lambda_{10}^2-1}}\left(\frac{\partial G}{\partial\lambda_{20}}\right)^{-1}.$$

Accordingly, $\delta\beta^2=-\lambda_{10}\,\delta\lambda_2+0(\delta\sigma)$, a result which shows that β^2 tends to its terminal value along the vertical. It is clear analytically and graphically that this mode persists as $p\to 0$, and must be identified with the only isotropic mode for this radius, namely TE_{11}. No other branches exist below $\sigma=\sigma_0$, since $G(\lambda_1,\sigma)$ and $G(T(\lambda_1),\sigma)$ have opposite signs except in the region just considered.

The two solution curves considered so far are not the only ones; in fact the infinity of sheets of the surface $G(\lambda_1,\sigma)$ in the region bounded by I_1, O_c and $(I_A')_T$, Fig. 9(a), intersect the transformed sheets $G(T(\lambda_1)\sigma)$ in infinitely many more curves. In the blank areas of that region the G-functions have equal sign, and all these areas must be carriers of solution curves, since in every one of them every single contour $G(\lambda_1,\sigma)=g$

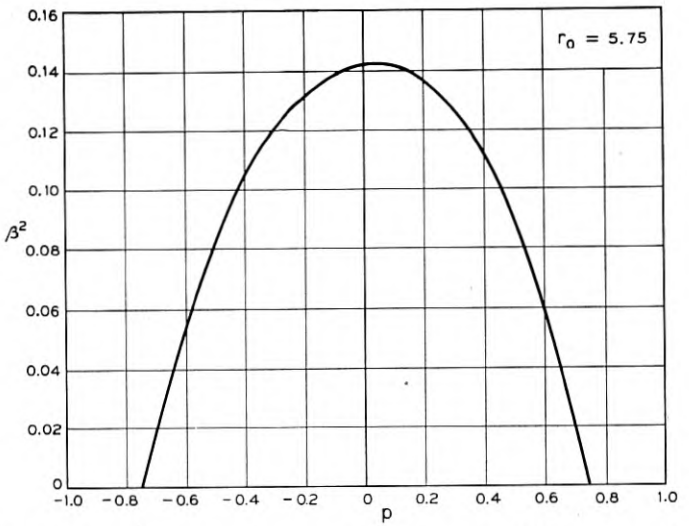


Fig. 10(c) — β^2 versus p for small values of σ — the TE_{12}-limit mode.

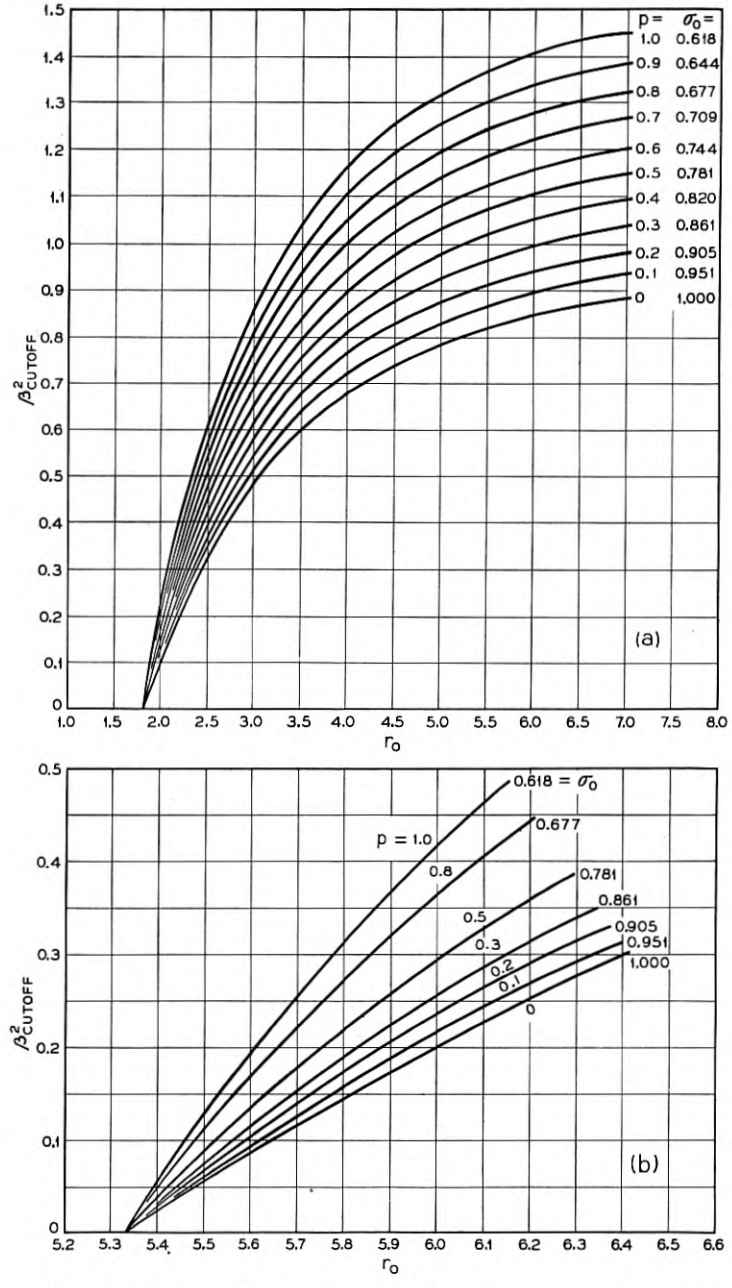


Fig. 11 — β^2 at the type 2′ cutoff as a function of r_0 for various p. $2_0'$ cutoff, TE_{11}-limit mode; (b), $2_1'$ cutoff, TM_{11}-limit mode. The presence of a curve for $p = 0$ is clarified in the text.

crosses all contours of $G(T(\lambda_1), \sigma)$, in particular $G(T(\lambda_1), \sigma) = g$. All the additional solution curves arising in this way start at $\sigma = 1, \lambda_1 = 1$; the n^{th} of them threads its way from one blank region to another, first through the intersection of I_{n+1} with $(I_B')_T$, then through the intersection of 0_n with $(0_1')_T$, and finally comes to an end at the intersection of I_n with $(I_A')_T$. At the end point ($\sigma = 1, \lambda_1 = 1$), λ_2 and, therefore, β^2 are infinite, (just as for the TE_{11} solution curve). At the end point (I_n, $(I_A')_T$), λ_2, and, therefore, β^2 are zero. The σ and λ_1 values corresponding to the latter are obtained from the equations

$$\frac{1 - \lambda_{1n}^2}{1 - \sigma_n \lambda_{1n}} = \frac{j_n^2}{r_0^2}; \qquad \lambda_{1n} = \sigma_n + p. \tag{40}$$

It is possible to derive the slope $\delta\beta^2/\delta\sigma$ of the $\beta^2 - \sigma$ curves at these cut-off points. Near cut-off, the infinity I_n of $G(\lambda_1, \sigma)$ is matched by the infinity I_A' of $G(\lambda_2, \sigma)$. The G-equation therefore degenerates to

$$\frac{1}{\lambda_2} F(r_0) = \frac{j_n}{r_0 \sqrt{\dfrac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} - j_n} \cdot \frac{1}{\lambda_{1n}} \cdot \frac{r_0^2}{j_n^2}.$$

Writing $\sigma = \sigma_n - x\lambda_2$, $\lambda_1 = \lambda_{1n} - y\lambda_2$, expansion of the right hand side of this equation to order $1/\lambda_2$ furnishes one relation between x and y; the Polder equation furnishes another. The two can be solved for x, and so, since to first order

$$\delta\beta^2 = -\lambda_{1n}\lambda_2 = \lambda_{1n} \frac{\sigma - \sigma_n}{x} = \lambda_1 \frac{\delta\sigma}{x}$$

$\delta\beta^2/\delta\sigma$ may be found. It is found that for convenience in computation, the results of this calculation are best presented parametrically. Equations (46–8) represent equations (40) and $\delta\beta^2/\delta\sigma$ in this way. Fig. 12 (a) and (b) show the result of some computations. Near $\sigma = 1$, $\beta^2 = \infty$, these added solution curves behave rather like the TE_{11} curve. The leading term in the expansion of β^2 in powers of $\dfrac{1}{\sigma - 1}$ is now

$$\frac{p\left(1 - \dfrac{z_{n+1}^2}{r_0^2}\right)}{\sigma - 1}$$

for the solution curve ending at $I_n - (I_A')_T$. Here z_{n+1} is the $(n + 1)^{\text{th}}$ root of $F(z) = 1$, not counting 0.

It will turn out later that the infinity of solution curves just discussed represents an incipient form of the whole mode spectrum; the reservoir

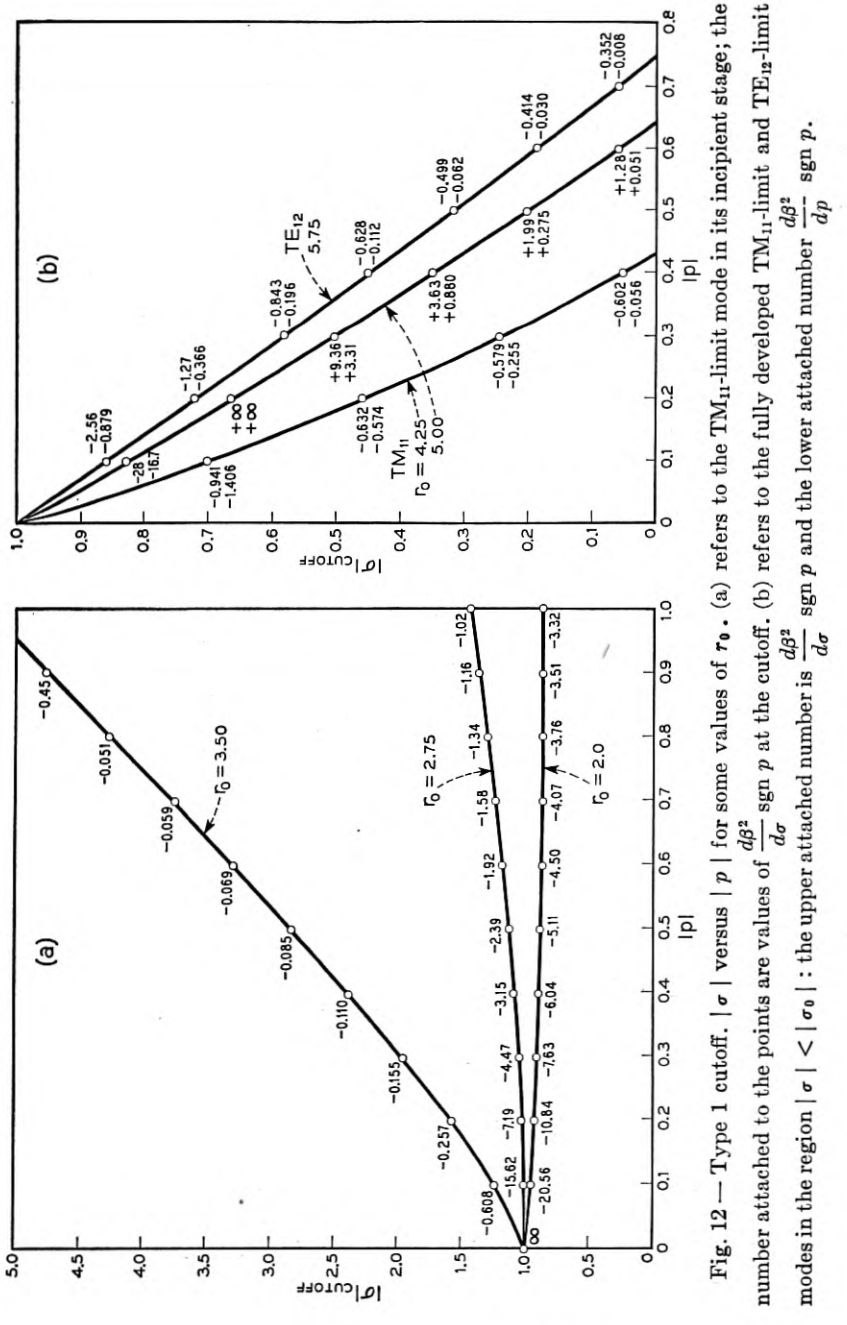


Fig. 12 — Type 1 cutoff. $|\sigma|$ versus $|p|$ for some values of r_0. (a) refers to the TM_{11}-limit mode in its incipient stage; the number attached to the points are values of $\frac{d\beta^2}{d\sigma}$ sgn p at the cutoff. (b) refers to the fully developed TM_{11}-limit and TE_{12}-limit modes in the region $|\sigma| < |\sigma_0|$: the upper attached number is $\frac{d\beta^2}{d\sigma}$ sgn p and the lower attached number $\frac{d\beta^2}{dp}$ sgn p.

from which higher modes are drawn as the guide radius is increased. That the propagation of modes which for larger guide radii correspond to higher TM and TE modes is possible for limited ranges of σ might be ascribed to the larger μ-values in those ranges, which cause the wave to see an effectively larger guide. This explanation is convincing only when $\sigma > 1$. When $\sigma_0 < \sigma < 1$, μ is negative, and the propagation must then be the result of an interplay between μ and κ. In passing we remark that we are here dealing with the propagation analog of so-called "shape resonances," which physicists sometimes encounter in resonance experiments on small spheres of ferrite in cavities.

We now turn to a discussion of the solution curves for $\sigma < 0$ which lie in the third quadrant. Fig. 9(b) shows the partition of the region allowed by the Polder relation (again for $p = \frac{3}{8}$) into positive and negative regions by the various I', $0'$, $(I)_T$ and $(0)_T$ curves. Regions in which $G(\lambda_2, \sigma)$ and $G(T(\lambda_2), \sigma)$ have opposite signs are shaded. For $\sigma < -\sigma_0$, the question whether a given region of like signs is the site of a solution curve may, with one exception, be answered by the same type of geometrical argument as used for $\sigma > 0$. The singular area is that part of the region bounded by I_B' and 0_c in which the G-functions are both positive. Here both $G(\lambda_2, \sigma)$ and $G(T(\lambda_2), \sigma)$ are zero on 0_c; $G(\lambda_2, \sigma)$ goes to ∞ on I_B', whereas $G(T(\lambda_2), \sigma)$ is finite throughout the region. No intersection can be predicted, then, by the earlier argument. It can indeed be shown (for all r_0) that there is no such intersection. For, in the case $p = 0$, the solution curves are I_n' or $0_n'$ curves as demonstrated in Section (4.11). The region under consideration contains no such curves, and hence no solution curves. Thus, for $p = 0$, since $G(\lambda_2, \sigma)$ goes to infinity on I_B', the surface $G(T(\lambda_2), \sigma)$ must lie entirely below the surface $G(\lambda_2, \sigma)$. Consider now, for fixed σ and increasing p, a point on the $G(T(\lambda_2), \sigma)$ surface whose height remains unchanged. For such a point $T(\lambda_2, p, \sigma)$ remains fixed and from the Polder relation this means an increasingly negative λ_2. Since it can be shown that $G(\lambda_2, \sigma)$ goes monotonically from 0 to ∞ as λ_2 becomes more negative, it follows that $G(T(\lambda_2), \sigma)$ continues to be below $G(\lambda_2, \sigma)$ for all p.

All other regions of common sign do carry solution curves. That corresponding to the TE_{11}-limit mode begins at $\sigma = -1$, $\lambda_2 = -1$, passes through the intersection of $(0_0)_T$ and $0_0'$ and persists for indefinitely large σ. The asymptotic formula (38), for β^2 at large σ also holds as $\sigma \to -\infty$, if the signs of both σ and p are taken to be negative. The behavior of β^2 near $\sigma = -1$ may be found by the same means used at $\sigma = 1$. The resulting expression* is to order $\frac{1}{\sigma + 1}$, essentially the same

* See Section 4.17 for a more exact formula.

as the earlier (37) except that the smallest root of $F(z) = -1$ replaces that of $F(z) = 1$. The remaining solution curves confined to the region bounded by $0_0'$, 0_c, $(I_A)_T$ portray the incipient modes already encountered in the first quadrant. Their behavior near $\sigma = -1$ also follows (58), associated with the higher roots of $F(z) = -1$. Their end points, the intersections of I_n' with $(I_A)_T$, are still given by the parametric representation (46–8), due regard being paid to the signs of σ_n and p.

The remaining branch of the TE_{11}-limit mode, lying above $\sigma = -\sigma_0$, is found in the triangle between $(I_A)_T$, $\sigma = 0$ and I_B. Its end points are given by (39) with p negative and by the intersection of I_B' with $(I_n)_T$ which is $\sigma = -(1 + p)$, $\lambda_2 = -1$, $\lambda_1 = 0$. Thus the cut-off in contrast to the analogous branch for $\sigma > 0$, is given by $\beta^2 = 0$, $\sigma = -(1 + p)$. (When $p < -1$, the branch does not exist at all.) We note that a left-circular plane wave is cut off at exactly the same value of σ as the TE-mode is in this particular case (see, however, the following sections). The slope at cut-off is determined by expanding the G functions near their infinities at I_n and I_B' and utilizing the Polder relation. The slope is found to be

$$\frac{d\beta^2}{d\sigma} = \frac{F(r_0)}{p(1 - F(r_0))}. \tag{41}$$

A further solution curve lies in the region between 0_c and $(0_0)_T$ for $\sigma > -\sigma_0$. It has no analogue in a guide with isotropic material and will be discussed later.

In the discussion of the mode spectrum for radii between u_1 and j_1 three distinct types of cut-off point have already been encountered. When larger radii are treated it is found that no other types arise.* In Section 4.17 formulas relevant to the three types are given. An examination of the field components in the neighborhood of the cut-off points is of some interest. Cut-off points of type one (intersections of I_n and $(I_A')_T$ or I_n' and $(I_A)_T$), at which $\beta^2 = 0$, have $E_z = 0$ and the field is of a pure TE-type. The medium behaves transversely as though it had a permeability, $\mu - \kappa^2/\mu$. Although the field is purely TE at cut-off the mode terminating at such a point may in the limit of vanishing magnetization be either a TE- or a TM-mode. This impartiality extends to cut-off points of the other types. Cut-off points of type II $[(0_n')_T - 0_c$ or $(0_n)_T - 0_c']$ occur at $\sigma = \pm\sigma_0$, where $\mu = 0$ and here β^2 does not vanish. In such cases one of the χ's is finite and the corresponding contributions to the field pattern quite normal. The other, however, tends

* There is an exception to this statement. This is the type designated in Section 4.17 as $2_{0\infty}$ which cuts off an isolated mode having no TE or TM analogue.

to an infinite imaginary value and the associated fields are confined very closely to the guide walls. The wall currents are very large and essentially longitudinal. Type III cut-off points $[I_B - (I_A)_T]$ at which $\mu = -\kappa$ have $\beta^2 = 0$, but the fields are not of a purely TE- or TM-type. They consist essentially of a rotating, transverse, H, which is uniform over the guide. The components H_z, E_θ and E_r are smaller by one order of $\sigma - \sigma_{\text{cut-off}}$ and E_z, two orders smaller.

It should be stressed again that, in general, the modes are never of pure TE or TM type. Nevertheless, for the sake of brevity, we shall refer to them as such; calling them TE-modes or TM-modes according to their limit as the magnetization is removed.

4.13. We now consider the behavior of the modes as a function of radius. The reader will be aided by Figs. 8(a) to (e) and 9(a) to (g). In preparation for this it is necessary to examine the movement of the I_n, I_n' and 0_n, $0_n'$ curves when r_0 is varied. It will be recalled that the equation for the I_n curves and their reflections, I_n', in the origin, is

$$\frac{1 - \lambda^2}{1 - \sigma\lambda} = \frac{j_n^2}{r_0^2}.$$

The contours $\chi(\lambda, \sigma) = C$, where $\chi^2 = (1 - \lambda^2)/(1 - \sigma\lambda)$ have already been plotted in Fig. 4. The I_n, I_n' curves are among these, and, clearly, for a fixed n, the associated χ^2 decreases as r_0 increases. The course of a given pair (I_n, I_n') may then be seen directly from Fig. 4. The qualitative behavior of the pair changes radically only when r_0 passes through the value j_n. Before it does so, I_n lies, for λ between 0 and 1, above $\sigma = \lambda$ and tends to $\sigma = \infty$ as λ tends to zero. At $r_0 = j_n$, the I_n and I_n' curves merge into the lines $\sigma = \lambda$ and $\lambda = 0$. Beyond j_n, I_n lies below $\sigma = \lambda$ for λ between 0 and 1 and goes to $-\infty$ as λ approaches zero. The I_n'-curve remains, throughout the reflection of I_n in the origin. As $r_0 \to \infty$, I_n tends to the line $\lambda = 1$, I_n' to $\lambda = -1$. No I_n curves ever enter the region $\lambda > 1$, $\sigma < 0$; no I_n' curves enter $\lambda < -1$, $\sigma > 0$. It is also important to relate the I_n, I_n' curves to the boundaries of the Polder regions. I_n curves cut the Polder boundary $\sigma = \lambda - p$, of the first quadrant in at most one point. As r_0 increases from 0 to j_n, this point moves from $\sigma = \sigma_0$ to $\sigma = \infty$. Thereafter, no intersection occurs at fixed p until r_0 equals $j_n/\sqrt{1 - p^2}$; it here reappears at $\sigma = 0$ and moves steadily to $\sigma = 1 - p$ as r_0 increases indefinitely. The only intersection with the other Polder boundary $\sigma = 1/\lambda$, is at $\lambda = 1$, $\sigma = 1$, regardless of r_0.

The 0_n, $0_n'$ curves are given by

$$\sigma = \frac{1}{\lambda}\left[1 - r_0^2 \frac{1 - \lambda^2}{(F^{-1}(\lambda))^2}\right]$$

if the n^{th} branch of $F^{-1}(\lambda)$ is used. Thus, as r_0 increases, the successive curves either all pass through a fixed point (which can only be $\lambda = \pm 1$, $\sigma = \pm 1$, $n > 0$) or move steadily up or down without further intersection. An 0_n curve starts from $\sigma\lambda = 1$ at $r_0 = 0$ and falls for $\lambda < 1$, rises for $\lambda > 1$, as r_0 increases. For large λ, since $\sigma \sim j_n^2/r_0^2$ ($\lambda \to 2$), $n > 0$, the 0_n and I_n curves move together with a constant separation. 0_0 is singular, since it does not pass through λ, $\sigma = 1$ and falls steadily for all λ; it tends to $\sigma = 0$ for large λ. The $0_n'$ curves rise from $\sigma\lambda = 1$ at $r_0 = 0$ for $-1 < \lambda$, fall for $\lambda < -1$. They run parallel to I_{n+1}' for $-\lambda$ very large. For small λ there is an expansion

$$\sigma = \left(1 - \frac{r_0^2}{u_n^2}\right)\frac{1}{\lambda} - \frac{2r_0^2}{u_n^2(u_n^2 - 1)} + 0(\lambda)$$

holding for 0_n and for $0_n'$. This indicates that for $r_0 < u_n$, 0_n goes to $+\infty$ and $0_n'$ to $-\infty$ for small λ, but at $r_0 = u_{n+1}$, 0_n and $0_n'$ merge momentarily at

$$\lambda = 0,$$

$$\sigma = -\frac{2r_0^2}{u_{n+1}^2(u_{n+1}^2 - 1)} < 0.$$

For larger r_0, 0_n goes to $-\infty$ and $0_n'$ to $+\infty$. Since the union of 0_n and $0_n'$ takes place at a negative σ, it is clear that 0_n curves, unlike I_n curves, may cross the line $\sigma = 0$ twice. Intersections of the 0_n, $0_n'$ curves with the Polder boundary are difficult to examine explicitly and this may lead to some obscure situations for $0 < |\lambda| < 1$. However, for $\sigma > \sigma_0$, since 0_n and I_n have a fixed separation for large $|\lambda|$, this pair escape intersection with the boundary at the same value of r_0, namely j_n. Similarly $0_n'$ and I_{n+1} escape together at $r_0 = j_{n+1}$ for $\sigma < -\sigma_0$.

We shall now examine the effect of varying r_0 upon the sequence of modes when $\sigma > \sigma_0$. When r_0 is less than u_1, a case in which the isotropic medium would not propagate, no part of $0_0'$ lies in the upper half plane and there is then no $(0_0')_T$ curve. The solution curve which in the previous discussion of Section 4.12 was assigned to TE_{11}, after passing the intersection $[I_1 - (I_B)_T]$ can no longer escape to infinity and terminates on $[I_1 - (I_A)_T]$. Thus, the TE_{11} mode at this radius has become an incipient mode with cut-off and other properties given by the formulae already quoted for such modes. As r_0 approaches u_1 from below, the $\beta^2 - \sigma$ curve is double valued between $\sigma_{\text{cut-off}}$ and some larger value. This is borne out by the fact that $d\beta^2/d\sigma$ becomes positive at cut-off, and by the observation that the solution curve bulges towards large σ between $I_1 - (I_B')_T$ and its terminus. The part of the $\beta^2 - \sigma$

curve along which $\frac{d\beta^2}{d\sigma} < 0$, will tend smoothly towards the $\beta^2 - \sigma$ curve for r_0 just greater than u_1. The course of the TE_{11} solution curve remains qualitatively unchanged for all $r_0 > u_1$.

When r_0 passes through u_1, and the TE_{11} solution curves escapes discontinuously to infinity, the solution curves below it disengage from their former end points $I_{n+1} - (I_A)_T$ and instead end at the point $I_n - (I_A)_T$. When r_0 exceeds j_1, the curves I_1 and 0_1 escape intersection with $\sigma = \lambda - p$ simultaneously, for $\sigma > \sigma_0$, and the curve $(I_1')_T$ makes its first appearance. From the asymptotic formulae (App. II) the latter runs to infinity between I_1 and 0_1, and now the solution curve which ended for $u_1 < r_0 < j_1$ at $I_1 - (I_A)_T$ is carried to infinity between I_1 and $(I_1)_T$. The asymptotic expression for β^2 versus σ, given in formula (56) indicates that β^2 tends to the isotropic value for the TM_{11} mode. No further qualitative changes will take place in behavior of this mode as r_0 increases.

As r_0 increases through u_2 (the value at which the isotropic medium supports the TE_{12} mode), the $(0_1')_T$ curve makes its appearance, an event accompanied by the escape of the uppermost incipient solution curve (the one ending at $(I_A)_T - I_2$) to infinity. The escape takes place in the same way as that of the TE_{11} solution curve as r_0 passed through u_1. The newly escaped curve, of course, represents the TE_{12}-limit mode. The end points of the remaining incipient solution curves also jump discontinuously to their next higher neighbors as they did at $r_0 = u_1$. The course of events as r_0 is increased further should now be abundantly clear, and is summarized in Table I on page 642.

We now turn to the region $0 < \sigma < \sigma_0$ and consider first the situation $0 < r_0 < u_1$. It is clear that in the area bounded by 0_∞, 0_0, $(I_A')_T$ and $(I'_B)_T$ both G functions are negative. There is no simple geometrical argument which determines the existence of a solution curve in this region. It is therefore necessary to use a type of analytic argument, which is useful in a number of other cases, although fully discussed only in the present instance.

We show that the least value attained by $G(\lambda_1, \sigma)$ in the admissible region for $p = 0$ (which contains all regions admissible for other p-values) is greater than the maximum value of $G(\lambda_2, \sigma)$ in the range $-1 < \lambda_2 < 0$, $\sigma > 0$. Consider the variation of $G(\lambda_1, \sigma)$ as the point $\lambda_1 = 1$, $\sigma = 1$ is approached along a line of constant χ^2 in the admissible region for $p = 0$ (see Fig. 4). We have the relation

$$G(\lambda,\sigma) = \frac{1}{\chi^2}\left[\frac{1}{\lambda} F(r_0\chi) - 1\right].$$

For χ^2 negative, $F(r_0\chi)$ is positive and thus, as λ approaches unity from above G decreases. Again for χ^2 positive and $r_0 < u_1$, $F(r_0\chi)$ is positive and thus, as λ approaches unity from below, G also decreases. Thus for any χ^2, $G(\lambda, \sigma)$ takes on its least value at $\sigma = 1, \lambda = 1$ and this value is

$$\frac{1}{\chi^2}\left[F(r_0\chi) - 1\right].$$

The minimum value of this limit in this region is $F(r_0) - 1$ and is greater than -1. In the region $-1 < \lambda_2 < 0$, $\sigma > 0$, χ^2 is between 0 and 1, and the χ^2 curves run from $\sigma = 0$ to $\sigma = \infty$. G will clearly decrease as σ increases from zero on any one of these curves. Thus G attains its maximum on $\sigma = 0$, where its value is

$$-\frac{1}{1-\lambda^2}\left[\frac{F(r_0\sqrt{1-\lambda^2})}{|\lambda|} + 1\right].$$

Since $F(r_0\sqrt{1-\lambda^2})$ is positive for $r_0 < u_1$ and $|\lambda| < 1$, G is clearly less than -1. In passing we note that for $r_0 = u_1$, both G functions may attain the value -1.

As r_0 passes through u_1, the $(0_0')_T$ curve appears in the region under discussion and together with $(I_B')_T$ delimits the region carrying the TE_{11}-solution curve already discussed at length. No qualitative changes occur in that curve as r_0 is increased indefinitely. When r_0 exceeds j_1, the $(I_1')_T$ curve appears between $(0_0')_T$ and $(I_A')_T$. Between $(I_A')_T$ and $(I_1')_T$ the G functions have a region of common sign, yet no solution curve arises there for a given p until r_0 reaches $j_1/\sqrt{1-p^2}$.* From then on, the I_1 curve cuts $(I_A')_T$, see Fig. 9(c), and a solution curve exists between $(I_1')_T$ and I_1. It is cut off at the intersection $(I_A')_T - I_1$; there, $\beta^2 = 0$ and σ, $\dfrac{d\beta^2}{d\sigma}$ are given by the same parametric formulae (46–8) applying to the cut-off of incipient modes, the parameter θ being negative. The curve begins at $\sigma = 0$, where it satisfies the usual equation, which for this radius has two solutions. The solution with the smaller λ, belonging to the present curve, tends to the isotropic TM_{11}-limit as $p \to 0$. At a fixed r_0, sufficiently below u_2, this mode does not exist at

* There are some exceptions to this statement. When $4.82 < r_0 < u_2 = 5.33$ and p exceeds $\sqrt{1 - j_1^2/r_0^2}$, a double-valued $\beta^2 - \sigma$ curve exists between two positive σ values. For values of r_0 still closer to u_2 further regions of common sign may arise as a result of the interplay of the $(0_1')T$ and $(I_A')_T$ curves. We have not examined these regions closely. Such dubious regions are confined to the immediate neighborhoods below the u_n.

all when

$$p > \sqrt{1 - \frac{j_1^2}{r_0^2}}.*$$

If r_0 is greater than u_2, the $(0_1')_T$ curve has appeared. A new region of like signs of the G's arises between it and $(I_1')_T$, see Fig. 9(e), and contains a solution curve. This ends at σ_0, λ_{10} and begins at $\sigma = 0$ at a value of λ_1 pertaining to the TM_{11}-mode. Thus, it is clear that as r_0 passed u_2, the end-point of the TM_{11} curve jumped discontinuously from $(I_A')_T - I_1$ to σ_0, λ_{10}. This jump is anticipated as r_0 approaches u_2; the $\beta^2 - \sigma$ curve first bulges beyond $(I_A')_T - I_1$ towards its later course and returns to that point with positive slope. As r_0 increases further no change occurs in the qualitative behavior of the mode. It may be noted that above u_2 the mode exists for all p.

Beyond $r_0 = u_2$, at least part of the area between I_1 and $(I_A')_T$ is an admissible region and does in fact contain the TE_{12} solution curve. It begins at $\sigma = 0$ and λ_1 given by that solution of eqn. (39) which is, in the limit $p = 0$, the TE_{12} solution. It is cut off with $\beta^2 = 0$ at $(I_A')_T - I_1$, the end point relinquished by the TM_{11}-solution curve. As r_0 passes j_2, the TE_{12} solution retains its cut-off point, but, beyond $r_0 = u_3$, it will transfer this point discontinuously to σ_0, λ_{10}. Thereafter its course remains essentially unaltered. Tables I, II and III show the progression of cut-off points of the various modes.

It may be recalled that in the analysis of $u_1 < r_0 < j_1$, the modes in $\sigma < -\sigma_0$ followed essentially the same course as in $\sigma > \sigma_0$. This is also true of their progress with changing radius and of the escape process. The singular character of the $(0_0)_T$ curve and the presence of I_B lead to some local changes in the progress of the modes but have no effect on their more salient features in this particular range of σ. The scheme of progression of the end points is shown in Table I.

In contrast with the state of affairs in the region just discussed, the mode structure in the area between $\sigma = 0$ and $\sigma = -\sigma_0$ is very markedly affected by the presence of $(0_0')_T$ and I_B'.

When $r_0 < u_1$, a solution curve exists between $\sigma = -1 - p$, and $\sigma = -\sigma_0$. It starts with $\beta^2 = 0$ at the intersection of $(I_A)_T$ and (I_B') with a slope given by (41). For sufficiently small r_0, β^2 tends to infinity as $\sigma \to \sigma_0$, since the solution curve approaches the line $0_c'$ or $(0_\infty)_T$. Its shape is then given by (52), see Section 4.17. As r_0 increases, 0_0 falls steadily. Eventually, for sufficiently large p, its minimum falls below

* See footnote on page 628.

$-\sigma_0$. $(0_0)_T$ now has two branches for $\sigma > -\sigma_0$, which pass through the point $\sigma = -\sigma_0$, $\lambda_2 = -\frac{1}{\sigma_0}$ making there a finite angle with each other. $(0_0)_T$ is completed by a loop in $\sigma < -\sigma_0$, Fig. 9(g), which does not affect the incipient modes appreciably. The mode in question now has two branches. The first starts as before and ends at $\sigma = -\sigma_0$, where the associated β^2 is given by $\beta_a{}^2 = \lambda_a/\sigma_0$ and λ_a is the smaller root of

$$F\left(r_0\sqrt{\frac{1-\lambda_a{}^2}{1-\sigma\lambda_a}}\right) = \lambda_a\,. \tag{42}$$

It resumes at $\sigma = -\sigma_0$ and $\beta^2 = \beta_b{}^2 = \lambda_b/\sigma_0$, where λ_b is the larger root of the above equation, progresses to smaller $|\sigma|$-values and then back to $\sigma = -\sigma_0$ where β^2 tends to infinity again in accordance with (52). Beyond $r_0 = u_1$, where 0_0 rises steadily from $\sigma = -\infty$ to $\sigma = 0$ with increasing λ, one branch of $(0_0)_T$ in $-\sigma_0 < \sigma < 0$ disappears and only the second branch of the mode remains. Neither branch has an analogue in ordinary waveguides; as $p \to 0$ each lies in a smaller and smaller neighborhood of $\sigma = 1$, and finally vanishes into $\sigma = 1$, $\lambda_2 = -1$.

For r_0 between u_1 and j_1 there is a single solution curve starting at $\sigma = 0$ and ending with $\beta^2 = 0$ at $(I_A)_T - I_B'$. This may be identified in the limit $p = 0$, with the TE_{11}-limit mode, and has already been fully discussed for $u_1 < r_0 < j_1$. No change in the formula for its cut-off point occurs up to $r_0 = u_2$. A useful spot-point $(I_B' - (I_1)_T)$ along its course can be found when

$$0 < p < 1 - \sqrt{1 - \frac{j_1{}^2}{r_0{}^2}}$$

and is given by (60).

In the range $j_1 < r_0 < u_2$, a further solution curve (corresponding to the TM_{11}-limit mode) can arise, provided

$$p < \sqrt{1 - \frac{j_1{}^2}{u_2{}^2}}$$

The radius at which it will then first appear is

$$r_0 = \frac{j_1}{\sqrt{1-p^2}}\,.$$

It begins on $\sigma = 0$, according to (39) and is cut off, with $\beta^2 = 0$, at $(I_A)_T - I_1'$.

As r_0 passes u_2 , the cut-off point of the TE_1 solution curve moves dis-

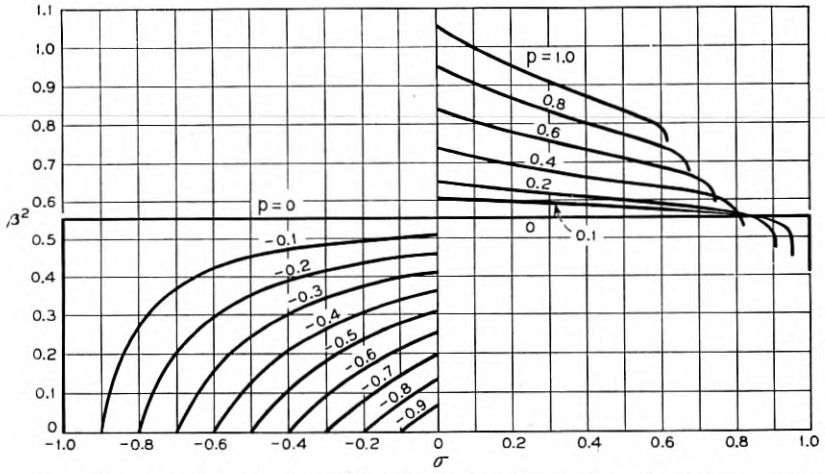


Fig. 13 — Approximate course of modes in the $\beta^2 - \sigma$ plane for various p values and at two values of r_0. Fig. 13(a), above, $r_0 = 2.75$, TE_{11}-limit modes; $-1 < \sigma < 1$. Fig. 13(b), $r_0 = 2.75$, TE_{11}-limit mode and incipient TM_{11}-limit mode, $\sigma > 1$. Fig. 13(c), $r_0 = 2.75$, TE_{11}-limit mode, $\sigma < -1$. Fig. 13(d), (e) and (f), $r_0 = 5.75$, TE_{11}-limit and TE_{12}-limit modes; Fig. 13(d), $-1 < \sigma < 1$; Fig. 13(e), $\sigma > 1$; Fig. 13(f), $\sigma < -1$. Fig. 13(g), 13(h) and 13(i), $r_0 = 5.75$, TM_{11}-limit modes; Fig. 13(g); $-1 < \sigma < 1$; Fig. 13(h), $\sigma > 1$; Fig. 13(i), $\sigma < 1$. It should be noted that a scale linear in $\frac{\beta^2}{1+\beta^2}$ is used for convenience when $|\sigma| > \sigma_0$.

continuously to $\sigma = -\sigma_0$, $\lambda_2 = -\frac{1}{\sigma_0}$ and, simultaneously, the cut-off point of the TM_{11} curve occupies the position relinquished by the former. The TM_{11} mode now exists for all p. A new solution curve (TE_{12}) appears in the region bounded by $0_1'$, $(0_1)_T$ and I_1', if p is not too large, terminating at $(I_A)_T - I_1'$, the point left by the TM_1 terminus. (If p exceeds $\sqrt{1 - j_1^2/u_3^2}$ this curve will not exist at all.)

Figs. 13(a) to (i) and 14 show the approximate course of the $\beta^2 - \sigma$ curves for the TE_{11} mode at $r_0 = 2.75$ and for the TE_{11}, TM_{11} and TE_{12} modes at $r_0 = 5.75$. The incipient TM_{11} mode at $r_0 = 2.75$ is shown for positive σ, p only. They were computed by the methods outlined above.

4.14. *Guides of large radius.* It is of some interest because of the high dielectric constant of ferrites to examine the behavior of the modes as the radius, r_0, is allowed to become very large. The two sides of the G-equation will remain determinate for unlimited r_0 provided

$$r_0 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} \left(\text{or } r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \right)$$

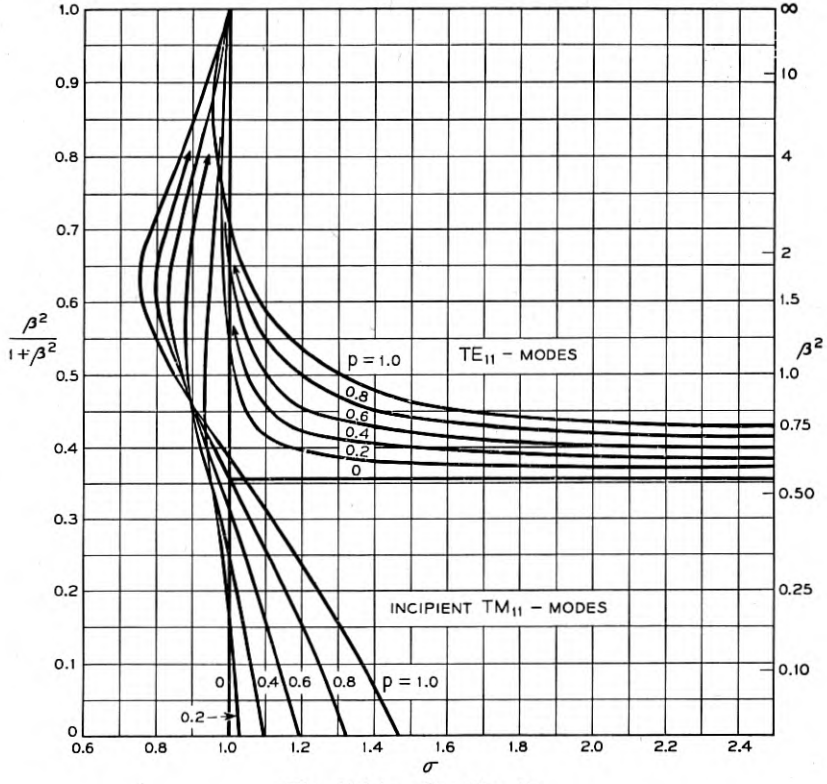


Fig. 13(b) — See Fig. 13.

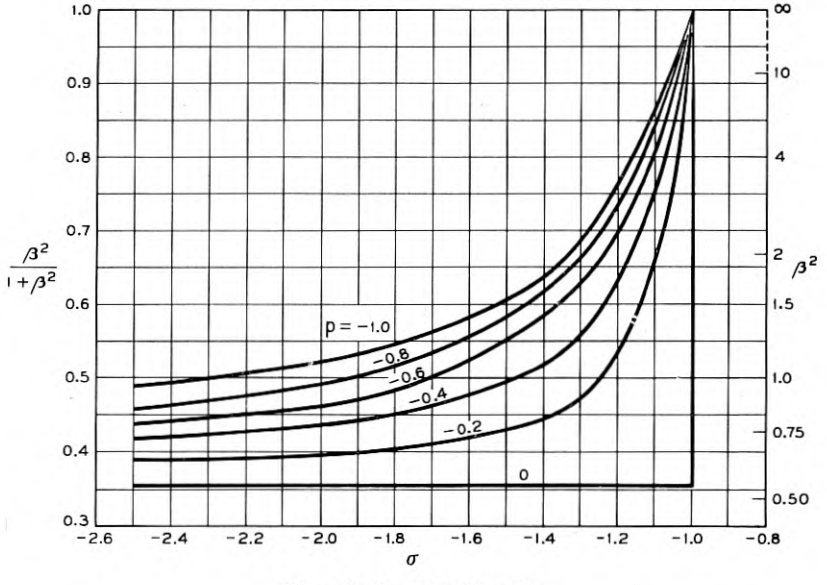


Fig. 13(c) — See Fig. 13.

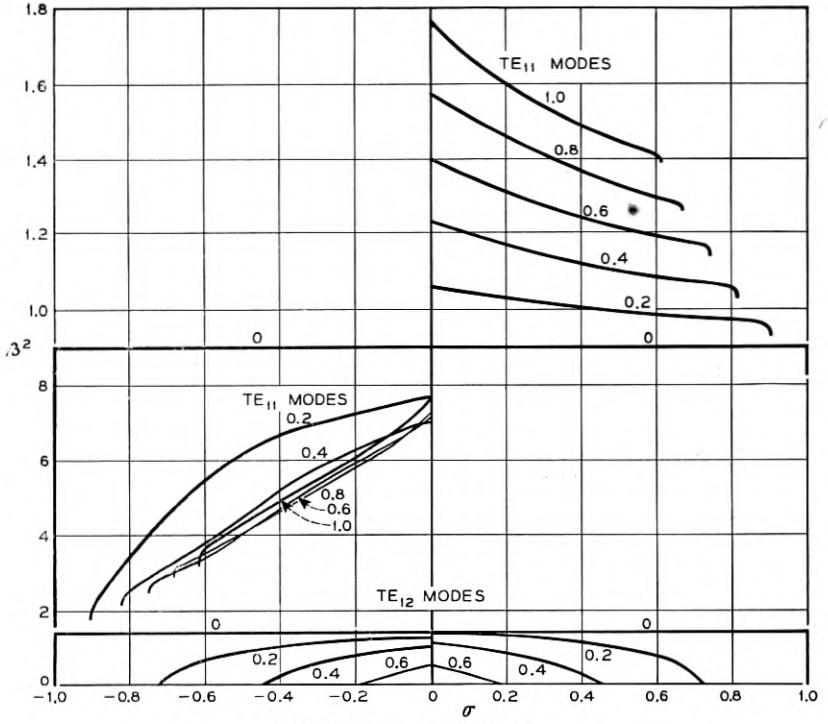


Fig. 13(d) — See Fig. 13.

remains finite, while

$$r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \left(\text{or } r_0 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} \right)$$

becomes infinite imaginary. Examining the solutions obtained under these conditions it is possible to find expansions for β^2 in inverse powers of r_0^2. These are as follows:

$$\text{for } p > 0, \sigma > 1 \text{ or } p < 0, -\sigma_0 < \sigma < 0$$

$$\beta^2 = 1 + \frac{p}{\sigma - 1} - \left(1 + \frac{p/2}{\sigma - 1}\right) \frac{x_n^2}{r_0^2}, \tag{43a}$$

where the x_n are the successive roots greater than zero of $F(x_n) = 1$ and the modes are associated with the x_n by the scheme: $TE_{11} \rightarrow x_1$, $TM_{11} \rightarrow x_2$, $TE_{12} \rightarrow x_3$, etc:

$$\text{for } p > 0, 0 < \sigma < \sigma_0 \text{ or } p < 0, \sigma < -1$$

$$\beta^2 = 1 + \frac{p}{\sigma - 1} - \left(1 + \frac{p/2}{\sigma + 1}\right) \frac{y_n^2}{r_0^2}, \tag{43b}$$

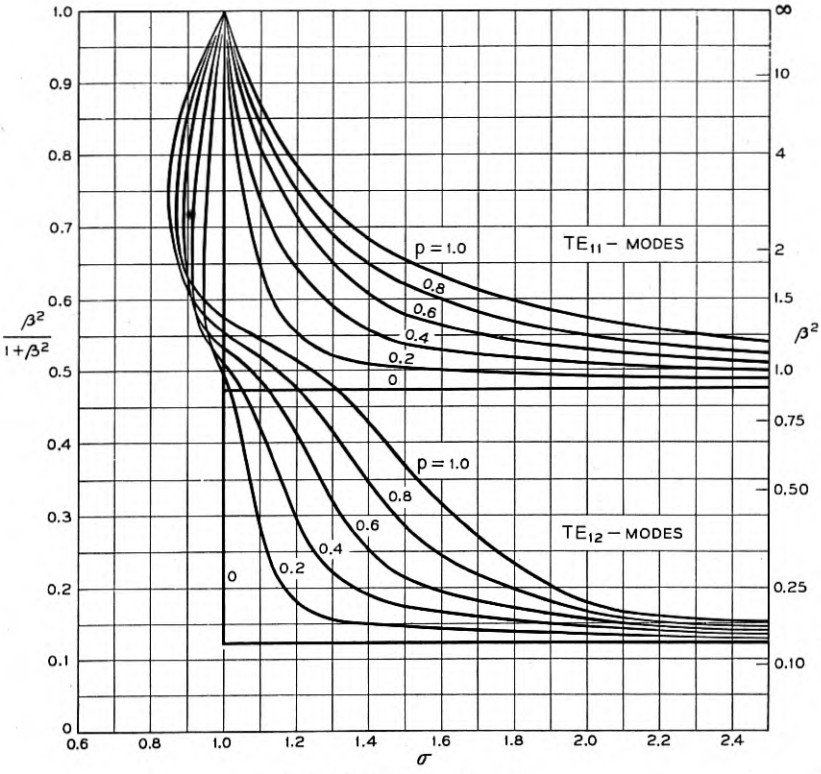


Fig. 13(e) — See Fig. 13.

with $F(y_n) = -1$ and $TE_{11} \rightarrow y_1$; $TM_{11} \rightarrow y_2$; $TE_{12} \rightarrow y_3$, etc.

$$\text{for } p > 0,\ \sigma_0 < \sigma < 1 \text{ and } p < 0,\ -1 < \sigma < -\sigma_0,$$

there are no solutions.

These formulae are valid for any p which is not itself so small as to be of order $1/r_0^2$. If they are applied to modes varying as $e^{jn\theta}$, where $n = \pm 1$ and σ, p are positive, they indicate the following results: for $\sigma > 1$, $n = \pm 1$, $\beta^2 \rightarrow 1 + \dfrac{p}{\sigma - 1}$; for $\sigma_0 < \sigma < 1$, no $n = \pm 1$ modes; for $0 < \sigma < \sigma_0$, $n = \pm 1$, $\beta^2 \rightarrow 1 + \dfrac{p}{\sigma + 1}$. These, in turn, may be classified in the following way. For $\sigma > 1$, $n = -1$, and for $0 < \sigma < \sigma_0$, $n = +1$ which correspond to μ and κ both positive, the propagation constant tends to the value for a plane wave whose direction of circular polarization coincides with that of the wave guide pattern. For $\sigma_0 < \sigma < 1$,

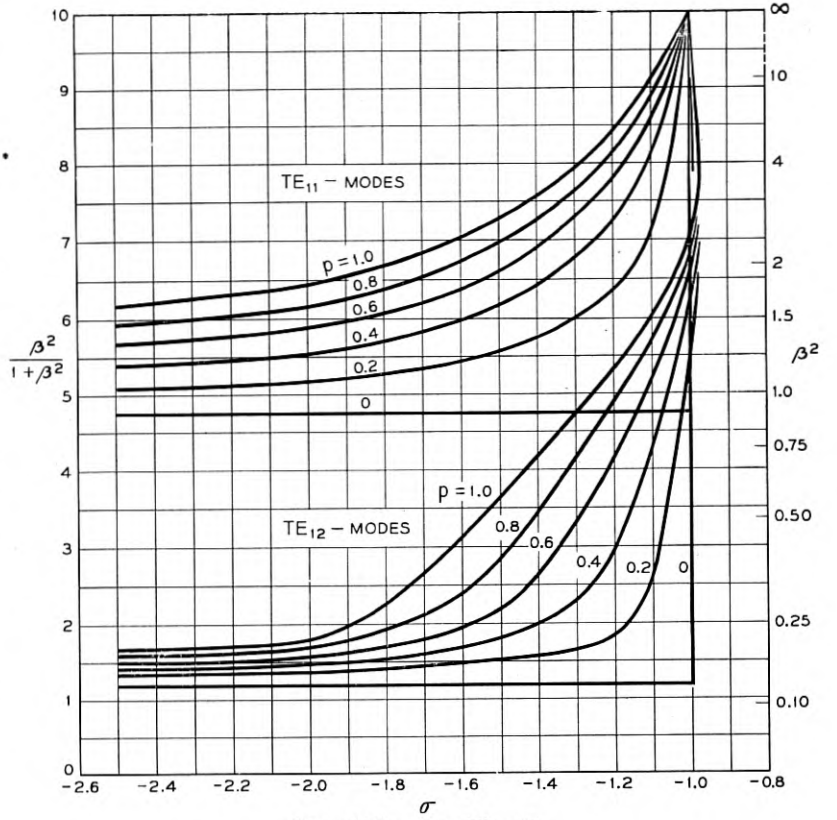


Fig. 13(f) — See Fig. 13.

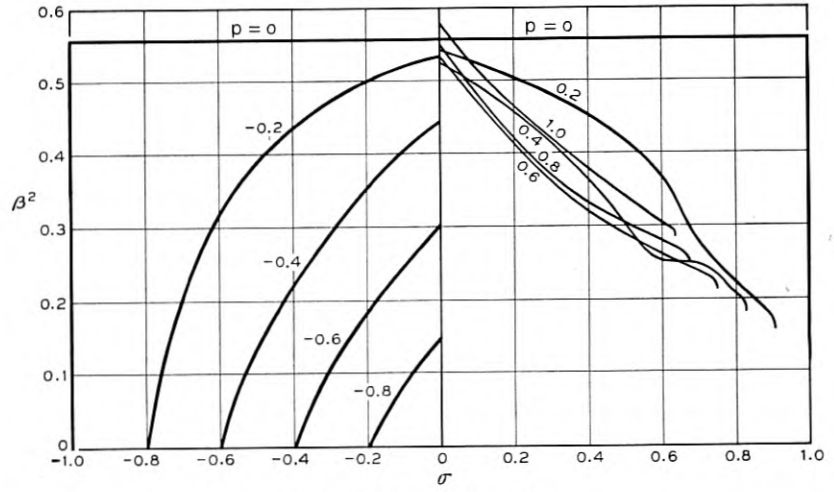


Fig. 13(g) — See Fig. 13.

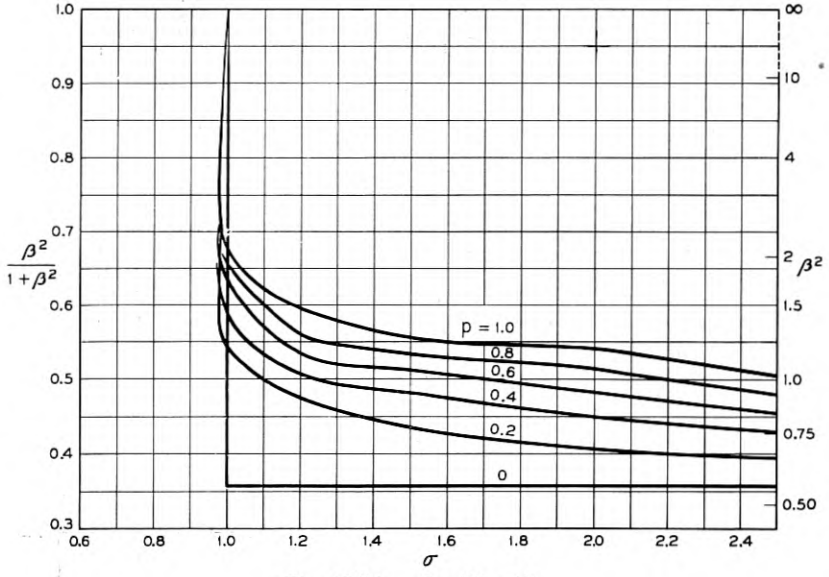


Fig. 13(h) — See Fig. 13.

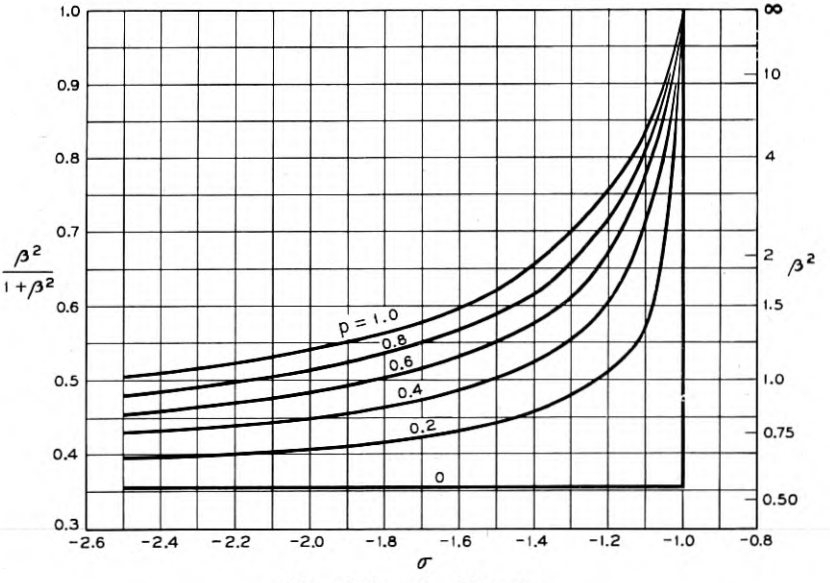


Fig. 13(i) — See Fig. 13.

where μ is negative, $n = \pm 1$, no modes exist for large enough guide. For $0 < \sigma < \sigma_0$, $n = -1$ and $\sigma > 1$, $n = +1$, the propagation constant tends to that for a plane wave whose polarization is in the opposite sense to that of the field pattern and here μ is positive, but κ is negative. An examination of the field pattern in this last case shows that most of the field energy is indeed associated with a circular polarization opposite to that of the pattern as a whole.

The discussion of the preceding sections shows that the complete structure of the mode spectrum for a guide filled with lossless ferrite is very complex. It is also clear that for some combinations of guide size and magnetic parameters the course of an individual mode in the $\beta^2 - \sigma$ plane may be quite involved. In particular, two values of β^2 associated with the same mode often occur at a given σ. The extent to which the complexity of the spectrum will be observed in practice will depend principally upon the loss of the real ferrite and upon the guide radius. The effect of loss near $\sigma = 1$, where the incipient modes are crowded will be to cause simultaneous excitation of many of these and consequently a confused z dependence of the guide excitation. For values of r_0 just below j_n, the point of escape of the TE modes, the latter exist over considerable ranges of σ, see Fig. 12(a), and would probably be observable. The TE modes near u_n also persist over a wide range, but are double-valued. Concerning such double-valued waves it may be observed that from the results of the subsequent treatment of losses, it is clear that if $\dfrac{d\beta^2}{d\,|\,\sigma\,|} > 0$, it is necessary to put the source of power at the opposite end of the guide.

4.15. *Losses, Faraday rotation and merit figure.* So far the analysis has been concerned with the loss-free medium. It is of some interest to determine the attenuation constant (the imaginary part of β) that arises when losses are taken into account. As long as these are small, this can be done rather easily; in fact, sufficiently far from resonance ($\sigma = 1$), for each formula giving β^2, we can establish one giving the attenuation constant.

If the losses are of magnetic origin we utilize the fact (already demonstrated in section 2) that to first order in α, the permeabilities μ, κ are functions of $\sigma + j\alpha \operatorname{sgn} p$, and of no other combination of σ, α. Since σ, α enter Maxwell's equations only through μ and κ, β^2, which is derived from them, must likewise depend on σ through $\sigma + j\alpha \operatorname{sgn} p$. Any formula for β^2 derived for the loss-free medium can, therefore, be generalized to the lossy case by replacing σ with $\sigma + j\alpha \operatorname{sgn} p$, to first order in α. To this order, then, we find

$$\beta^2 = \beta'^2 + ja \operatorname{sgn} p \frac{\partial(\beta')^2}{\partial\sigma},$$

where β' is the propagation constant for the loss-free case. Thus

$$j\beta = j\beta' - \frac{\alpha}{2\beta'} \frac{\partial(\beta')^2}{\partial \mid \sigma \mid}, \tag{44}$$

and the last term on the right, (multiplied by our scaling variable $\beta_0 = \omega\sqrt{\mu_z\epsilon_0}$) is the attenuation in nepers per meter. The present convention is that the waves propagate in the positive z direction, as exp $(-j\beta z)$. It follows that they will decrease in that direction only if $\partial(\beta')^2/\partial \mid \sigma \mid < 0$. Occasionally this is not the case, and presumably indicates that the direction of the power flow opposes that of the phase velocity.

For small dielectric loss, too, it is possible to derive formulae for the attenuation constant from those already obtained; obviously the latter depend on ϵ only through $\epsilon = \epsilon_0 - j\epsilon_1$, and can therefore be expanded. But it must now be remembered that β was defined as $\beta_{\text{actual}}/\omega\sqrt{\mu_z\epsilon}$. and r_0 as $r_{\text{actual}}\ \omega\sqrt{\mu_z\epsilon}$ so that the scaling parameter $\omega\sqrt{\mu_z\epsilon}$ will make contributions to the imaginary part. It is then readily verified that

$$\frac{\beta_{\text{actual}}}{\omega\sqrt{\mu_z\epsilon_0}} = \beta' - j\frac{\epsilon_1}{2\epsilon_0}\frac{1}{\beta'}\frac{\partial}{\partial(r_0^2)}(r_0^2\beta^2). \tag{45}$$

A few words may be said about the relation of Faraday rotation to the $\beta^2 - \sigma$ curves. A linearly polarized plane wave traveling in the unbounded medium along the magnetizing field can be regarded as the

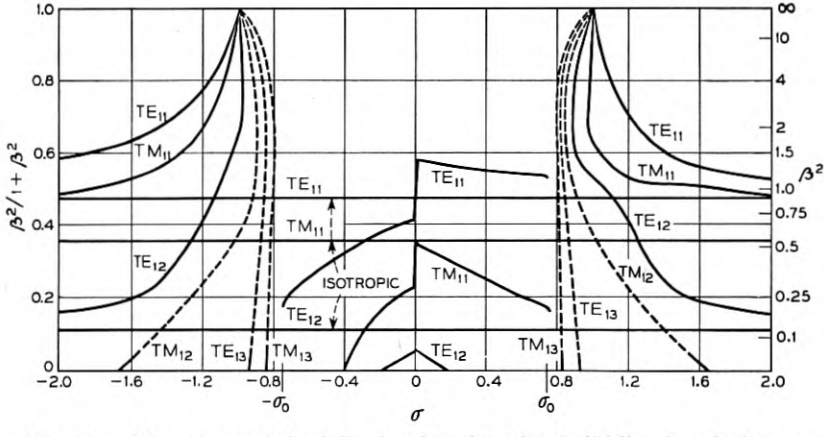


Fig. 14 — The course of the fully developed modes (solid lines) and of some of the lower incipient modes (dotted lines) as a function of σ for $r_0 = 5.75$ and $\mid p \mid = 0.6$.

sum of right and left circular components which travel with different propagation constants. If these are β_+ and β_- (measured in units of β_0) the plane of polarization of the resultant will appear to rotate by $(\beta_+ - \beta_-)/2$ radians per reduced wavelength $\frac{\lambda}{2\pi} = \frac{1}{\beta_0}$.

In the filled waveguide, on the other hand, it is no longer true that right and left circularly polarized modes add up to a plane polarized mode, as is readily seen by reference to the field components given in Appendix IV. To define Faraday rotation in a simple way it is therefore necessary to neglect changes in the field pattern due to the magnetization and consider only the changes in the propagation constants. Then the rotation of a mode with azimuthal mode number n will be $\frac{1}{2n}(\beta_+ - \beta_-)$. In the present case $n = 1$, and the β_+, β_- are found from the curves of β^2 versus σ, Figs. 13(a) to (i), for positive (σ, p) and negative (σ, p) respectively.

The merit figure is defined as the ratio — radians rotation per neper loss — and is independent of path-length. For small losses, (neglecting terms $0(\alpha^2)$), this ratio is

$$\frac{Rl(\beta_+ - \beta_-)}{Im(\beta_+ + \beta_-)} = \frac{1}{\alpha} \frac{\beta_+' - \beta_-'}{\frac{\partial \beta_+'}{\partial |\sigma|} + \frac{\partial \beta_-'}{\partial |\sigma|}},$$

in the notation of the present Section.

4.16. *Formulae for the ferrite.*

I. *Cut-off points*

Cut-off points will be classified into three types, 1, 2 and 3, according to the nature of the intersecting curves which generate them. All points of a given type may be assigned an index which further identifies the generating curve. This will be written as a subscript.

Type 1. Intersections of $I_n - (I_A')_T$, $\sigma > 0$, written as 1_n and of $I_n' - (I_A)_T$, $\sigma < 0$, written as $1_n'$

$\beta^2 = 0$

There is a parametric representation:

$$\begin{aligned} |\lambda_{1,2}| &= e^{\theta}, \\ |p| &= 2 \sinh \theta \left(1 - \frac{r_0^2}{j_n^2}\right) \text{ and} \\ |\sigma| &= \frac{r_0^2}{j_n^2} e^{\theta} + \left(1 - \frac{r_0^2}{j_n^2}\right) e^{-\theta}. \end{aligned} \tag{46}$$

The slope at cut-off:

$$\left(\frac{\partial\beta^2}{\partial\sigma}\right)_p = \frac{\frac{j_n^2}{r_0^2}\left(\frac{j_n^2}{r_0^2} - 1\right)\coth\theta}{\left[\frac{j_n^2}{r_0^2} - 1 + \frac{2}{F(r_0)}\right]e^{-\theta} - e^{\theta}} \cdot \operatorname{sgn}\sigma, \tag{47}$$

and

$$\left(\frac{\partial\beta^2}{\partial p}\right)_\sigma = -\frac{\sigma}{p\coth\theta}\left(\frac{\partial\beta^2}{\partial\sigma}\right)_p. \tag{48}$$

Type 2. Intersections of

$$(0_n')_T - 0_c \text{ at } \sigma = \sigma_0 = \sqrt{\frac{p^2}{4} + 1} - \frac{p}{2}\text{ ; type } 2_n'$$

$$(0_n)_T - 0_c' \text{ at } \sigma = -\sigma_0\text{ ; type } 2_n.$$

$\beta^2 \neq 0$

Define λ_0 as a root of

$$\sigma_0 = \frac{-1}{|\lambda_0|}\left[1 - (1 - \lambda_0^2)\left(\frac{r_0}{F^{-1}(\lambda_0)}\right)^2\right], \tag{49}$$

with the following convention: for $2_n'$, λ_0 is negative and the n^{th} branch of $F^{-1}(\lambda_0)$ is used; for 2_n, λ_0 is positive and the n^{th} branch of $F^{-1}(\lambda_0)$ is used. Now

$$\beta^2 = \frac{|\lambda_0|}{\sigma_0}. \tag{50}$$

Near cut-off

$$\beta^2 = \frac{|\lambda_0|}{\sigma_0} + \frac{r_0}{A}\sqrt{\sigma_0\frac{1+\sigma_0^2}{1-\sigma_0^2}}\sqrt{\frac{\sigma_0 - |\sigma|}{1 + |\lambda_0|\sigma_0}}. \tag{51}$$

with

$$(1 - \lambda_0^2)A = \frac{1}{2|\lambda_0|}\left(1 - \frac{r_0^2}{1 + |\lambda_0|\sigma_0}\right)(\sigma_0 + 2|\lambda_0| + \sigma_0\lambda_0^2) + \frac{1 + \sigma_0|\lambda_0|}{\lambda_0}.$$

For 2_0 a special situation arises for $1 < r_0 < u_1$ where there are two positive solutions for λ_0. We write 2_{01} and 2_{02} for the points corresponding to the smaller and greater of these. A special cut-off point will be labeled $2_{0\infty}$ and arises from $0_c'$ and $(0_0)_T$ as λ_1 goes to infinity. For this point

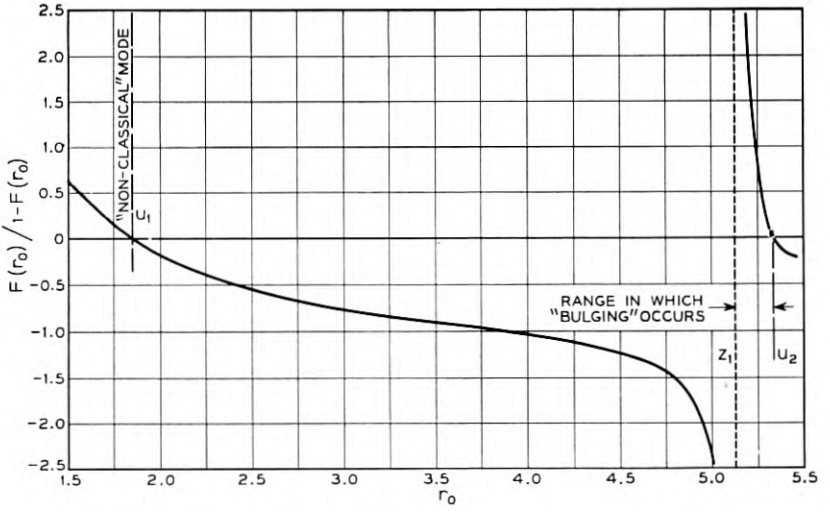


Fig. 15 — The function $\dfrac{F(r_0)}{1 - F(r_0)}$, related to cutoff of Type 3.

$\tau = \sigma_0$, $\beta^2 \to \infty$ and near $-\sigma_0$ we have

$$\beta^2 = \frac{|\, p \,|}{r_0^2(|\, \sigma_0 \,| - |\, \sigma \,|)} \frac{1}{(2 - |\, p \,||\, \sigma_0 \,|)} . \tag{52}$$

Type 3. Intersections of $I_B - (I_A)_\tau$; for $-\sigma_0 < \sigma < 0$ only; no subscript is needed.

$$\beta^2 = 0,$$

$$\sigma = -1 - p,$$

and

$$\frac{(\partial\beta^2)}{(\partial\sigma)_p} = \frac{(\partial\beta^2)}{(\partial p)_\sigma} = \frac{1}{p} \frac{F(r_0)}{1 - F(r_0)} \qquad \text{(see Fig. 15)} \tag{54}$$

The cut-off points of the modes follow various schemes in different ranges of σ as indicated below.

For $\sigma > \sigma_0$ we have Table I. When $\sigma < -\sigma_0$, $1_n'$ replaces 1_n in the Table I. "None" indicates that the mode exists, but has no cut-off.

For $0 < \sigma < \sigma_0$ we have Table II. "N.P." in Table II indicates that the mode is not propagated. For $-\sigma_0 < \sigma < 0$ we have Table III. In this range of σ one has also the mode without classical analogue. For $r_0 < u_1$ this is cut-off at 3 and 2_{01} and may have a second branch from 2_{02} to $2_{0\infty}$. For $r_0 > u_1$ the second branch only exists.

TABLE I

Mode / Radius	TE_{11}	TM_{11}	TE_{12}	TM_{12}	TE_{13} etc.
$r_0 < u_1$	1_1	1_2 ↘	1_3	1_4	1_5
$u_1 < r_0 < j_1$	None	1_1	1_2 ↓	1_3	1_4
$j_1 < r_0 < u_2$	None	None	1_2 ↘	1_3	1_4
$u_2 < r_0 < j_2$	None	None	None	1_2	1_3
$j_2 < r_0 < u_3$	None	None	None	None	1_3
etc.					

II. *Asymptotes, genesis of the modes and spot points*

For $|\sigma| \to \infty$ there are asymptotic formulae:

For TE_{1n}-modes

$$\beta^2 \to \left(1 - \frac{u_n^2}{r_0^2}\right)\left(1 + \frac{p}{\sigma}\right). \tag{55}$$

For TM_{1n}-modes

$$\beta^2 \to \left(1 - \frac{j_n^2}{r_0^2}\right) + \frac{p}{\sigma}. \tag{56}$$

For $|\sigma| > \sigma_0$, all modes have their origin in the points $\sigma = 1, \lambda_1 = 1$ or $\sigma = -1, \lambda_2 = -1$, where $\beta^2 \to \infty$. The variation of β^2 with σ in the neighborhood of these points is described by the two expansions:

for $\sigma \sim 1$

$$\beta_n^2 = \frac{p}{a_n + 1}\left[\frac{1}{\sigma - 1} + \left\{\frac{1 - a_n}{p} + \frac{2a_n^2}{x_n^2}\left(\frac{2}{p} + 1\right) + \frac{a_n^2}{2}\right\} + 0(\sigma - 1)\right], \tag{57}$$

where

$$\frac{a_n}{a_n + 1} = \frac{x_n^2}{2r_0^2},$$

and

$$F(x_n) = 1 \qquad x_n > 0,$$

with the scheme

n	1	2	3	4 etc.
Mode	TE_{11}	TM_{11}	TE_{12}	TM_{12} etc.

TABLE II

Mode / Radius	TE_{11}	TM_{11}	TE_{12}	TM_{12}	TE_{13} etc.
$0 < r_0 < u_1$	N.P.	N.P.	N.P.	N.P.	N.P.
$u_1 < r_0 < j_1$	$2_0'$	N.P.	N.P.	N.P.	N.P.
$j_1 < r_0 < u_2$	$2_0'$	1_1 ↘	N.P.	N.P.	N.P.
$u_2 < r_0 < j_2$	$2_0'$	$2_1'$	1_1 ↓	N.P.	N.P.
$j_2 < r_0 < u_3$	$2_0'$	$2_1'$	1_1 ↘	1_2	N.P.
$u_3 < r_0 < j_3$ etc.	$2_0'$	$2_1'$	$2_2'$	1_1	1_2

TABLE III

Mode / Radius	TE_{11}	TM_{11}	TE_{12}	TM_{12}	TE_{13} etc.
$0 < r_0 < u_1$	N.P.	N.P.	N.P.	N.P.	N.P.
$u_1 < r_0 < j_1$	3 ↓	N.P.	N.P.	N.P.	N.P.
$j_1 < r_0 < u_2$	3 ↘	$1_1'$	N.P.	N.P.	N.P.
$u_2 < r_0 < j_2$	2_1	3 ↓	$1_1'$	N.P.	N.P.
$j_2 < r_0 < u_3$	2_1	3 ↘	$1_1'$	$1_2'$	N.P.
$u_3 < r_0 < j_3$ etc.	2_1	2_2	3	$1_1'$	$1_2'$

for $\sigma \sim -1$

$$\beta_n^2 = \frac{p}{a_n + 1}\left[\frac{1}{\sigma + 1} + \left\{\frac{1 - a_n}{p} + \frac{2a_n^2}{y_n^2}\left(1 - \frac{2}{p}\right) - \frac{a_n^2}{2}\right\} + 0(\sigma + 1)\right], \tag{58}$$

where

$$\frac{a_n}{a_n + 1} = \frac{y_n^2}{2r_0^2},$$

and

$$F(y_n) = -1,$$

with the same identification as above.

For $\sigma > \sigma_0$ a spot-point is given by $I_n - (I_B)_T$ with

$$\beta^2 = \lambda_1 = 1 + \frac{p}{1+\sigma},$$

and (59)

$$\frac{1-\lambda_1^2}{1-\sigma\lambda_1} = \frac{j_n^2}{r_0^2},$$

The identification scheme is again that shown above.

For $-\sigma_0 < \sigma < 0$ an isolated identifiable point arises from I_B'-$(I_n)_T$ which is expressible in inverse form by the relations

$$\sigma = \frac{1-\frac{r_0^2}{j_n^2}}{\beta^2} + \frac{r_0^2}{j_n^2}\beta^2 \qquad \text{and}$$
$$p = (\beta^2-1)\left[1 + \frac{r_0^2}{j_n^2}\beta^2 + \frac{1-\frac{r_0^2}{j_n^2}}{\beta^2}\right] \tag{60}$$

for

$$1 - \frac{j_n^2}{r_0^2} < \beta^2 < \sqrt{1-\frac{j_n^2}{r_0^2}}.$$

The identification of the modes with n proceeds as in the earlier parts of this section.

III. *Small p.*

To order p^2 there exist the following expansions for β^2: for the TE_{1n}-mode

$$\beta^2 = \beta_0^2 + \frac{\beta_0^2}{1-\sigma^2}\left[\frac{2}{u_n^2-1} - \sigma\right]p + \frac{2\beta_0^2}{(1-\sigma^2)^2} \cdot \frac{\frac{\beta_0^2}{(u_n^2-1)^2} + \frac{5+5u_n^2-2u_n^4}{4(u_n^2-1)}}{(\beta_0^2-1)(1-u_n^2)}p^2\ldots, \tag{61}$$

where

$$\beta_0^2 = 1 - \frac{u_n^2}{r_0^2}.$$

For the TM_{1n}-modes

$$\beta^2 = \beta_0^2 - \frac{\sigma}{1-\sigma^2}p + \frac{1}{(1-\sigma^2)^2}\frac{3\beta_0^2-2}{2(1-\beta_0^2)}p^2\ldots, \tag{62}$$

where

$$\beta_0^2 = 1 - \frac{j_n^2}{r_0^2}.$$

The radius of convergence of these series is not known. It is clear that it will depend on σ and will become smaller as $\sigma^2 \to 1$.

4.2. *The Plasma* ($\rho_H = 0$, $\nu_H = 1$). The characteristic equation (26) may now be written

$$\frac{1}{\chi_1^2}[\lambda_1 F_n(\chi_1 r_0) - n] = \frac{1}{\chi_2^2}[\lambda_2 F_n(\chi_2 r_0) - n], \tag{63}$$

where (in contrast with the ferrite case) $\lambda_{1,2} = \beta\Lambda_{1,2}$. The λ satisfy

$$\lambda_{1,2}^2 - \frac{(\nu_E - 1)(1 - \beta^2/\nu_E) - \nu_E\rho_E^2}{\rho_E}\lambda_{1,2} - \beta^2 = 0, \tag{64}$$

and the χ's are given by

$$\chi_{1,2}^2 = (1 - \beta^2/\nu_H) - \rho_H\lambda_{1,2} \tag{65}$$

From the equations for ρ_E and ν_E in terms of σ, q given in Section 2, equation (64) may be written

$$\lambda_{1,2}^2 - \left(\frac{\sigma}{1 - q^2} - \sigma\beta^2\right)\lambda_{1,2} - \beta^2 = 0, \tag{66}$$

or

$$\lambda_1\lambda_2 = -\beta^2 \tag{67a}$$

$$\begin{aligned}\lambda_1 + \lambda_2 &= \frac{\sigma}{1 - q^2} - \sigma\beta^2, \\ &= \frac{\sigma}{1 - q^2} + \sigma\lambda_1\lambda_2.\end{aligned} \tag{67b}$$

Elimination of β^2 between equations (64) and (65) enables us to express $\chi_{1,2}^2$ solely in terms of $\lambda_{1,2}$, ρ_H and ν_H:

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \dfrac{1 - 1/\nu_E}{\rho_E}\lambda_{1,2}},$$

which, from the plasma formulae for ρ_E, ν_E, can be written

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \sigma\lambda_{1,2}}. \tag{68}$$

With these expressions for the χ, the characteristic equation (63) takes the form

$$H(\lambda_1, \sigma, r_0) = H(\lambda_2, \sigma, r_0),$$

where

$$H(\lambda, \sigma, r_0) = \frac{1 - \sigma\lambda}{1 - \lambda^2}\left[\lambda F_n\left(r_0\sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}\right) - n\right]. \tag{69}$$

For given σ, and q, equations (67b) and (69) are simultaneous equations for λ_1, λ_2. When λ_1, λ_2 have been found, $\beta^2 = -\lambda_1\lambda_2$ is known. Since β^2 must be positive, λ_1, λ_2 must have opposite signs. As in the ferrite, the convention $\lambda_1 > 0$, $\lambda_2 < 0$ will be adopted. Equation (67b) will hereafter be called the plasma relation. The transformation

$$\lambda_1 \to -\lambda_2, \qquad \lambda_2 \to -\lambda_1, \qquad \sigma \to -\sigma$$

leaves the plasma relation unchanged and changes n to $-n$ in the H-equation. As more fully explained in connection with the ferrite section, it is therefore necessary to consider positive n only if σ is allowed to take on negative as well as positive values. As before, only the first azimuthal mode number ($n = \pm 1$) is considered in this paper.

The method of analysis is the same as that used for the ferrite. Here we shall only sketch the most important steps; the reader will have no difficulty in completing the analysis by referring to Section 4.11. For fixed r_0, a contour map of H is drawn in the λ, σ plane (see Fig. 16 drawn for $r_0 \sim 2.2$). The gross features of this map are determined by the lines $H = 0$, $H = \pm\infty$. For greater detail recourse is had to the lines $\dfrac{1 - \lambda^2}{1 - \sigma\lambda} = \text{constant}$, along which values of H are readily generated. Further help is obtained from a knowledge of the location of the saddle point of H. The infinity curves are given by the same formulae as for the ferrite, except that the line $\lambda = 0$ is no longer an infinity line: along $\lambda = 0$, $H = -1$. Zero curves are given by

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda^{-1})]^2}.$$

The branches of $\sigma\lambda = 1$ are also zero curves in the same restricted sense as for the G function. In the same notation as for the ferrite, all I_n curves pass through $\sigma = 1$, $\lambda = 1$; all I_n' curves through -1, -1. The same is true for all 0_n, $0_n'$ curves ($n > 0$). The only exception is denoted by 0_0, it arises from that branch of F^{-1} along which $F^{-1}(1) = 0$.

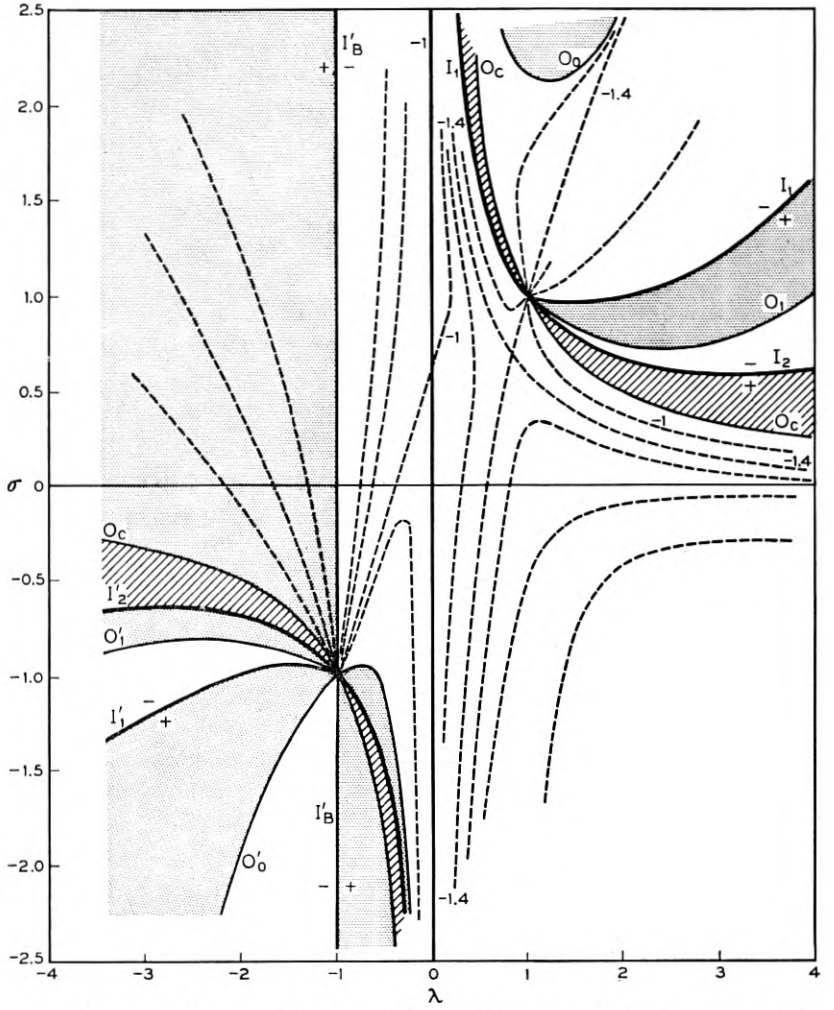


Fig. 16 — The division of the $\lambda - \sigma$ plane into regions of positive and negative H by the first few O and I curves. Dotted regions are positive. Cross-hatched regions contain the higher O and I curves.

Along all lines, $\sigma = \lambda c + d$, H tends to infinity except when $c = r_0^2/u_n^2$ (the slopes of the linear asymptotes of 0 curves). Along $\sigma = \frac{r_0^2}{u_n^2}\lambda + d$, H tends to a value depending on d. More about the general behavior of H can be derived by means entirely analogous to those employed in the study of G.

λ-pairs determined from the H-diagram do not necessarily solve the problem, since for a fixed q, they may not satisfy the plasma relation (67b). To take it into account, we interpret it as a transformation of the whole of the second quadrant onto part of the first, and of the whole fourth quadrant on part of the third. Writing (67b) in the form

$$\lambda_1 = \frac{1}{\sigma} + \frac{1}{\sigma}\frac{\dfrac{\sigma^2}{1-q^2} - 1}{1 - \sigma\lambda_2} = T(\lambda_2),$$

we see that the curves $\lambda_2 = \text{const.}$ transform into a bundle of hyperbola passing through the intersection of $\sigma^2 = 1 - q^2$ with $\sigma = 1/\lambda$, that is, through

$$\lambda_{10} = \frac{1}{\sqrt{1-q^2}}, \qquad \sigma_0 = \sqrt{1-q^2}.$$

These hyperbolae have vertical asymptotes $\lambda_1 = -\dfrac{1}{\lambda_2}$, and cut the line $\sigma = 0$ in $-\lambda_2$. For a fixed positive $\sigma < \sigma_0$, λ_1 decreases from $1/\sigma$ to $\sigma/(1-q^2)$ as λ_2 increases from $-\infty$ to 0, but, when $\sigma > \sigma_0$, λ_1 increases from $1/\sigma$ to $\sigma/(1-q^2)$. Thus the second quadrant transforms into the region between $\sigma = \lambda(1-q^2)$ and $\sigma = 1/\lambda$ in the first quadrant. Similarly, the inverse transformation $\lambda_2 = T(\lambda_1)$ transforms the fourth quadrant into the region between $\sigma = \lambda(1-q^2)$ and $\sigma = 1/\lambda$ in the third quadrant. Points outside these regions cannot be site of acceptable solutions of the H equation. In order to locate acceptable solutions, the $H =$ equation is now written in the form

$$H(\lambda_1, \sigma, r_0) = H(T(\lambda_1), \sigma, r_0)$$

when $\sigma > 0$, and in the form

$$H(\lambda_2, \sigma, r_0) = H(T(\lambda_2), \sigma, r_0)$$

when $\sigma < 0$. These equations represent the curves of intersection of the H-surfaces. Along each such curve, both H-equation and plasma relation are satisfied. Their projections onto the first (or third) quadrant give λ_1 (or λ_2) as a function of σ, and hence λ_2 (or λ_1) from the plasma relation. Thus $\beta^2 = -\lambda_1\lambda_2$ is known along each solution curve. The rough locating of the solution curves, and the establishment of precise analytical formulae near special points on them proceeds in complete analogy with the ferrite case. Here we shall consider only the radius $r_0 \sim 2.2$, as typical of radii large enough to permit propagation of the TE_{11} mode

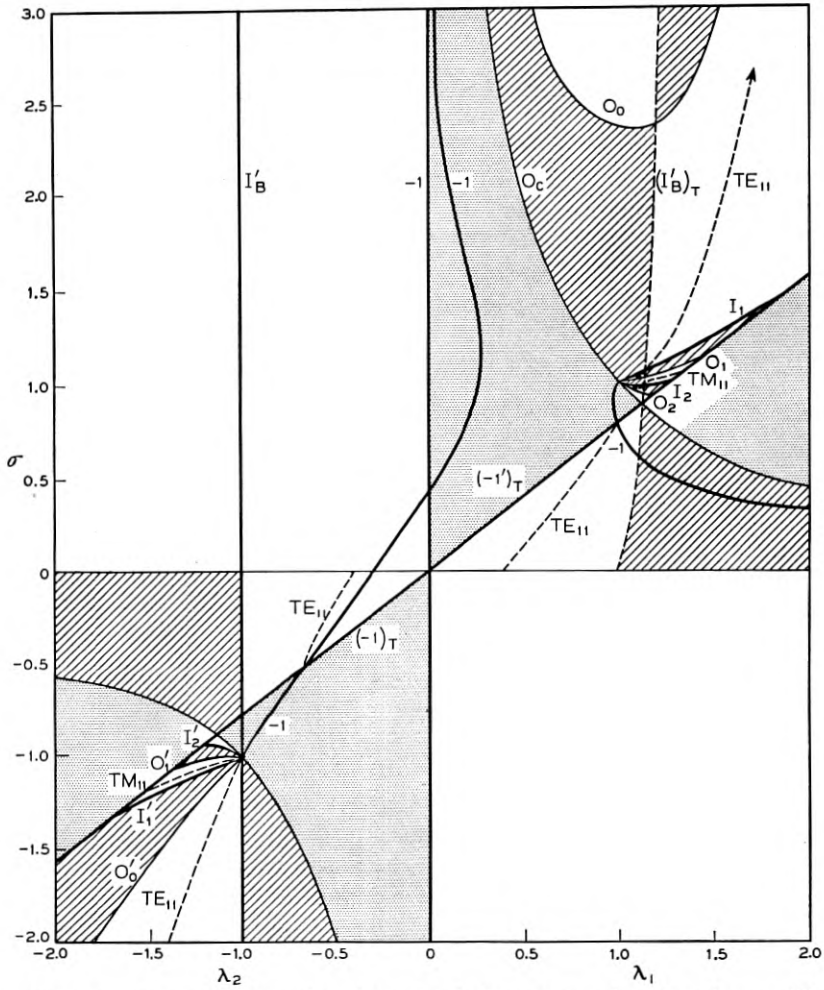


Fig. 17 — Geometrical exploration of solution curves for the plasma. The dotted regions are excluded by the plasma relation; the cross-hatched regions are those in which $H(\lambda, \sigma)$ and $H(T(\lambda), \sigma)$ have unlike sign. Solution curves may lie only in unshaded parts of the 1st and 3rd quadrants. $q^2 = 0.25$, $r_0 \sim 2.0$.

through the unmagnetized plasma, but too small to admit higher modes. The solution curves for $r_0 \sim 2.2$ are indicated roughly in Fig. 17.

In the first quadrant, for $\sigma > \sigma_0$, a solution curve starts at $\lambda_1 = 1$, $\sigma = 1$, passes through the intersection of I_1, $(I_B')_T$ and proceeds to infinity as indicated. The formulae in Section 4.21 describe the corresponding curve near $\sigma = 1$, and at $\sigma \to \infty$, showing that the solution

curve describes the TE_{11}-limit mode. Incipient modes also exist, just as in the ferrite; their end-points on $\sigma = (1 - q^2)\lambda_1$ (or, briefly, on $(\lambda_2 = 0)_T$) are now the points for which $H(\lambda_1, \sigma) = -1$ and, simultaneously, $\sigma = (1 - q^2)\lambda_1$.

Below σ_0, there is only one solution curve for $r_0 \sim 2.2$. It begins at $\sigma = 0$, $\lambda_1 = \beta_{iso}$ $(= -\lambda_2$, by the plasma relation), where β_{iso} is the propagation constant of the TE_{11} mode in the unmagnetized plasma. (In contrast with the ferrite, the plasma becomes isotropic as $\sigma \to 0$). It is cut off at the intersection of the contour $H(\lambda_1, \sigma) = -1$ in that region with $(\lambda_2 = 0)_T$. At that point $\beta^2 = 0$ and σ is best stated, thus:

$$\sigma = (1 - q^2)\sqrt{(1 + y^2)/[1 + (1 - q^2)y^2]},$$

where y is the (unique) real root of

$$F(jyr_0) = \sqrt{(1 + y^2)[1 + (1 - q^2)y^2]}.$$

Alternatively these two equations may (by varying y) be used to generate r_0's and the corresponding cut-off values of σ. Of course, the two equations are merely a re-statement of the equations $H(\lambda_1, \sigma) = -1$, $\sigma = \lambda_1(1 - q^2)$, heed being paid to the fact that the argument of F is imaginary in the region considered for the radius under discussion.

In the third quadrant for $\sigma < -\sigma_0$, we also find the TE_{11}-limit mode. Its solution curve begins at $\lambda_2 = -1$, $\sigma = -1$, and proceeds to $\sigma = -\infty$ without passing through any easily computed intersections of I curves. Formulae pertaining to the TE_{11} mode in this range are stated in Section 4.22. Again the incipient modes are found in their usual region. For $0 > \sigma > -\sigma_0$, the solution curve corresponding to the TE_{11} mode begins at $\sigma = 0$, $\lambda_2 = -\beta_{iso}(= -\lambda_1)$ and is cut off at the intersection of $H(\lambda_2, \sigma) = -1$ with $\lambda_2(1 - q^2) = \sigma$ (or $(\lambda_1 = 0)_T$). At that point $\beta^2 = 0$, and σ is given by

$$\sigma = -(1 - q^2)\sqrt{(1 - y^2)/[1 - (1 - q^2)y^2]},$$

where y is the least real root of

$$F(r_0 y) = -\sqrt{(1 - y^2)[1 - (1 - q^2)y^2]}.$$

Alternatively, this equation can be used to generate r_0, and the associated σ, if y is regarded as a parameter, which for $u_1 < r_0 < j_1$ is between zero and unity.

At a fixed r_0 the higher roots of the last equation with sign reversed and the corresponding σ are associated with the cut-offs of the incipient

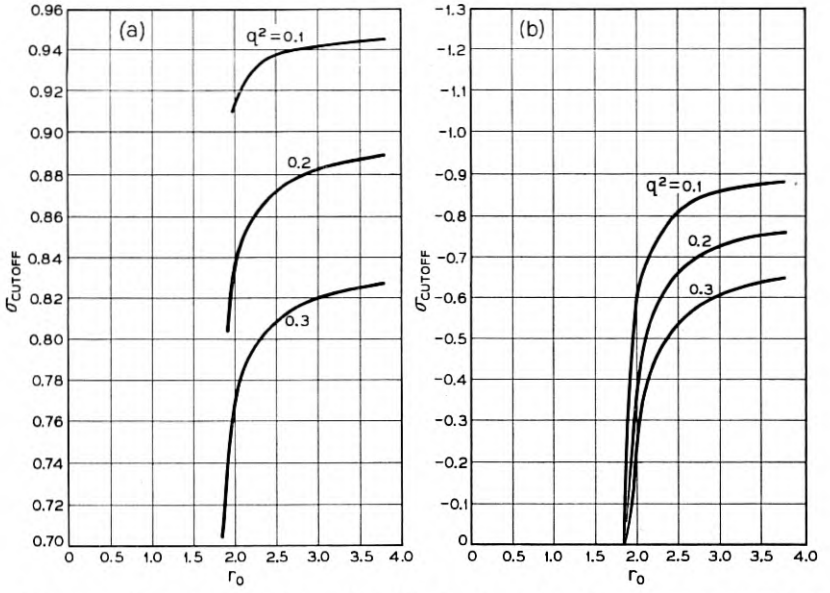


Fig. 18 — Cutoff value of σ for the TE_{11}-limit mode in the plasma as a function of r_0 for various q^2.

modes in the lower half plane. Similarly the real roots of

$$F(r_0 y) = -\sqrt{(1 - y^2)[1 - (1 - q^2)y^2]},$$

and the corresponding

$$\sigma = +(1 - q^2)\sqrt{(1 - y^2)/[1 - (1 - q^2)y^2]},$$

are associated with the cut-offs of the incipient modes in the upper half-plane. These equations have been solved for the TE_{11} mode and their solutions shown in Figs. 18(a) and 18(b).

4.21. *Some formulae relating to the plasma (chiefly for TE-modes).*

The formulas given here employ dimensionless variables (Section 3) except where otherwise stated.

Approximations for extreme values of σ *or* q^2

(a) small σ, q^2 not near unity

TE_{1m} mode:

$$\beta^2 = \beta_m{}^2 + A_m\sigma + B_m\sigma^2 + \cdots \qquad (70)$$

where

$$\beta_m^2 = 1 - \frac{u_m^2}{r_0^2},$$

$$A_m = \frac{2}{u_m^2 - 1}\frac{q^2}{1 - q^2}, \qquad \text{and}$$

$$B_m = -\frac{q^2}{(1 - q^2)^2}\left[1 + \frac{q^2}{u_m^2 - 1}\left\{\left(\frac{1}{2u_m^2} - 1\right)\beta_m^2 + \frac{4}{(u_m^2 - 1)^2}\left(1 - \frac{1}{2u_m^2}\right)\right\}\right].$$

TM modes show no first order variation with σ.

(b) small q^2, σ^2 not near unity. Here β_a and r_a are the actual propagation constant and radius, without scaling factors:

TE_{1m} mode:

$$\beta_a^2 = (1 - q^2)\omega^2\mu_0\epsilon_0 - \frac{u_m^2}{r_a^2} + \omega^2\epsilon_0\mu_0\frac{\sigma q^2}{1 - \sigma^2}\left(\frac{2}{u_m^2 - 1} - \sigma\right) + 0(q^4). \tag{71}$$

TM_{1m} modes:

$$\beta_a^2 = (1 - q^2)\omega^2\mu_0\epsilon_0 - \frac{j_m^2}{r_0^2} - \omega^2\epsilon_0\mu_0\frac{\sigma^2 q^2}{1 - \sigma^2}\left(1 - \frac{j_m^2}{\omega^2\mu_0\epsilon_0 r_a^2}\right). \tag{72}$$

(c) Approximation for large σ; q^2 not near unity.

TE_{1m} mode:

$$\beta^2 = \frac{1}{1 - q^2} - \frac{u_m^2}{r_0^2} - \frac{2q^2}{\sigma(1 - q^2)}\frac{1}{(u_m^2 - 1)}. \tag{73}$$

All formulae in (a), (b), (c) apply to both positive and negative σ (right and left circular waves). Formulae 70 and 73 show that the first order changes in β^2, whether due to very large or very small σ, have coefficients that differ only in sign.

The TE_{11} mode near resonance ($q^2 < 1$)

Near $\sigma = +1$,

$$\beta^2 = -\frac{q^2}{1 - q^2}\frac{\left(\frac{j_1^2}{2r_0^2} - 1\right)}{\sigma - 1} \tag{74}$$

Near $\sigma = -1$,

$$\beta^2 = \frac{-q^2}{1-q^2}\frac{1}{1+\sigma}. \tag{75}$$

Cut-off of the TE_{11} mode ($q^2 < 1$) ($u_1 < r_0 < j_1$)

σ positive:

$$\beta^2 = 0; \qquad \sigma = (1-q^2)\sqrt{(1+y^2/[1+(1-q^2)y^2]}, \tag{76}$$

where y is the only real root or the smallest imaginary root of

$$F(jr_0y) = \sqrt{(1+y^2)[1+(1-q^2)y^2]}.$$

σ negative:

$$\beta^2 = 0; \sigma = -(1-q^2)\sqrt{(1-y^2)/[1-(1-q^2)y^2]}, \tag{77}$$

where y is the smallest real root of

$$F(r_0y) = -\sqrt{(1-y^2)[1-(1-q^2)y^2]}.$$

(See section 4.2 for further explanation).

APPENDIX I. THE F-FUNCTION

The function $F(x)$ has been defined by the equation

$$F(x) = x\frac{J_1'(x)}{J_1(x)}.$$

Using the infinite product for $J_1(x)$ and differentiating logarithmically one finds

$$F(x) = 1 - 2\sum_{n=1}^{\infty}\frac{x^2}{j_n^2 - x^2}, \tag{78}$$

where $J_1(j_n) = 0$. Near one of its poles, j_n, $F(x)$ behaves as $j_n/(x - j_n)$. It is also useful to know the form of $F(x)$ near one of its zeros, u_n, which are also zeros of $J_1'(x)$. Such an expansion may conveniently be found by using the Ricatti equation satisfied by $F(x)$, which is

$$x\frac{dF}{dx} = 1 - x^2 - F^2. \tag{79}$$

The expansion near u_n is then

$$F(u_n + y) = y\left[\frac{1}{u_n} - u_n\right] - \frac{y^2}{2}\left[\frac{1}{u_n^2} + 1\right] + \text{higher terms.} \tag{80}$$

TABLE IV

Equation Number	Asymptote	Polder Transform of Asymptote	Range of Validity
83	$\sigma = -g\left[\lambda + F\left(\frac{r_0}{\sqrt{-g}}\right)\right]$	$\lambda = \frac{1+g}{\sigma} + O\left(\frac{1}{\sigma^2}\right)$	σ large, g finite and not equal to zero or $-r_0^2/j_n^2$
84	$\sigma = \frac{r_0^2}{j_n^2}\left(\lambda - \frac{2}{1 + \frac{j_n^2}{r_0^2} g}\right)$ $n = 1,2,3\cdots$	Not required	σ large, g unrestricted
85	$\sigma = \frac{1}{\lambda} + \frac{g^2}{r_0^2}\lambda$	Not required	Large σ Small λ $g > 0$ for $\lambda < 0$ $g < 0$ for $\lambda > 0$ g finite
86	$\sigma = \left(1 - \frac{r_0^2}{u_n^2}\right)\frac{1}{\lambda} + \frac{2\left(1 + \frac{u_n^2}{r_0^2} g\right)}{1 - u_n^2}\frac{r_0^2}{u_n^3}$ $n = 1,2,3\cdots$	$\sigma = \lambda\frac{r_0^2}{u_n^2} - p - \frac{2\left(1 + \frac{u_n^2}{r_0^2} g\right)}{u_n(1 - u_n^2)}\left(1 - \frac{r_0^2}{u_n^2}\right)$	σ large g finite

The equation may also be used to furnish an expansion near $x = 0$. This is

$$F(x) = 1 - \frac{x^2}{4} - \frac{x^4}{96} + \text{higher terms.} \tag{81}$$

Finally, putting $x = jy$, one finds from Eq. (79) for large y

$$F(jy) = y - \frac{1}{2} + \frac{3}{8}\frac{1}{y} + \text{higher terms.} \tag{82}$$

APPENDIX II. INFORMATION PERTAINING TO THE CONSTRUCTION OF G-DIAGRAMS

The accurate construction of the contours $G = \text{const.}$ is conveniently based on the contours $(1 - \lambda^2)/(1 - \sigma\lambda) = \text{const.}$ along any one of which G is a function of λ alone. These contours are shown in Fig. 4. Their asymptotic properties are almost self-evident.

The curves $G = g = \text{const.}$ have various asymptotes. These, together with their range of validity, and their Polder transforms where needed are stated in Table IV.

The formulas given in Table IV show that the curves $G = \text{const.}$ generally have two kinds of asymptotes; linear and hyperbolic. Formula (83) shows the behavior of G along a line of constant finite slope unequal to r_0^2/j_n^2, the asymptotic slope of the I_n curves. Parallel to a line of slope r_0^2/j_n^2 *all* G contours must be found, not just the restricted range given by the first formula. Writing $\sigma = (r_0^2/j_n^2)\lambda + x$ in the equation $G = g$, and expanding F near its pole j_n, we find x in terms of g and obtain (84) which holds for all g, from $-\infty$ to $+\infty$. When $g = 0$ it also gives the linear asymptotes of 0_n, $0_n'$ curves except 0_0, as is readily verified from the equation

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda)]^2},$$

for the zero curves.

Formula (85) shows how the G-contours tend towards $\sigma\lambda = 1$ from the side $\sigma\lambda > 1$ as $\lambda \to 0$.

Formula (86) relates the asymptotic behavior of the curves $G = g$ to the zero curves, $g = 0$, for small λ. All G curves approach zero curves arbitrarily closely as $\lambda \to 0$. The only exceptions are the infinity curves whose form near $\lambda = 0$ is

$$\sigma = \frac{1}{\lambda}\left(1 - \frac{r_0^2}{j_n^2}\right) \tag{87}$$

When $r_0 = u_{n+1}$, the 0_n curve merges with the $0_n'$ curve at

$$\sigma = \frac{-2}{(u_{n+1}^2 - 1)} \frac{r_0^2}{u_{n+1}^3}.$$

Similarly all of the contours $G = g$ of the sheet to which 0_n belongs merge with the corresponding $G = g$ of the sheet to which $0_n'$ belongs, at

$$\sigma = -\frac{2(1+g)}{(u_{n+1}^2 - 1)} \frac{r_0^2}{u_{n+1}^3}. \tag{88}$$

These remarks apply to all 0_n, $0_n'$ curves, 0_0 included.

The 0_0-curve, for large λ, behaves as

$$\sigma = \frac{1 - r_0^2}{\lambda},$$

and so tends to σ from above for $r_0 < 1$ and from below when $r_0 > 1$. (In fact, for $r_0 < 1$, 0_0 lies wholly in the first quadrant; when $u_1 > r_0 > 1$, 0_0 cuts $\sigma = 0$ once.)

The saddle points of G are most easily found by considering G in the coordinate net formed by the curves $\chi^2 = \text{const.}$ and $\lambda = \text{const.}$ At a saddle point

$$\frac{\partial G}{\partial \chi} = \frac{\partial}{\partial \chi}\left[\frac{1}{\chi^2}\left(\frac{1}{\lambda} F(r_0\chi) - 1\right)\right] = 0$$

and simultaneously

$$\frac{\partial G}{\partial \lambda} = 0$$

The only saddle points that might be missed in this way are points at which the two derivatives are not independent, that is points where the χ^2 contours have vertical tangents, and it is easily verified that no saddle points exist there.

Proceeding with the differentiations, we find that $\frac{\partial G}{\partial \lambda} = 0$ gives

$$F(r_0\chi) = 0$$

or

$$r_0\chi = u_n \tag{89}$$

and so $\frac{\partial G}{\partial \chi} = 0$ gives

$$\frac{2}{\chi} + \frac{r_0}{\lambda} F'(u_n) = 0$$

or

$$\lambda_{ns} = \frac{u_n^2 - 1}{2} \tag{90}$$

The corresponding σ_{ns} are given by

$$\frac{1 - \lambda_{ns}^2}{1 - \sigma_{ns}\lambda_{ns}} = \frac{u_n^2}{r_0^2} \tag{91}$$

and are all positive. Thus all saddle points lie in the first quadrant. At a saddle point $G = -r_0^2/u_n^2$ and therefore it is the intersection of two contours $G = -r_0^2/u_n^2$. For $n > 1$, one of these obeys the asymptotic formula (83), the other is asymptotic to I_n and I_{n-1} (see Fig. 5), and obeys (84), with n and $n - 1$, near those curves. For $n = 1$, one of them still follows formula (83), but the two "arms" of the other are asymptotic to $\sigma = 1/\lambda$ and I_1, and so follow (85) and (84) with $n = 1$ respectively.

Three further facts useful in the construction of G-diagrams are:

Along a curve $\dfrac{1 - \lambda^2}{1 - \sigma\lambda} = \dfrac{u_n^2}{r_0^2}$, G equals $-\dfrac{r_0^2}{u_n^2}$; thus the zero curves of F are contours of constant G.

Along $\lambda = +1$,

$$G = \frac{1 - \sigma}{2} - \frac{r_0^2}{4}. \tag{92}$$

As $\sigma, \lambda \to 1$ along $(\sigma - 1) = \alpha\,(\lambda - 1)$;

$$G \to \frac{\alpha + 1}{2}\left[F\left(r_0\sqrt{\frac{2}{\alpha + 1}}\right) - 1\right] \tag{93}$$

As $\sigma, \lambda \to -1$ along $(\sigma + 1) = \alpha(\lambda + 1)$;

$$G \to -\frac{\alpha + 1}{2}\left[F\left(r_0\sqrt{\frac{2}{\alpha + 1}}\right) + 1\right]$$

As for $G(T(\lambda), \sigma, r_0)$ we have, in addition to the asymptotic formulas in the table:

$$I_B' \text{ transforms into } = 1 + \frac{p}{\sigma + 1}.$$

The intersection of $(I_n)_T$ or $(I_n')_T$ with $\sigma = 0$ is given by

$$\lambda_{1,2} = p \pm \sqrt{1 - \frac{j_n^2}{r_0^2}}.$$

APPENDIX II. THE FIELD COMPONENTS

The field components are given here for the ferrite and for the plasma. They are normalized in such a way that E_z takes a simple form. It should be noted that the λ's appearing in these equations are those defined in Secs. 4.11 and 4.2 for the ferrite and plasma respectively and have a different significance in the two cases.

We write

$$A_1(r) = \frac{J_n(\chi_1 r)}{J_n(\chi_1 r_0)} \qquad A_2(r) = \frac{J_n(\chi_2 r)}{J_n(\chi_2 r_0)}$$

Then, for the ferrite,

$$E_z = [A_1(r) - A_2(r)]e^{jn\theta},$$

$$E_r = -j\beta\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\left\{F_n(\chi_1 r) - \frac{n}{\lambda_1}\right\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\left\{F_n(\chi_2 r) - \frac{n}{\lambda_2}\right\}\right]e^{jn\theta},$$

$$E_\theta = -\beta\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\left\{\frac{F_n(\chi_1 r)}{\lambda_1} - n\right\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\left\{\frac{F_n(\chi_2 r)}{\lambda_2} - n\right\}\right]e^{jn\theta},$$

$$H_z = j\beta\left[\frac{1}{\lambda_1}A_1(r) - \frac{1}{\lambda_2}A_2(r)\right]e^{jn\theta},$$

$$H_r = -\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\left\{n - \frac{1-\chi_1^2}{\lambda_1}F_n(\chi_1 r)\right\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\left\{n - \frac{1-\chi_2^2}{\lambda_2}F_n(\chi_2 r)\right\}\right]e^{jn\theta}, \text{ and}$$

$$H_\theta = -j\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\left\{F_n(\chi_1 r) - \frac{n}{\lambda_1}(1-\chi_1^2)\right\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\left\{F_n(\chi_2 r) - \frac{n}{\lambda_2}(1-\chi_2^2)\right\}\right]e^{jn\theta}$$

and, for the plasma,

$$E_z = [A_1(r) - A_2(r)]e^{jn\theta},$$

$$E_r = -\frac{j}{\beta}\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\{(1-\chi_1^2)F_n(\chi_1 r) + n\lambda_1\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\{(1-\chi_2^2)F_n(\chi_2 r) + n\lambda_2\}\right]e^{jn\theta},$$

$$E_\theta = \left[\frac{A_1(r)}{(r)}\frac{1}{\chi_1^2}\{\lambda_1 F_n(\chi_1 r) + n(1-\chi_1^2)\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\{\lambda_2 F_n(\chi_2 r) + n(1-\chi_2^2)\}\right] e^{jn\theta},$$

$$H_z = -\frac{j}{\beta}[\lambda_1 A_1(r) - \lambda_2 A_2(r)]\, e^{jn\theta},$$

$$H_r = -\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\{\lambda_1 F_n(\chi_1 r) + n\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\{\lambda_2 F_n(\chi_2 r) + n\}\right] e^{jn\theta},$$

and

$$H_\theta = -j\left[\frac{A_1(r)}{r}\frac{1}{\chi_1^2}\{F_n(\chi_1 r) + n\lambda_1\} - \frac{A_2(r)}{r}\frac{1}{\chi_2^2}\{F_n(\chi_2 r) + n\lambda_2\}\right] e^{jn\theta}$$

REFERENCES

1. Goldstein, L., Lampert M. A., and Heney, J. F., Phys. Rev., **82,** p. 956, 1951.
2. Polder, D., Phil. Mag., **40,** p. 99, 1949.
3. Hogan, C. L., B.S.T.J., **31,** p. 1, 1952.
4. Suhl H., and Walker, L. R., Phys. Rev., **86,** p. 122, 1952.
5. Kales, M. L., J. Appl. Phys., **24,** p. 604, 1953.
6. Gamo, H. J., J. Phys. Soc. Japan, **8,** pp. 176–182, 1953.
7. Cook, J. S., R. Compfner and H. Suhl, Letter to the Editor, I.R.E. Proc., to be published.

Coupled Wave Theory and Waveguide Applications

By S. E. MILLER

(Manuscript received February 2, 1954)

Some theory describing the behavior of two coupled waves is presented, and it is shown that this theory applies to coupled transmission lines. A loose-coupling theory, applicable when very little power is transferred between the coupled waves, shows how to taper the coupling distribution to minimize the length of the coupling region. A tight-coupling theory, applicable when the coupling is uniform along the direction of wave propagation, shows that a periodic exchange of energy between coupled waves takes place provided that the attenuation and phase constants (α and β respectively) are both equal, or provided that the phase constants are equal and the difference between the attenuation constants ($\alpha_1 - \alpha_2$) is small compared to the coefficient of coupling c. Either ($\alpha_1 - \alpha_2$)/c or ($\beta_1 - \beta_2$)/c being large compared to unity is sufficient to prevent appreciable energy exchange between the coupled waves. Experimental work has confirmed the theory. Applications include highly efficient pure-mode transducers in multi-mode systems, and frequency-selective filters.

INTRODUCTION

This paper describes some theoretical relations in coupled transmission lines, and the use of coupled lines as circuit elements. In order to illustrate the points of interest in the theoretical material, several applications will be stated first. Detailed discussion of experimental models will be given after the theoretical sections.

The theory of coupled transmission lines may be used to determine many properties of a multi-mode transmission system in which there is distributed coupling between modes. In round pipe, for example, the individual modes of propagation can be considered as separate transmission lines which in the perfect waveguide are completely independent. Geometric imperfections in the waveguide, if distributed over many wavelengths, cause a transfer of power between modes which in general

form is predicted by coupled transmission line theory. As a consequence, analysis of the mode-conversion effects associated with circular-electric-wave transmission in commercial round pipe has been aided materially by applying the coupled-transmission-line concept.[1] In another problem, the transmission of the circular-electric waves through bends,[2] the coupled-wave theory of subsequent sections has also provided valuable insight.

Coupled transmission lines can be employed as circuit elements to exchange power between one mode of a multi-mode line and a designated mode of another transmission line. Consider Fig. 1, which shows a rectangular waveguide having entries 1 and 2 coupled through a series of apertures to a parallel round waveguide having entries 3 and 4. The rectangular guide may be made single mode for convenience, and for the configuration shown may be made to couple to any TE mode of the round guide. Input power at entry 1 may be transferred in whole or in part to the selected mode at entry 4, the remaining portion of the power appearing at entry 2. Very little power in any mode will appear at entry 3 for excitation at 1, and very little power in undesired modes will appear at entry 4. Thus the structure has the hybrid property in addition to being mode selective. A matched impedance is presented at all entries to all modes over a very broad frequency band.

Recently, coupled transmission lines have found use as input and output circuits for travelling-wave tubes. In this instance a helical input (or output) line was electromagnetically coupled to the travelling-wave-tube helix, with conditions adjusted for complete energy transfer between the helices. The result is an input-output circuit requiring no metallic connection to the tube helix and requiring no connection through

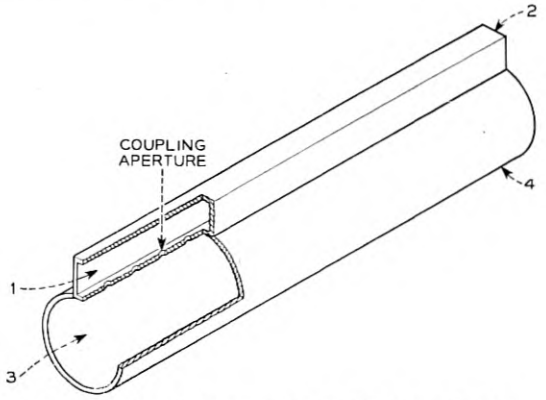


Fig. 1 — Coupled transmission line transducer.

the vacuum seal. R. Kompfner conceived this form of connection to travelling-wave tubes while working with the Admiralty in England, and demonstrated the usefulness of the idea here at the Laboratories. Similar work was done by the group at the Electronics Research Laboratory at Stanford University, and was described by S. T. Kaisel at the August, 1953, West Coast I.R.E. Convention. Both groups requested pre-publication copies of this paper for use in their research.

LOOSE COUPLING THEORY

On the assumption that negligible power is abstracted from the driven line of two coupled transmission lines, the magnitude and mode content of the forward and backward waves in the side line may be written. With reference to Fig. 2, there is assumed coupling between two uniform lines in the interval $-L/2$ to $+L/2$ along the axis of propagation, and no coupling elsewhere. On the basis of loose coupling a normalized voltage wave on line 2 may be written

$$E_2 = 1.0\epsilon^{-i(2\pi/\lambda_2)(x+L/2)}, \tag{1}$$

in which the phase reference is taken as $x = -L/2$. The forward current I_f in the side line at the point $x = L/2$ is

$$I_f = KFM \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x)\epsilon^{i2\pi(1/\lambda_1 - 1/\lambda_2)x}\, dx, \tag{2}$$

where

$$F = \frac{\epsilon^{-i\pi L(1/\lambda_1 + 1/\lambda_2)}}{Z_{10}}.$$

$\phi(x)$ = a coupling function. More precisely, $1/\phi(x)$ is the ratio of the voltage on line 2, $E_2(x)$, to the equivalent voltage generator in series with line 1 at x.

K = fraction of the transferred current which travels in the forward direction.

M = the transfer constant for the various modes which can propagate, relative to the mode for which $\phi(x)$ is defined. The backward current I_b at the point $x = -L/2$ is

$$I_b = (1 - K)FM \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x)\epsilon^{-i2\pi(1/\lambda_1 + 1/\lambda_2)x}\, dx. \tag{3}$$

If the coupling mechanism is non-directive (sending equal waves forward and backward) and has the same value for all modes, then $K = \frac{1}{2}$ and $M = 1.0$. For simplicity these values are assumed in writing the remainder of the expressions. However, the theory is applicable if the coupling mechanism is mode selective and/or directive provided that these properties do not change over the length of the coupling interval.

The mode discriminating property of the coupled lines is the ratio of the forward current for $\lambda_1 = \lambda_2$ to the forward current for $\lambda_1 \neq \lambda_2$. This ratio is

$$\text{Discrimination} = \left| \frac{I_f\,(\lambda_1 = \lambda_2)}{I_f\,(\lambda_1 \neq \lambda_2)} \right| = \frac{\int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x)\,dx}{\int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x) \epsilon^{i\frac{2\theta}{L}x}\,dx}, \tag{4}$$

where $\theta = \pi L(1/\lambda_1 - 1/\lambda_2) = L(\beta_1 - \beta_2)/2$ and the β's are the phase constants of the two transmission lines.

The directivity of the coupling arrangement is defined as the ratio of the forward current for $\lambda_1 = \lambda_2$ to the backward current; this ratio is also given by equation (4) provided

$$\theta = -\pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right) = -\frac{L}{2} (\beta_1 + \beta_2).$$

Thus, in the loose coupling case, the critical performance characteristics are given by the discrimination function, equation (4), for appropriate values of the parameter θ.

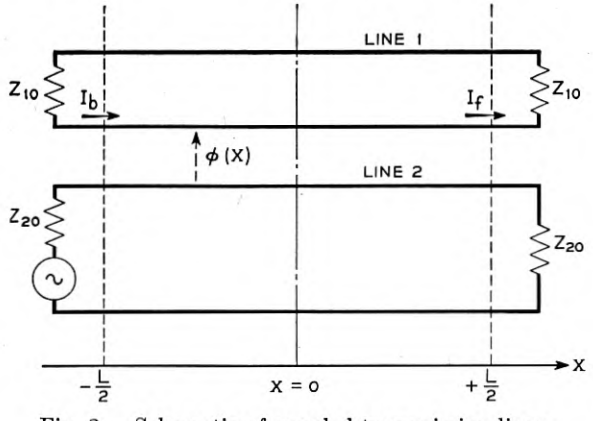


Fig. 2 — Schematic of coupled transmission lines.

A simplified example will illustrate the application of these relations. Suppose the coupling function $\phi(x)$ is constant in the interval $-L/2$ to $L/2$ and zero for other values of x. Then the discrimination function is, from (4)

$$\text{Uniform Coupling Discrimination} = \frac{\theta}{\sin \theta}. \tag{5}$$

Let us further assume, in the hypothetical example, that line 2 (Fig. 2) is a single-mode line having a guide wavelength λ_2 equal to $1.2\lambda_0$, and that line 1 is the three-mode line having guide wavelengths λ_1, λ_2, and λ_3 equal to $1.1\lambda_0$, $1.2\lambda_0$, and $1.3\lambda_0$ respectively. Assume the coupling length L equals $20\lambda_0$. For equal coupling to all modes in a differential unit of length, the relative current waves travelling in the forward direction in the three modes of line 1 are obtained from (4). For the ratio of the λ_2 forward current to the λ_1 forward current,

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.1\lambda_0} - \frac{1}{1.2\lambda_0} \right) = 1.52\pi$$

for which (5) gives a discrimination of about 13.5 db. For the ratio of the λ_2 forward current to the λ_3 forward current

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.2\lambda_0} - \frac{1}{1.3\lambda_0} \right) = 1.28\pi,$$

corresponding to a discrimination of about 14 db. For the ratio of the λ_2 forward current to the λ_2 backward current,

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.2\lambda_0} + \frac{1}{1.2\lambda_0} \right) = 33.3\pi,$$

corresponding to a discrimination of about 43 db. The backward currents in modes λ_1 and λ_3 can similarly be verified to be very small compared to the forward-travelling λ_2 current.

Thus, directivity and mode purity in a simplified case have been shown to be of the desired form.

It may be noted that the denominator of (4) is the Fourier transform of the coupling function $\phi(x)$. Since the numerator of (4) is independent of θ, the discrimination is maximized by minimizing the denominator. An analogous problem exists in the time versus frequency domain relations, and experience with the latter can be used to predict the discriminations to be expected using various coupling distributions.

In the simple example cited above, a length of coupling interval of $20\lambda_0$ yielded a discrimination between the desired versus undesired for-

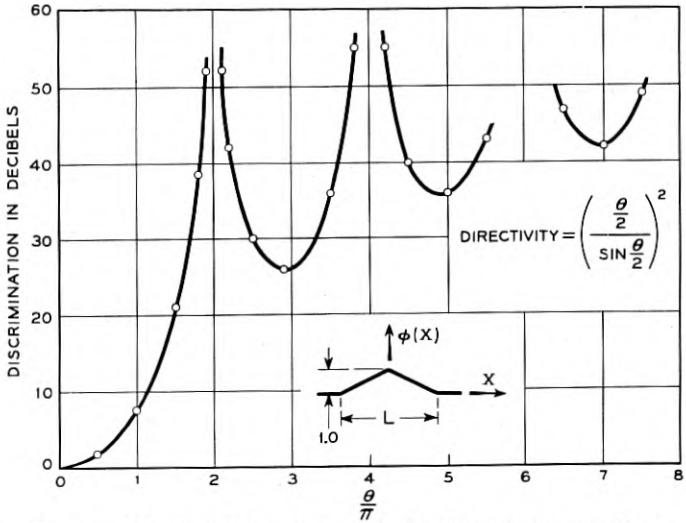


Fig. 3 — Discrimination versus θ/π for linear taper coupling.

ward wave components of about 13 db. How can this discrimination be improved? If the difference between the wavelengths of the desired and undesired wave types is increased, the value of θ is increased and greater discrimination results. In practical cases, however, there frequently is very little that can be done about the wavelength difference because it is inherent in a structure which is fixed by other considerations. By increasing the length of the coupling interval L the value of θ is also increased; in the case of uniform coupling, (5) shows that a value of θ/π equal to about 8 is required to get 30 db discrimination. In the above example this corresponds to L approximately equal to $125\lambda_0$. The latter coupling length is probably impractical, and is certainly inconvenient. The final alternative is to alter the distribution of coupling between the lines, and considerable can be done in this manner.

Suppose a linear taper of the strength of coupling is used, as sketched in Fig. 3. Then the discrimination becomes

$$\text{Linear Taper Discrimination} = \left(\frac{\theta/2}{\sin \theta/2}\right)^2. \tag{6}$$

which is plotted in Fig. 3. The first peak in discrimination occurs at θ/π equals two, compared to a value of θ/π equals one for the first peak using uniform coupling; however, for all values of θ/π greater than about 3, the linear taper provides superior discrimination. This illustrates a general trend; tapering the coupling distribution improves the discrimina-

tion for large θ/π values at the expense of an increased θ/π value for the first discrimination peak.

The first two lines of Fig. 4 give the discrimination functions for two forms of cosine taper; Fig. 5 shows a plot of the first function and Fig. 6 shows a plot of the second function for a particular case. These figures illustrate the importance of the slope at the ends of the coupling distribution. Comparing Fig. 5 with Fig. 3, Fig. 5 has a larger end-slope, shows a lower value of θ/π for the first peak in discrimination, but provides

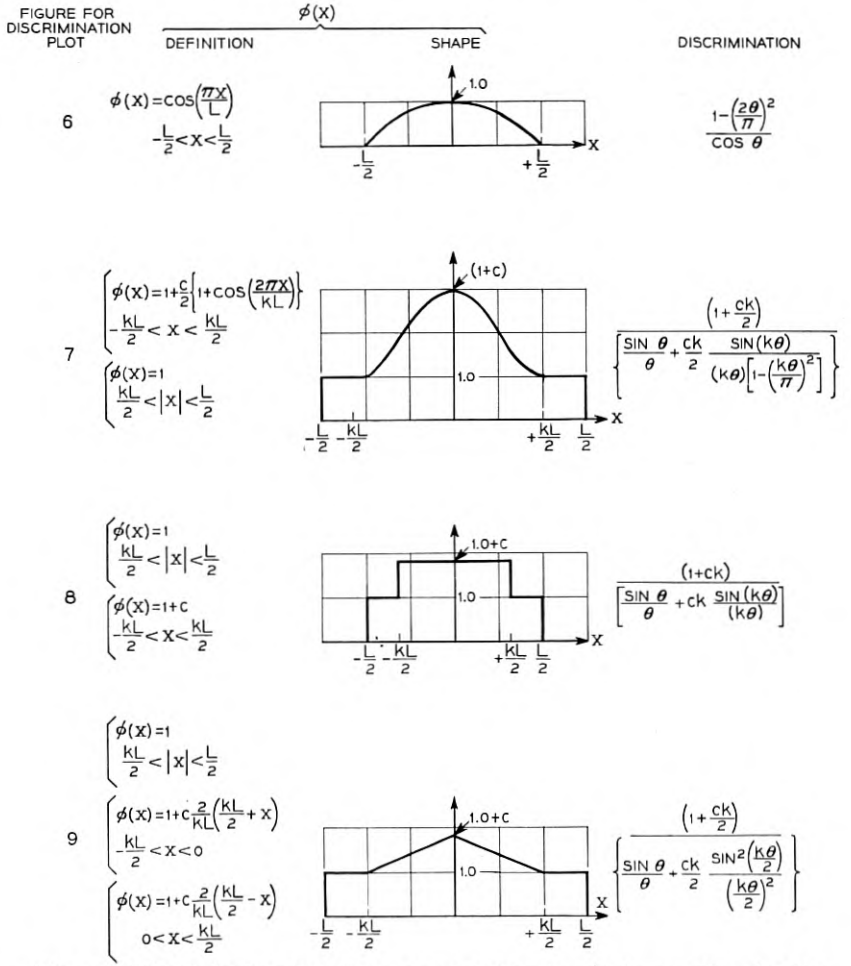


Fig. 4 — Discrimination functions corresponding to certain coupling distributions.

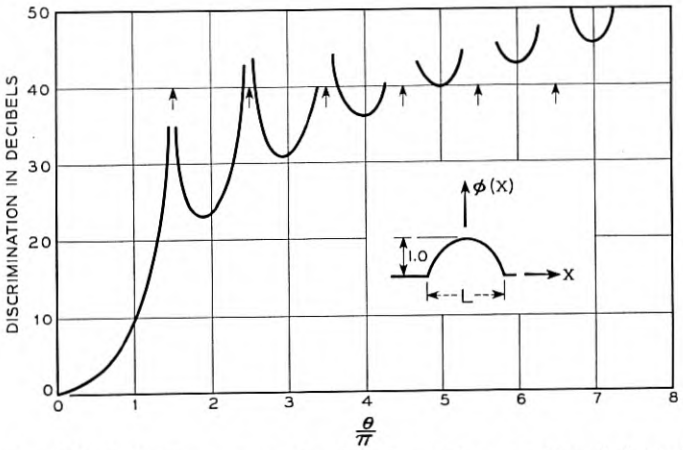


Fig. 5 — Discrimination versus θ/π for the cosine coupling disbribution.

poorer discrimination at values of θ/π slightly above the first peak. In a similar way Fig. 6 shows better discrimination than Fig. 5.

Linear superposition of forward or backward currents may be employed to advantage when designing a coupling distribution. The second line of Fig. 4 gives the discrimination for a coupling function composed of a raised cosine plus uniform coupling. For a value of $c = 22.4$ and $k = 1$,

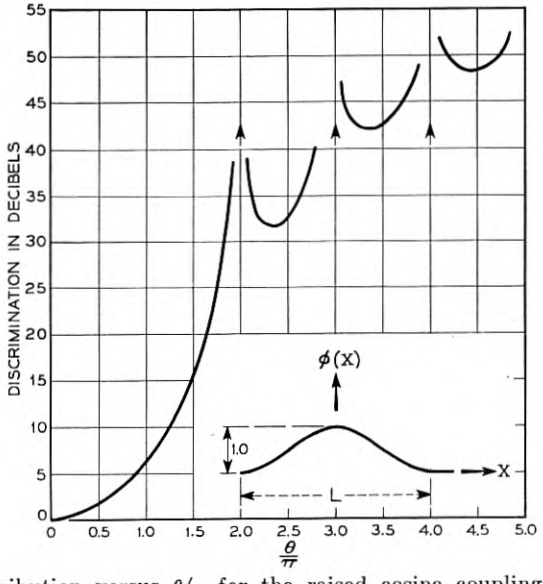


Fig. 6 — Distribution versus θ/π for the raised cosine coupling distribution.

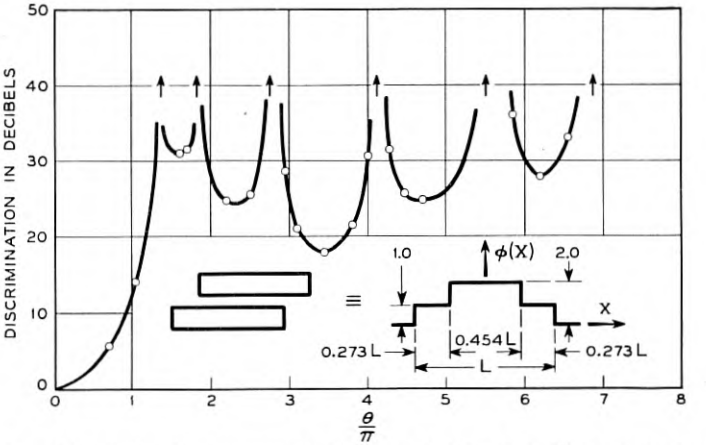


Fig. 7 — Discrimination versus θ/π for two uniform couplings superposed.

the discrimination is greater than 38 db for θ/π between 1.95 and 3.0, and is greater than 50 db for θ/π larger than 3. Below $\theta/\pi = 1.95$ the discrimination is similar to that shown in Fig. 6.

Linear superposition of two uniform coupling distributions yields a structure which is easy to fabricate and, in cases where the requirements are not too complex, may provide satisfactory discrimination. The third line of Fig. 4 gives the general relation, and Fig. 7 shows the discrimination plot for a case of interest. Discriminations on the order of 30 db are available in a broad region between θ/π equal to 1.3 to 2, an attractive abscissa value compared to the $\theta/\pi = 8$ required for simple uniform coupling.

Linear superposition of a linear taper and uniform coupling also yields a structure which is easy to fabricate, and the theoretical discrimination plot for an interesting set of conditions is shown in Fig. 8. High discriminations are provided over greater ranges of θ than for the case of two uniform coupling functions superposed.

The general relations involved in the superposition of coupling functions may be summarized as follows: Let $\phi_1(x), \phi_2(x) \cdots \phi_n(x)$ be known coupling functions and let

$$\phi_T = \phi_1 + \phi_2 + \cdots \phi_n . \tag{7}$$

Let the maximum length of the coupling interval be L. Then, designating the transforms of $\phi_1, \phi_2 \cdots \phi_n$ by F_1 and $F_2 \cdots F_n$ respectively, where

$$F_n = \int_{-L/2}^{L/2} \phi_n(x) \epsilon^{i(2\theta/L)x}\, dx, \tag{8}$$

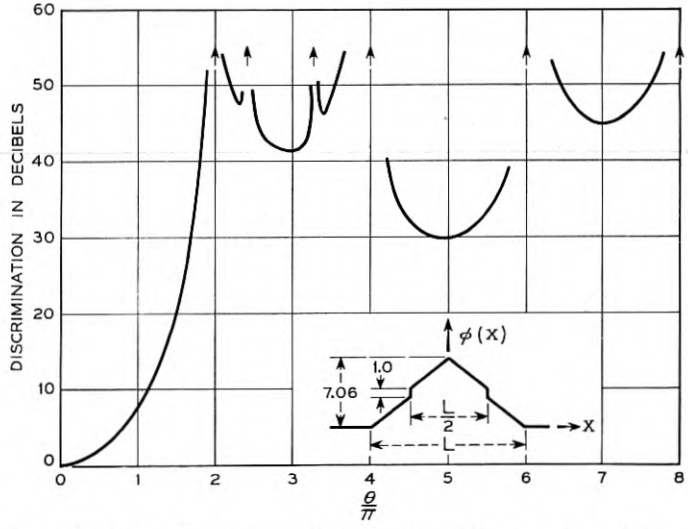


Fig. 8 — Discrimination versus θ/π for a linear taper and uniform coupling superposed.

and letting

$$F_T = F_1 + F_2 + \cdots F_n , \tag{9}$$

the discrimination function for the composite coupling distribution $\phi_T(x)$ is given by

$$\text{Discrimination} = \frac{F_T(\theta = 0)}{F_T} . \tag{10}$$

Another useful theoretical approach to the employment of multiple distributed coupling functions is illustrated in Fig. 9. The top sketch represents any coupling function $\phi_1(x)$. The lower sketch shows a new coupling function $\phi_2(x)$ formed by locating a $\phi_1(x)$ at $\pm d/2$ on the "x" axis. Using F_1 to denote the transform for $\phi_1(x)$, and F_2 to denote the transform corresponding to $\phi_2(x)$,

$$F_2 = 2F_1 \cos \theta', \tag{11}$$

wherein

$$\theta' = d\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)$$

for the forward wave discrimination and

$$\theta' = d\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)$$

for the directivity as defined earlier in connection with (4).

The discrimination function for the composite coupling function $\phi_2(x)$ is

$$\text{Discrimination} = \frac{F_2(\theta = 0, \theta' = 0)}{F_2} = \frac{F_1(\theta = 0)}{F_2}\frac{1}{\cos\theta'}. \quad (12)$$

The factor $1/\cos\theta'$ is the discrimination function associated with two point couplings, and the overall discrimination is the *product* of that discrimination and the discrimination associated with a single distributed coupling function $\phi_1(x)$. This line of thought may be extended to show that use of the same *distributed* coupling function in place of each point coupling in the multi-element distributions described in the following section results in multiplying the discrimination of the multi-element coupling function by the discrimination associated with the distributed coupling function.

In many cases of interest it is either inconvenient or impossible to use absolutely continuous coupling between transmission lines. In the waveguide case illustrated in Fig. 1, for example, a continuous slot cut in the common wall would not provide coupling of the distributed form due to a wave which would oscillate back and forth in the slot itself. We know, however, that the effects of the continuous coupling distribution can be

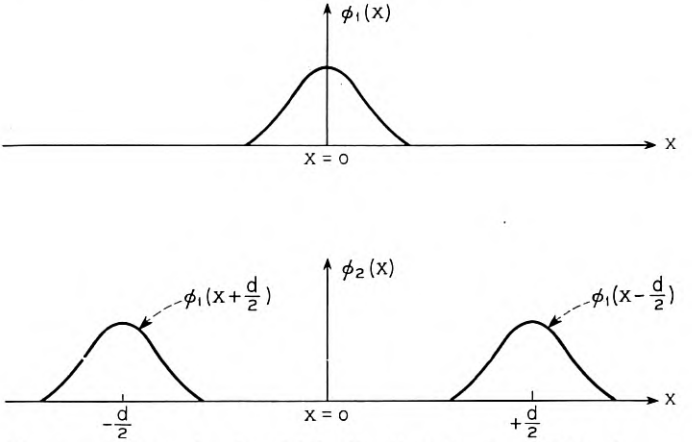


Fig. 9 — Schematic of multiple distributed coupling functions.

simulated closely by using closely spaced point couplings. In order to do this intelligently we need a theory for multi-element point couplings.

The most general symmetrical point coupling distribution for parallel coupled lines is illustrated in Fig. 10. The letters a_0, a_1, $a_2 \cdots a_n$ designate the strength of the couplings, and d_1, d_2, $\cdots L$ represents the spacings between them. The transform for the total coupling distribution is

$$F_T = \int_{-L/2}^{L/2} \phi_T \epsilon^{i(2\theta x/L)}\, dx. \tag{13}$$

$F_T = a_0 + 2a_1 \cos \gamma_1 + 2a_2 \cos \gamma_2 + 2a_3 \cos \gamma_3 + \cdots 2a_n \cos \theta$ in which

$$\gamma_k = \pi d_k \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \text{ or } \pi d_k \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right),$$

and

$$\theta = \pi L \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \text{ or } \pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right),$$

depending on whether forward wave discrimination or directivity is required. The discrimination function is then

$$\text{Discrimination} = \frac{F_T(\gamma_k = 0, \theta = 0)}{F_T}. \tag{14}$$

Let us take as an example the familiar 1–3–3–1 binomial distribution of amplitudes for equally spaced couplings. In the terminology of equation (13), $a_0 = 0$, $a_1 = 3$, $a_2 = 1$, $a_k = 0$ for $k > 2$, $d_1 = L/3$, and $d_2 = L$. Then (14) yields

$$\text{Discrimination} = \frac{8}{6 \cos \theta/3 + 2 \cos \theta} = \frac{1}{\cos^3 \theta/3}. \tag{15}$$

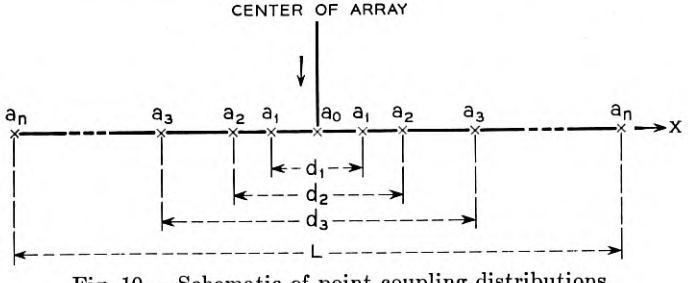


Fig. 10 — Schematic of point coupling distributions.

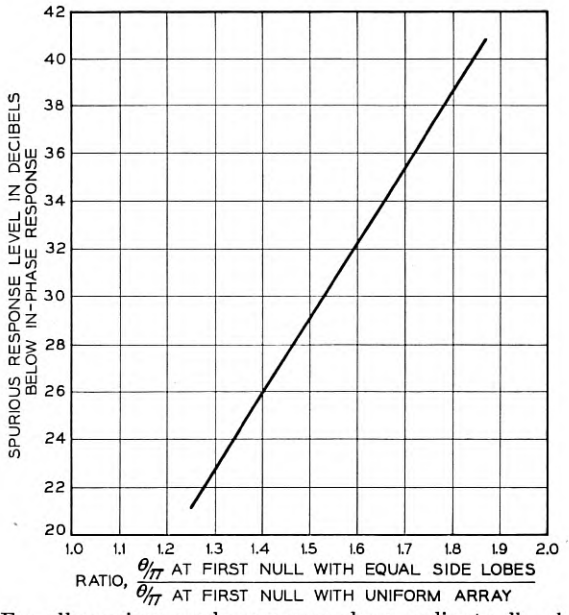


Fig. 11 — For all spurious mode responses down ordinate db, abscissa is the ratio of the required coupling length to the length required for constant amplitude coupling to produce a first null at the same value of $(1/\lambda_1 - 1/\lambda_2)$. (After C. L. Dolph, Reference 4).

which is the relation given by Mumford.[3] The approach is perfectly general, and henceforth the coupling distribution only will be given with the understanding that the corresponding discrimination function can be obtained from (13) and (14).

For the case of tapered amplitudes and an even number of equally spaced couplings, (13) can be simplified to

$$F_T = 2a_1 \cos\left(\frac{\theta}{2n-1}\right) + 2a_2 \cos\left(\frac{3\theta}{2n-1}\right) + \cdots 2a_n \cos 6. \quad (16)$$

This case is of interest because a solution has been worked out for the analogous antenna problem to bring the spurious responses (the peaks of the side lobes in the antenna case, or the peaks of the undesired mode responses in the wave selector case) to the same level relative to the desired response. *This makes the total length of the coupling array a minimum for a given required degree of discrimination.* The solution[4] includes specification of the Tchebysheff distribution of coupling strengths a_1, a_2, $a_3 \cdots a_n$ that are required to achieve various levels of spurious response, and the resulting increase in total array length required to place the first null

in undesired mode response at the same value of

$$\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \text{ or } \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)$$

as for uniform strength couplings equally spaced. Fig. 11 shows the latter relation, a very useful yardstick with which to evaluate the extra coupling length required by less ideal but more easily constructed coupling distributions.

An important practical question is "What is the smallest number of point couplings which will satisfy requirements in a given stiuation?", for it is time-consuming and expensive to fabricate the coupling holes or probes in some circumstances. The large range of possible mode conditions and discrimination requirements makes it difficult to give an answer in closed form, but the general restrictions involved may be stated. In the case of n equally spaced couplings (of any amplitude taper) the discrimination vanishes at $\theta/\pi = (n - 1)$. This is illustrated by the discrimination plot of Fig. 12.

Moreover, it is found that equally spaced couplings produce discriminations which are periodic in θ/π on the interval $(n - 1)$, and which are symmetrical about $\theta/\pi = (n - 1)/2$.

The implication of the discrimination zero at $\theta/\pi = (n - 1)$ is that a large number of point couplings are required to get good directivity and good forward wave discrimination. In the simple case cited above in which $L = 20\lambda_0$, the θ/π value for directivity was shown to be 33.3.

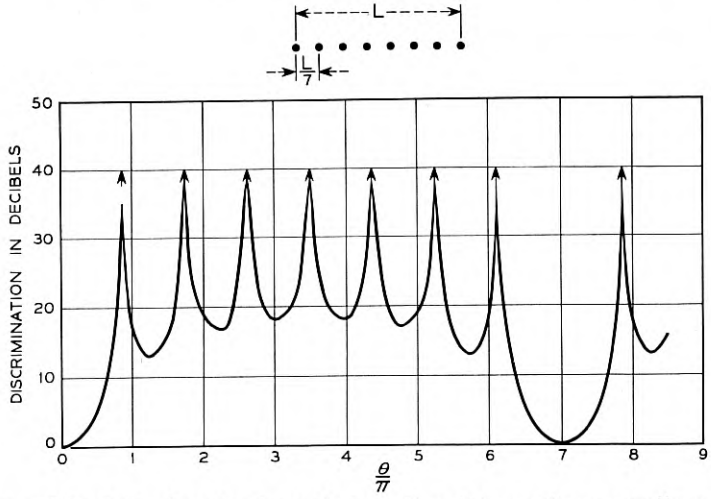


Fig. 12 — Discrimination for 8 equal-strength point couplings equally spaced.

Thus, something on the order of 50 or 60 equally spaced couplings might be needed.

Simulation of continuous coupling functions with equal strength couplings may be carried out as follows: the coupling amplitude versus distance plot may be divided along the distance axis into a number of intervals of equal area, and a point coupling placed at the center of each interval. The more efficient continuous coupling functions require more point couplings to get a good simulation in this manner. For example, the function of line 2, Fig. 4, with $c = 22.4$ and $k = 1$ has been simulated with 12 and 40 equal strength couplings (as described above) and the exact discrimination plotted using (13) and (14). The results are given in Figs. 13 and 14. The original continuous coupling function yields discriminations greater than 38 db for all values of θ/π greater than 2; the 40-point simulation approximates this well in the region of $\theta/\pi = 1.7$ to 4.5, but thereafter begins to fail. The 12-point simulation (Fig. 13) never matches the original but does best in the region of small θ/π.

It is more efficient to seek high discriminations by tapering the strength of equally spaced couplings than by tapering the spacing between equal strength couplings. However, when low discriminations are acceptable, the relative efficiency of tapering the spacing between con-

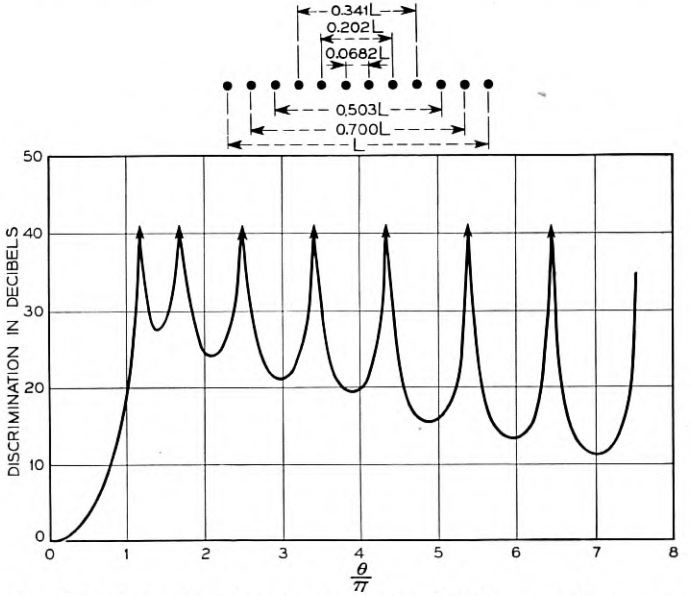


Fig. 13 — Discrimination for 12 equal-strength point couplings arranged to simulate the continuous distribution of Fig. 4, line 2.

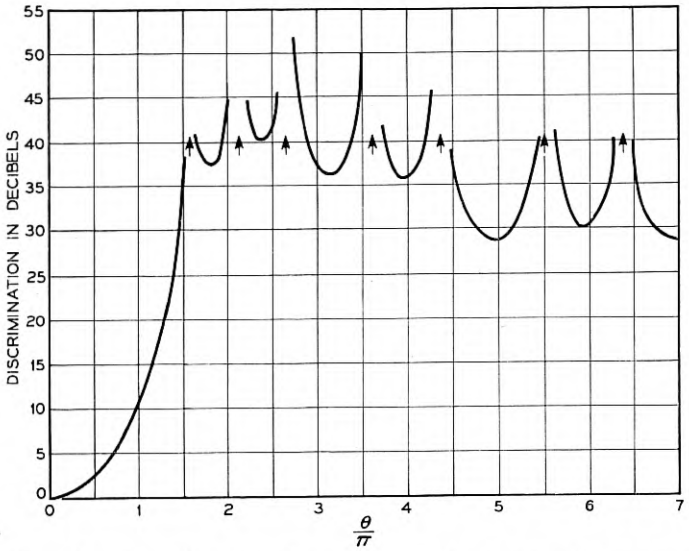


Fig. 14 — Discrimination for 40 equal-strength point couplings arranged to simulate the continuous distribution of Fig. 4, line 2.

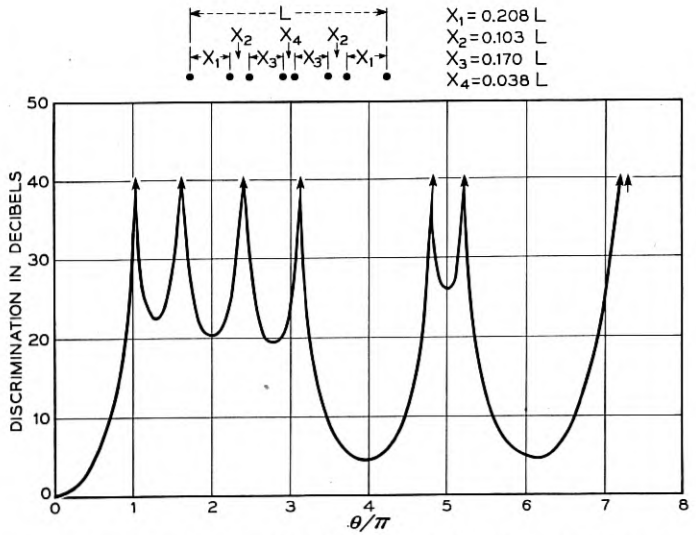


Fig. 15 — Discrimination function for 8 equal strength couplings arranged to maximize the bandwidth (θ/π range) for moderate discrimination.

stant strength couplings is much greater than when high discriminations are required. Fig. 15 shows a distribution which produces about 20 db discrimination from $\theta/\pi = 1$ to 3.25. Eight couplings arranged with the Tchebysheff amplitude taper for 20 db discrimiuation would produce that discrimination from $\theta/\pi = 1.05$ to 5.95.

It is possible to obtain directivity or mode discrimination at smaller θ/π values than made available with uniform coupling. This situation is analogous to the superdirectivity problem in antenna design, with similar results — the lobes of spurious response are increased. In particular, if the coupling near the ends of the third array of Fig. 4 is made larger than the coupling in the center region, making "c" a negative quantity, the first peak in discrimination occurs at θ/π less than one, and the first minimum in discrimination becomes less than 13 db.

By implication, emphasis has been placed on obtaining both mode discrimination and directivity simultaneously. However, by employing a relatively short coupling length it is apparent that the discrimination associated with

$$\theta = \pi L \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right)$$

may be kept small when the directivity associated with

$$\theta = \pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)$$

is in suitable range for good discrimination. Consequently, one can design a directional coupler with little mode discrimination. Conversely, when using a relatively small number of point couplings, the mode discrimination in the forward wave may be good when the directivity is poor.

TIGHT COUPLING THEORY*

We now consider the case in which a significant amount of power is taken from the driven transmission line by the line coupled to it. To simplify the problem the coupling is assumed uniform along the length

* An analysis of coupled transmission lines was given by W. J. Albersheim,[11] and the effects of coupling between waves on certain particular forms of transmission media were analyzed by Meyerhoff[9] and Krasnushkin and Khokhlov.[10] The treatment given here is intended to be more general and is believed to describe the effects of wave coupling under a greater variety of conditions.

axis. The space variation of the wave amplitude may be written

$$\frac{dE_1}{dx} = -(\Gamma_1 + k_{11})E_1 + k_{21}E_2 , \tag{17}$$

and

$$\frac{dE_2}{dx} = k_{12}E_1 - (\Gamma_2 + k_{22})E_2 , \tag{18}$$

in which k_{11} , k_{22} represent the reaction of the coupling mechanism on lines 1 and 2 respectively

k_{21} , k_{12} represent the transfer effects of the coupling mechanism

$\Gamma_{1,2}$ are the uncoupled propagation constants of line 1 and 2 respectively;

$E_{1,2}$ are the complex wave amplitudes on lines 1 and 2,

and are so chosen that $| E_1 |^2$ and $| E_2 |^2$ represent the power carried by lines 1 and 2 respectively at the input or output of the coupling region. The usual transmission-line equations are of this general form, except for second derivatives in place of the first. The first derivatives appear here because we deal only with the forward travelling waves, which the preceding section has shown are the only significant waves when small coupling per wave length is employed. Limiting our interest to the cases for which reciprocity holds and noting that there is always a transverse plane of symmetry midway between the ends of any pair of uniformly coupled lines, we may transform the wave amplitudes to make $k_{12} = k_{21} = k$. We may further simplify the equations without loss of essential generality by submerging the differences $(k_{11} - k)$ and $(k_{22} - k)$ into a modified propagation constant for lines 1 and 2 respectively, yielding

$$\frac{dE_1}{dx} = -(\gamma_1 + k)E_1 + kE_2 , \tag{19}$$

and

$$\frac{dE_2}{dx} = kE_1 - (\gamma_2 + k)E_2 , \tag{20}$$

in which

$$\begin{aligned} \gamma_1 &= \Gamma_1 + k_{11} - k, \qquad \text{and} \\ \gamma_2 &= \Gamma_2 + k_{22} - k. \end{aligned} \tag{20'}$$

For some cases $k_{11} = k_{22} = k$ and for all cases of interest here γ_n differs very little from Γ_n since we are concerned only with loose coupling per wavelength.

The solution, for $E_1 = 1.0$ and $E_2 = 0$ at $x = 0$, is

$$E_1 = \left[\frac{1}{2} - \frac{(\gamma_1 - \gamma_2)}{2\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}}\right] \epsilon^{r_1 x} + \left[\frac{1}{2} + \frac{(\gamma_1 - \gamma_2)}{2\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}}\right] \epsilon^{r_2 x}, \tag{21}$$

and

$$E_2 = \frac{k}{\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \epsilon^{r_1 x} - \frac{k}{\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \epsilon^{r_2 x}, \tag{22}$$

where

$$r_1 = -\tfrac{1}{2}(2k + \gamma_1 + \gamma_2) + \tfrac{1}{2}\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}, \tag{23}$$

$$r_2 = -\tfrac{1}{2}(2k + \gamma_1 + \gamma_2) - \tfrac{1}{2}\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}. \tag{24}$$

The nature of the coupling coefficient k is the first thing to investigate. Assume no dissipation in either the transmission line or in the coupling mechanism. Then it follows that for any value of x,

$$|E_1|^2 + |E_2|^2 = \text{constant} \tag{25}$$

on the basis of energy conservation. It may be determined that (25) leads to the requirement that the coupling constant k be purely imaginary. This is a very important result. In all of the following discussion k is taken to be purely imaginary. Even where dissipation in the transmission lines themselves is important, it is still assumed that the coupling mechanism is non-dissipative.

The simplest case is $\gamma_1 = \gamma_2 = \gamma$, coupling between identical transmission lines. Then (21) and (22) reduce to

$$E_1 = \cos cx \; \epsilon^{-(ic+\gamma)x}, \tag{26}$$

and

$$E_2 = i \sin cx \; \epsilon^{-(ic+\gamma)x}, \tag{27}$$

where $k = ic$. The exponential of (26) and (27) shows that the coupling modifies the average phase constant, and that the attenuation in the driven line (E_1) is the same as in the uncoupled case for cx (coupling length times coupling strength) equal to $n\pi$ radians. The amplitude and phase variations due to the coupling are plotted in Fig. 16. Complete power transfer between lines takes place cyclically, with a period of $cx = \pi$, and with suitable choice of the product cx, an arbitrary division of power between the lines may be selected.

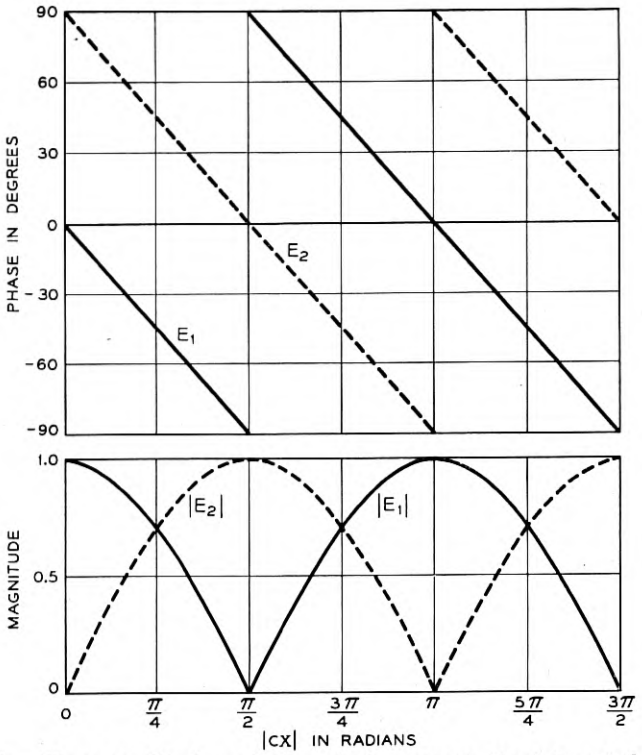


Fig. 16 — Wave amplitude and phase factors versus the integrated coupling strength cx for tightly coupled transmission lines having identical propagation constants.

Let us now assume that the phase constants of the two lines are unequal, but the attenuation constants are the same. Then

$$\alpha_1 = \alpha_2 = \alpha, \qquad \text{and}$$
$$(\gamma_1 - \gamma_2) = i(\beta_1 - \beta_2), \tag{28}$$

and equations (21) and (22) reduce to

$$E_1 = \epsilon^{-[\alpha + i(c + (\beta_1 + \beta_2)/2)]x} E_1^*, \tag{29}$$

where

$$E_1^* = \cos\left[\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}\, cx\right]$$
$$- \frac{i(\beta_1 - \beta_2)}{2c} \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \sin\left[\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}\, cx\right] \tag{30}$$

$$E_2 = \epsilon^{-[\alpha + i(c + (\beta_1 + \beta_2)/2)]x} E_2^* \tag{31}$$

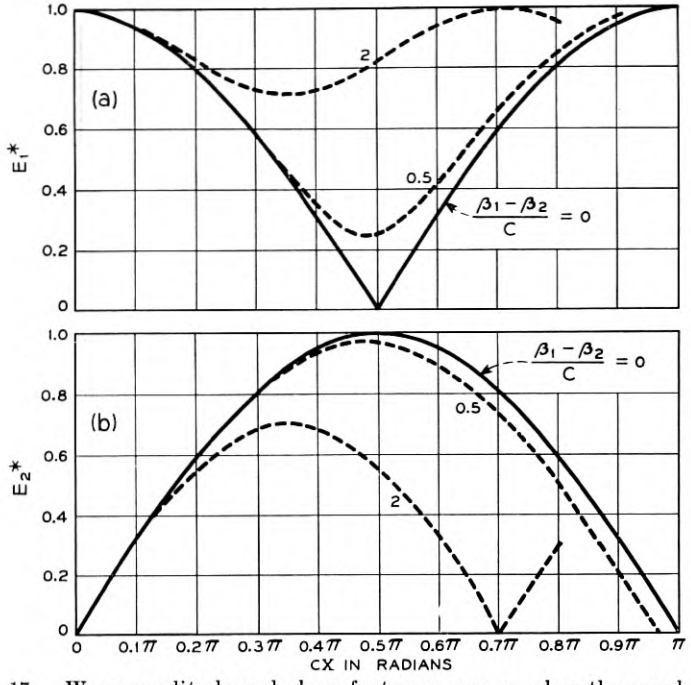


Fig. 17 — Wave amplitude and phase factors versus cx when the coupled lines have equal attenuation constants but unequal phase constants.

where

$$E_2^* = \frac{1 \quad i}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \sin\left[\sqrt{\frac{(\beta_1 + \beta_2)^2}{4c^2} + 1}\, cx\right] \tag{32}$$

The major effects of coupling in this case are represented by E_1^* and E_2^*, which are plotted in Fig. 17 for several values of $(\beta_1 - \beta_2)$. As $(\beta_1 - \beta_2)$ becomes different from zero, the maximum power transferred from the driven line to the undriven line decreases, and the period of the cyclical variation in amplitude is reduced. The latter period is the value of cx given by

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}\, cx = \pi \tag{33}$$

The driven and undriven-line wave amplitudes E_1^* and E_2^* at the maximum power transfer point, namely, at

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}\, cx = \frac{\pi}{2} \tag{34}$$

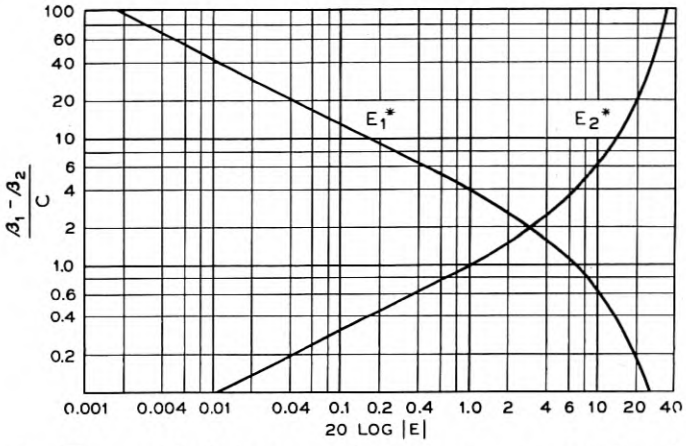


Fig. 18 — Wave amplitude factors at the maximum power transfer value of $\underline{cx}$ versus $(\beta_1 - \beta_2)/c$ when the coupled lines have equal attenuation constants.

are plotted in Fig. 18 as a function of the ratio $(\beta_1 - \beta_2)/c$. It is evident that this maximum energy transfer may be made very small for suitably large values of $(\beta_1 - \beta_2)/c$. The behavior of E_1^* and E_2^* as a function of coupling length x is shown with greater accuracy in the wide amplitude range of interest in Figures 19 and 20 respectively.

Consider now the case in which the coupled lines have identical phase constants, $\beta_1 = \beta_2 = \beta$, and unequal attenuation constants so that $(\gamma_1 - \gamma_2) = (\alpha_1 - \alpha_2)$. Then (21) and (22) reduce to

$$E_1 = \epsilon^{-[\alpha_1 + i(c+\beta)]x} \left\{ \left[\frac{1}{2} - \frac{(\alpha_1 - \alpha_2)}{2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \right] \epsilon^{\left[(\alpha_1 - \alpha_2)/2 + \frac{1}{2}\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}\right]x} + \left[\frac{1}{2} + \frac{(\alpha_1 - \alpha_2)}{2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \right] \epsilon^{\left[(\alpha_1 - \alpha_2)/2 - \frac{1}{2}\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}\right]x} \right\}, \tag{35}$$

$$E_1 = \epsilon^{-[\alpha_1 + i(c+\beta)]x} E_1^{**}, \tag{35'}$$

and

$$E_2 = \epsilon^{-[\alpha_1 + i(c+\beta)]x} \frac{ic}{\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \left\{ \epsilon^{\left[(\alpha_1 - \alpha_2)/2 + \frac{1}{2}\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}\right]x} - \epsilon^{\left[(\alpha_1 - \alpha_2)/2 - \frac{1}{2}\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}\right]x} \right\}, \tag{36}$$

or

$$E_2 = \epsilon^{-[\alpha_1 + i(c+\beta)]x} E_2^{**}. \tag{36'}$$

The amplitude factors E_1^{**} and E_2^{**} have been defined in such a way as to reflect the principal effects of *attenuation difference* in the two lines; for the case in which the driven line attenuation constant α_1 is negligible, note that E_1^{**} and E_2^{**} contain all the amplitude variations of E_1 and E_2 respectively. In general, E_1^{**} and E_2^{**} are the ratios of the wave amplitudes actually present in lines 1 and 2 respectively to the wave amplitude which would exist in line 1 at the same value of x in the absence of coupling.

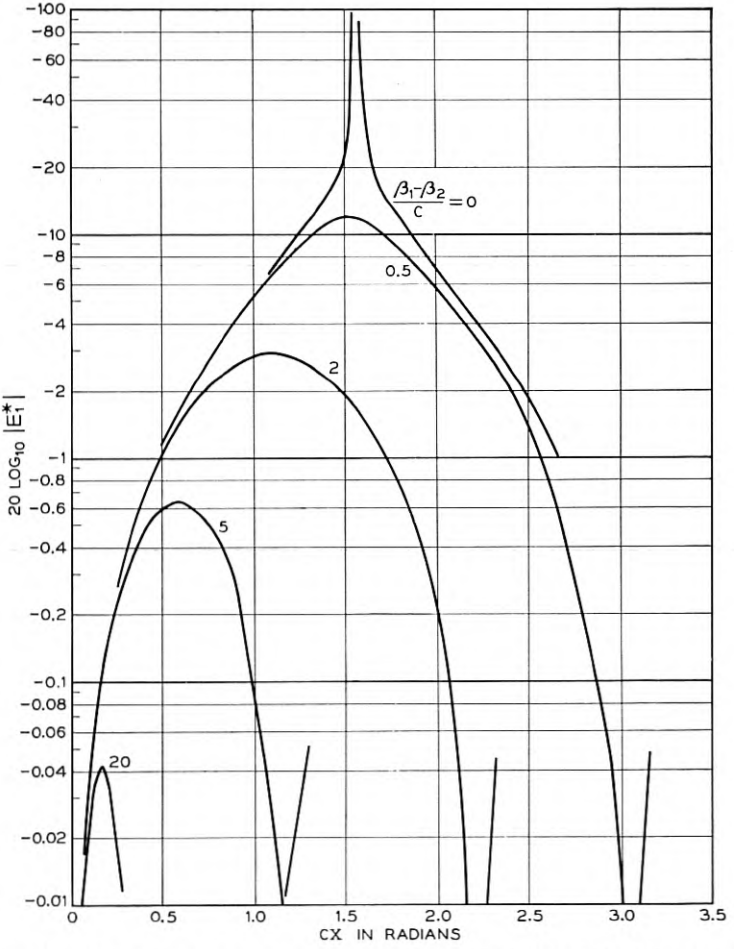


Fig. 19 — Driven line wave amplitude versus cx with unequal phase constants and equal attenuation constants. The curves are periodic for larger values of cx.

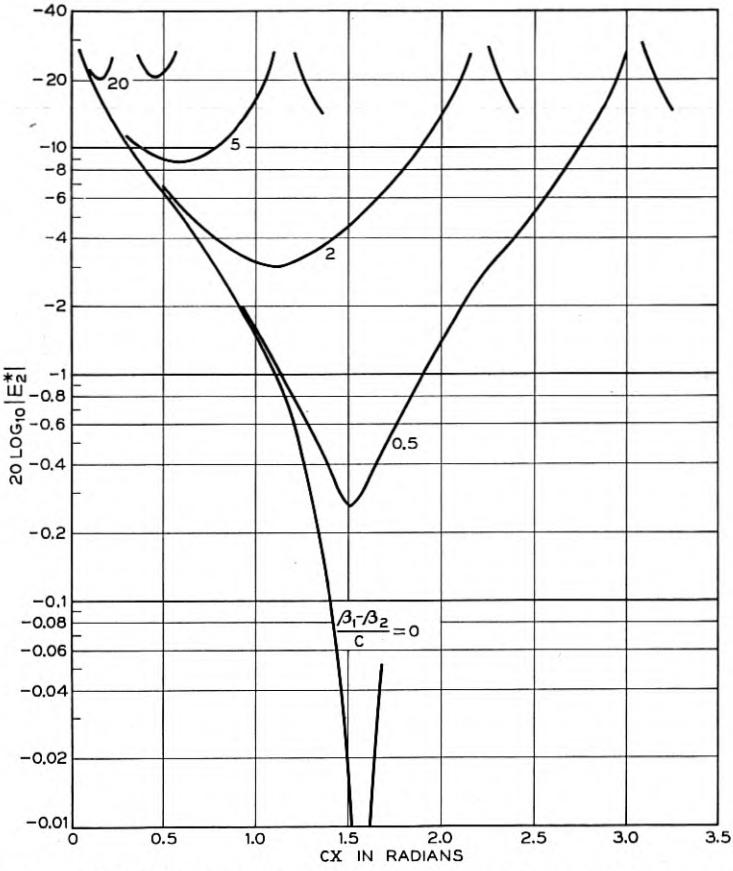


Fig. 20 — Undriven line wave amplitude versus $\underline{cx}$ with unequal phase constants and equal attenuation constants. The curves are periodic for larger values of $\underline{cx}$.

We consider first the case of $(\alpha_1 - \alpha_2)$ negative, i.e., a lower attenuation constant inthe driven line than in the undriven line. The effects of unequal attenuation constants may be illustrated at the integrated coupling strength $cx = \pi/2$ which, as Fig. 16 shows, results in complete transfer of power to the undriven line when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. Fig. 21 shows that the driven line wave amplitude E_1^{**} is very small when $(\alpha_1 - \alpha_2)/c$ is small, but is only ¼ db below unity when $(\alpha_1 - \alpha_2)/c$ is about 55. Fig. 22 illustrates the way the undriven line wave amplitude E_2^{**} decreases as $(\alpha_1 - \alpha_2)/c$ increases.

For integrated coupling strengths less than $\pi/2$, the effects of unequal attenuation constants are not pronounced at small $(\alpha_1 - \alpha_2)/c$, but again for large $(\alpha_1 - \alpha_2)/c$, E_1^{**} approaches unity and E_2^{**} becomes small.

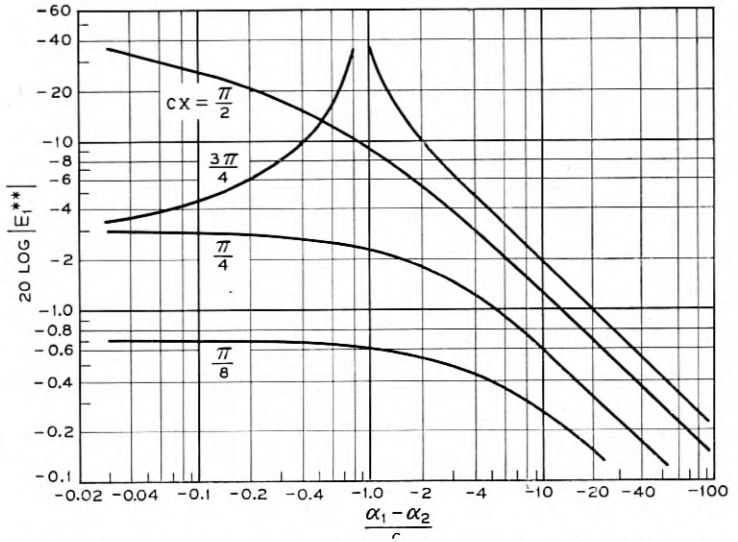


Fig. 21 — The effect of unequal attenuation constants on the driven line wave amplitude for equal phase constants, and $\underline{cx}$ constant. Negative $(\alpha_1 - \alpha_2)$ indicates that the undriven line has the larger attenuation constant.

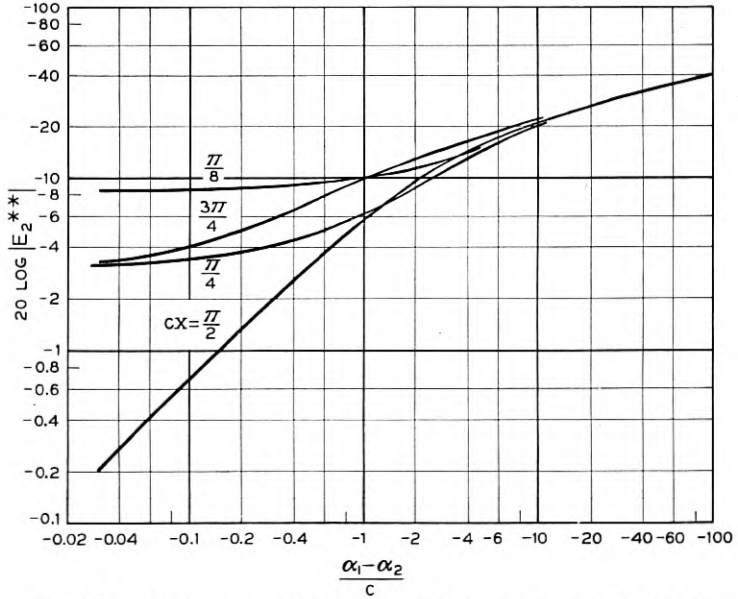


Fig. 22 — The effect of unequal attenuation constants on the undriven line wave amplitude for equal phase constants and $\underline{cx}$ constant.

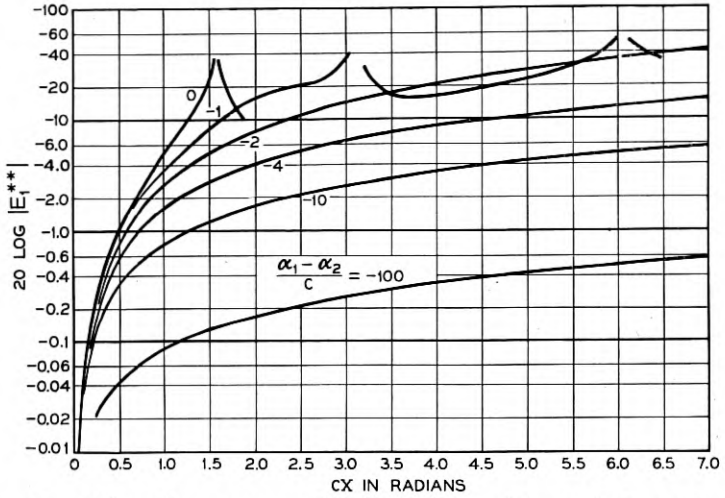


Fig. 23 — Driven line wave amplitude versus cx with equal phase constants and $(\alpha_1 - \alpha_2)/c$ as a parameter. The curve for $(\alpha_1 - \alpha_2)/c = 0$ is periodic.

For integrated coupling strengths greater than $\pi/2$ the effect of small values of $(\alpha_1 - \alpha_2)/c$ is to increase the loss to E_1**, as shown by the curve for $cx = 3\pi/4$ in Fig. 21. However, for sufficiently large values of $(\alpha_1 - \alpha_2)/c$ the loss to E_1** is made small.

The variation in E_1** and E_2** as a function of coupling strength (cx) is given in Figs. 23 and 24. The periodicity of E_1** is removed for

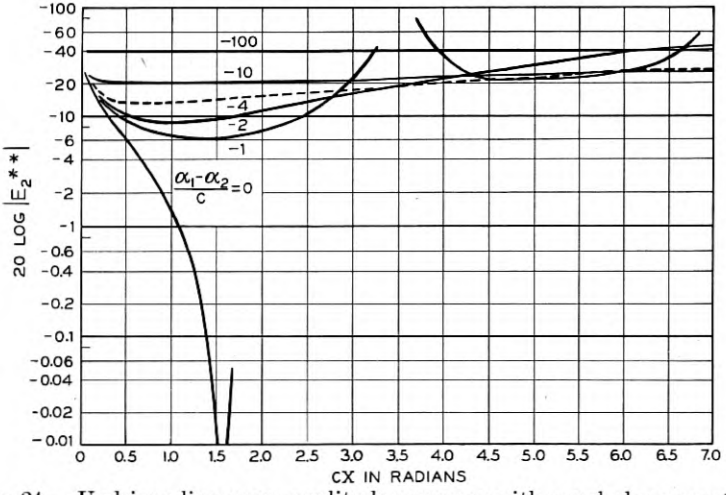


Fig. 24 — Undriven line wave amplitude versus cx with equal phase constants and $(\alpha_1 - \alpha_2)/c$ as a parameter. The curve for $(\alpha_1 - \alpha_2)/c = 0$ is periodic.

$(\alpha_1 - \alpha_2)/c$ as small as -1, but a value as large as -10 or more is required in order to reduce the loss to E_1^{**} to a moderate value for large integrated coupling ($\underline{c}\underline{x}$) values.

When $(\alpha_1 - \alpha_2)$ is positive, the attenuation constant for the undriven line is less than that for the driven line, and under these circumstances E_1^{**} can exceed unity. Physically this means that the power loss line is carrying the energy for a distance and returning it to the driven line at a more distant point. The curves of Fig. 25 and Fig. 26 show the variation of E_1^{**} and E_2^{**} versus positive $(\alpha_1 - \alpha_2)/c$ values, at fixed values of integrated coupling strength $\underline{c}\underline{x}$. For $\underline{c}\underline{x}$ equal to $\pi/4$, the driven line wave magnitude E_1^{**} decreases as the ratio $(\alpha_1 - \alpha_2)/c$ assumes small positive values and goes through a balanced type of null near $(\alpha_1 - \alpha_2)/c = 3.5$ (see Fig. 25). Again this is the resultant of the lower loss undriven wave carrying power for a distance and returning it to the driven wave in the proper phase to cause cancellation of the straight-through component of the driven wave. For $\underline{c}\underline{x}$ between $\pi/4$ and $\pi/2$ the null would move from $(\alpha_1 - \alpha_2)/c$ near 3.5 toward $(\alpha_1 - \alpha_2)/c = 0$.

Figures 27 and 28 show the variation of E_1^{**} and E_2^{**} versus the integrated coupling strength $\underline{c}\underline{x}$ at fixed values of $(\alpha_1 - \alpha_2)/c$. In these

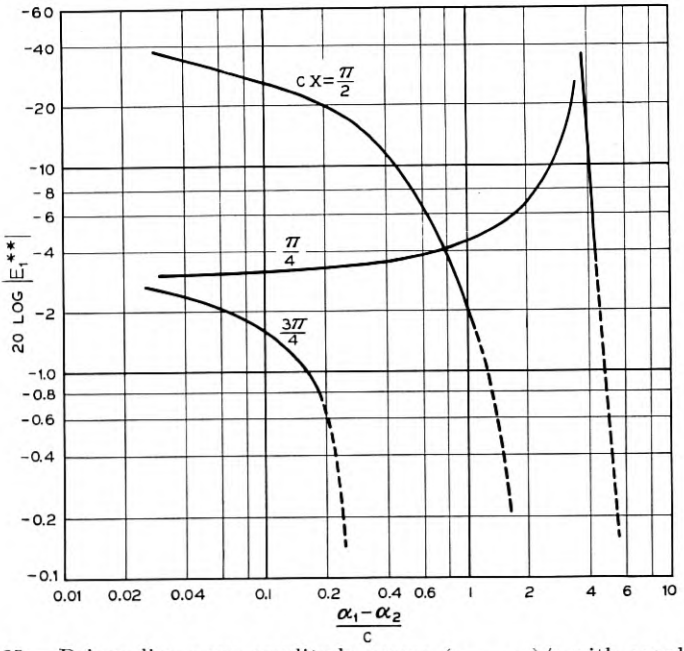


Fig. 25 — Driven line wave amplitude versus $(\alpha_1 - \alpha_2)/c$ with equal phase constants and $\underline{c}\underline{x}$ constant. Positive $(\alpha_1 - \alpha_2)$ indicates the undriven line has the smaller attenuation constant.

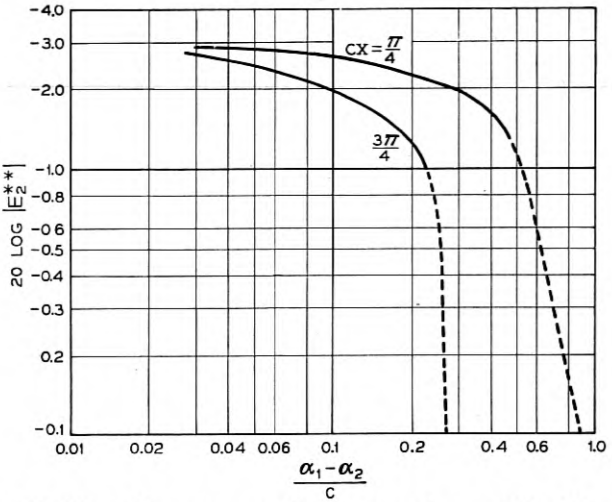


Fig. 26 — Undriven line wave amplitude versus $(\alpha_1 - \alpha_2)/c$ with equal phase constants and $\underline{cx}$ constant.

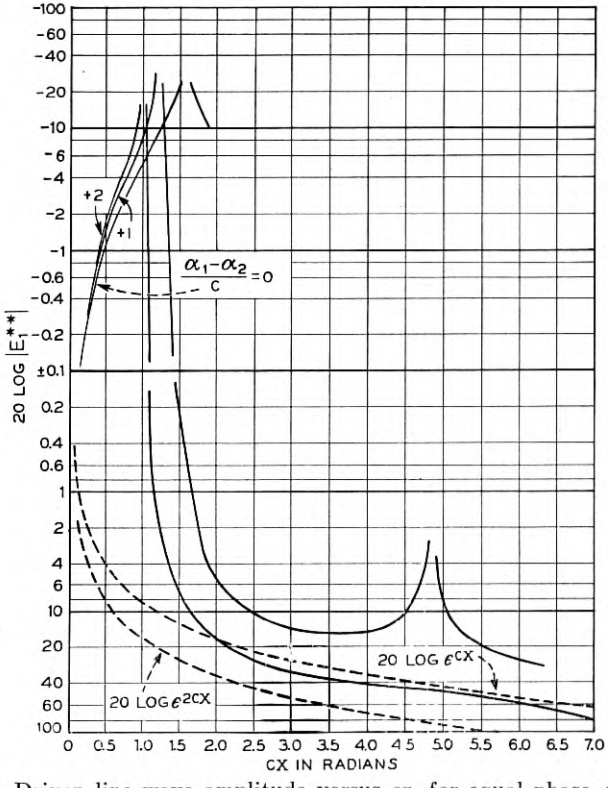


Fig. 27 — Driven line wave amplitude versus $\underline{cx}$, for equal phase constants; $(\alpha_1 - \alpha_2)/c$ as a parameter.

figures a double logarithmic scale is used on the ordinate to represent amplitude variations from 50 db below unity to amplitudes 50 db above unity. An arbitrary break in the scale has been made at ±0.1 db which for practical purposes will be assumed to correspond to amplitudes of unity. With reference to Figure 27, small positive values of $(\alpha_1 - \alpha_2)/c$ move the first null in E_1^{**} from $cx = \pi/2$ toward lower values of $c\underline{x}$. For abscissa values greater than $\pi/2$, E_1^{**} exceeds unity. For $(\alpha_1 - \alpha_2)/c = 1$, E_1^{**} again has a minimum in the vicinity of $cx = 3\pi/2$ but this second null has disappeared for $(\alpha_1 - \alpha_2)/c = +2$ and presumably also for larger positive values. With reference to Figure 28, E_2^{**} grows at a more rapid rate as a function of $c\underline{x}$ when $(\alpha_1 - \alpha_2)/c$ takes on positive values. The null in the vicinity of $cx = \pi$ is still present for $(\alpha_1 - \alpha_2)/c = 1$ but has disappeared at $(\alpha_1 - \alpha_2)/c = 2$. For $(\alpha_1 - \alpha_2)/c$ equal to +2 (and presumably for larger positive values) the undriven wave amplitude E_2^{**} is greater than E_1^{**} for $c\underline{x}$ larger than about 0.5.

The question comes to mind in connection with this case in which the

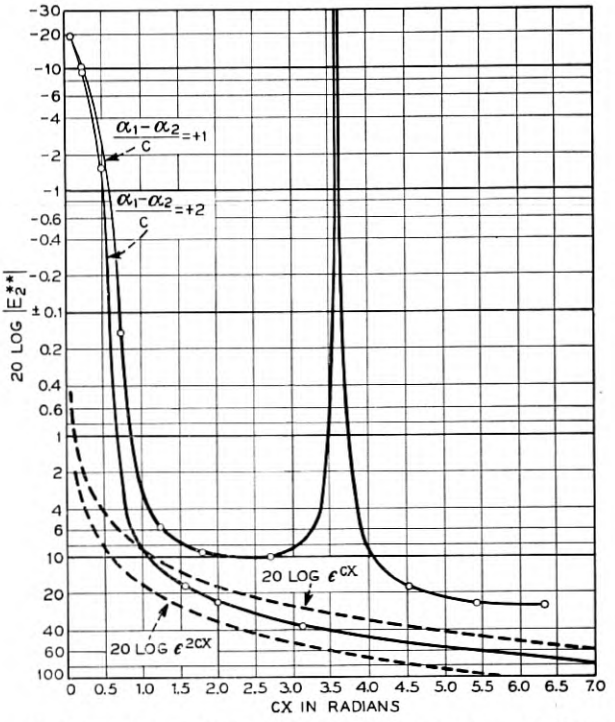


Fig. 28 — Undriven line wave amplitude versus $c\underline{x}$, for equal phase constants; $(\alpha_1 - \alpha_2)/c$ as a parameter.

undriven wave has a smaller attenuation coefficient than the driven wave, "How much less is the undriven line wave amplitude than would have existed at the same value of x if the same incident wave had been launched in the lower loss line and in the absence of coupling to the higher loss line?" This amplitude difference for the condition $(\alpha_1 - \alpha_2)/c = 1$ is represented in Fig. 28 by the *difference* between the curve for E_2** and the curve labeled 20 log e^{cx}. Similarly, for the condition $(\alpha_1 - \alpha_2)/c = 2$, this amplitude difference is represented by the difference between the curve for E_2** and the curve labeled 20 log e^{2cx}.

The general case of $\gamma_1 \neq \gamma_2$ is important both in interpreting undesired mode coupling effects in multi-mode systems as well as in evaluating

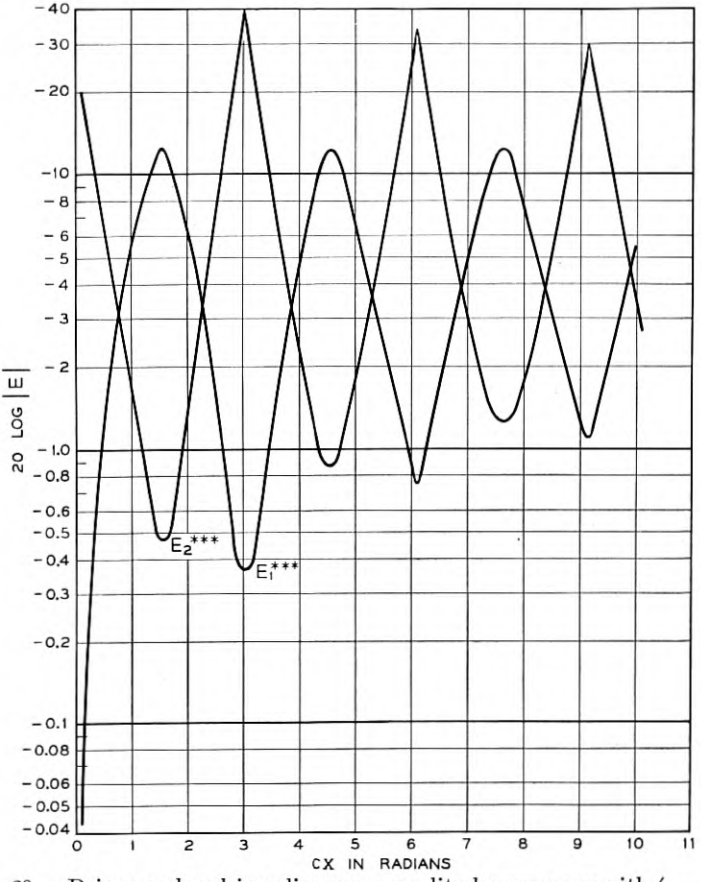


Fig. 29 — Driven and undriven line wave amplitudes versus cx with $(\alpha_1 - \alpha_2)/c = 0.03$ and $(\beta_1 - \beta_2)/c = 0.5$.

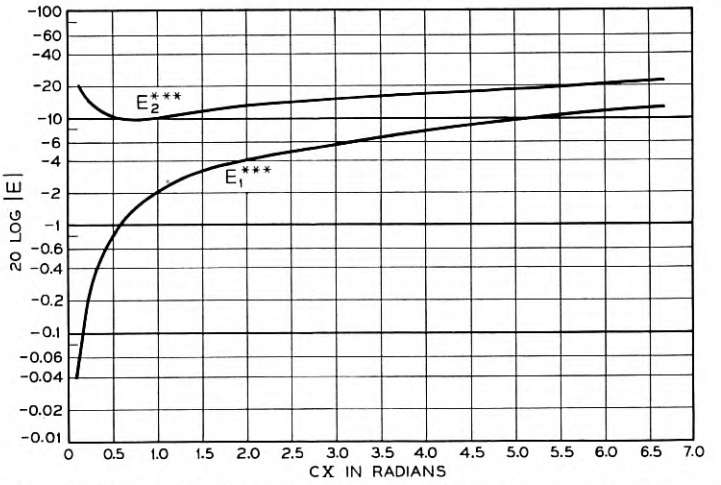


Fig. 30 — Driven and undriven line wave amplitudes versus $\underline{cx}$ with $(\alpha_1 - \alpha_2)/c = -2$ and $(\beta_1 - \beta_2)/c = 2$.

errors in construction of devices intended to produce $\gamma_1 = \gamma_2$. To facilitate discussion of this case we define

$$E_1 = E_1^{***}\epsilon^{-[\alpha_1 + i(c+(\beta_1+\beta_2)/2)]x}, \tag{37}$$

and

$$E_2 = E_2^{***}\epsilon^{-[\alpha_1 + i(c+(\beta_1+\beta_2)/2)]x}. \tag{38}$$

where E_1 and E_2 are defined by (21) through (24). The relation between E_1^{***} and E_1 (or E_2^{***} and E_2) is the same as described in connection with (35) and (36).

Small deviations from $\gamma_1 = \gamma_2$ are represented in Fig. 29, which shows E_1^{***} and E_2^{***} versus $\underline{cx}$ for $(\alpha_1 - \alpha_2)/c = -0.03$ and $(\beta_1 - \beta_2)/c = 0.5$. At $\underline{cx} = \pi/2$ radians, the first complete power transfer point in the $\gamma_1 = \gamma_2$ case, the above values correspond to a phase difference $(\beta_1 - \beta_2)x = \pi/4$ or 45°, and an attenuation difference $(\alpha_1 - \alpha_2)x = 0.03\ \pi/2$ or 0.047 nepers (0.41 db) for the path length of the coupling distance. In the absence of the dissipation difference, but for the same difference in phase constants, Fig. 20 shows that E_2^* reaches a maximum at −0.26 db near $\underline{cx} = \pi/2$, whereas the value including the dissipation difference (Fig. 32) is −0.46 db. The latter two values differ by 0.2 db or one-half of $(\alpha_1 - \alpha_2)x$; when $(\alpha_1 - \alpha_2)/c$ is small compared to unity, this is a general result.

More sizeable deviations from $\gamma_1 = \gamma_2$ are represented in Fig. 30, which shows E_1^{***} and E_2^{***} versus $\underline{cx}$ for $(\alpha_1 - \alpha_2)/c = -2$ and $(\beta_1 - \beta_2)/c =$

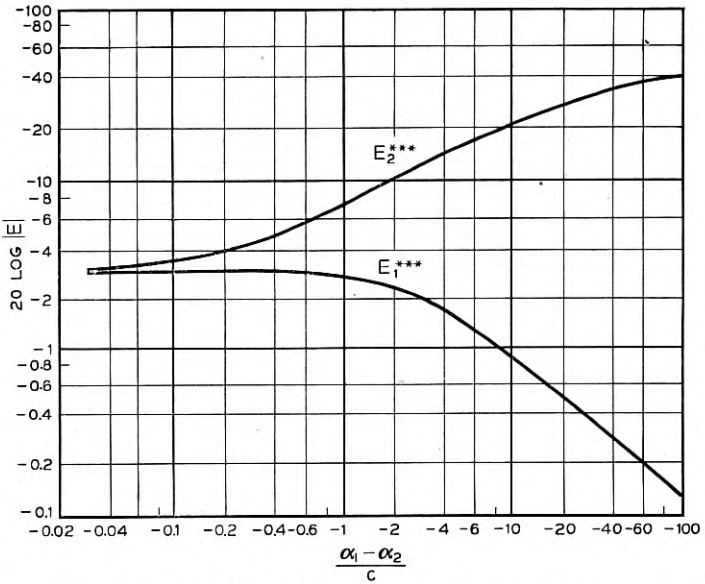


Fig. 31 — Driven and undriven line wave amplitudes versus $(\alpha_1 - \alpha_2)/c$ for $cx = \pi/2\sqrt{2}$ and $(\beta_1 - \beta_2)/c = 2$.

2. At $cx = \pi/2$, the phase difference is therefore π radians and the attenuation difference π nepers. The result is appreciable attenuation for E_1^{***} and only a moderate ratio of E_1^{***}/E_2^{***}.

Fig. 31 shows the way dissipation differences counteract the coupling forces when there is a phase constant difference $(\beta_1 - \beta_2)/c = 2$. This may be compared with Fig. 21 which represents the case of $(\beta_1 - \beta_2) = 0$. Very little change in E_1^{***} occurs until $(\alpha_1 - \alpha_2)/c$ exceeds $(\beta_1 - \beta_2)/c$; this is again a general result.

Finally, we may inquire as to how much power is dissipated in the system when attenuation constant differences are utilized to mitigate the effects of coupling. A measure of the power preserved is

$$| E_1^{***} |^2 + | E_2^{***} |^2$$

and this quantity is plotted in Fig. 32 for cases previously discussed in connection with Figs. 21 and 31. Either in the absence or presence of a phase constant difference, the attenuation constant difference shows a maximum effect in reducing the available power at $(\alpha_1 - \alpha_2)/c = 2$. This is probably a general result brought on by the factor

$$\sqrt{(\gamma_1 - \gamma_2)^2 - 4c^2}$$

found in the exponent of terms describing E_1 and E_2.

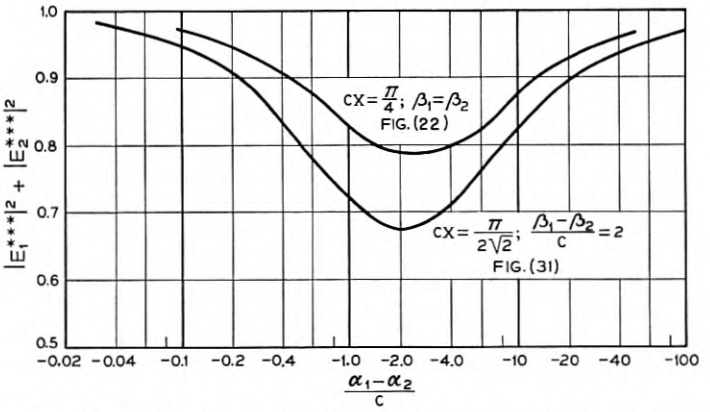


Fig. 32 — Available power versus $(\alpha_1 - \alpha_2)/c$ for several cases of interest.

TIGHT COUPLING EFFECTS OF MULTIPLE DISCRETE COUPLINGS

In practice it is convenient under some conditions to produce the desired coupling between transmission lines using multiple discrete couplings. It is then of interest to know the relation between the total power transferred and the number and strength of the individual couplings. It is the purpose of this section to state these relations.

We assume two transmission lines having identical propagation constants, with coupling units located at intervals along the lines as shown schematically in Fig. 33. A coupling unit may be a single point coupling, or an array of point couplings, but is always assumed to have the property of low reflection in the driven line and low back-wave transmission in the undriven line. If there are

n_1 couplings of magnitude α_1,

n_2 couplings of magnitude α_2,

and

n_k couplings of magnitude α_k

located along the lines in any order whatsoever, the wave amplitudes in

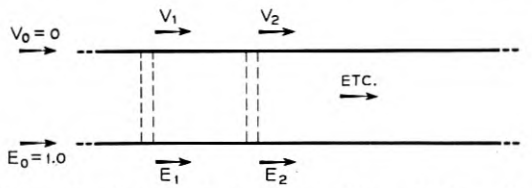


Fig. 33 — Schematic of transmission lines with multiple point couplings.

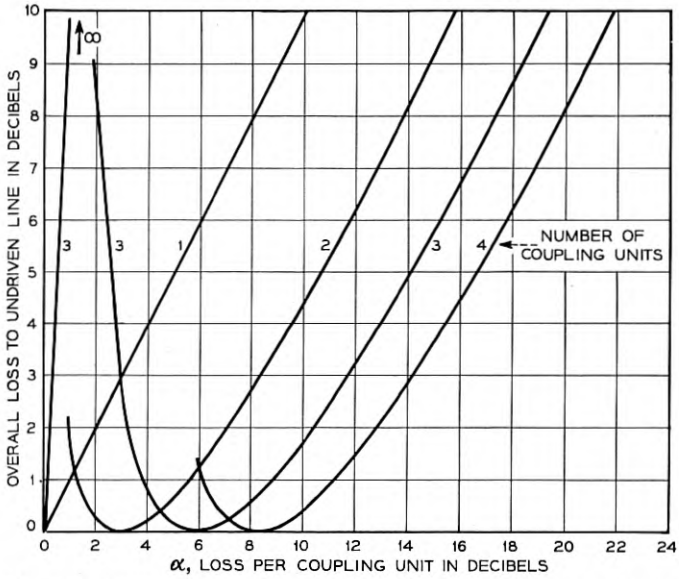


Fig. 34 — Overall loss to the undriven line versus loss per coupling unit, with the number of coupling units as a parameter.

the driven and undriven lines respectively are

$$E = \cos [n_1 \sin^{-1} a_1 + n_2 \sin^{-1} \alpha_2 + \cdots n_k \sin^{-1} \alpha_k], \qquad (39)$$

and

$$V = \sin [n_1 \sin^{-1} \alpha_1 + n_2 \sin^{-1} \alpha_2 + \cdots n_k \sin^{-1} \alpha_k]. \qquad (40)$$

These are amplitude factors due to coupling, and the normal attenuation effects in the uncoupled lines must be added separately. For complete power transfer we set the bracketed quantity of (39) and (40) equal to $\pi/2$, which gives the desired information about number and strength of point couplings. Other transfer losses may similarly be prescribed or determined.

For multiple coupling units of the same coupling strength, Fig. 34 shows the overall transfer loss to the undriven line versus loss per coupling units as a parameter. The shape of these curves from the complete transfer point toward higher losses is very nearly the same. Fig. 35 shows the loss per coupling unit versus number of coupling units, with overall transfer loss to the undriven line as a parameter.

SOME RESULTS OF EXPERIMENTS IN DOMINANT-MODE WAVEGUIDE

In a previous paper on dominant-mode waveguide directional couplers,[5] complete power transfer between dominant-mode rectangular

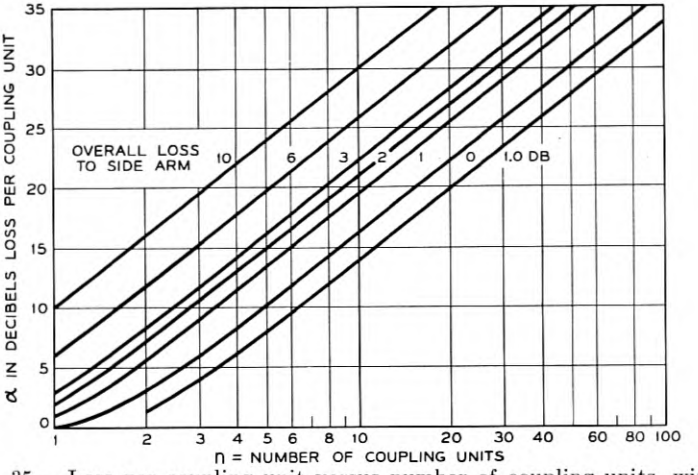


Fig. 35 — Loss per coupling unit versus number of coupling units, with the desired transfer loss as a parameter.

waveguides was shown to be possible in a coupling interval two wavelengths long, and very broad band directivity characteristics of a shape prescribed to meet given requirements were shown to be achievable.

The following paragraphs report on experiments which have been carried out with the objective of developing other useful devices and with the ancillary aim of verifying other predictions of the theory.

Experimental work was done to verify the cyclical nature of energy transfer between coupled lines, to determine the magnitude of losses which accompany such transfer in the waveguide case, and to determine desirable coupling distribution shapes in the tight coupling case. These experiments were carried out by R. W. Dawson in the 3.1 to 3.5 cm band using the 0.4″ x 0.9″ I.D. jig shown in Fig. 36, consisting of two wave-

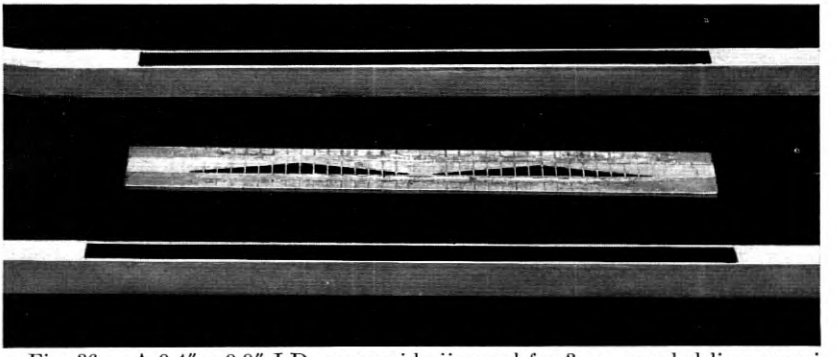


Fig. 36 — A 0.4″ x 0.9″ I.D. waveguide jig used for 3 cm coupled line experiments. The long waveguides on one side of the coupling insert were required to accommodate low-reflection terminations for directivity measurements.

guides having one wall cut away to accept a coupling insert. In one set of observations, the insertion loss in the driven-line and the transfer loss to the undriven line were recorded for a variable number of No. 22 copper wires dividing a coupling aperture $11\frac{1}{4}''$ long and linearly tapered from 0.030″ height at the ends to 0.33″ height at the center. The results are recorded in Fig. 37. At 102 holes, negligible power was abstracted from the driven line, and the transfer loss to the undriven line forward wave $|E_2|$ was about 18 db. Note that more coupling was observed at

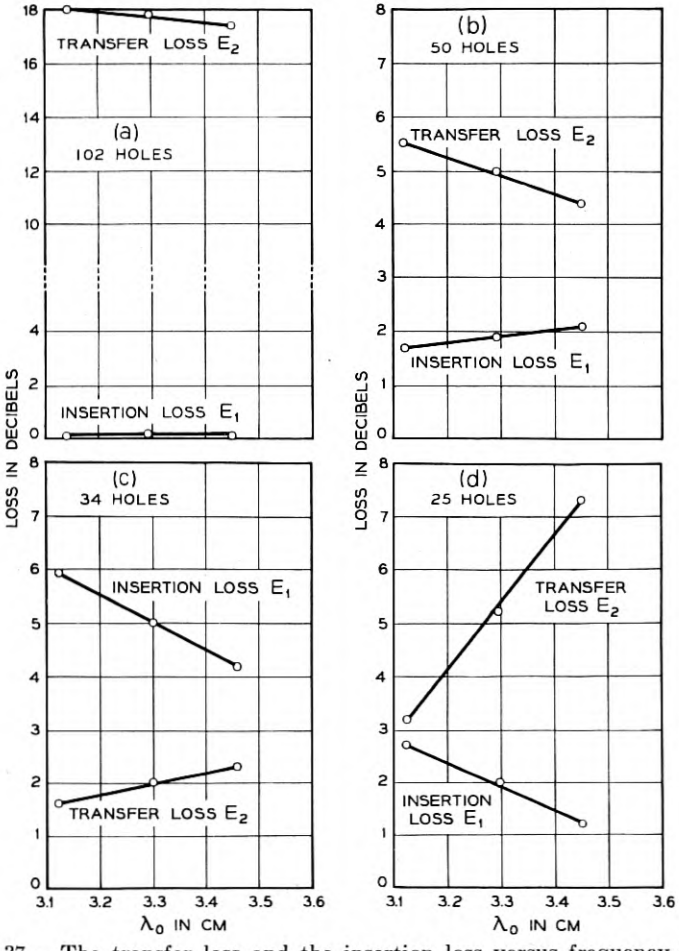


Fig. 37 — The transfer loss and the insertion loss versus frequency for the coupled waveguides of Fig. 36, showing the cyclical exchange of power as the coupling was increased by reducing the number of dividing wires in the fixed coupling aperture.

increasing wavelength values, a general result for small holes in the side wall. As the number of wires in the given aperture was reduced, markedly increased coupling resulted. This was due to the fact that the coupling loss per hole varied approximately as the fourth power of the hole dimension perpendicular to the electric vector, whereas the overall power loss varied only as the square of the number of holes in the loose coupling region. (Equations (39) and (40) describe the effects of number of coupling points more precisely.) At 50 holes, the transfer loss was about 5 db and the wave in the driven line was reduced by about 2 db; the slope of the $|E_2|$ versus λ_0 plot was the same as for 102 holes. At 34 holes, the transfer loss was about 2 db and the wave in the driven line was reduced by about 5 db; in this case, however, the undriven line wave loss increased with increasing λ_0. Since coupling increases with increasing λ_0 we deduced that the total coupling was greater than required for complete power transfer and the bracketed expression of (39) and (40) was greater than $\pi/2$. On the diagram of Fig. 16, the presumed operating point was near $cx = 2.2$ radians. At 25 holes, Fig. 37(d), the transfer loss was about 5 db and the wave in the driven line was reduced by about 2 db; as in the 34 hole case, the undriven line wave amplitude decreased with increasing λ_0 and hence with increasing coupling. Again the integrated coupling appeared to be in the region between $\pi/2$ and π. The driven line wave loss was headed for a low value at the long-wave end of Fig. 37(d), and it seems clear that periodic energy exchange is realized in practice.

The losses associated with this energy exchange may be inferred by comparing the total power output of the undriven and driven lines to the input power. Assuming that the forward waves in the driven and undriven lines contain all the output power, (i.e. neglecting reflection, back wave in the undriven line and waveguide losses) the following table gives the losses observed in the above described experiments:

Number of Holes	Coupling Mechanism Loss
	db
50	0.16
34	0.23
25	0.33

These losses may be due to circulating currents in the wires, in which case the loss would be expected to increase with increasing coupling.

Good agreement between the observed and theoretical directivities has been found in the loose coupling case,[5] but when appreciable power is

abstracted from the driven line it is clear that the theory given above does not apply. Dawson has obtained experimental data of interest in this connection. For a $6\lambda_g$ long linear-taper aperture of the form given above (0.33″ height at the center and 0.030″ height at the ends), the loose coupling theory predicts directivities in excess of 45 db for the wavelength band 3.1 to 3.5 cm. When using sufficient number of wires to obtain 18 db transfer loss, directivities in the range 36 to 48 db were observed. The reason for the 36 db observation being lower than the 45 db theoretical value may be inaccuracy of fabrication (jig per Fig. 36) or inapplicability of the loose coupling theory at 18 db transfer loss. At 3 db transfer loss, the observed directivity of a similarly shaped but 5.5 λ_g long coupling array is shown at the top of Fig. 38; again loose coupling theory predicts more than 45 db directivity. The reason the observed values are in the 24–33 db range rather than above 45 db is presumed to

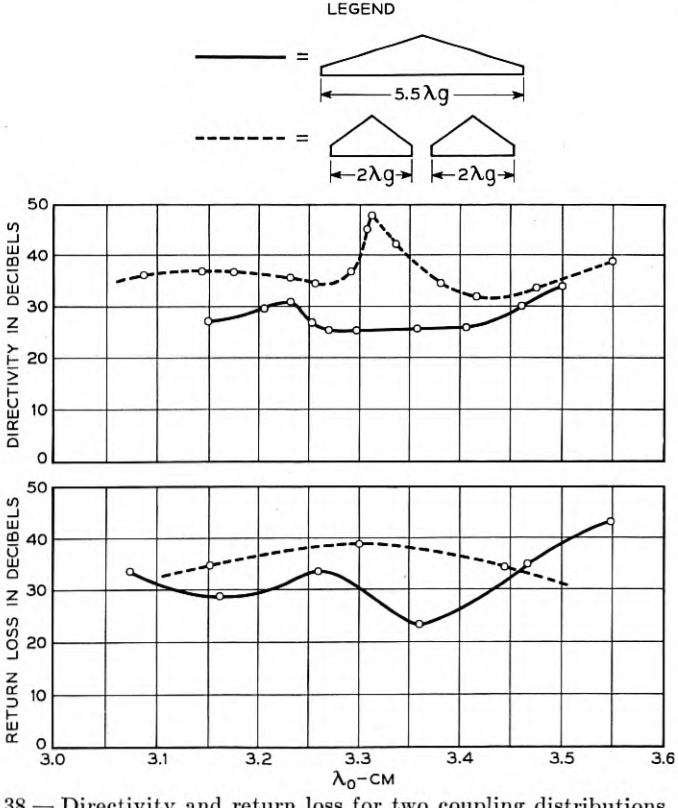


Fig. 38 — Directivity and return loss for two coupling distributions, each of which produced 3 db transfer loss.

TABLE I

Array	Transfer Loss		
	λ_0 = 3.15 cm	λ_0 = 3.30 cm	λ_0 = 3.45 cm
	db	db	db
Single $2\lambda_g$ array	3.2	3.0	2.8
Two cascaded arrays	4.3	4.2	4.0
Single $5.5\lambda_g$ array	3.5	2.9	2.4
Array	Straight Through Loss		
	λ_0 = 3.15 cm	λ_0 = 3.30 cm	λ_0 = 3.45 cm
	db	db	db
Single $2\lambda_g$ array	3.0	3.2	3.5
Two cascaded arrays	2.6	2.8	3.0
Single 5.5 λ_g array	2.9	3.6	4.3

be inapplicability of the theory. Loose coupling theory predicts better than 35 db directivity over a broad frequency band at a coupling length of about $2\lambda_g$; therefore one might expect to obtain better overall results by using two cascaded arrays each about $2\lambda_g$ long, and each having a transfer loss of 8.4 db to get the 3 db net transfer loss. Observed directivities for such a coupling array are also given in the top of Fig. 38; in this case values in the 32–37 db region were obtained. The destructive interference associated with addition of backward wave components is more nearly of the form computed by loose coupling theory because the exciting wave is more nearly constant over the length of one of the arrays. The observed return loss at any one of the four waveguide entries, when the others are terminated, is given for the $5.5\lambda_g$ and cascaded $2\lambda_g$ coupling arrays at the bottom of Fig. 38. The cascaded $2\lambda_g$ combination is again superior to the single long taper. The characteristic of being inherently matched at all terminals makes the coupled-line type of 3 db hybrid attractive at the very high frequencies where lumped element matching becomes difficult if not impracticable.

Where space is at a premium, or where more constant transfer loss values are to be desired a shorter array composed of larger holes is attractive. A single linear taper of the shape outlined above and $2\lambda_g$ long was observed to have better than 22 db directivity and better than 25 db return loss over the 3.1 to 3.5 cm band. The observed loss values of the three coupling arrays discussed above are given in Table I. The coupling arrays composed of larger holes have less slope in the loss versus frequency characteristic for side-wall coupling.

SOME LOOSELY COUPLED TRANSDUCERS IN MULTIMODE WAVEGUIDE

In connection with research on low-loss circular-electric-wave transmission,[1] there developed a need for means with which to measure the power present in any one of the modes of a multi-mode round waveguide. In particular, it was known that the circular electric wave in round waveguide converts readily to the TM_{11} wave due to curvature of the line,[2] and a direct measurement of the effect was needed. The TM_{11} wave will not exist in the round waveguide without the presence of at least four other modes, and in the waveguide size used for the experiments five other modes could propagate. In designing a transducer for this application, therefore, it was necessary to evaluate the discrimination function, equation (4), with regard to mode discrimination between five different pairs of modes as well as to insure directivity. Moreover, the TM_{11} wave is degenerate with the circular electric wave TE_{01}, i.e., they have the same phase constant. Therefore, mode discrimination against TE_{01} could not be obtained through the phase difference effects described by (4). This discrimination was obtained using geometric balance in the individual coupling orifices, which were narrow slits on the center line of the wide side of the rectangular guide, as shown in Fig. 39. The shape of the coupling distribution employed was that described in connection with Fig. 14 except that 80 point couplings were used to simulate the raised-cosine coupling distribution (instead of 40 as in Fig. 14) in order to assure good directivity for the very long coupling length that was required. The round guide diameter was two inches, the rectangular guide width 0.820 inches, calculated to produce the same cut-off frequency in the rectangular guide as exists for the TM_{11} wave in the round waveguide. The coupling length was about 17 inches.

One simple method for evaluating the mode content of such a transducer is to measure the azimuthal distribution of electric field at the round guide wall using the radial probe technique described by M. Aronoff.[7] If the power in a single mode is a great deal larger than the

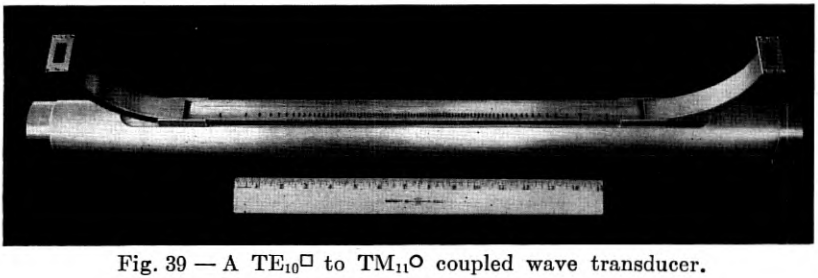


Fig. 39 — A TE_{10}□ to TM_{11}○ coupled wave transducer.

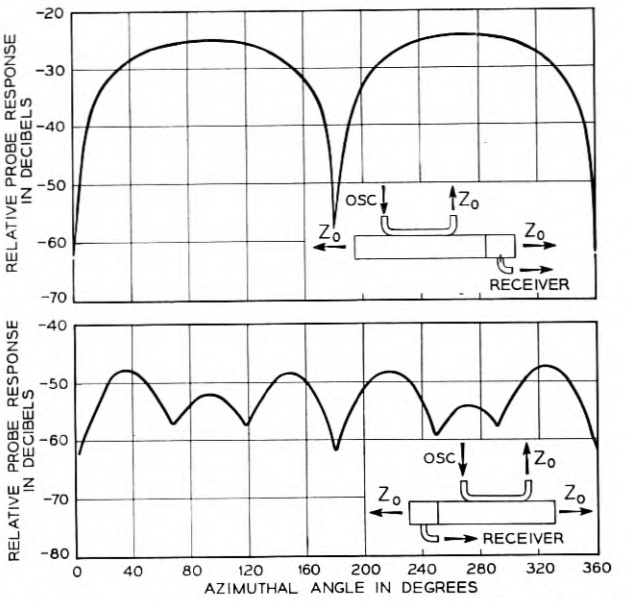


Fig. 40 — Distribution of radial electric field at the guide wall for the forward and backward waves of the transducer of Fig. 39.

power in any other mode of the multi-mode guide, the radial probe technique narrows down the possible mode types to a very few. Measurements of this type, recorded in Fig. 40, indicate that the forward wave has the radial electric field distribution to be expected for the TM_{11} wave. However, the forward wave might have the same radial field distribution at the wall and actually be the TE_{11} wave instead of TM_{11}. The TE_{11} wave is very simply generated from a dominant mode rectangular guide, by means of a long taper transition along the axis of propagation from the rectangular cross section to the circular cross section. Such a transducer was used to measure the output wave of the TM_{11} transducer and it was found that the TE_{11} component was down on the order of 30 db below the value which would be present if the radial field intensity observed at the top of Fig. 40 had been due to TE_{11}. By a process of elimination, therefore, and by virtue of the fact that we have a pure pattern suggesting the presence of a single mode, we have established that the mode generated is actually TM_{11}. Other checks can of course be made, such as measurement of the phase constant of the output wave.

The backward wave shown at the bottom of Fig. 40 has a maximum field more than 20 db below the maximum field of the forward wave and

has a six-peaked variation with angle which indicates the presence of TE_{31}.

The transfer loss of the TM_{11} transducer was derived by (1) calibrating the receiving probe on a known amount of power in the TE_{11} wave, (2) inserting this same amount of power in the rectangular waveguide of the coupled wave transducer and, with the probe at the transducer output, observing the change in the receiver response, and (3) correcting the observed loss using the theoretical difference in the radial electric field at the wall for the TE_{11} versus TM_{11} waves in the known waveguide diameter. (This technique is described in more detail by Aronoff.[7]) The result gave a transfer loss of about 25 db to the TM_{11} wave. The insertion loss for the rectangular guide of the transducer was less than 0.2 db.

Coupled-wave devices of the type shown in Fig. 39 were built for several of the modes in 2″ round waveguide. The one built for the TE_{31} mode in 2″ waveguide (mechanically similar to the TM_{11} model of Fig. 39) has several characteristics worthy of mention. Fig. 41 shows the

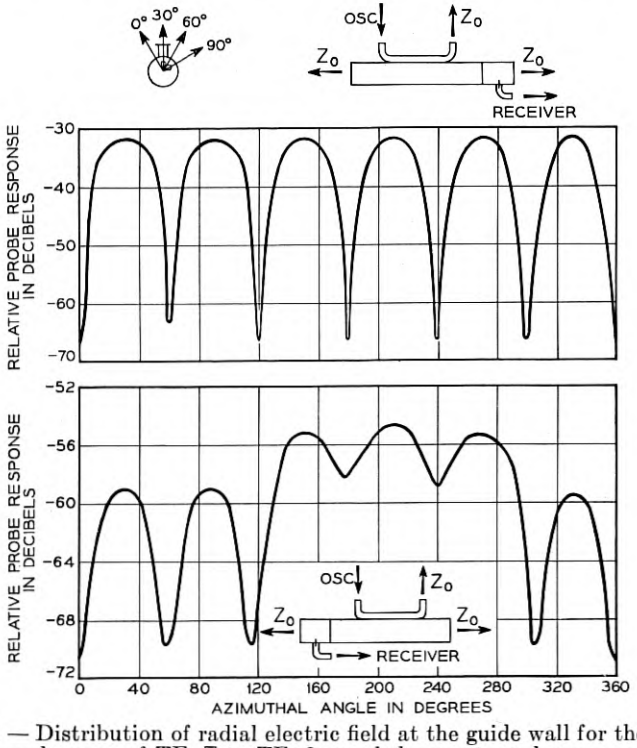


Fig. 41 — Distribution of radial electric field at the guide wall for the forward and backward waves of TE_{10}□ to TE_{31}○ coupled wave transducer.

TABLE II

Ratio of Forward Traveling TM_{11} Power to	Observed Discriminations		
	λ_0 = 3.1 cm	λ_0 = 3.3 cm	λ_0 = 3.5 cm
	db	db	db
TM_{11} Backward	>20	>20	>20
TE_{11} Forward	28.5	28	26
TE_{11} Backward	37	35	39
TM_{01} Forward	46	46	41.5
TM_{01} Backward	49	51	45.5
TE_{21} Forward	24.5	21	23
TE_{21} Backward	29.5	35	31
TE_{31} Forward	14	26	21.5
TE_{01} Forward	45	46	45
TE_{01} Backward	64	69	67

measured forward and backward wave patterns in the round guide, for excitation in one of the rectangular guides of the transducer. Only TE_{31} of the six modes possible in the 2″ pipe at 3.3 cm has a six-lobed pattern of azimuthal distribution of radial electric field at the wall, and hence the clean pattern with equally spaced deep nulls indicates the presence of a rather pure TE_{31} mode. The six maxima of the forward wave were equal within ±0.15 db. The backward wave had a peak electric field at least 23 db down on the peak electric field of the forward wave.

Using coupled transmission line techniques and the familiar geometric taper techniques, transducers were built for all of the six modes possible in 2″ diameter pipe at 3.3 cm for use in the circular electric wave research program.[1] These transducers were used to measure the forward wave and backward wave output of the TM_{11} transducers, as given is Table II. In reality, imperfections in either one of the two transducers involved in a measurement could result in the recorded values of discrimination. For example, if the TM_{11} transducer were perfect and the TE_{01} output transducer contained some TM_{11} . then the insertion loss measurement involving the two transducers face to face would produce an indication of mode impurity. Since we do not have independent information on the mode purity of any one of the transducers at the level of the observed wave impurities, we can only state that both transducers involved in a discrimination measurement are probably at least as good as the number tabulated.

It should be noted that very high discriminations between TE_{01} and TM_{11} were achieved, despite the fact that this one discrimination depends solely on the mode-selective nature of the coupling orifice. Similar discriminations can be employed effectively to augment the wave-inter-

ference discrimination even in cases where there is difference between the desired and undesired modes' phase constants, to achieve very large discriminations. In the TM_{11} discriminations listed above, the values for TE_{31} are not great but are consistent with computed values for the coupling length and the coupling function employed; longer coupling lengths would produce better TM_{11} versus TE_{31} discriminations.

A TIGHTLY COUPLED $TE_{10}^{\square}$ TO TE_{01}° WAVE TRANSDUCER*

A highly efficient means of transferring power from dominant-mode rectangular waveguide to one of the higher modes of a multi-mode waveguide would be essential in a waveguide transmission system.[1] When several modes can propagate in one or both of the guides, the problem of achieving complete power transfer is more difficult and requires some new techniques. This section describes these techniques and gives experimental data for a circular-electric-wave ($TE_{10}^{\square} - TE_{01}^{\circ}$) transducer.

The desired transducer was required to make the wave transformation between a single-mode rectangular waveguide and the circular electric mode (TE_{01}°) of an 0.875″ round waveguide at a nominal frequency of 24,000 mc. The 0.875″ round waveguide at this frequency will support 10 modes of which the circular electric mode and its degenerate partner TM_{11}° are the fourth and fifth in order of appearance.

The minimum length of the coupling interval required to achieve mode discrimination may be estimated using loose coupling theory (equation 4). The mode nearest to TE_{01}° in phase constant is the TE_{31}° and for this mode a coupling length of about 0.18 meters is required in order to produce a value of θ/π equal to unity. As shown by equation (5) for uniform coupling, it is necessary to have θ/π equal to unity or greater in order to develop discrimination against the undesired mode.

The maximum coupling coefficient permissible for a given amount of mode impurity at the complete power transfer point may be estimated using the tight coupling theory of the preceding sections. For example, equations (31) and (32) show that for the ratio $(\beta_1 - \beta_2)/c$ equal to 10, the transfer loss to the undesired wave will always be greater than 14 db (regardless of the length of the coupling interval), corresponding to an energy loss for the desired wave of less than 0.2 db. For the TE_{01}° and TE_{31}° modes the calculated values of β_1 and β_2 lead to the conclusion that the coupling coefficient c between TE_{31}° and $TE_{10}^{\square}$ must be less

* When discussing the modes of hollow metallic waveguides of different cross-sectional shapes, it has been found convenient to use a superscript to designate the shape of the cross section. (See G. C. Southworth, *Principles and Applications of Waveguide Transmission*, D. Van Nostrand Co., 1950). Thus, $TE_{10}^{\square}$ refers to the TF_{10} mode in rectangular waveguide.

than 3.45 radians per meter. If the coupling coefficient for $TE_{10}^{\square}$ to TE_{01}° is equal to that for $TE_{10}^{\square}$ to TE_{31}° it follows that the total coupling length must be greater than 0.455 meters, because complete power transfer requires that the product of coupling-length times coupling-coefficient be exactly $\pi/2$ (see Fig. 17). Actually, the $TE_{10}^{\square} - TE_{31}^{\circ}$ coupling may be greater than the $TE_{10}^{\square} - TE_{31}^{\circ}$ coupling which leads to the requirement for longer coupling intervals. It is evident that the shorter coupling intervals may be employed at the sacrifice of greater mode impurities. The preceding calculations were made for the $TE_{10}^{\square} - TE_{31}^{\circ}$ and $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer ratios as though only one mode of the multi-mode waveguide were present at a time, i.e., using a theory based on coupling between two waves instead of a theory for the simultaneous coupling between a plurality of waves. It is felt that this is probably

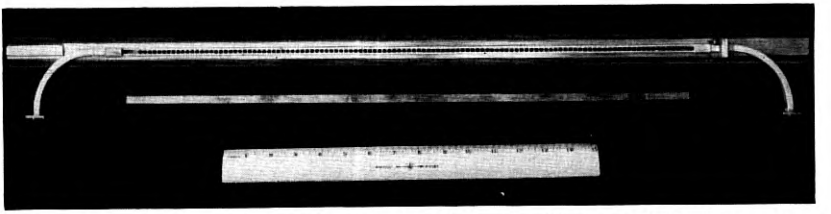


Fig. 42 — An experimental circular electric wave ($TE_{10}^{\square}$ to TE_{01}°) transducer for 24,000 mc.

justified provided that the coupling per unit length is weak and only one mode in each guide carries an appreciable amount of power.

Fig. 42 shows a photograph of one of the models used to obtain experimental data. The coupling holes were located in the narrow wall of the rectangular waveguide, thus avoiding coupling to all of the TM modes of the round waveguide. The total coupling length was 0.55 meters. The coupling orifices were spaced about 0.3 wavelengths in the dominant-mode rectangular waveguide, which assured reasonable directivity in the transfer of power between waveguides, provided that two or more coupling elements were employed.

The transfer loss between the rectangular waveguide and the circular electric mode of the round waveguide was measured as a function of the number of coupling elements, using the structure of Fig. 42 with the addition of a movable thin-walled metallic cylinder. The latter could be moved inside the transducer in such a way as to cover up a variable number of coupling holes, and contained a long wooden termination so that all the power entering the movable cylinder was absorbed. The inner diameter of the movable cylinder was large enough to propagate the

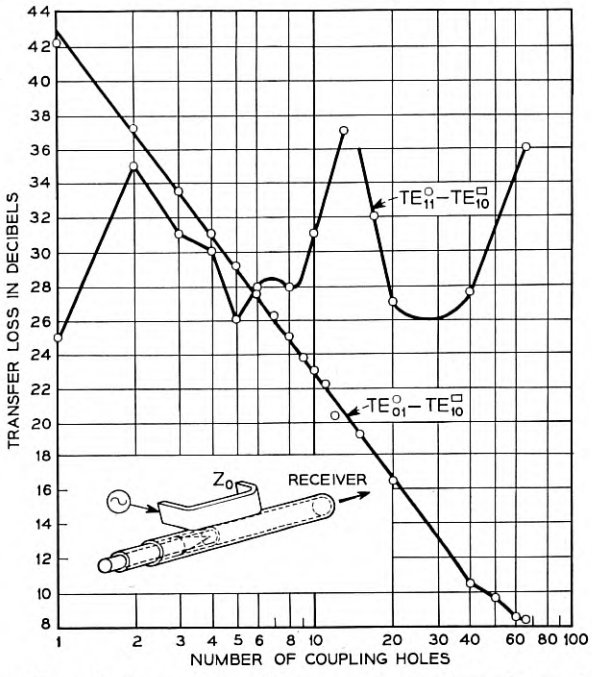


Fig. 43 — Transfer losses versus frequency for the transducer of Fig. 42.

circular electric wave but did cut off some of the waves which could propagate in the round guide of the transducer itself. The measured transfer loss under these conditions is recorded in Fig. 43. It is seen that the $TE_{10}^{\square}$ — TE_{01}° coupling was so weak as to be in the region where power from successive coupling elements should add inphase all the way up to 40 coupling elements. The observations show the inphase addition for less than 30 coupling elements but show a marked deviation in the vicinity of 40 to 66 coupling elements. This is evidence of inequality of the phase constants for the TE_{01}° and $TE_{10}^{\square}$ waves. More will be said about this matter presently. The transfer loss between the rectangular waveguide and the TE_{11} mode of round waveguide, is also recorded in Fig. 43. As expected, the power from successive coupling elements did not add inphase and no appreciable build-up of power in the TE_{11} mode took place.

One way of evaluating the total power in all modes other than the circular electric mode, is to measure the value of the transverse magnetic intensity at the wall of the round waveguide. The circular electric wave has no such field component and all other waves do possess such a field

component. Thus the total value of the transverse magnetic intensity at the round waveguide wall is a measure of the impurity associated with the circular electric wave. (This is very similar to the radial probe technique described by M. Aronoff.[7]) Using this method of evaluation, the mode impurities present at the output of the transducer were measured as a function of the number of coupling elements, and the results are recorded in Fig. 44. The absolute calibration of the ordinate relates the observed magnetic intensity to that which the same power input used at the rectangular guide would have produced if placed in the round waveguide in the TE_{11} mode. These measurements show that for all of the modes other than the circular electric mode, the energy components from successive coupling elements suffer destructive interference. Although curves are shown only for one and for 66 coupling elements, the patterns for intervening numbers of coupling elements were similar in shape and never exceeded an intensity value greater than about 6 db above that given for the 66 coupling element case; thus the mode discriminating property of the coupled wave transducer was verified experimentally.

Returning to the question of $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer loss, it is clear from Fig. 43 that the rectangular waveguide has a phase constant which is not equal to that of the circular electric mode in the round waveguide. One reason for this inequality lies in the fact that the coupling elements disturb the phase constant in the two waveguides unequally, a consequence of the fact that some of the power transferred to the round wave-

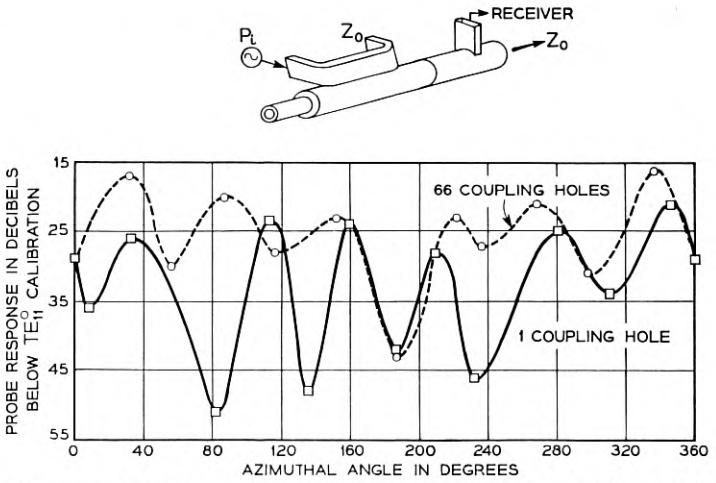


Fig. 44 — Distribution of transverse magnetic intensity at the wall for the transducer of Fig. 42.

guide on a single coupling element basis, appears in modes other than TE_{01}. Thus, the total coupling to $TE_{10}^{\square}$ is greater than to TE_{01}°. The total coupling modifies the phase constant of each line, per (20′), and since the total coupling coefficient is unequal for the $TE_{10}^{\square}$ and the TE_{01}° modes, the perturbed phase constants should be expected to be unequal when the unperturbed phase constants are made equal. A method of determining the magnitude of this phase-constant disturbance has been suggested by S. A. Schelkunoff. In this method the reflected wave from a single coupling orifice is measured in the dominant waveguide and in the single mode of interest in the multi-mode waveguide. Having defined the ratio of the incident to the reflected power in the same mode by the sympol p, Schelkunoff determines that the disturbed phase constant β', is related to the undisturbed phase constant β by the relation

$$\beta' = \beta + \sqrt{\frac{p}{d}}, \tag{41}$$

in which "d" is the distance between the coupling orifices in the coupling arrangement which one wishes to evaluate. This relation may be used to evaluate the change in the phase constant for the circular electric mode and for the wave in the dominant waveguide, and the change of wave-

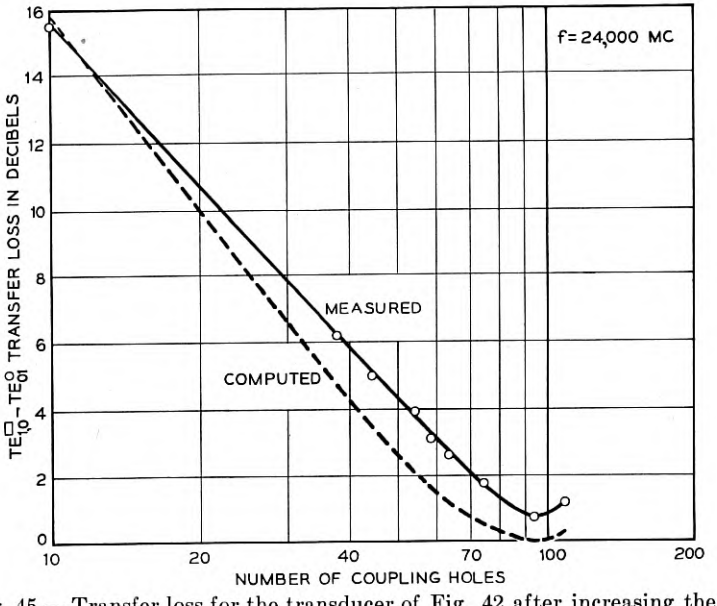


Fig. 45 — Transfer loss for the transducer of Fig. 42 after increasing the coupling hole sizes and correcting the phase constant.

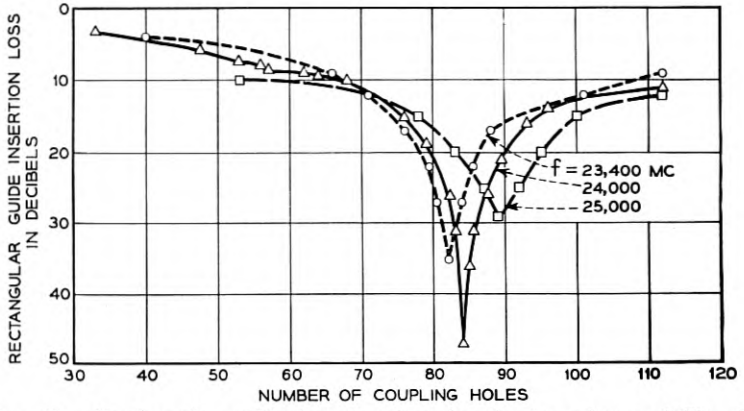


Fig. 46 — Rectangular guide insertion loss for the transducer of Fig. 42.

guide dimensions required to correct this phase constant difference may be computed as though the coupling elements were not present.

For the small phase constant disturbances which are associated with the weak couplings employed, this procedure was found very accurate. The reflection measurements and associated calculations for the model of Fig. 42 indicated that the rectangular guide width should be 0.340″ for equality of phase constants instead of 0.359″ as computed neglecting coupling effects. The measured value of the transfer loss when the individual coupling holes had been enlarged and the rectangular guide width had been altered to the 0.340″ value is shown in Fig. 45. It is evident that the theoretical value of 0 db transfer loss was approached, and that the shape of the transfer loss versus number of coupling elements, was reproduced very well. The 0.75 db minimum transfer loss consisted of no more than 0.3 db heat loss, the remaining loss being due to power present in other modes.

The measured insertion loss in the rectangular waveguide is shown as a function of the number of coupling holes at the three frequencies in Fig. 46. Complete power transfer would, of course, correspond to an infinite insertion loss in the rectangular waveguide. It is interesting to note that at 24,000 mc the peak in the rectangular guide insertion loss occurred at 85 coupling elements whereas the maximum in the $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer loss characteristic occurred at about 96 coupling elements (Fig. 45). This difference is likely to be the result of power transferred back to the rectangular waveguide from round waveguide modes other than circular electric. Additional evidence of deviations due to the coupling between a plurality of waves was obtained; the rectangular-guide insertion loss as a function of number of coupling elements did not increase

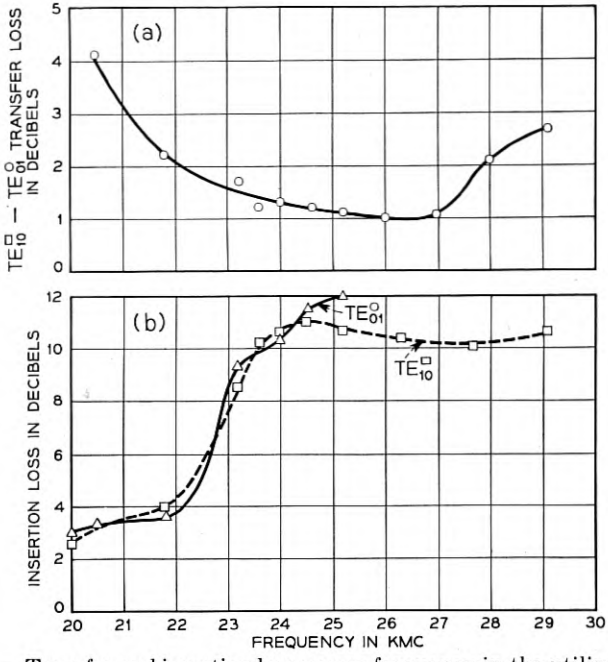


Fig. 47 — Transfer and insertion loss versus frequency in the utilized modes of the transducer of Fig. 42 using all (112) coupling holes.

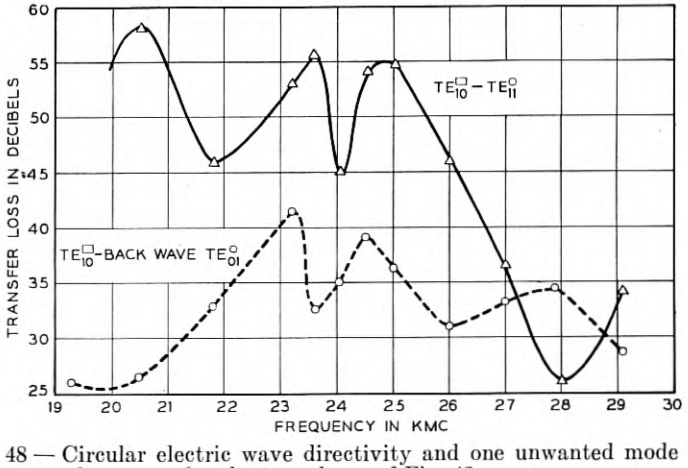


Fig. 48 — Circular electric wave directivity and one unwanted mode (TE_{11}° output versus frequency for the transducer of Fig. 42.

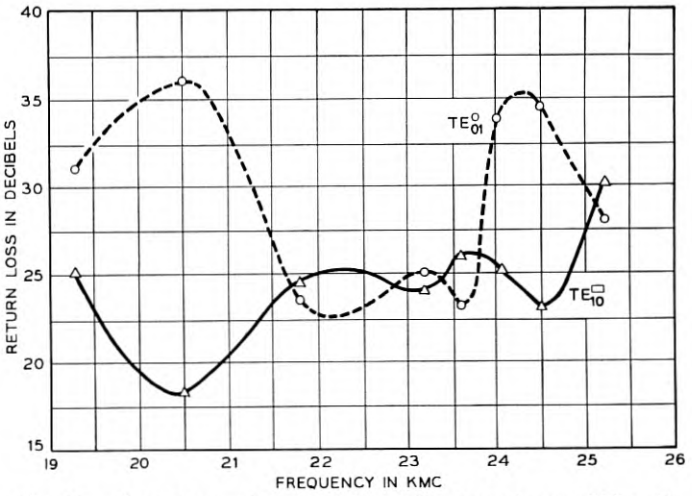


Fig. 49 — Impedance characteristic of the transducer of Fig. 42.

smoothly according to a cosine amplitude function as would be expected for two coupled waves of identical phase constant, but instead exhibited ripples. The remarkable thing about the data of Figs. 45 and 46 is that it agrees with the theory for two coupled waves as well as it does.

The coupling per individual orifice decreases with increasing frequency and this is verified by the observation (Fig. 46) that a greater number of coupling elements are required to reach the maximum insertion loss in the rectangular guide at the higher frequency.

Some indication of the overall bandwidth of this first experimental model is given in Figs. 47, 48 and 49 which show respectively the $TE_{10}^{\square}$ – TE_{01}° transfer loss, the insertion losses in the $TE_{10}^{\square}$ and TE_{01}° modes, the $TE_{10}^{\square}$ – TE_{11}° and $TE_{10}^{\square}$ – backward wave TE_{01}° transfer losses, and the $TE_{10}^{\square}$ and TE_{01}° return losses in the frequency range 20,000 to 30,000 mc. No one of these characteristics represents the degree of excellence which is achievable but they do demonstrate that good impedance match, low transfer losses to the desired mode, and appreciable discrimination against unwanted modes, can be achieved over frequency ratios on the order of 1.5.

FREQUENCY SELECTIVITY

In the case wherein the coupling is so weak as to not affect the total phase constant appreciably, all modes of hollow conductor waveguides of any cross section have the same phase constant at all frequencies provided that these modes have the same cut-off frequency. This results

in very broad band mode-selective characteristics, as has been demonstrated.

The transfer loss characteristics are in general a function of frequency, since the individual coupling holes are somewhat frequency selective. There may be applications wherein less variation in transfer loss as a function of frequency is required. One approach to this problem is to make the coupling holes individually have less coupling variation with frequency; since the total coupling loss between two identical transmission lines is a function only of number of coupling holes and the loss per hole (equations (39) and (40)) constant coupling per hole will produce constant coupling overall. Riblet and Saad[6] have reported on this approach.

There is another approach to obtaining flat coupling versus frequency despite variations in the coupling per hole, and that is to intentionally create a difference between the phase constants of the two coupled lines. Fig. 17 illustrates the transfer characteristic when the coupled lines have unequal phase constant, and either identical or negligible attenuation constants. Near the maximum for the transferred wave $| E_2^* |$ there is a region wherein the transfer loss is independent of coupling strength, and the transfer loss in this flat-loss region is under control of the ratio $(\beta_1 - \beta_2)/c$. Hence for a given transfer loss there is an optimum ratio of phase constant difference to coupling strength in order to minimize the overall transfer loss variation. For the distributed coupling case, equations (31) and (32) represent the transferred wave amplitude and show that the transferred wave goes through a maximum as a function of integrated coupling strength $\underline{cx}$, when

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}\, cx = \frac{\pi}{2} + n\pi. \tag{42}$$

The transferred amplitude at this maximum point is

$$E_{2\max}^* = \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}}. \tag{43}$$

The integrated coupling strength at the maximum point is

$$c_0 x_0 = \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \cdot \frac{\pi}{2}. \tag{44}$$

For the important case of an optimum 3 db transfer loss coupler, E_2^* is 0.707. Then $(\beta_1 - \beta_2)/c$ equals 2 and $\underline{c_0 x_0}$ equals $\pi/2\sqrt{2}$ from (43)

and (44). Assuming a coupling length x_0 of two wavelengths in the line with the smaller phase constant, it follows that β_1/β_2 is about 1.18 showing that a phase-constant difference of 18 % is required. This phase-constant difference is quite readily attainable in the waveguide structure of Fig. 50(a). The two modes coupled together are given slightly different cut-off wavelengths in the coupling region, and may be tapered to the standard waveguide size outside the coupling region. The desired phase-constant difference can also be obtained in two identical metallic guides by inserting a piece of dielectric into one of the guides in the coupling region as sketched in Fig. 50(b). Although rectangular waveguides are used in Fig. 50 to illustrate the method of obtaining frequency independent transfer characteristics, the approach is general and may be applied to any form of single or multi-mode transmission line.

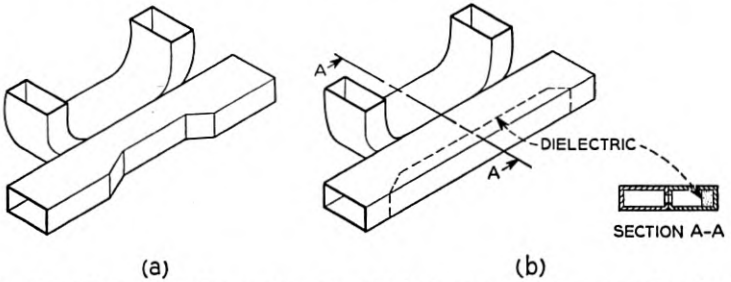


Fig. 50 — Examples of structures in which flat transfer loss may be obtained despite coupling loss variations.

In either dominant-mode directional couplers or in multi-mode coupled-wave devices such as the one illustrated in Fig. 1, one may obtain much more frequency selectivity than occurs incidentally due to the frequency sensitivity of the coupling elements used. This may be done by coupling two transmission lines which have the same phase constant at one frequency, but unequal phase constants at other frequencies. Then, as shown by equation (31), the midband transfer loss may be set at any desired value by adjusting the integrated coupling strength $\underline{cx}$ at midband (where $\beta_1 - \beta_2 = 0$), and at other frequencies where $(\beta_1 - \beta_2) \neq 0$, the transfer loss will increase. For the particular case of $\underline{cx} = \pi/2$ (fixed) for which complete power transfer occurs when $\beta_1 = \beta_2$ (and assuming $\alpha_1 = \alpha_2$ or both α's are negligible), Fig. 51 shows the shape of the filter characteristic, E_2^* versus $(\beta_1 - \beta_2)/2c$. This plot is valid for any form of transmission line.

A very simple configuration for realizing such a frequency-selective filter involves coupling between two hollow conductor waveguides, one

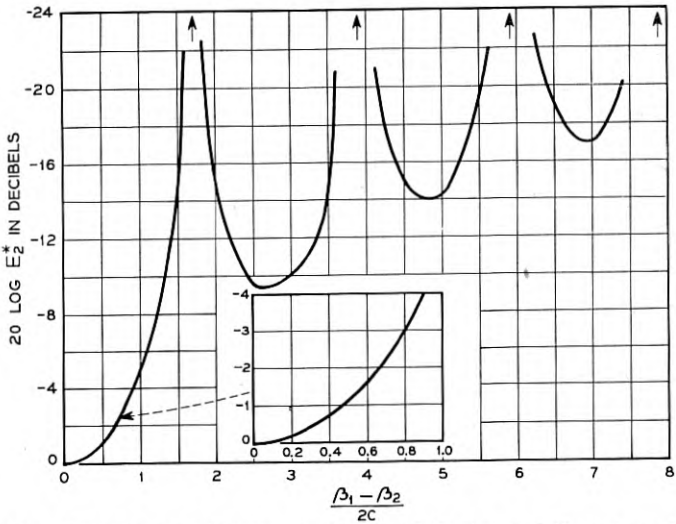


Fig. 51 — Transfer loss E_2^* versus $(\beta_1 - \beta_2)/2c$ for coupling strength $cx = \pi/2$, the value required for complete power transfer.

of which is air-filled and the other of which is filled with a material of dielectric-constant ϵ. The phase constants for these waveguides have the form sketched in Fig. 52, in which β_0 is the phase constant in free space. At the frequency f_m the two waveguides have identical phase constants and, in a typical case, negligible loss constants so that complete power transfer can be obtained. For the case $\epsilon = 2.55$, Fig. 53 shows the computed frequency characteristic on the assumption that the integrated coupling is set for complete transfer ($cx = \pi/2$) and is independent of frequency. (Actually the usual coupling mechanisms are somewhat frequency sensitive and would increase the selectivity somewhat.) This filter

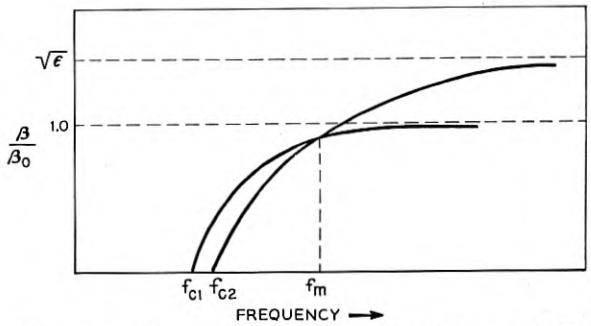


Fig. 52 — The general form of the phase constants for two hollow conductor waveguides, one of which is filled with a dielectric.

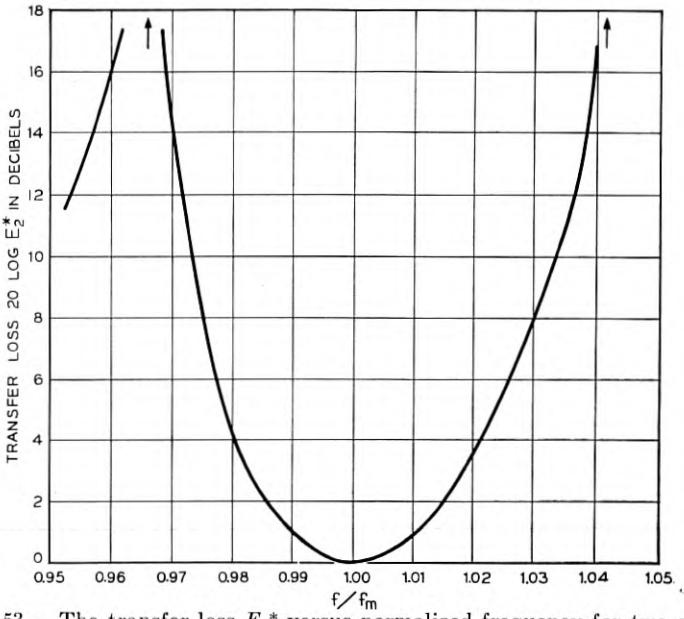


Fig. 53 — The transfer loss E_2^* versus normalized frequency for two coupled hollow conductor waveguides, one of which is air filled and has a guide wavelength $\sqrt{2}$ times the free space wavelength at f_m , and the other of which is filled with a material of dielectric constant 2.55 with dimensions chosen for equality of phase constant with the air-filled guide at f_m . Coupling $\underline{cx}$ assumed constant at $\pi/2$.

characteristic applies regardless of the shapes of the hollow conductor waveguidess (which may be dissimilar) and regardless of the modes selected.

It is apparent that frequency selectivity in the transfer characteristic E_2^* can also be obtained without requiring that the phase constants be unequal by using coupling elements which are frequency sensitive.

DIELECTRIC WAVEGUIDE CONFIGURATIONS

The coupled-wave approach to circuit design is applicable using any form of transmission line, the only important variant associated with different forms of line being the physical structure associated with introducing the desired coupling between lines. In a recent publication[8] A. G. Fox showed that dielectric waveguides are very attractive for use in the millimeter wavelength range, and this section points out how dielectric waveguides can be used in various forms of coupled wave devices. Fox showed that dielectric waveguides arranged in the configuration sketched in Fig. 54 are coupled by the electric field components only, and that

periodic energy exchange of the type described by equations (26) and (27) is observed. Moreover, he also showed that if one line were made very lossy the energy exchange phenomena disappeared and, despite sufficient coupling to cause complete power transfer when both lines were loss-free, power passed through the coupling region in the low-loss line with less than 0.25 db attenuation. This verified the predictions of equations (35) and (36).

Other implications of the coupled wave theory can also be utilized in dielectric waveguides. If the two lines (Fig. 54) are made of materials having different dielectric constants and their cross-sectional dimensions set so as to secure identical phase constants at a frequency f_m, then a frequency-selective coupled-wave filter results and the selectivity characteristic of Fig. 53 applies. As an alternative to using materials having different dielectric constants, the same dielectric may be used for both lines by making one line solid and the other hollow.

If both lines are made of the same material and the cross-sectional dimensions are set so as to obtain a known difference between their phase constants, the result is a directional coupler having a region of flat transfer loss (of any desired magnitude) and equations (42), (43) and (44) apply.

Both of the preceding applications can be carried out in dielectric waveguides having arbitrary cross-sectional shapes.

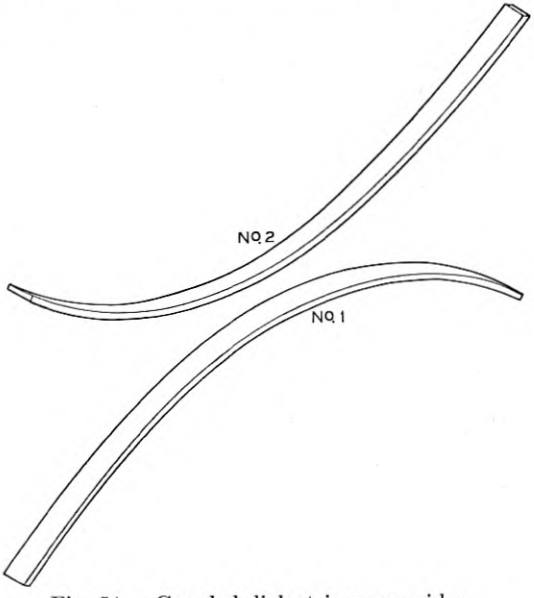


Fig. 54 — Coupled dielectric waveguides.

If one of the transmission lines (Fig. 54) is round and the other is rectangular and if their cross-sectional dimensions are set for equal phase constants, then the power in one of the two polarizations of the round line may be transferred to any desired extent to the rectangular guide, and power in the other polarization of the round guide will pass the coupling region undisturbed. Two such rectangular-rod to round-rod coupling configurations arranged in cascade along the round-rod, with the two rectangular rods coupled in planes at 90° to each other, constitutes a means for independently connecting to the two polarizations of the round-rod. This type of device depends upon the fact that the phase constants of the two polarizations of round-rod are identical, whereas the two phase constants for the rectangular rod are different. Thus a wave interference occurs in the transfer characteristic for one of the polarizations, and for suitable values of $(\beta_1 - \beta_2)/c$ (see Fig. 18) the power transferred in this polarization can be made small.

SUMMARY

Two approaches to a theoretical description of the behavior of two coupled waves have been presented. One, based on the assumption of negligibly small coupling, is applicable in cases where very little power is transferred between the coupled waves. The other, a solution based on uniform coupling between waves in the coordinate of propagation, is valid for any magnitude of total coupling.

The loose coupling theory shows how to taper the coupling distribution in order to minimize the length of the coupling interval required for a given degree of directivity and/or for a given magnitude of mode impurity. In particular, it is possible to shape the coupling distribution so as to discriminate sharply against one or more undesired modes in a coupled-wave arrangement involving just a few modes. (See Figs. 7 and 15 for examples).

The theory indicates that significant exchange of power takes place provided that the attenuation and phase constants of the coupled waves are equal, or provided that the difference between the attenuation constants and the difference between the phase constants are small compared to the coefficient of coupling. A suitable difference between either the attenuation constants or the phase constants of two coupled waves is sufficient to prevent appreciable energy exchange (equations 29–32 and 35–36).

It follows that substantially single-mode propagation is possible in a multi-mode structure even though geometrical effects tending to cause coupling between modes are present. A gradual transition in the boundary

of a multi-mode waveguide will not cause an appreciable exchange of power between modes provided that the quantity $(\beta_1 - \beta_2)/c$ is sufficiently large for the modes which are coupled by the boundary change. Similarly, for disturbances in the coupled-wave system which takes place over a large number of wavelengths in the direction of propagation, the coupled-wave theory indicates that all conversion will take place in the forward direction and very little reflection in any mode will result.

The tight coupling theory shows that for the case of identical complex propagation constants, a periodic exchange of energy between waves takes place along the coordinate of propagation. The only effect of the existence of an attenuation constant for both waves (compared to the dissipationless case) is to add the same exponential attenuation factor (to the periodic energy exchange phenomenon) which would have existed for a wave traveling on one of the lines in the uncoupled state.

When the phase constants of the two coupled waves are not equal (and the attenuation constants are either equal or negligibly small compared to the coupling coefficient), the exchange of energy between waves is no longer complete but remains periodic (Fig. 17). The quantity $(\beta_1 - \beta_2)/c$ determines the fraction of the total energy which is exchanged, and also modifies the period of the energy exchange phenomenon along the axis of propagation.

When the phase constants of the two lines are equal but the attenuation constants are unequal, the energy transfer phenomenon differs only slightly from that associated with equal propagation constants provided that the quantity $(\alpha_1 - \alpha_2)/c$ is less than about -0.1. For $(\alpha_1 - \alpha_2)/c$ more negative than about -1, the periodicity of the energy transfer phenomenon has largely disappeared (Fig. 23) and as $(\alpha_1 - \alpha_2)/c$ becomes on the order of -10 or more, the principal effect of the coupling for the low loss line is a minor alteration of the phase and attenuation constants. The wave amplitude for unit input on the low-loss line becomes [from (33) for $|(\alpha_1 - \alpha_2)|/c \gg 1$]

$$E_1 = \epsilon^{-[\alpha_1 - c^2/(\alpha_1 - \alpha_2) + i(c+\beta)]x}. \tag{45}$$

Through proper choice of the phase constants relative to the coupling coefficient in two coupled transmission lines, it is possible to make directional couplers having an arbitrary transfer loss that is independent of frequency despite variations in coupling strength with frequency (equations 43–44). It was also shown that the coupled-wave approach may be utilized to create highly frequency-selective filters which may operate between single-mode media or between selected individual modes of a multi-mode system.

The experimental data given for two dominant-mode rectangular waveguides showed that the periodic energy exchange theoretically predicted for a coupled-wave system can be achieved in coupled transmission lines.

Performance characteristics were given for some loosely coupled transducers between a dominant-mode rectangular waveguide and one mode of a six-mode waveguide. A tapered coupling distribution was used to achieve the mode selectivity in a limited length interval.

The problems associated with a coupled-wave transducer for transferring all of the power from a dominant-mode rectangular waveguide to the circular electric mode in a ten mode waveguide, were discussed and the observed characteristics of an experimental model were given.

The application of coupled-wave techniques to other types of transmission systems was illustrated by pointing out analogous structures using coupled dielectric waveguides.

ACKNOWLEDGMENT

The writer is indebted to W. W. Mumford for helpful discussions during the early stages of this work; to R. W. Dawson who made most of the measurements on the models of Figs. 36 through 44, and to G. D. Mandeville who made measurements on the model of Fig. 39.

BIBLIOGRAPHY

1. S. E. Miller and A. C. Beck, Low-Loss Waveguide Transmission, Proc. I.R.E., **41,** pp. 348–358, Mar., 1953.
2. S. E. Miller, Notes on Methods of Transmitting the Circular Electric Wave Around Bends, Proc. I.R.E., **40,** pp. 1104–1113, Sept., 1952.
3. W. W. Mumford, Directional Couplers, Proc. I.R.E., **35,** pp. 160–165, Feb., 1947.
4. C. L. Dolph, A Current Distribution for Broadside Arrays Which Optimizes the Relation Between Beam Width and Side Lobe Level, Proc. I.R.E., **34,** pp. 335–348, June 1946.
5. S. E. Miller and W. W. Mumford, Multi-Element Directional Couplers, Proc. I.R.E., **40,** pp. 1071–1078, Sept., 1952.
6. H. J. Riblet and T. S. Saad, A New Type of Waveguide Directional Coupler, Proc. I.R.E., **36,** pp. 61–64, Jan., 1948.
7. M. Arnoff, Radial Probe Measurements of Mode Conversion in Large Round Waveguide with TE_{01} Excitation, (submitted to Proc. I.R.E.).
8. A. G. Fox, New Guided Wave Techniques for the Millimeter Wavelength Range, given orally at the March, 1952, I.R.E. National Convention. To be submitted to the Proceedings.
9. Alan A. Meyerhoff, Interaction Between Surface-Wave Transmission Lines, Proc. I.R.E., **40,** pp. 1061–1064, Sept., 1952.
10. P. E. Krasnushkin and R. V. Khokhlov, Spatial Beating in Coupled Waveguides Zh. Tekh. Fiz, **19,** pp. 931–942, Aug., 1949, (in Russian).
11. W. J. Albersheim, Propagation of TE_{01} Waves in Curved Wave Guides, B.S.T.J., **27,** pp. 1–32, Jan., 1949.

Theoretical Fundamentals of Pulse Transmission — I

By E. D. SUNDE

(Manuscript received September 23, 1953)

A compendium is presented of theoretical fundamentals relating to pulse transmission, for engineering applications. Emphasis is given to the consideration of various imperfections in transmission systems and resultant transmission impairments or limitations on transmission capacity.

In Part I of this paper, Sections 1 to 11, fundamental properties of transmission-frequency characteristics are discussed, together with general relations between frequency and pulse transmission characteristics and special transmission characteristics of importance in pulse systems. This is followed by a presentation of engineering methods of evaluating pulse distortion from various types of gain and phase deviations.

In Part II, Sections 12–16, transmission limitations imposed by characteristic distortion will be discussed.

Part I

Part II

INTRODUCTION

Pulse transmission is a basic concept in communication theory and certain methods of modulating pulses to carry information approach in their characteristics the ideal performance allowed by nature. In certain applications, such as telegraphy, pulse signalling and data transmission, it has the advantage of great accuracy, since the information is transmitted in digital form by "on-off" pulses. This at the same time facilitates regeneration of pulses to avoid accumulation of distortion from noise and other system imperfections, together with the storing, automatic checking and ciphering of messages, as well as their translation into different digital systems or transmission at different speeds, as may be required in extensive communication systems. Another characteristic of pulse systems is that improved signal-to-noise ratio can be secured in exchange for increased bandwidth, as in pulse code, pulse position and certain other methods of pulse modulation. Finally, pulse modulation systems permit multiplexing of communication channels on a time division basis, which under appropriate conditions may have appreciable advantages over frequency division in the design of multiplex terminals.

In pulse modulation systems, pulses are applied at the transmitting end in various combinations, or in varying amplitude, duration or position, depending on the type of system. Pulses thus modulated to carry information may be transmitted in various ways, or undergo a second modulation process suitable to the transmission medium. The received pulses will differ in shape from the transmitted pulses because of bandwidth limitations, noise and other system imperfections. The performance of the system in the absence of noise can be predicted if the "pulse transmission characteristic" is known, that is, the shape of a received pulse for a given applied pulse.

Although the pulse-transmission characteristic suffices for determination of system performance it is customary for various reasons to relate it to the "transmission-frequency characteristic," that is, the steady-state transmission response expressed as a function of frequency. For one thing the transmission-frequency characteristics of various existing facilities and their components are known, and for new facilities can be determined more readily by calculation or measurements than the pulse-transmission characteristic. But the more fundamental reason is that the transmission-frequency characteristics of various system components connected in tandem or parallel can readily be combined to obtain the over-all transmission characteristic, while this is not the case for pulse transmisssion characteristics. It is thus possible to analyze complicated systems with the transmission-frequency characteristic as a basic

parameter, and to specify requirements that must be imposed on the transmission-frequency characteristic of the system and its components for a given transmission performance.

A fundamental problem in pulse modulation systems is transmission distortion of pulses by system imperfections in the form of phase and gain deviations over the transmission band or a low-frequency cut-off, usually referred to as "characteristic distortion," which may give rise to excessive interference between pulses and resultant crosstalk noise or errors in reception, depending on the type of system. Because of such interference, characteristic distortion limits the number of pulse amplitudes permissible in the transmission of information or messages over a given channel, and may reduce the rate at which pulses can be transmitted in systems employing only two pulse amplitudes, the minimum number. It thus places a limitation on channel capacity which, unlike signal distortion by noise, cannot be overcome by increasing the signal power.

Characteristic distortion is an important consideration particularly in wire systems where there is a low-frequency cut-off caused by transformers, and where the transmission band may extend over several octaves with substantial variation in attenuation and phase shift, or may be sharply confined by filters. In wire systems there are also fine structure deviations from a smooth attenuation and phase characteristic of a more or less random nature, resulting from small random impedance variations and mismatches along the lines. Gain and phase deviations remaining even after fairly elaborate equalization may be appreciable and difficult to overcome, especially in systems comprising a large number of repeater sections.

The purpose of this paper is to present a compendium of theoretical fundamentals on pulse transmission in a form suitable for engineering applications, both from the standpoint of design of new pulse transmission systems and pulse transmission over existing facilities. Emphasis is placed on considerations of various system imperfections, because of their importance from the standpoint of transmission performance, and since literature on this question is rather limited. Certain fundamental properties of transmission-frequency characteristics are discussed, together with general relations between frequency and pulse transmission characteristics and special transmission characteristics of importance in pulse systems. This is followed by a presentation of methods of evaluating pulse distortion from various types of gain and phase deviations, together with resultant transmission impairments or limitations on pulse transmission rates in low-pass, symmetrical and asymmetrical sideband

systems. Conversely, these methods may be used in the design of pulse modulation systems to evaluate requirements imposed on the transmission characteristics for a given transmission performance.

Transmission impairments may result from system imperfections other than characteristic distortion, which require a different theoretical approach and are not considered here. Among them are erratic timing of pulses, thermal and other noise within the transmission system and interference from outside sources, such as other communication systems or atmospheric disturbances.

1. PROPERTIES OF TRANSMISSION-FREQUENCY CHARACTERISTICS

A basic parameter of transmission systems is the transmission-frequency characteristic

$$T(i\omega) = A(\omega)e^{-i\psi(\omega)}, \tag{1.01}$$

in which $\omega = 2\pi f$ is the radian frequency, $A(\omega)$ is the amplitude and $\psi(\omega)$ the phase characteristic. The transmission-frequency characteristic may designate the ratio of received voltage to transmitted current, of received current to transmitted voltage, of received to transmitted current or of received to transmitted voltage. The two latter ratios are not the same except for symmetrical networks with impedance matching at both ends. For symmetrical structures having appreciable attenuation, such as transmission lines between repeaters, the ratios are virtually the same with impedance matching at the receiving end. In the following, $T(i\omega)$ will designate any of the above ratios, as the case may be.

When a number of networks are connected in series, as is usually the case in transmission systems, the resultant transmission characteristic is

$$\begin{aligned} T(i\omega) &= T_1(i\omega)\ T_2(i\omega) \cdots T_n(i\omega), \\ &= (A_1\ A_2 \cdots A_n)\ e^{-i(\psi_1+\psi_2+\ \ldots\ +\psi_n)}, \end{aligned} \tag{1.02}$$

where T_1, $T_2 \cdots T_n$ are the transmission characteristics of the individual networks with the same impedance terminations as encountered in the series arrangement, i.e. as measured in place or with equivalent terminations.

The phase characteristic ψ can in general be regarded as the sum of three components. The first is the minimum phase shift component, ψ^0, which has a definite relation to the amplitude characteristic of the system, and is of particular interest in connection with phase distortion with different types of amplitude characteristics. The second is a

linear component $\omega\tau_d$, which represents a constant transmission delay τ_d for all frequencies, as in the case of an ideal delay network. Ladder type structures and transmission lines have phase characteristics which can be represented by the above two components. The third component can be represented by a lattice structure with constant amplitude characteristic but varying phase. Such a network component may be present in a transmission system or may be inserted intentionally for phase equalization, i.e. to supplement the first component above so as to secure a linear phase characteristic without altering the amplitude characteristic of the system.

The following discussion is concerned with the relationship of the first component to the amplitude characteristic of the system, or conversely.

The natural logarithm of the transmission-frequency characteristic given by (1.01) is

$$\ell nT(i\omega) = \ell nA(\omega) - i\psi(\omega). \tag{1.03}$$

The component $\ell nA(\omega)$ is referred to as the attenuation characteristic, and when expressed in decibels equals $8.69\ lnA(\omega)$.

The following relations exist between the attenuation and phase characteristics of minimum phase shift systems or system components:[1,2]

$$\ell nA(\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\psi^0(u)}{\omega - u}\,du = \frac{2}{\pi}\int_0^{\infty}\frac{u\psi^0(u)}{u^2 - \omega^2}\,du, \tag{1.04}$$

and

$$\psi^0(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\ell nA(u)}{\omega - u}\,du = -\frac{2}{\pi}\int_0^{\infty}\frac{\omega\ell nA(u)}{u^2 - \omega^2}\,du. \tag{1.05}$$

In the evaluation of these integrals, the principal values are to be used, i.e., results of the form $\ell n(-u)$ are to be taken as $ln\,|-u|$ rather than $\ell n\,|u| + i\pi$.

As an example consider an attenuation characteristic as shown in Fig. 1, with $A(\omega) = A_0$ between $\omega = 0$ and ω_c and A_1 between $\omega = \omega_c$ and ∞. Equation (1.05) then becomes

$$\begin{aligned}\psi^0(\omega) &= -\frac{2\omega}{\pi}\left[\ell nA_0\int_0^{\omega c}\frac{du}{u^2 - \omega^2} + \ell nA_1\int_{\omega_c}^{\infty}\frac{du}{u^2 - \omega^2}\right],\\ &= \frac{1}{\pi}\,\ell n(A_0/A_1)\ell n\left|\frac{\omega_c + \omega}{\omega_c - \omega}\right|.\end{aligned} \tag{1.06}$$

In Fig. 1 is shown the phase characteristic for $A_0/A_1 = 100$, corresponding to a 40 db cutoff at $\omega = \omega_c$.

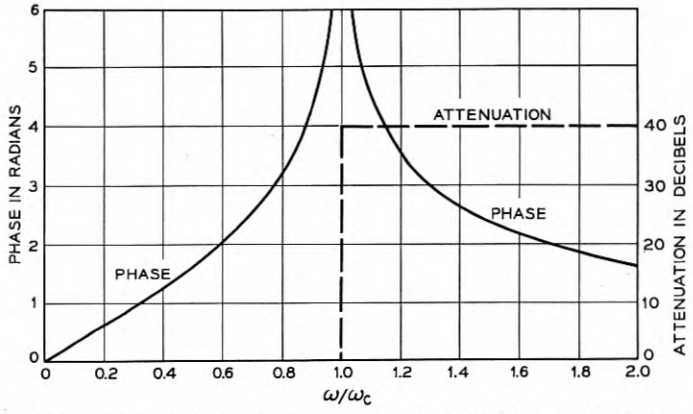


Fig. 1 — Low-pass transmission frequency characteristic with sharp cut-off.

In Fig. 2 the attenuation and phase characteristics are shown as a function of ω/ω_c for $\omega < \omega_c$ and as a function of the inverse ratio ω_c/ω for $\omega > \omega_c$. It will be noticed that for the above case the phase characteristic is infinite for $\omega/\omega_c = 1$ and has even symmetry about this point, while the attenuation characteristic has odd symmetry with respect to the midpoint of the amplitude discontinuity. The phase characteristic may be modified by a gradual cutoff in the attenuation characteristic, as illustrated in the figure. It is possible to shape the attenuation characteristic to obtain a linear phase characteristic in the transmission band, i.e. between $\omega/\omega_c = 0$ and 1. Since transmission systems with a

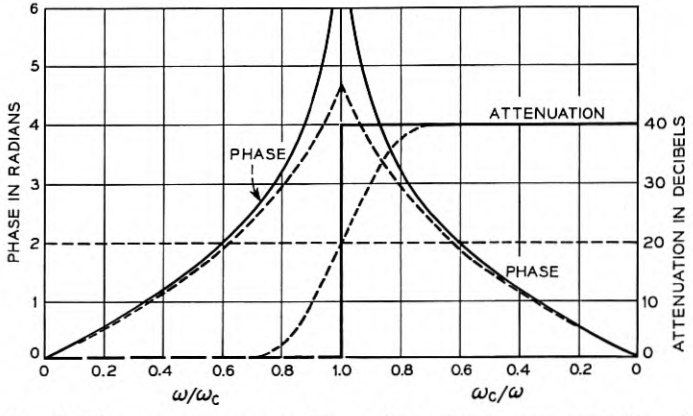


Fig. 2 — Solid curves same as in Fig. 1, but with inverse scale for $\omega/\omega_c > 1$. Dashed curves illustrate modification in phase characteristic with gradual cut-off in attenuation (not computed).

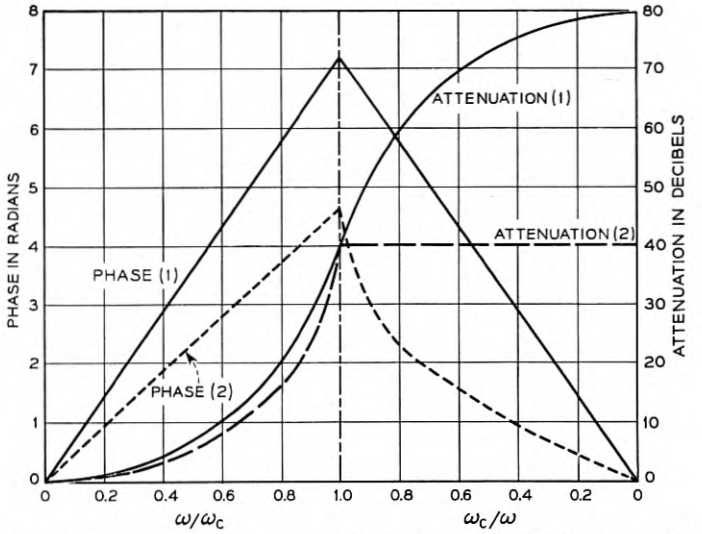


Fig. 3 — Low-pass transmission frequency characteristics with natural linear phase shift for $\omega/\omega_c < 1$.

linear phase characteristic in this range are of particular importance in pulse transmission, this case will be considered further.

It will be assumed that the phase characteristic has even symmetry when expressed in the scales of Fig. 2, in which case the phase characteristic as shown by the solid lines in Fig. 3 is given by

$$\begin{aligned} \psi^0(\omega) &= \omega\tau & \omega/\omega_c < 1, \\ &= \omega_c^2\tau/\omega & \omega/\omega_c > 1. \end{aligned} \tag{1.07}$$

With these expressions in (1.04) the attenuation characteristic becomes:

$$\ell nA(\omega) = \frac{2\omega_c\tau}{\pi}\left[1 + \frac{1}{2}\left(\frac{\omega_c}{\omega} - \frac{\omega}{\omega_c}\right)\ell n\,\frac{1 + \omega/\omega_c}{1 - \omega/\omega_c}\right].$$

For $\omega = 0$, the latter expression approaches the limit $\ell nA(0) = 4\omega_c\tau/\pi$, so that

$$\ell nA(\omega)/A(0) = -\frac{2\omega_c\tau}{\pi}\left[1 + \frac{1}{2}\left(\frac{\omega}{\omega_c} - \frac{\omega_c}{\omega}\right)\ell n\,\frac{1 + \omega/\omega_c}{1 - \omega/\omega_c}\right]. \tag{1.08}$$

which is the attenuation characteristic shown in Fig. 3.

Other attenuation characteristics with a linear phase characteristic between $\omega/\omega_c = 0$ and 1 are possible with other types of variations in the attenuation or phase characteristic for $\omega/\omega_c > 1$ than assumed above. For example, the attenuation characteristics may be assumed

constant for $\omega/\omega_c > 1$, in which case the attenuation characteristic will be somewhat different for $\omega/\omega_c < 1$ and the phase characteristic different for $\omega/\omega_c > 1$, as illustrated in Fig. 3. (The solution for the latter case is given in Reference 2.) It will be noticed that there is a comparatively minor difference between the attenuation characteristics for $\omega/\omega_c < 1$ in the above cases, so that the attenuation characteristic for $\omega/\omega_c > 1$ has a relatively minor effect, provided there is no discontinuity near $\omega/\omega_c = 1$. The transmission loss characteristics shown in Fig. 3 represent a close approximation to the type of characteristic employed in pulse transmission systems, as will be shown later.

In the above examples low-pass characteristics were assumed. For high-pass characteristics the algebraic sign of the phase is reversed with respect to the amplitude characteristic as indicated in Fig. 4, which also illustrates relationships for band-pass characteristics. The band-pass characteristics are obtained by connecting low-pass and high-pass networks in tandem. The resultant attenuation and phase characteristics are obtained by adding the low and high-pass attenuation and phase characteristics, as illustrated in the figure. In the second case shown in the figure, the band-pass characteristic is assumed to have a linear phase characteristic in the transmission band, in which case the attenuation characteristic will not be symmetrical about the midband frequency, unless the latter is high in relation to the bandwidth. The third case illustrates the type of band-pass characteristic encountered in wire systems with a low-frequency cutoff. There will then be phase distortion at the low end of the band, since it is not feasible with a fairly sharp low-frequency cutoff to obtain a linear phase characteristic in the transmission band.

If the amplitude or attenuation characteristic of a transmission system is modified, it will be accompanied by a modification in the phase characteristic. Of basic importance are cosine modifications in the attenuation and amplitude characteristics. Let the modified amplitude characteristic be of the form

$$A(\omega) = A_0(\omega)\, e^{a \cos \omega\tau}, \tag{1.09}$$

where $A_0(\omega)$ is the original amplitude characteristic. The modified attenuation characteristic is then

$$\ell nA(\omega) = \ell nA_0(\omega) + a \cos \omega\tau. \tag{1.10}$$

In accordance with (1.05) the modified phase characteristic becomes,

$$\begin{aligned} \psi^0(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ell nA_0(\omega)}{\omega - u}\, du + \frac{a}{\pi} \int_{-\infty}^{\infty} \frac{\cos \omega\tau}{\omega - u}\, du, \\ &= \psi_0(\omega) + a \sin \omega\tau, \end{aligned} \tag{1.11}$$

where $\psi_0(\omega)$ is the phase characteristic of the original amplitude characteristic $A_0(\omega)$.

Thus, for any consine modification in the attenuation characteristic there is a corresponding sine modification in the phase characteristic, and for any sine modification in the phase characteristic a corresponding cosine modification in the attenuation characteristic. In general any modification in the attenuation characteristic may be represented by a Fourier cosine series, in which case the modification in the phase characteristic will be the corresponding Fourier sine series.

With a cosine modification in the amplitude rather than in the at-

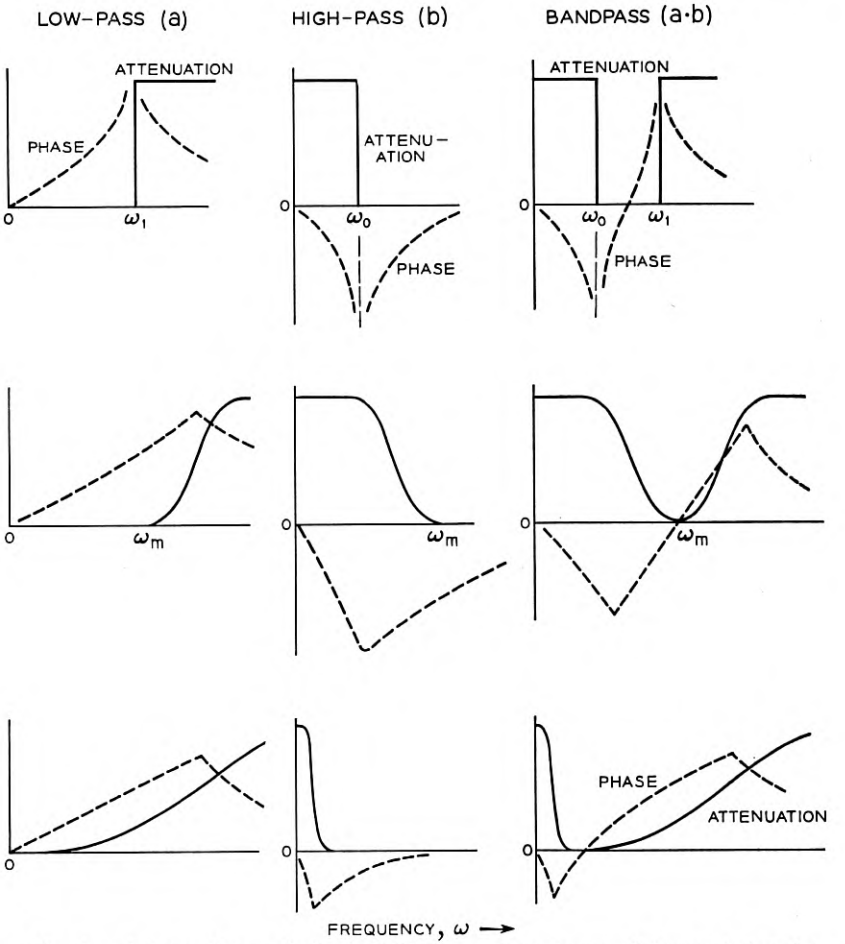


Fig. 4 — Attenuation and phase shift for various types of transmission frequency characteristics.

tenuation characteristic

$$A(\omega) = A_0(\omega)\,[1 + a \cos \omega\tau], \tag{1.12}$$

and the corresponding phase characteristic becomes

$$\begin{aligned} \psi^0(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ell n A_0(u)[1 + a \cos u\tau]}{\omega - u}\, du, \\ &= \psi_0(\omega) + 2 \tan^{-1} \frac{r \sin \omega\tau}{1 + r \cos \omega\tau}, \\ &= \psi_0(\omega) + 2[r \sin \omega\tau + \frac{r^2}{2} \sin 2\,\omega\tau \\ &\qquad + \frac{r^3}{3} \sin 3\,\omega\tau + \ldots \end{aligned} \tag{1.13}$$

where

$$r = \frac{1}{a}\,[1 \mp \sqrt{1 - a^2}], \tag{1.14}$$

and the minus sign is to be used.

Thus, a cosine modification in the amplitude characteristic is accompanied by an infinite series of sine deviations in the phase characteristic. For sufficiently small values of a, $r \cong a/2$ and (1.13) reduces to (1.11).

2. FREQUENCY AND IMPULSE TRANSMISSION CHARACTERISTICS

In dealing with pulse transmission, it is customary to consider three basic types of time variations of currents and electromotive forces, a cisoidal variation, a unit impulse and a unit step. The cisoidal variation, $e^{i\omega t}$, is basic in the solution of network and transmission problems in terms of complex impedances and admittances. The unit impulse is a current or electromotive force of very high intensity and short duration, such that the area under the impulse is unity. The unit step is a current or electromotive force which is zero for $t < 0$ and unity thereafter.

The time responses of networks or transmission systems to these three basic time functions are interrelated so that each may be obtained when one of the others is known. Furthermore, the time responses for electromotive forces or currents of arbitrary wave shape may be obtained from the response characteristic for any one of these basic time functions.

The pulses applied in pulse systems can usually be approximated by impulses. Furthermore, with impulses certain simple relationships can be established which are either obscured or more complicated when a

unit step is assumed. For these reasons, only the transmission characteristic for impulses will be considered here, or for pulses of sufficiently short duration to be regarded as impulses.

Corresponding to any transmission-frequency characteristic is an impulse transmission characteristic, $P(t)$, which designates the received pulse as a function of time for a transmitted unit impulse. The impulse and transmission frequency characteristics are interrelated by the following Fourier integral relations

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T(i\omega) e^{i\omega t} \, d\omega, \tag{2.01}$$

$$T(i\omega) = \int_{-\infty}^{\infty} P(t) e^{-i\omega t} \, dt. \tag{2.02}$$

The transmission characteristic for an applied pulse or signal of arbitrary shape $G(t)$ is given by

$$H(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T(i\omega) S(i\omega) e^{i\omega t} \, d\omega, \tag{2.03}$$

where $S(i\omega)$ is the frequency spectrum of the applied pulse and is given by

$$S(i\omega) = \int_{-\infty}^{\infty} G(t) e^{-i\omega t} \, dt. \tag{2.04}$$

In the case of a symmetrical pulse $S(i\omega)$ is a real function.

In view of (1.01), expression (2.03) may also be written

$$H(t) = \frac{1}{\pi} \int_{0}^{\infty} A(\omega) S(\omega) \cos [\omega t - \psi(\omega)] \, d\omega, \tag{2.05}$$

where the relations $A(-\omega) = A(\omega)$, $S(-\omega) = S(\omega)$, $\psi(-\omega) = -\psi(\omega)$ have been used, and it is assumed that $S(i\omega) = S(\omega)$ is a real function, as for a symmetrical pulse.

In most pulse transmission systems, the applied pulses can be approximated by short rectangular pulses. Rectangular pulses of unit amplitude and duration δ have a frequency spectrum

$$S(\omega) = \delta \frac{\sin \omega\delta/2}{\omega\delta/2}. \tag{2.06}$$

The same pulse transmission characteristic as when an impulse is applied is obtained with a rectangular pulse if $A(\omega)$ is modified by the factor $(\omega\delta/2)/\sin(\omega\delta/2)$. In the following it will be assumed that the applied pulses are of sufficiently short duration to be regarded as im-

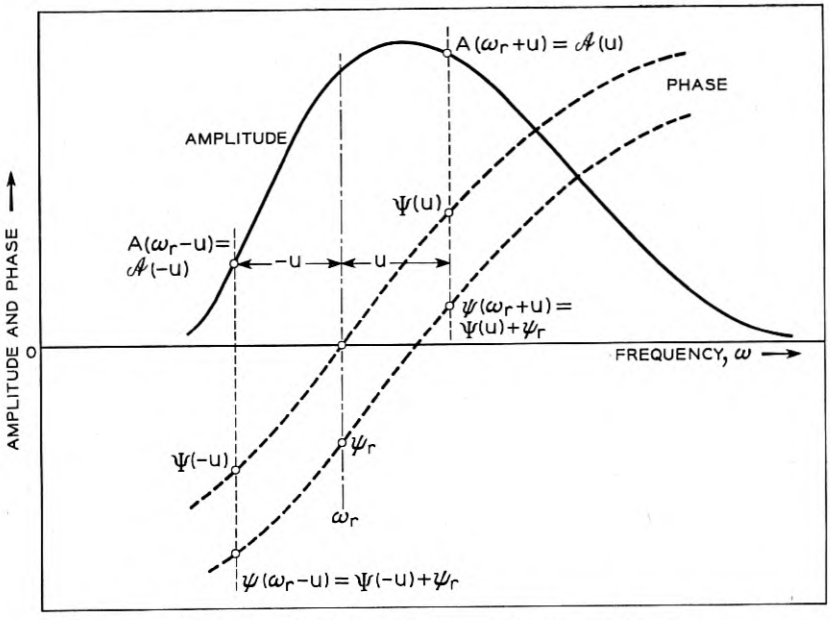


Fig. 5 — Transfer of reference frequency from $\omega = 0$ to $\omega = \omega_r$.

pulses or that otherwise the above modification is applied, in which case

$$P(t) = \frac{\delta}{\pi} \int_0^\infty A(\omega) \cos [\omega t - \psi(\omega)]\, d\omega. \tag{2.07}$$

In the latter equation $A(\omega)$ can also be regarded as the frequency spectrum of a pulse applied to a transmission system having a constant amplitude characteristic and a phase characteristic $\psi(\omega)$ over the band of the pulse spectum.

Equation (2.07) applies to any type of transmission-frequency characteristic and is convenient in this form for low-pass characteristics. For band-pass characteristics as shown in Fig. 5 however, it is convenient from the standpoint of general analysis as well as for numerical evaluation to use a reference frequency ω_r within the tranmsission band, that is, to employ the transformation $\omega = \omega_r + u \quad d\omega = du$.

With the notation

$$\begin{aligned} \mathfrak{A}(u) &= A(\omega) = A(u + \omega_r), \\ \Psi(u) &= \psi(\omega) - \psi(\omega_r) = \psi(\omega) - \psi_r, \end{aligned} \tag{2.08}$$

equation (2.07) can be written:

$$P(t) = \cos(\omega_r t - \psi_r)[R_-(t) + R_+(t)] + \sin(\omega_r t - \psi_r)[Q_-(t) - Q_+(t)]. \tag{2.09}$$

$$R_- = \frac{\delta}{\pi}\int_0^{\omega_r} \mathfrak{A}(-u)\cos[ut + \Psi(-u)]\,du, \tag{2.10}$$

$$R_+ = \frac{\delta}{\pi}\int_0^{\infty} \mathfrak{A}(u)\cos[ut - \Psi(u)]\,du,$$

$$Q_- = \frac{\delta}{\pi}\int_0^{\omega_r} \mathfrak{A}(-u)\sin[ut + \Psi(-u)]\,du, \quad \text{and} \tag{2.11}$$

$$Q_+ = \frac{\delta}{\pi}\int_0^{\infty} \mathfrak{A}(u)\sin[ut - \Psi(u)]\,du.$$

The envelope $\overline{P}(t)$ of the impulse transmission characteristic is given by

$$\overline{P}(t) = [(R_- + R_+)^2 + (Q_- - Q_+)^2]^{1/2}. \tag{2.12}$$

Comparison of (2.09) with (2.07) shows that R_- and R_+ can be identified with the impulse characteristics of low-pass systems having the same frequency characteristics as the bandpass system below and above ω_r. The impulse characteristics Q_- and Q_+ which arise from asymmetry in the transmission characteristic with respect to ω_r are not present in low-pass systems, since by definition the amplitude characteristic has even symmetry and the phase characteristic odd symmetry with respect to zero frequency.

The first and second components of (2.09) are referred to as the in-phase and quadrature components of the impulse characteristic of band-pass systems.[3] The transmission-frequency characteristic may correspondingly be regarded as made up of a component with even symmetry and another component with odd symmetry about ω_r, as indicated in Fig. 6. These two components, together with the in-phase and quadrature components, will depend on the choice of ω_r. However, $P(t)$ as given by (2.09) and the envelope as given by (2.12), will remain the same, since a single impulse characteristic is associated with a given transmission-frequency characteristic.

With the customary pulse transmission methods, the reference frequency ω_r may be identified with a modulating or carrier frequency, which has a special significance when the envelope of a sequence of received pulses is considered. Although for a single pulse the envelope

is always the same, for a sequence of pulses the resultant envelope of the received pulse train will depend on the in-phase and quadrature components.[4] The reason for this is that one has even and the other odd symmetry about the peak amplitude of the envelope for a single pulse, when the phase characteristic is linear.

In order to compare the transmission performance as the reference or carrier frequency is changed, it is necessary to determine the in-phase and quadrature components for each carrier frequency under considera-

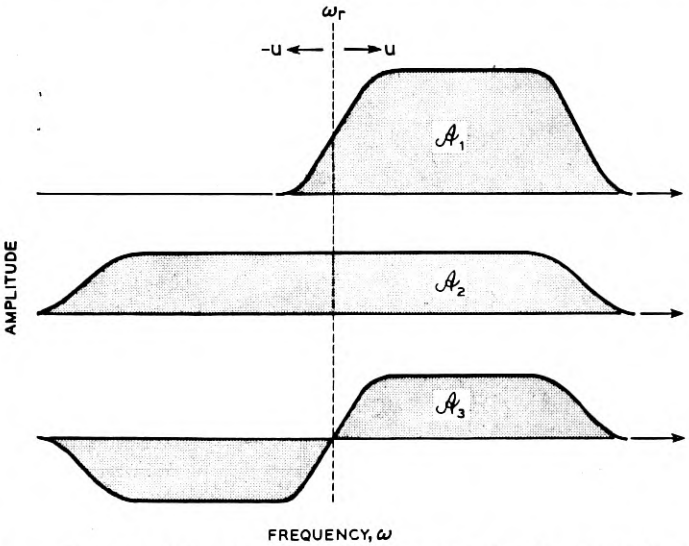


Fig. 6 — Decomposition of amplitude characteristic $\mathcal{A}_1$ asymmetrical with respect to ω_r into a component $\mathcal{A}_2$ of even symmetry and a component $\mathcal{A}_3$ of odd symmetry about ω_r. When the phase shift is linear, $\mathcal{A}_1 = \mathcal{A}_2 + \mathcal{A}_3$.

tion. One method is to evaluate integrals (2.10) and (2.11) for each carrie frequency, which may be facilitated by resolving the transmission-frequency characteristic into symmetrical and anti-symmetrical components as indicated in Fig. 6. This, however, is a rather elaborate procedure which can be avoided with the aid of a simple translation from one reference or carrier frequency to another, as shown below, provided the in-phase and quadrature components or the envelope has been determined for one reference frequency.

Equation (2.09) may also be written, with $\varphi = \varphi(t)$:

$$\begin{aligned} P(t) &= \cos(\omega_r t - \psi_r - \varphi)\, \overline{P}(t), \\ &= \cos(\omega_r t - \psi_r) \cos\varphi\, \overline{P}(t) + \sin(\omega_r t - \psi_r) \sin\varphi\, \overline{P}(t). \end{aligned} \tag{2.13}$$

Comparison of (2.13) with (2.09) shows that:

$$\begin{aligned} R_- + R_+ &= \cos\varphi\, \overline{P}(t), \\ Q_- - Q_+ &= \sin\varphi\, \overline{P}(t), \end{aligned} \tag{2.14}$$

$$\tan\varphi = (Q_- - Q_+)/(R_- + R_+). \tag{2.15}$$

To find the corresponding components when ω_r is changed to ω_r', equation (2.13) may be written

$$\begin{aligned} P(t) &= \cos[\omega_r' t - \psi_r' - (\omega_r' - \omega_r)t + (\psi_r' - \psi_r) - \varphi]\, \overline{P}(t) \\ &= \cos(\omega_r' t - \psi_r' - \varphi')\, \overline{P}(t), \end{aligned} \tag{2.16}$$

where $\varphi' = \varphi'(t)$ is given by:

$$\begin{aligned} \varphi' &= \varphi + (\omega_r' - \omega_r)t - (\psi_r' - \psi_r), \\ &= \varphi + \omega_y t - \psi_y\,. \end{aligned} \tag{2.17}$$

Thus, when the reference frequency is changed by ω_y and its phase by ψ_y, the corresponding in-phase and quadrature components become:

$$\begin{aligned} R_-' + R_+' &= \cos(\varphi + \omega_y t - \psi_y)\, \overline{P}(t), \qquad \text{and} \\ Q_-' - Q_+' &= \sin(\varphi + \omega_y t - \psi_y)\, \overline{P}(t) \end{aligned} \tag{2.18}$$

To summarize, when the in-phase and quadrature components have been determined for any reference frequency ω_r from (2.10) and (2.11), and the envelope $\overline{P}$ together with the function φ from (2.12) and (2.14), the in-phase and quadrature components for another reference frequency ω_r' can readily be determined with the aid of (2.18). In the particular case where the amplitude characteristic has even and the phase characteristic odd symmetry with respect to the midband frequency, the quadrature component disappears with respect to the midband frequency, so that $\varphi = 0$ and (2.18) simplifies to

$$\begin{aligned} R_-' + R_+' &= \cos(\omega_y t - \psi_y)\, \overline{P}(t), \qquad \text{and} \\ Q_-' - Q_+' &= \sin(\omega_y t - \psi_y)\, \overline{P}(t). \end{aligned} \tag{2.19}$$

The above relations (2.18) and (2.19) facilitate comparison of transmission performance as the reference or carrier frequency is changed, for example the comparison of double with vestigial sideband transmission, as illustrated in section 14.

3. IDEALIZED CHARACTERISTICS WITH SHARP CUTOFF

In pulse transmission theory, particularly in dealing with transmission capacity of idealized transmission systems, an ideal low-pass transmission frequency characteristic is ordinarily assumed, with constant amplitude and delay in the transmission band together with an abrupt cutoff at the top frequency and zero amplitude beyond, as shown in Fig. 7. As is evident from Fig. 1, this type of characteristic is an abstraction which cannot be physically realized since it will have phase distortion and infinite transmission delay. It can, however, be approached with sufficiently elaborate phase equalization.

For the above type of characteristic, $A(\omega) = 1$ between $\omega = 0$ and ω_1, while $\psi(\omega) = \omega\tau_d$, where τ_d is the transmission delay. With these values in (2.07):

$$P(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_0}{\omega_1 t_0}, \tag{3.01}$$

where $t_0 = t - \tau_d$ is the time referred to the peak amplitude of the received pulse.

The resultant pulse transmission characteristic is shown in Fig. 7, with the factor $\delta\omega_1/\pi$ omitted. The peak amplitude is attained after an infinite time, since the above type of characteristic can be realized only with $\tau_d \to \infty$. The impulse characteristic is zero when $\omega_1 t_0 = \pm n\pi$, or $t_0 = \pm\tau_1, \pm 2\tau_1, \cdots \pm n\tau_1$ where

$$\tau_1 = \frac{1}{2f_1}. \tag{3.02}$$

Impulses can thus be transmitted at the latter intervals without

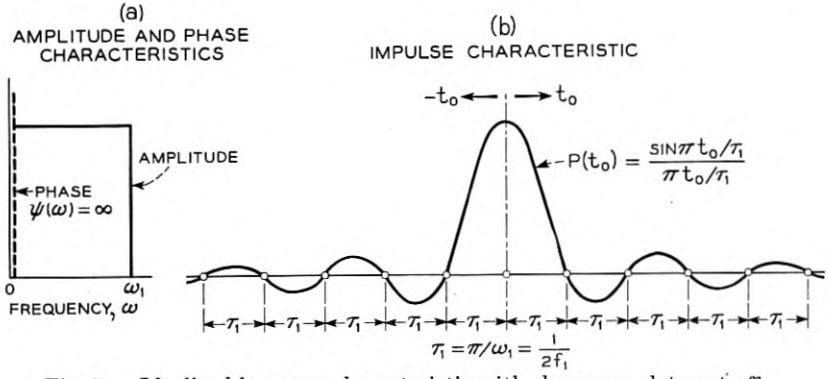


Fig. 7 — Idealized low-pass characteristic with sharp complete cut-off.

mutual interference between the peaks of the received pulses. This is a basic theorem underlying the determination of the transmission capacity of idealized systems.[3]

For an idealized bandpass characteristic between ω_0 and ω_1, it follows from (2.09) with $\Psi(u) = u\tau_d$ and $\Psi(-u) = -u\tau_d$ that the impulse characteristic with respect to the midband frequency $\omega_r = \omega_m$ is

$$P(t) = 2\cos[\omega_m t_0 - \psi_0]\,\overline{P}(t), \tag{3.03}$$

where $\overline{P}(t)$ is given by (3.01) and $\psi_0 = \Psi_m - \omega_m\tau_d$ is the phase intercept at zero frequency. For the transmission characteristic to be ideal in the sense that the peak pulse amplitude occurs when $t_0 = t - \tau_d = 0$, it is necessary that $\psi_0 = \pm n\pi$, where n is an integer. This is not necessary if the bandwidth is small in relation to the midband frequency. There will then be a large number of cycles of the modulating frequency ω_m within the envelope $\overline{P}(t)$, and the latter can be recovered by envelope detection regardless of the phase of the modulating frequency.

With $\psi_0 = \pm n\pi$,

$$P(t) = \frac{2\omega_s\delta}{\pi}\cos\omega_m t_0\,\frac{\sin\omega_s t_0}{\omega_s t_0}, \tag{3.04}$$

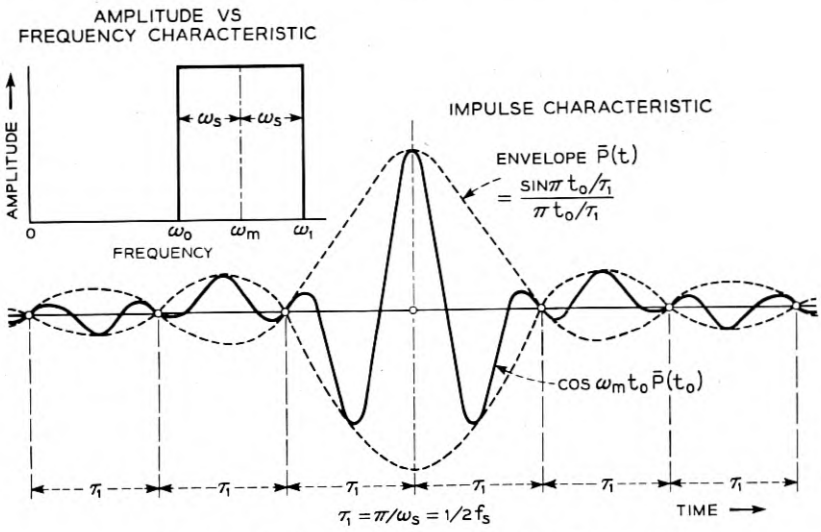
$$= \frac{\omega_1\delta}{\pi}\,\frac{\sin\omega_1 t_0}{\omega_1 t_0} - \frac{\omega_0\delta}{\pi}\,\frac{\sin\omega_0 t_0}{\omega_0 t_0}, \tag{3.05}$$

where $\omega_m = (\omega_0 + \omega_1)/2$ and $\omega_s = (\omega_1 - \omega_0)/2$.

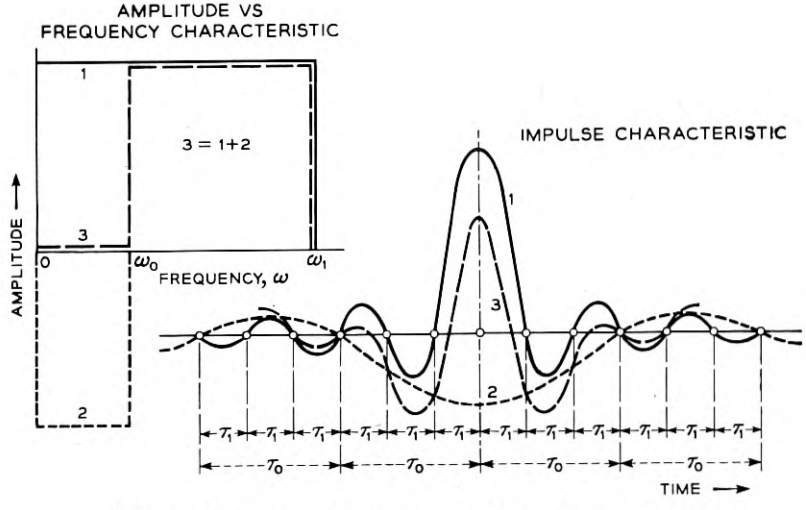
The shape of the impulse characteristic as given by (3.04) is illustrated in the upper half of Fig. 8. Alternately the impulse characteristic may be regarded as made up of two components in accordance with (3.05). The first component corresponds to a low-pass characteristic of bandwidth ω_1, the second component to a negative low-pass characteristic of bandwidth ω_0, as indicated in the lower part of the Fig. 8.

The factor $\sin\omega_s t_0/\omega_s t_0$ in (3.04) is zero at the same intervals as for a low-pass characteristic of bandwidth ω_s, as shown in Fig. 8, so that pulses may be transmitted at the same rate without mutual interference between pulse peaks. The bandwidth in the present case, however, is $2\omega_s = \omega_1 - \omega_0$, so that for the same bandwidth the pulse transmission rate is half as great as for a low-pass characteristic.

An exception to this is the particular case when $\omega_1 = 2\omega_0$, so that the total bandwidth is ω_0. The factor $\sin\omega_0 t_0/\omega_0 t_0$ in (3.05) is then zero at intervals $\tau_0 = 1/2f_0$, while the factor $\sin\omega_1 t_0/\omega_1 t_0$ is zero at intervals $1/2f_1 = 1/4f_0$, as shown in Fig. 9. Pulses may accordingly in principle be



(a) REPRESENTATION OF IMPULSE CHARACTERISTIC AS ENVELOPE MODULATED BY MIDBAND FREQUENCY



(b) REPRESENTATION OF AMPLITUDE VS FREQUENCY AND IMPULSE CHARACTERISTICS, 3, AS THE SUM OF A POSITIVE LOW-PASS CHARACTERISTIC, 1, AND A NEGATIVE LOW-PASS CHARACTERISTIC, 2

Fig. 8 — Idealized band-pass characteristics and corresponding impulse transmission characteristics.

transmitted without mutual interference at the same rate as for a low-pass characteristic of bandwidth ω_0, or at the same rate as with single sideband transmission over a band-pass system of bandwidth ω_0. More generally, pulses can in principle be transmitted without mutual interference between pulse peaks at the same rate as for a low-pass characteristic of bandwidth $\omega_1 - \omega_0 = 2\omega_s$ if ω_0 is a multiple of $\omega_1 - \omega_0$. It should be noted however, that this pulse transmission rate cannot actually be realized since the phase characteristic will have infinite slope, so that the transmission delay will be infinite. In addition, the zero frequency phase intercept ψ_0 must be $\pm n\pi$, a condition which cannot be attained or remain stable in view of the infinite slope of the phase characteristic.

With the envelope given by the factor $\sin \omega_s t_0/\omega_s t_0$ in (3.04), the in-phase and quadrature components for any reference frequency can be determined with the aid of (2.19). If the lower band-edge is selected, i.e. $\omega_r = \omega_0$, then $\omega_y = \omega_s$. With a linear phase characteristic $\psi_y = \omega\tau_d$, so that in (2.19) $\omega_y t - \psi_y = \omega_s t_0$. The in-phase and quadrature components are accordingly obtained by multiplying the envelope by $\cos \omega_s t_0$ and $\sin \omega_s t_0$, respectively.

As an alternate method, the two components can be obtained from

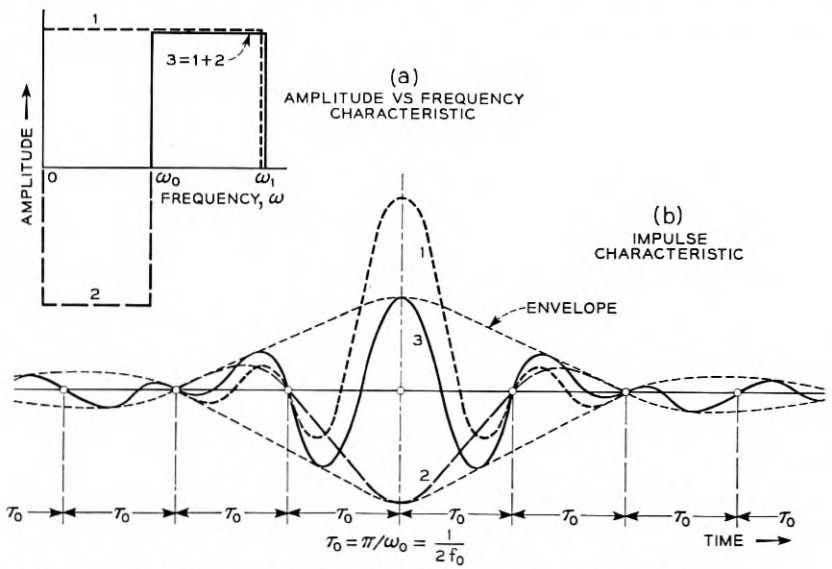


Fig. 9 — Special case of idealized band-pass characteristic in which $\omega_1 = 2\omega_0$ and resultant impulse characteristic is zero at intervals $\tau_0 = \dfrac{1}{2f_0}$.

(2.09), which with $R_- = 0$, $Q_- = 0$ becomes:

$$P(t) = \cos \omega_0 t_0 R_+(t) + \sin \omega_0 t_0 Q_+(t), \tag{3.06}$$

with

$$\begin{aligned} R_+ &= \frac{\delta}{\pi} \int_0^{\omega_b} \cos ut_0 \, du, \\ &= \frac{\delta\omega_b}{\pi} \frac{\sin \omega_b t_0}{\omega_b t_0} = \frac{2\delta\omega_s}{\pi} \cos \omega_s t_0 \frac{\sin \omega_s t_0}{\omega_s t_0}, \quad \text{and} \end{aligned} \tag{3.07}$$

$$\begin{aligned} Q_+ &= \frac{\delta}{\pi} \int_0^{\omega_b} \sin ut_0 \, du, \\ &= \frac{\delta\omega_b}{\pi} \frac{1 - \cos \omega_b t_0}{\omega_b t_0} = \frac{2\delta\omega_s}{\pi} \sin \omega_s t_0 \frac{\sin \omega_s t_0}{\omega_s t_0}, \end{aligned} \tag{3.08}$$

where $\omega_b = 2\omega_s$ is the bandwidth. It will be noticed that R_+ and Q_+ are obtained by multiplying the envelope by $\cos \omega_s t_0$ and $\sin \omega_s t_0$ in accordance with (2.19).

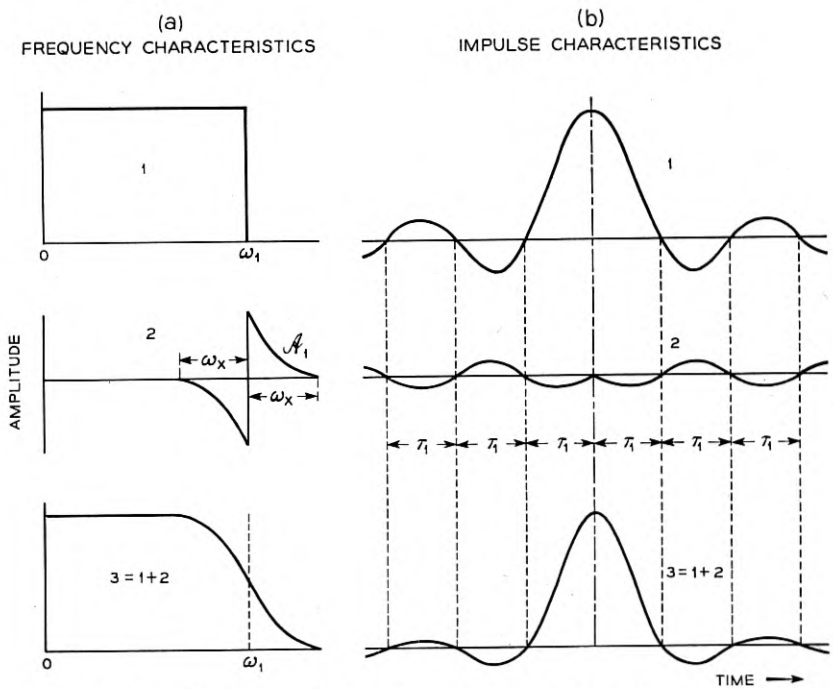


Fig. 10 — Idealized transmission characteristic with gradual cut-off, 3, obtained by superposition of characteristic with sharp cut-off, 1, and characteristic, 2, with odd symmetry about ω_1. Linear phase shift assumed.

4. IDEALIZED CHARACTERISTICS WITH GRADUAL CUTOFF

The idealized transmission characteristics discussed above are of principal interest in that they indicate the physical limitations on pulse transmission rates for a given bandwidth. Even if these impulse characteristics could be realized without undue difficulties from the standpoint of phase equalization, they would be impracticable in most applications. Their oscillatory nature would entail the use of discrete pulse positions and precise synchronized sampling at fixed intervals, and would preclude certain methods of pulse modulation and detection.

The non-linearity in the phase characteristic as well as the oscillations in the impulse characteristic can be reduced with a gradual rather than a sharp cut-off, as illustrated in Fig. 10. It is assumed that an ideal characteristic with a sharp cutoff is supplemented by an amplitude characteristic $\mathfrak{A}_1$ which has odd symmetry about the cutoff frequency ω_1, i.e., $\mathfrak{A}_1(-u) = -\mathfrak{A}_1(u)$.

If the latter component alone is considered, and a linear phase characteristic assumed, it follows from (2.09) with $\omega_1 = \omega_r$ that the effect of this component on the pulse transmission characteristic is given by

$$P_1(t) = -Q_1 \sin \omega_1 t_0 , \tag{4.01}$$

where $t_0 = t - \tau_d$ and

$$Q_1 = \frac{2\delta}{\pi} \int_0^{\omega_x} \mathfrak{A}_1(u) \sin ut_0 \, du. \tag{4.02}$$

The function $P_1(t)$ will be zero at the same points as the original pulse transmission characteristic with a sharp cut-off at ω_1 and under certain conditions also at other points. It will modify the original impulse characteristic by reducing the oscillatory tail, as illustrated in Fig. 10, but the zero points remain unchanged.[3]

With the above modification, the resultant impulse characteristic obtained by superposition of (3.01) and (4.02) becomes

$$\begin{aligned} P(t) &= \frac{\delta}{\pi} \sin \omega_1 t_0 \left(\frac{1}{t_0} - 2 \int_0^{\omega_x} \mathfrak{A}_1(u) \sin ut_0 \, du \right), \\ &= \frac{\delta}{\pi} \sin \omega_1 t_0 F(t), \end{aligned} \tag{4.03}$$

where

$$F(t) = \left[\frac{1}{t_0} - 2 \int_0^{\omega_x} \mathfrak{A}_1(u) \sin ut_0 \, du \right]. \tag{4.04}$$

In the following the expression for $F(t)$ is given for the case when the

band-edge is modified by a supplementary characteristic of the form

$$\begin{aligned} \mathfrak{A}_1(u) &= \tfrac{1}{2}(1 - \sin \pi u/2\omega_x) \qquad u < \omega_x , \\ &= 0 \qquad\qquad\qquad\qquad u > \omega_x . \end{aligned} \tag{4.05}$$

This form of $\mathfrak{A}_1$ represents a close approximation to actual modifications of band-edges by a gradual cutoff and also results in rather simple expressions for the modified impulse characteristic

With (4.05) in (4.04),

$$\begin{aligned} F(t) &= \frac{1}{t_0} - \int_0^{\omega_x} (1 - \sin \pi u/2\omega_x) \sin ut_0 \, du, \\ &= \frac{1}{t_0} - \omega_x \left[\frac{1 - \cos \omega_x t_0}{\omega_x t_0} + \frac{\cos \omega_x t_0}{\pi + 2\omega_x t_0} - \frac{\cos \omega_x t_0}{\pi - 2\omega_x t_0} \right], \\ &= \omega_x \cos \omega_x t_0 \left[\frac{1}{\omega_x t_0} + \frac{1}{\pi - 2\omega_x t_0} - \frac{1}{\pi + 2\omega_x t_0} \right], \\ &= \frac{1}{t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2} . \end{aligned} \tag{4.06}$$

The impulse characteristic obtained from (4.03) is

$$P(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_0}{\omega_1 t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2} \tag{4.07}$$

For the particular case shown in Fig. 11 the value of ω_x is taken to be $\omega_1/2$.

For a symmetrical bandpass characteristic, as shown in Fig. 12,

$$P(t) = 2 \cos (\omega_m t_0 - \psi_0) \, \overline{P}(t). \tag{4.08}$$

$\overline{P}(t)$ is obtained by replacing ω_1 by ω_s in (4.07), and ψ_0 is the phase intercept at zero frequency as in connection with (3.03). This gives

$$\overline{P}(t) = \frac{\delta\omega_s}{\pi} \frac{\sin \omega_s t_0}{\omega_s t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2} . \tag{4.09}$$

For the particular case shown in Fig. 12, the value of ω_x is taken to be $\omega_s/2$.

The in-phase and quadrature components with respect to any frequency are obtained from (2.19) with $\psi_y = \omega\tau_d$ and are shown in Fig. 12 for the particular case in which the reference frequency is displaced from the midband frequency by $\omega_y = \omega_s$.

5. IDEALIZED CHARACTERISTICS WITH NATURAL LINEAR PHASE SHIFT

With the type of amplitude characteristics discussed above it is necessary to employ phase equalization to obtain a linear phase characteristic. Furthermore, oscillations of appreciable amplitude remain in the impulse characteristic. A virtually linear phase characteristic together with a reduction of these oscillations can be attained by a further extension of the gradual cut-off in Fig. 10, such that $\omega_x = \omega_1$. An amplitude characteristic of this type, together with the corresponding impulse

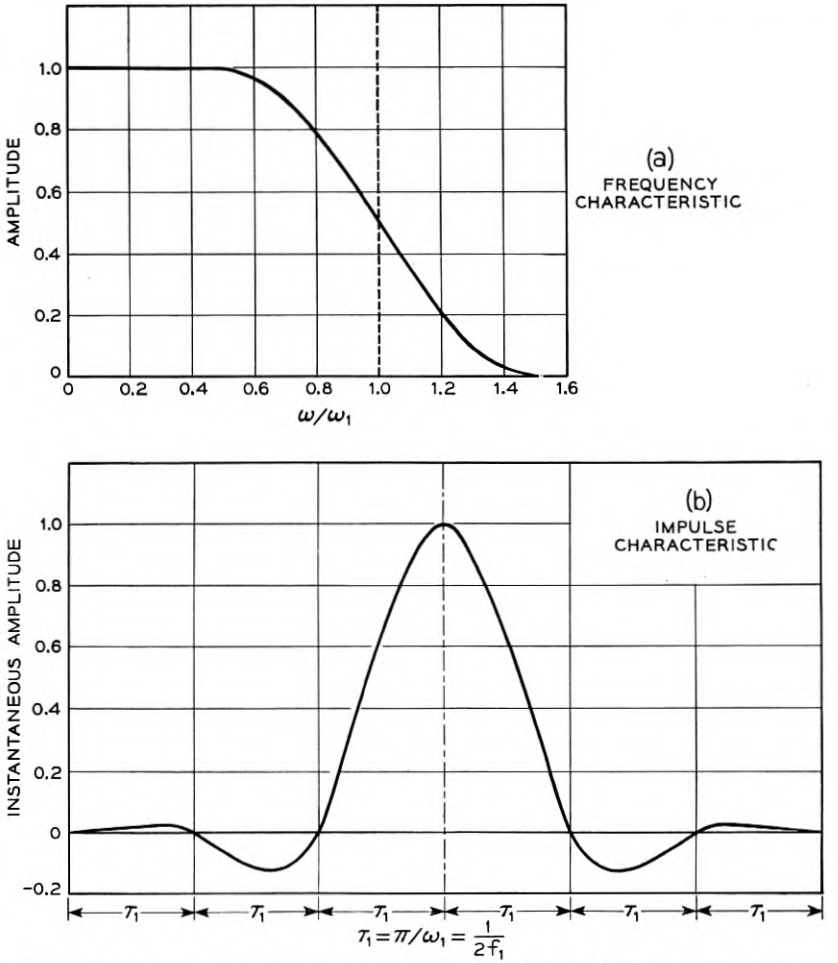
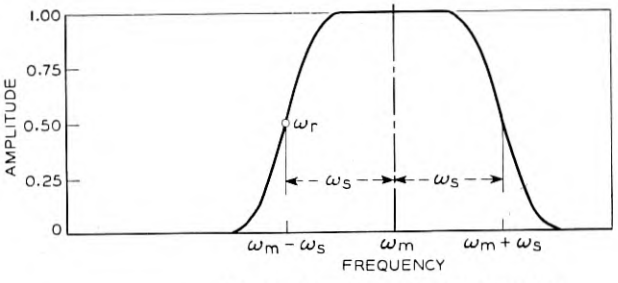
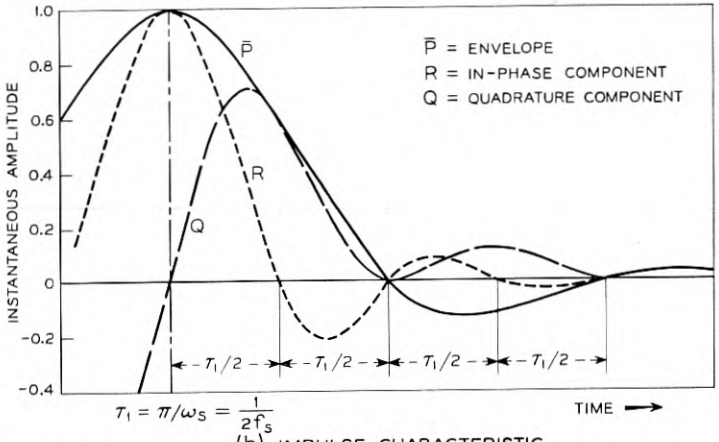


Fig. 11 — Low-pass characteristic with gradual cut-off and associated impulse characteristic. Linear phase characteristic assumed.



(a) FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTIC

Fig. 12 — Symmetrical band-pass characteristic with gradual cut-off and associated impulse characteristic. In-phase and quadrature components shown with respect to $\omega_r = \omega_m - \omega_s$.

characteristic is shown in Fig. 13. The supplementary amplitude characteristic and the impulse characteristic are obtained by making $\omega_x = \omega_1$ in (4.05) and 4.07).

The resultant amplitude characteristic between $\omega = 0$ and $\omega = 2\omega_1$ in this case becomes

$$A(\omega) = \frac{1}{2}\left[1 + \cos\frac{\pi\omega}{2\omega_1}\right] = \cos^2\frac{\pi\omega}{4\omega_1}, \tag{5.01}$$

and the impulse characteristic:

$$P(t) = \frac{\delta\omega_1}{\pi}\,\frac{\sin 2\omega_1 t_0}{2\omega_1 t_0[1 - (2\omega_1 t_0/\pi)^2]}, \tag{5.02}$$

where ω_1 is the bandwidth to the half-amplitude point on the trans-

mission frequency characteristic and $2\omega_1$ the bandwidth to the point of zero amplitude.

In Fig. 13 is also shown the amplitude characteristic given by (1.08), which will have a linear phase characteristic in the transmission band, i.e. from $\omega = 0$ to $2\omega_1$. Because of the close approximation of (5.01) to the proper type of amplitude characteristic as regards phase linearity, the phase characteristic associated with (5.01) may for practical purposes be regarded as linear.

For a symmetrical band-pass characteristic as shown in Fig. 14, the impulse characteristic is given by (4.08) and the envelope by (4.09) with $\omega_x = \omega_s$, or

$$\overline{P}(t) = \frac{\hat{\imath}\omega_s}{\pi} \frac{\sin 2\omega_s t_0}{2\omega_s t_0[1 - (2\omega_s t_0/\pi)^2]}. \tag{5.03}$$

The in-phase and quadrature components shown in Fig. 14 with

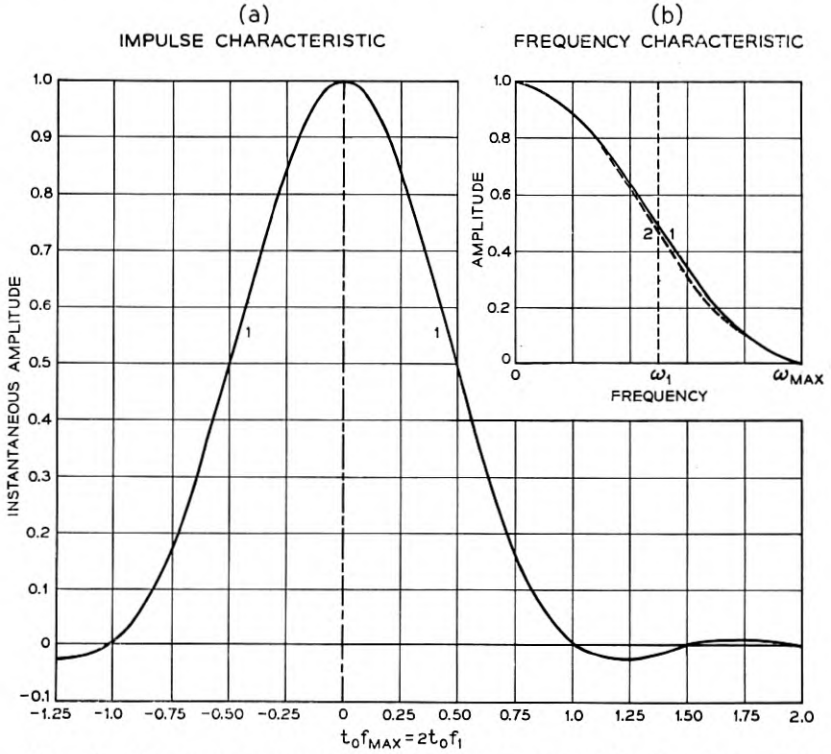


Fig. 13 — Low-pass transmission frequency characteristic, 1, and associated impulse characteristic. Frequency characteristic, 2, is same as shown by solid lines in Fig. 3 and has a linear phase characteristic between $\omega = 0$ and ω_{max}.

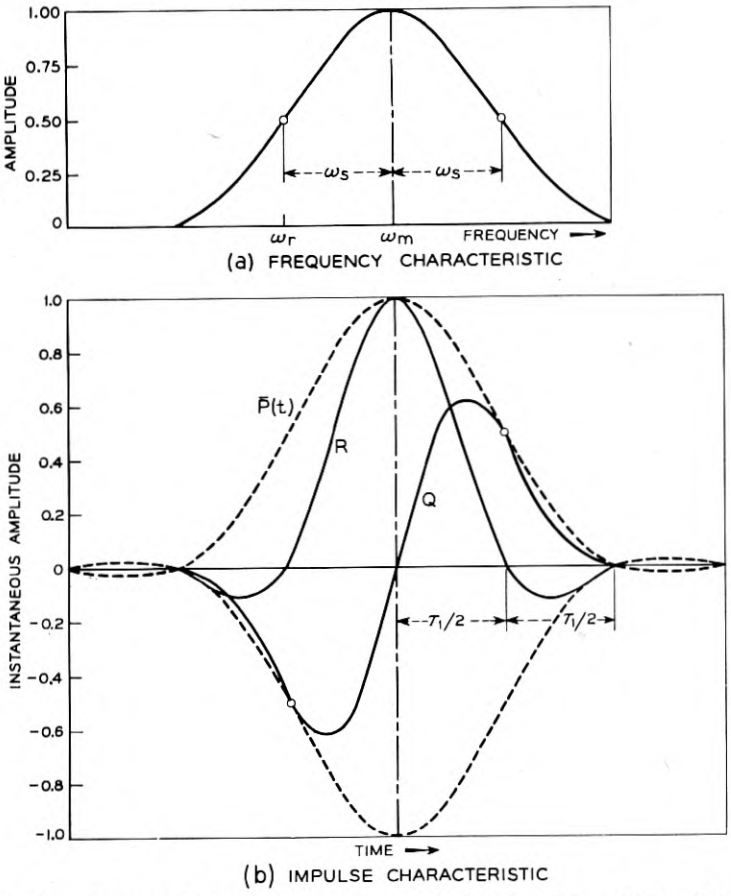


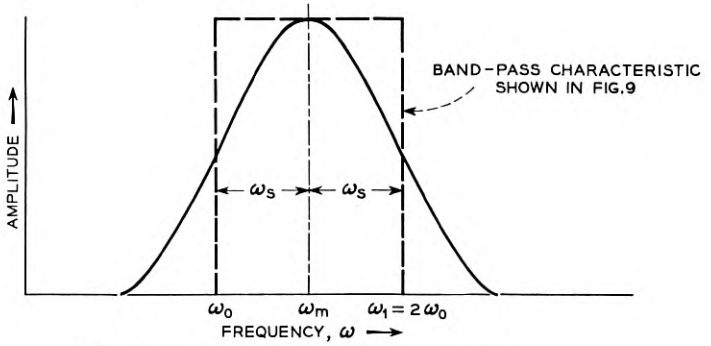
Fig. 14 — Symmetrical band-pass characteristic with linear phase shift and corresponding impulse characteristic. In-phase and quadrature components shown with respect to $\omega_r = \omega_m - \omega_s$.

respect to a reference frequency at the midpoint of the band-edge are obtained from (2.19) with $\omega_y = \omega_s$. This gives $\overline{P}(t) \cos \omega_s t_0$ for the in-phase and $\overline{P} \sin \omega_s t_0$ for the quadrature component.

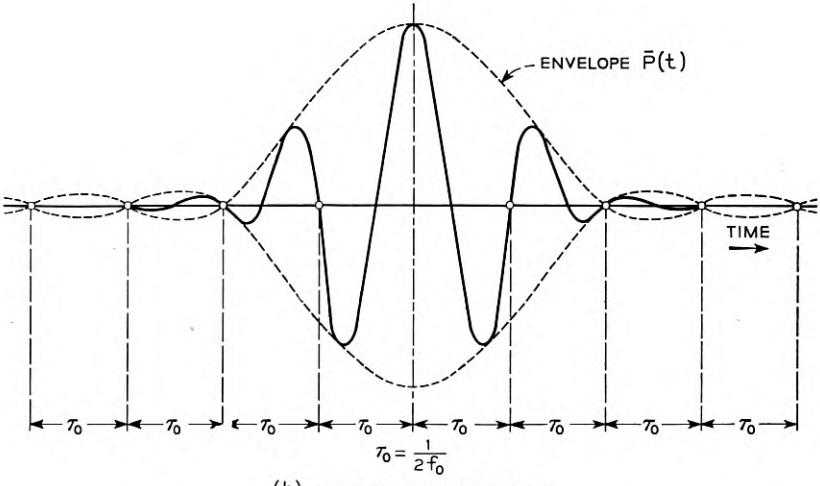
In Fig. 15 is shown a special case of a band-pass characteristic, which corresponds to that illustrated in Fig. 9 with $\omega_1 = 2\omega_0$, shown for comparison by dashed lines in Fig. 15. In this particular case $\omega_m = 3\ \omega_0/2$ and $\omega_x = \omega_s = \omega_0/2$. With $\psi_0 = n\pi$, equation (4.08) in conjunction with (4.09) gives

$$P(t) = \frac{\delta\omega_0}{2\pi} \cos(\omega_0 t_0/2) \frac{\sin 2\omega_0 t_0 - \sin \omega_0 t_0}{\omega_0 t[1 - (\omega_0 t_0/\pi)^2]}. \tag{5.04}$$

This expression is zero when $\sin 2\omega_0 t_0 - \sin \omega_0 t_0 = 0$, and also when $\cos \omega_0 t_0/2 = 0$. Zero points in the impulse characteristic will occur at uniform intervals $\tau_0 = \pi/\omega_0 = 1/2f_0$. Pulses can accordingly be transmitted at these intervals without mutual interference, or at the same rate as for a low-pass characteristic with the bandwidth to the half-amplitude point equal to ω_0. This is the same pulse transmission rate as is possible in principle with an ideal band-pass characteristic as shown



(a) AMPLITUDE VS FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTIC

Fig. 15 — Particular case of band-pass characteristic with gradual cut-off in which impulse characteristic is zero at intervals $\tau_0 = \dfrac{1}{2f_0}$.

by the dashed lines in Fig. 15. With a gradual cut-off, however, the phase characteristic will be nearly linear and have a finite slope, so that the above pulse transmission rate can be realized provided $\psi_0 = \pm n\pi$. The same pulse transmission rate can also be attained with vestigial side-band transmission, discussed in section 14.

Another particular case of interest is that shown in Fig. 16, in which $\omega_m = 2\omega_s$. In this case (4.08) becomes with $\psi_0 = \pm n\pi$ and with $\bar{P}(t)$ as

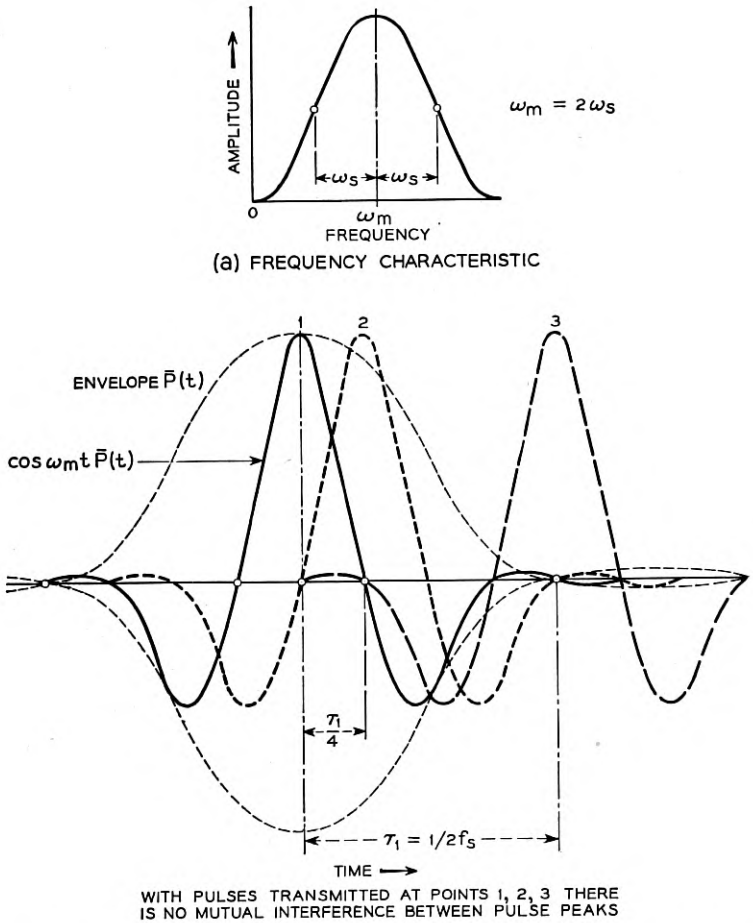


Fig. 16 — Particular case of symmetrical band-pass characteristic for which $\omega_m = 2\omega_s$.

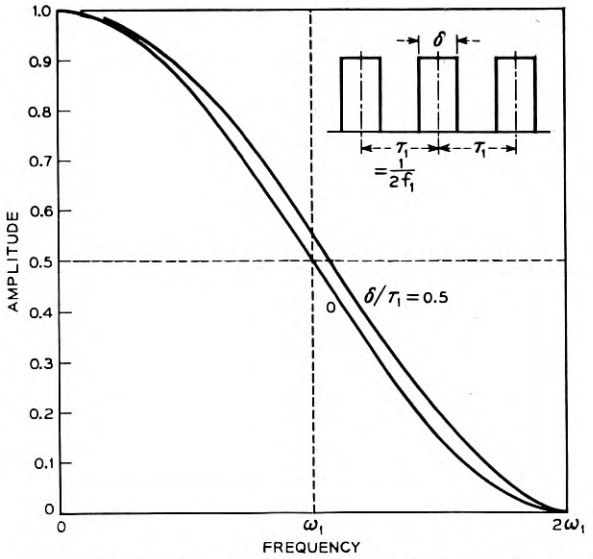


Fig. 17 — Modification of frequency characteristic to obtain same response as for impulses, when pulse duration is prolonged to half the pulse interval.

given by (5.03)

$$P(t) = \frac{\delta\omega_m}{\pi} \cos \omega_m t_0 \frac{\sin \omega_m t_0}{\omega_m t_0[1 - (\omega_m t_0/\pi)^2]},$$
$$= \frac{\delta\omega_m}{\pi} \frac{\sin 2\omega_m t_0}{2\omega_m t_0[1 - (\omega_m t_0/\pi)^2]}. \tag{5.05}$$

Pulses can in this case be transmitted without mutual interference between the pulse peaks at the points shown in the above figure. The effective pulse transmission rate is the same as for a low-pass characteristic between $\omega = 0$ and $\omega = 2\omega_m$ with half amplitude at ω_m.*

As mentioned in Section 2, when pulses of finite duration are employed, the same response as for impulses is obtained if the amplitude characteristic is modified by the factor $(\omega\delta/2)/\sin(\omega\delta/2)$. In Fig. 17 is shown the resultant minor modification in the amplitude characteristic (5.01) when the duration of the pulses is equal to half the pulse interval.

The low-pass and band-pass amplitude characteristics considered above can also be regarded as the spectra of pulses applied to a transmission system having a constant amplitude characteristic over the

* W. R. Bennett and C. B. Feldman originally proposed this type of characteristic in an unpublished memorandum, as a means of matching the bandwidth economy of baseband transmission without inclusion of frequencies near zero.

band of the spectra. If the phase characteristic of the system is linear over this band, the received pulses will have the same shape as the impulse characteristics. It should be recognized, however, that there may be appreciable phase distortion within the transmission band or pulse spectrum, if there are amplitude discontinuities beyond the band resulting from a sharp cut-off by filters. Nevertheless, the type of amplitude characteristic or frequency spectrum considered above has decisive advantages from the standpoint of transmission distortion of the pulses, as shown later, since appreciable phase distortion will ordinarily be confined to the edges of the band where the frequency components of the pulse spectrum have low amplitudes.

Another type of amplitude characteristic resembling that shown in Fig. 13 and frequently considered in connection with pulse transmission is a Gaussian characteristic:

$$A(\omega) = e^{-\sigma\omega^2}. \tag{5.06}$$

The corresponding impulse characteristic is

$$P(t) = \frac{\delta}{2(\pi\sigma)^{1/2}} e^{-t_0^2/4\sigma}. \tag{5.07}$$

If it is assumed that the amplitude is reduced to 1 per cent of the peak value after an interval $t_0 = \pi/\omega_1$, corresponding to the first zero point of an ideal impulse characteristic, it is necessary that $t_0^2/4\sigma = 4.6$, or $\sigma = .54/\omega_1^2$. The corresponding amplitude and impulse characteristics are

$$A(\omega) = e^{-0.54(\omega/\omega_1)^2}, \tag{5.08}$$

and

$$P(t) = \frac{\delta\omega_1}{0.83\pi} e^{-0.46(t_0\omega_1)^2}. \tag{5.09}$$

In Fig. 18 a comparison is made of the two frequency characteristics (5.01) and (5.08) considered above, and of the corresponding impulse characteristics (5.02) and (5.09). The comparison shows that for the same pulse transmission rate and with negligible intersymbol interference, a somewhat wider band must be provided with a Gaussian amplitude characteristic. This is a disadvantage, particularly when the band is restricted within prescribed limits by considerations of interference in adjacent transmission bands, as radio pulse systems.

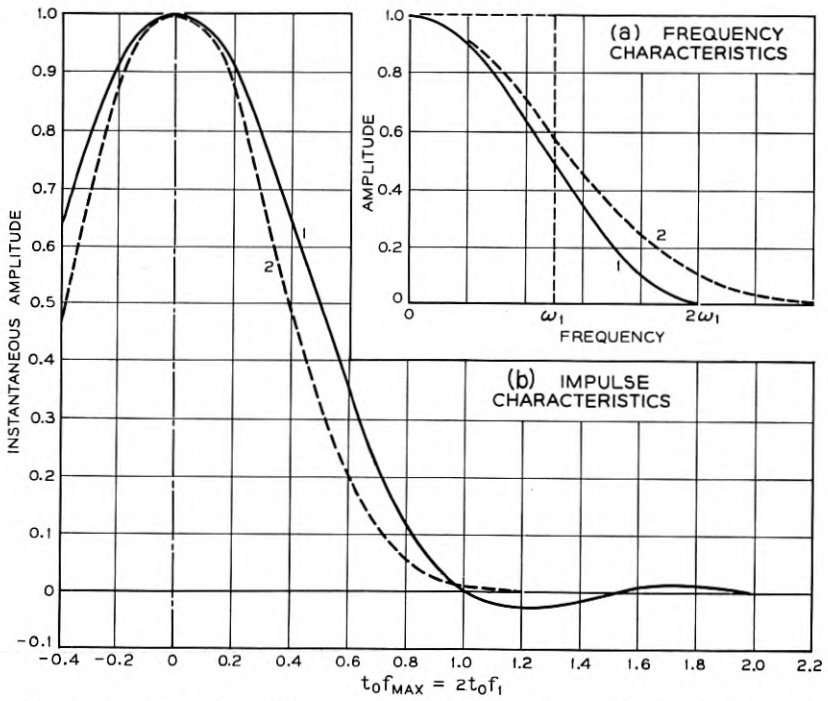


Fig. 18 — Comparison of two representative frequency and impulse transmission characteristics. Frequency characteristic 1: $T(\omega) = \frac{1}{2}[1 + \cos \pi\omega/2\omega_1]$. Frequency characteristic 2: $T(\omega) = \exp - 0.54(\omega/\omega_1)^2$.

Amplitude characteristic 1 of Fig. 18 has certain properties, aside from the linearity of the associated phase characteristic, which makes it preferable to a Gaussian as well as other types of amplitude characteristics for most pulse systems. The corresponding impulse characteristic has zero points at intervals $\tau_1 = 1/2f_1$ with the minimum possible oscillation consistent with this property for a given bandwidth. This permits the use of this impulse characteristic for pulse systems with discrete pulse positions with minimum intersymbol interference and considerable tolerance on synchronization. Since the oscillation in the impulse characteristic is inappreciable, it can also be used for pulse systems without discrete pulse positions and with other methods of detection than synchronized instantaneous sampling. In view of these attributes, an amplitude characteristic of the above type, rather than a constant amplitude characteristic with sharp cut-off, may be regarded as ideal when various physical requirements for practicable pulse systems are taken into consideration.

6. PULSE ECHOES FROM PHASE DISTORTION

For any transmission—frequency characteristic the corresponding impulse characteristics can be determined from the Fourier integral relation (2.01). This, however, may involve the evaluation of complicated integrals, which in general would require numerical integration and would be a rather elaborate procedure. A preferable method of sufficient accuracy in most engineering applications is to employ the theoretical solutions given previously for various ideal transmission characteristics with a linear phase shift as a point of departure or first approximation. A satisfactory second approximation can in many instances be secured by evaluating the transmission distortion resulting from a sinusoidal deviation in the phase characteristic. Furthermore, any type of deviation in the phase characteristic can in principle be represented by a Fourier series in terms of harmonic sinusoidal deviations.

Aside from the circumstance that in many cases a sine deviation in the phase characteristic affords a fairly satisfactory approximation to actual phase distortion it has the advantage in theoretical formulation that it permits determination of the resultant pulse distortion by the method of "paired echoes." In the usual application of this method only small phase deviations are considered resulting in a single pair of pulse or signal echoes of small amplitude, and the method is then particularly simple.[5, 6] When delay distortion is appreciable, however, as is frequently the case in wire circuits, it becomes necessary to consider a large number of pulse or signal echoes of considerable amplitude. Since the amplitudes of the pulse echoes may be obtained from available tables of Bessel Functions, the determination of the echoes is, nevertheless, simple in procedure and the determination of the shape of the distorted pulses or other signals not too elaborate.

A given amplitude characteristic within the transmission band may be associated with various phase characteristics, depending on the shape of the amplitude characteristic outside the transmission band and also on whether or not a minimum phase shift system is involved. It is therefore permissible to consider the effect of various departures from a given phase characteristic independent of the amplitude characteristic within the transmission band.

With a sinusoidal departure from a given phase characteristic $\psi_0(\omega)$ as shown in Fig. 19, the modified phase function becomes

$$\psi(\omega) = \psi_0(\omega) - b \sin \omega\tau. \tag{6.01}$$

With

$$T_0(i\omega) = A_0(\omega)\, e^{-i\psi_0(\omega)}$$

the modified transmission-frequency characteristic becomes

$$T(i\omega) = T_0(i\omega)\, e^{ib \sin \omega\tau}, \tag{6.02}$$

which, inserted in (2.01) gives

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T_0(i\omega) e^{ib \sin \omega\tau} e^{i\omega t}\, d\omega. \tag{6.03}$$

The following relation (Jacobi's expansion) in which J_1, $J_2 \cdots$ are Bessel Functions in their usual notation can now be employed[7]

$$\begin{aligned} e^{ib \sin \omega\tau} = {} & J_0(b) + J_1(b)[e^{i\omega\tau} - e^{-i\omega\tau}] \\ & + J_2(b)[e^{2i\omega\tau} + e^{-2i\omega\tau}] \\ & + J_3(b)[e^{3i\omega\tau} - e^{-3i\omega\tau}] \\ & + J_4(b)[e^{4i\omega\tau} + e^{-4i\omega\tau}] + \cdots \end{aligned} \tag{6.04}$$

Let $P_0(t)$ designate the shape of the received pulse or other signal for a transmission frequency characteristic $T_0(i\omega)$ obtained from (6.03) with $b = 0$. In view of (6.04) the solution of (6.03) may then be written

$$\begin{aligned} P(t) = {} & J_0(b)P_0(t) + J_1(b)[P_0(t + \tau) - P_0(t - \tau)] \\ & + J_2(b)[P_0(t + 2\tau) + P_0(t - 2\tau)] \\ & + J_3(b)[P_0(t + 3\tau) - P_0(t - 3\tau)] \\ & + J_4(b)[P_0(t + 4\tau) + P_0(t - 4\tau)] + \cdots \end{aligned} \tag{6.05}$$

The shape of the received pulse or signal $P(t)$ is thus obtained by superposing an infinite sequence of pulses or signals of shape $P_0(t)$. The peak amplitudes of the pulse or signal echoes and the times at which they occur with respect to $t = 0$ are given in the following table. The reference point $t = 0$ is arbitrarily selected to coincide with the peak of the pulse $P_0(t)$:

Time	-3τ	-2τ	$-\tau$	0	τ	2τ	3τ
Amplitude	$J_3(b)$	$J_2(b)$	$J_1(b)$	$J_0(b)$	$-J_1(b)$	$J_2(b)$	$-J_3(b)$

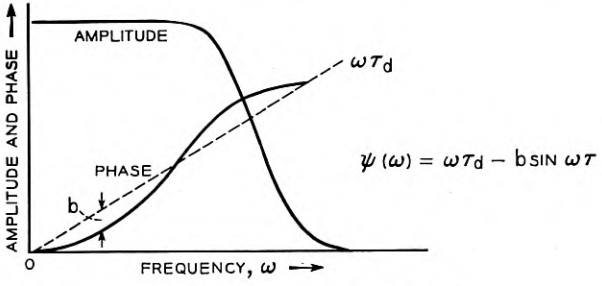
A sufficient number of echoes must be considered until their peak amplitudes become negligible.

The superposition of echoes to obtain the resultant pulse is illustrated in Fig. 20. Instead of plotting the various echoes and combining them

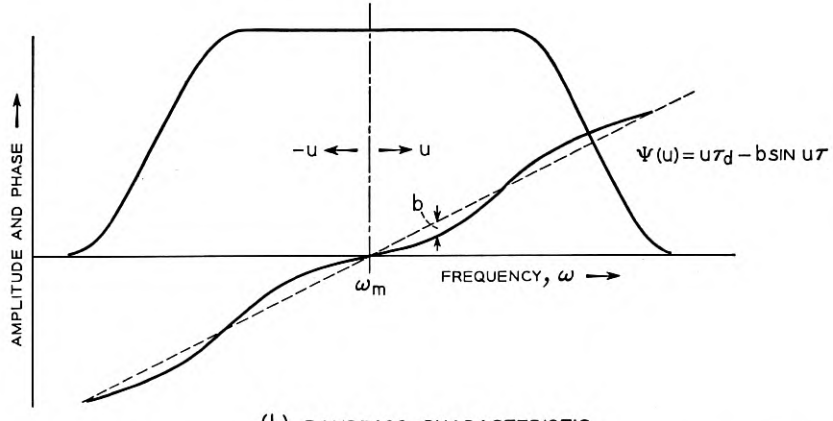
into a resultant pulse or signal as in Fig. 20(c) the equivalent and less laborious method shown in (d) can be employed. With the latter method the pulse P_0 is plotted with reversed time scale and its peak made to coincide with the point for which the amplitude of the resultant pulse P is to be determined. The amplitude of P is determined as indicated in the figure. In the particular case where the original phase characteristic ψ_0 is linear, the pulses $P_0(t)$ will be symmetrical with respect to their peak amplitude, and this assumption will be made in the following applications.

For amplitudes $b \ll 1$, the Bessel Functions become negligible except for J_0 and J_1, which are given by $J_0(b) \cong 1$ and $J_1(b) \cong b/2$, so that (6.05) becomes

$$P(t) = P_0(t) + \frac{b}{2} P_0(t + \tau) - \frac{b}{2} P_0(t - \tau). \tag{6.06}$$



(a) LOW-PASS CHARACTERISTIC



(b) BANDPASS CHARACTERISTIC

Fig. 19 — Low-pass and band-pass characteristics with sinusoidal phase distortion.

For amplitudes $b > 1$, it is necessary to consider a greater number of echoes, as will be evident from Table I for $b = 1, 2, 5, 10$ and 15 radians.[8]

The preceding equations apply to low-pass characteristics and also to symmetrical bandpass characteristics, as shown in Fig. 19. In the latter case $\mathfrak{A}(u) = \mathfrak{A}(-u)$ and $\Psi(-u) = -\Psi(u)$ in (2.10) and (2.11) so that $R_+ = R_-$ and $Q_+ = Q_-$ and (2.09) becomes

$$P(t) = \cos(\omega_r t - \psi_r) R(t), \tag{6.07}$$

where $R(t) = R_+ + R_-$ and $\omega_r = \omega_m =$ midband frequency. The envelope $R(t)$ is accordingly obtained by replacing $P_0(t)$ in (6.05) by $R_0(t)$, the envelope in the absence of phase distortion.

In Fig. 21 is shown a particular case of a sine deviation in the phase characteristic and the corresponding delay distortion, which approximates that encountered in many instances. For a low-pass system the phase and delay distortion would be as shown for $u > 0$. In this particu-

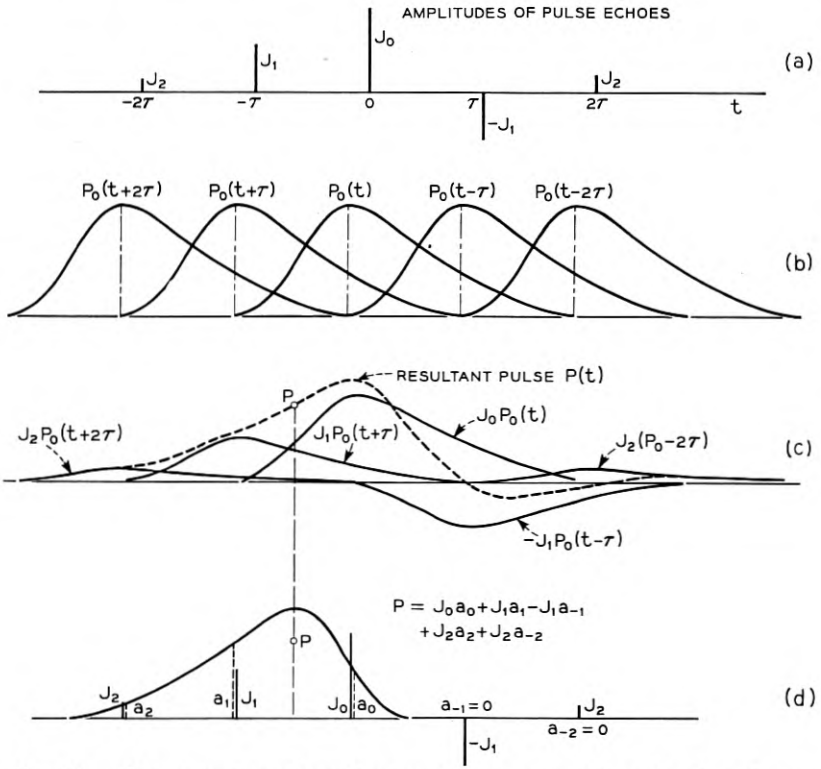


Fig. 20 — Determination of resultant pulse by superposition of pulse echoes.

TABLE I — AMPLITUDES OF ECHOES, $J_n(b)$

	$b = 1$	2	5	10	15
$n = 0$	0.7652	0.2239	−0.1776	−0.2459	−0.0142
1	0.4401	0.5767	−0.3276	0.0434	0.2051
2	0.1149	0.3528	0.0466	0.2546	0.0416
3	0.0196	0.1289	0.3648	0.0584	−0.1940
4		0.0340	0.3912	−0.2196	−0.1192
5			0.2611	−0.2341	0.1305
6			0.1310	−0.0145	0.2061
7			0.0534	0.2167	0.0345
8			0.0184	0.3179	−0.1740
9			0.0055	0.2919	−0.2200
10				0.2075	−0.0901
11				0.1231	0.1000
12				0.0634	0.2367
13				0.0290	0.2787
14				0.0120	0.2464
15				0.0045	0.1813
16					0.1162
17					0.0665
18					0.0346
19					0.0166
20					0.0073

lar case the maximum amplitude b is at the maximum frequency $\omega_{max} = 2\omega_1$, so that $\sin \omega\tau = 1$, or $\omega\tau = \pi/2$, for $\omega = 2\omega_1$. Hence the interval between pulse echoes is $\tau = \pi/4\omega_1 = 1/8f_1$. The interval τ is accordingly ¼ the interval $\tau_1 = 1/2f_1$ required for the pulse $P_0(t)$ to reach zero amplitude in the absence of phase distortion.

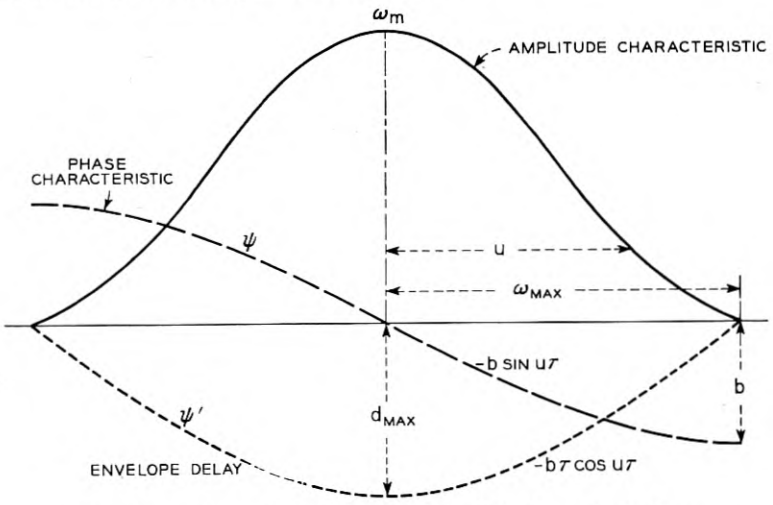


Fig. 21 — Particular case of sinusoidal phase deviation.

For the particular case illustrated in the above figure, the delay distortion is given by

$$d\psi/d\omega = -b\tau \cos \omega\tau.$$

When $\omega = 0$

$$d\psi/d\omega = -b\tau = -d_{\max}. \tag{6.08}$$

When $\omega\tau = \tau\omega_{\max} = \pi/2$

$$d\psi/d\omega = 0.$$

Hence

$$d\psi/d\omega = -d_{\max} \cos (\omega\pi/2\omega_{\max}). \tag{6.09}$$

With $\tau = \pi/2\omega_{\max}$ and $b\tau = d_{\max}$, the following relation is obtained

$$b = \frac{2}{\pi}\omega_{\max} d_{\max} = 4f_{\max} d_{\max}. \tag{6.10}$$

In Fig. 22 are shown the positions of the pulse echoes for the above case on a numerical scale $t \cdot f_{\max}$, together with their amplitudes for $b = 5$ radians. On this scale the interval between pulse echoes $\tau = 1/4f_{\max}$ is $\frac{1}{4}$. In the same figure is shown an assumed pulse shape in the absence of phase distortion, which is the same as shown in Fig. 13, except that the small tail has been neglected. The peak of the pulse is taken at $tf_{\max} = -0.75$, and the amplitude of the resultant pulse at the

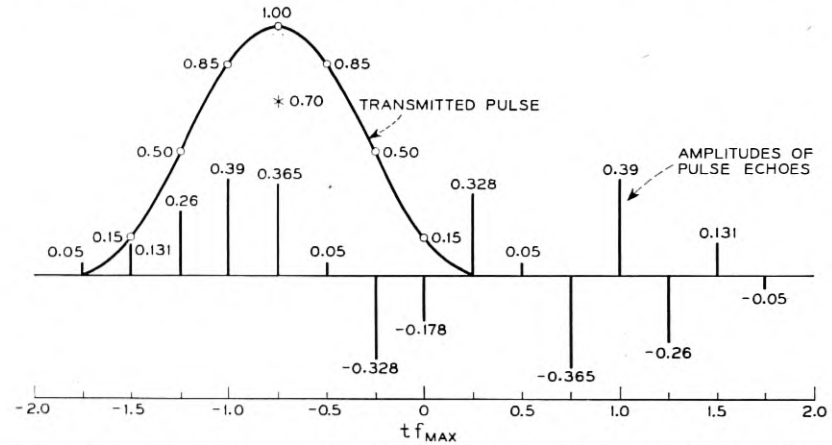


Fig. 22 — Illustrative example of calculation of impulse characteristic shown in Fig. 23, by method illustrated in Fig. 20(d).

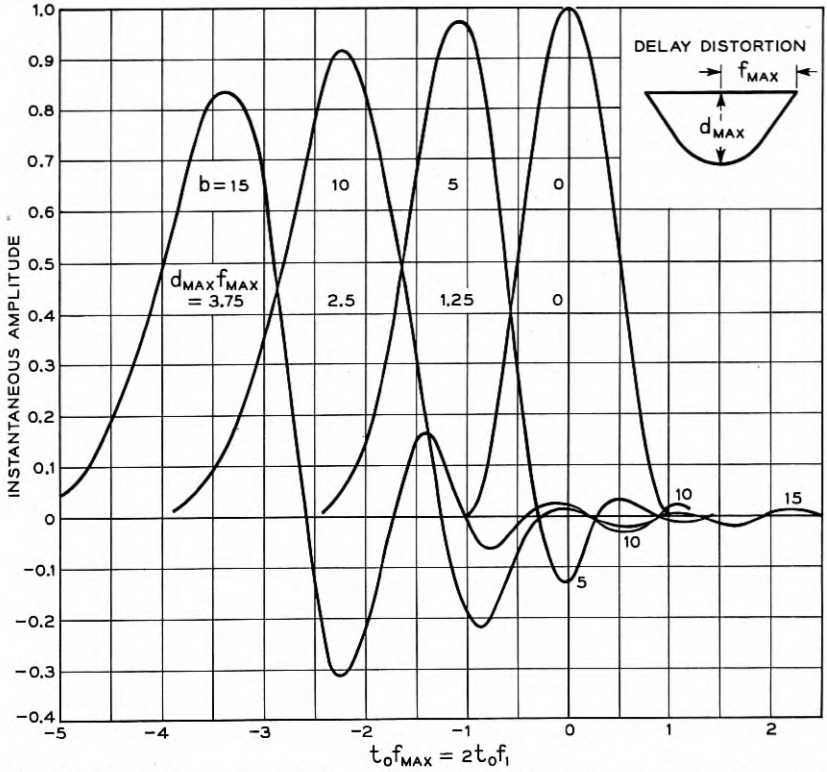


Fig. 23 — Impulse transmission characteristics with cosine variation in delay.

corresponding point obtained by the method illustrated in *d* of Fig. 20 is

$$P = 1 \cdot 0.365 + 0.85\ (0.39 + 0.05) + 0.5\ (0.26 - 0.328) + 0.15\ (0.131 - 0.178)$$

$$= 0.70.$$

In Fig. 23 are shown the resultant pulses obtained by the above method for various values of b and the corresponding values of $d_{max}f_{max}$. Since the interval between pulse echoes is small in relation to the duration of the pulse $P_0(t)$, as seen from Fig. 22, the individual pulse echoes cannot be discerned in the resultant pulses shown in Fig. 23. It will be noticed that as b increases, the pulses are received with decreasing transmission delay, which is due to the choice of reference delay in the delay distortion curve. That is, as d_{max} or b is increased, the delay becomes increas-

ingly negative with respect to $d_{max} = 0$ used for reference. The curves apply to a band-pass system as indicated in the figure, and also to a low-pass system having the delay distortion shown above the midband frequency of the band-pass system.

An improved approximation to phase distortion is sometimes obtained by considering two sine deviations in the phase characteristic.

If the phase characteristic is given by

$$\psi(\omega) = \psi_0(\omega) - b' \sin \omega\tau - b'' \sin \omega\tau', \tag{6.11}$$

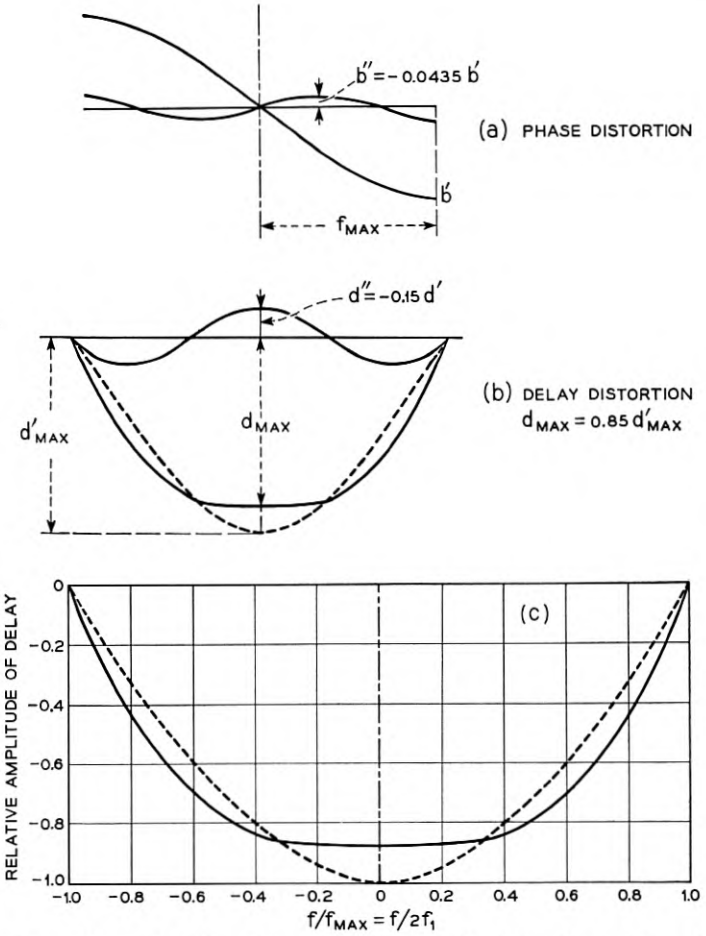


Fig. 24 — Shape of delay distortion with combined fundamental and third harmonic cosine variation in delay.

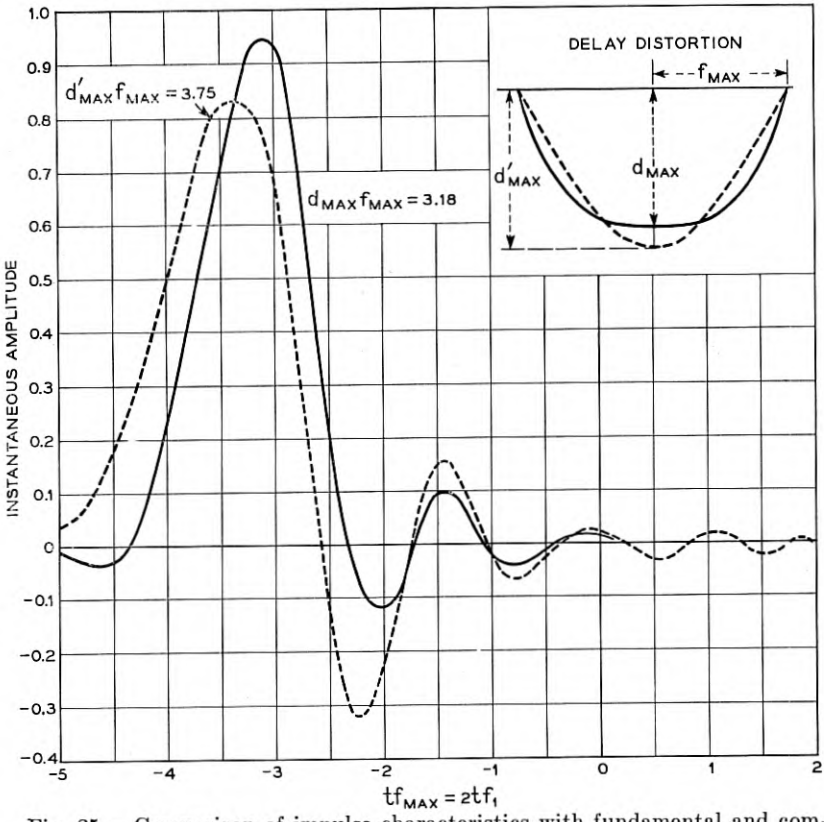


Fig. 25 — Comparison of impulse characteristics with fundamental and combined fundamental and third harmonic cosine variation in delay as in Fig. 24.

the combined effect of the two sine deviations is obtained by first determining the effect of $-b'$ sin $\omega\tau'$ from (6.05). The value of $P(t) = P_1(t)$ thus obtained from (6.05) with $b = b'$ and $\tau = \tau'$ is next substituted for $P_0(t)$ in (6.05), with $b = b''$ and $\tau = \tau''$ to evaluate the effect of $-b''$ sin $\omega\tau''$. That is, the system is considered to consist of a tandem arrangement of two components, the first with a phase distortion $-b'$ sin $\omega\tau$ and the second with phase distortion $-b''$ sin $\omega\tau''$.

In Fig. 24 is shown a particular case in which the second component is a triple harmonic of the first with amplitude $b'' = -0.0435b'$. This results in an improved approximation to the delay distortion encountered in certain wire facilities, where the band is sharply confined by filters. In Fig. 25 is shown the pulse shape for this case with $b' = 15$ radians, together with that for a single sine deviation of $b = 15$ radians.

It will be recognized from the above that as the number of sine com-

ponents required to represent a given phase distortion increases, the determination of the resultant pulse becomes rather laborious, unless the sine deviations are all small in amplitude. In the latter case each sine deviation corresponds in a first approximation to a single pair of echoes, so that the effect of a number of sine deviations can be obtained by direct superposition.

7. PULSE ECHOES FROM AMPLITUDE DISTORTION

Departures from a given amplitude characteristic may in certain cases be approximated by a single cosine variation, as illustrated in Fig. 26. Since the amplitude characteristic is an even function of ω, any departure from a given amplitude characteristic may be represented by a cosine Fourier series. The effect of a cosine variation in the amplitude characteristic is therefore of basic interest.

A cosine variation will in general be accompanied by a change in the phase characteristic, as discussed in Section 1, but it will first be assumed that phase correction is employed to maintain a fixed phase characteristic.

Let $A_0(\omega)$ be the original amplitude characteristic and let the modified amplitude characteristic be of the form

$$A(\omega) = A_0(\omega)[1 + a \cos \omega\tau]. \tag{7.01}$$

Equation (2.01) for the impulse transmission characteristic then becomes, with $T_0(i\omega) = A_0(\omega)e^{-i\psi_0(\omega)}$,

$$\begin{aligned} P(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} T_0(i\omega)\left[1 + \frac{a}{2}\,(e^{i\omega\tau} + e^{-i\omega\tau})\right] e^{i\omega t}\, d\omega, \\ &= P_0(t) + \frac{a}{2}\,P_0(t + \tau) + \frac{a}{2}\,P_0(t - \tau). \end{aligned} \tag{7.02}$$

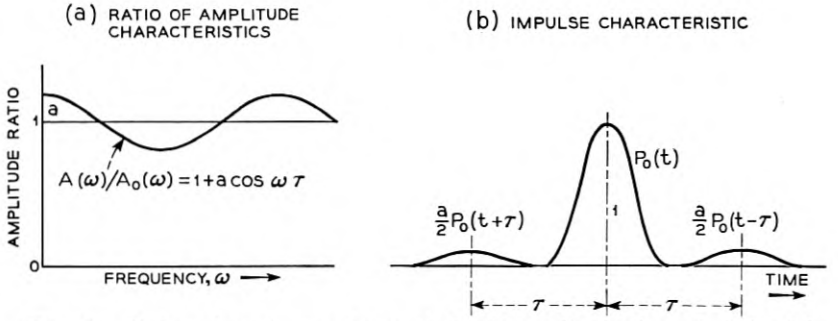


Fig. 26 — Pulse echoes from cosine variation in amplitude characteristic without change in phase characteristic.

There will thus be pulse or signal echoes of amplitude $a/2$ at the time τ before and after the main pulse as illustrated in Fig. 26.

With a cosine variation in the attenuation rather than in the amplitude characteristic, the modified amplitude characteristic becomes

$$A(\omega) = A_0(\omega)\, e^{a \cos \omega\tau}, \tag{7.03}$$

and the modified impulse characteristic

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T_0(i\omega) e^{a \cos \omega\tau} e^{i\omega t}\, d\omega. \tag{7.04}$$

The expansion corresponding to (6.04) is in this case:

$$\begin{aligned} e^{a \cos \omega\tau} = {} & I_0(a) + I_1(a)(e^{i\omega\tau} + e^{-i\omega\tau}) \\ & + I_2(a)(e^{2i\omega\tau} + e^{-2i\omega\tau}) \\ & + I_3(a)(e^{3i\omega\tau} + e^{-3i\omega\tau}), + \cdots \end{aligned} \tag{7.05}$$

where I_1, $I_2 \cdots$ are Bessel functions for imaginary arguments in their usual notation.

The resultant modified impulse characteristic in this case becomes

$$\begin{aligned} P(t) = {} & I_0(a)P_0(t) + I_1(a)[P_0(t + \tau) + P_0(t - \tau)] \\ & + I_2(a)[P_0(t + 2\tau) + P_0(t - 2\tau)] \\ & + I_3(a)[P_0(t + 3\tau) + P_0(t - 3\tau)] + \cdots \end{aligned} \tag{7.06}$$

which can be interpreted in a similar way as discussed for (6.05). For small values of a, $I_0\,(a) \cong 1$, $I_1\,(a) \cong a/2$ and the remaining terms in (7.06) negligible, so that (7.02) is obtained.

As discussed in Section 1, when the amplitude characteristic is modified in accordance with (7.01), the resultant modification in the phase characteristic is in accordance with (1.13)

$$\psi_1 = 2 \tan^{-1} \frac{r \sin \omega\tau}{1 + r^2 \cos \omega\tau}. \tag{7.07}$$

The modified transmission-frequency characteristic is in this case

$$T(i\omega) = T_0(i\omega)(1 + a \cos \omega\tau)e^{-i\psi_1}, \tag{7.08}$$

which can be transformed into

$$\begin{aligned} T(i\omega) &= T_0(i\omega) \frac{1}{1 + r^2} (1 + re^{-i\omega\tau})^2, \\ &= T_0(i\omega) \frac{1}{1 + r^2} (1 + 2re^{-i\omega\tau} + r^2 e^{-2i\omega\tau}). \end{aligned} \tag{7.09}$$

Thus, with a cosine variation in the amplitude characteristic in accordance with (7.01), accompanied by a minimum phase shift change in the phase characteristic in accordance with (7.07), the modified impulse characteristic becomes

$$P(t) = \frac{1}{1 + r^2} [P_0(t) + 2rP_0(t - \tau) + r^2P_0(t - 2\tau)], \qquad (7.10)$$

where

$$r = \frac{1}{a} [1 - \sqrt{1 - a^2}]. \qquad (7.11)$$

The received pulse or signal $P(t)$ will thus consist of three components each having the same shape as the pulse or signal $P_0(t)$, but differing in amplitude and displaced in time, as indicated in Fig. 27.

For small values of the amplitude a of the cosine deviation, $r \leqq a/2$ and $1 + r^2 \leqq 1$, so that

$$P(t) = P_0(t) + aP_0(t - \tau) + \frac{a^2}{4} P_0(t - 2\tau). \qquad (7.11)$$

The solution for a somewhat similar case given elsewhere,[9] has an infinite number of echoes, with the second echo given by $a^2P_0(t - 2\tau)$ rather than $(a^2/4)P_0(t - 2\tau)$ as above. In the case referred to, the amplitude deviation is in a first approximation $a \cos \omega\tau$, but there are additional terms in $\cos 2\omega\tau$, $\cos 3\omega\tau$ etc, which are responsible for the different amplitude of the second echo and for the infinite sequence of echoes.

With a cosine modification in the attenuation characteristic as given by (7.03), there will be a corresponding sine modification in the phase characteristic in accordance with (1.11). The modified transmission-

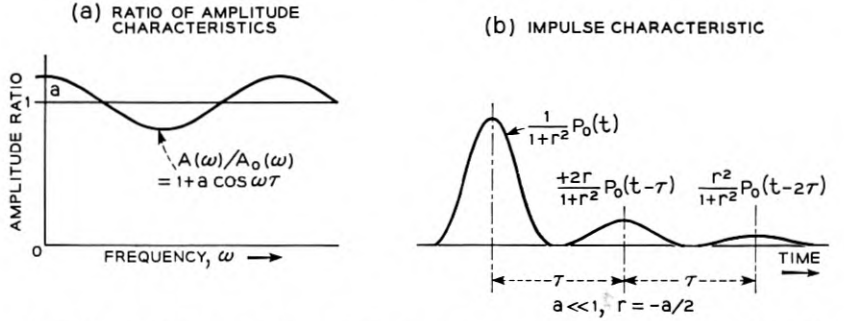


Fig. 27 — Pulse echoes from cosine variation in amplitude characteristic with associated minimum phase shift variation in phase characteristic.

frequency characteristic is in this case

$$\begin{aligned} T(i\omega) &= T_0(i\omega)e^{a(\cos\omega\tau - i\sin\omega\tau)}, \\ &= T_0(i\omega)e^{ae^{-i\omega\tau}}, \\ &= T_0(i\omega)\left[1 + ae^{-i\omega\tau} + \frac{a^2}{2!}e^{-2i\omega\tau} + \frac{a^3}{3!}e^{-3i\omega\tau} + \cdots\right]. \end{aligned} \tag{7.12}$$

The modified impulse characteristics is in this case

$$P(t) = P_0(t) + aP_0(t-\tau) + \frac{a^2}{2!}P_0(t-2\tau) + \frac{a^3}{3!}P_0(t-3\tau) + \cdots \tag{7.13}$$

For small values of a both (7.11) and (7.13) give for the modification in the impulse characteristic resulting from a small cosine deviation in the amplitude or attenuation characteristics accompanied by changes in the phase characteristic:

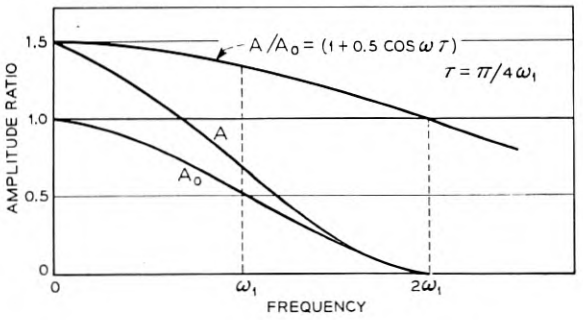
$$P(t) = P_0(t) + a\,P_0(t-\tau). \tag{7.14}$$

In certain applications it is convenient to regard $P_0(t)$ as a pulse or signal applied to a transmission line and $P(t)$ as the received pulse or signal with a cosine deviation in the amplitude characteristic of the transmission line.

In the lower part of Fig. 28 is shown the modification in the received pulses resulting from a slow pronounced cosine deviation in the amplitude characteristic shown at the top. In Fig. 29 is shown the effect of positive and negative cosine variations when the amplitude at zero frequency is held constant, a condition which may be approximated in wire systems as a result of variation in attenuation over the transmission band with temperature. Curve 1 would correspond to a 3.5 db smaller loss at the maximum frequency $2\omega_1$ than at zero frequency, and curve 2 to a 6 db greater loss at the maximum frequency. It will be noticed that pulse distortion as well as the variation in the peak amplitude of the pulses is greater under the first condition, i.e. curve 1. Pulse overlaps can in both cases be avoided by a moderate increase in pulse spacing, and in the first case can be substantially reduced also by a decrease in pulse spacing.

8. FINE STRUCTURE IMPERFECTIONS IN TRANSMISSION CHARACTERISTICS

As a result of imperfections in the transmission medium and in equalization there may be fine structure departures from a nominal transmission characteristic, as illustrated in Fig. 30. They are often caused by



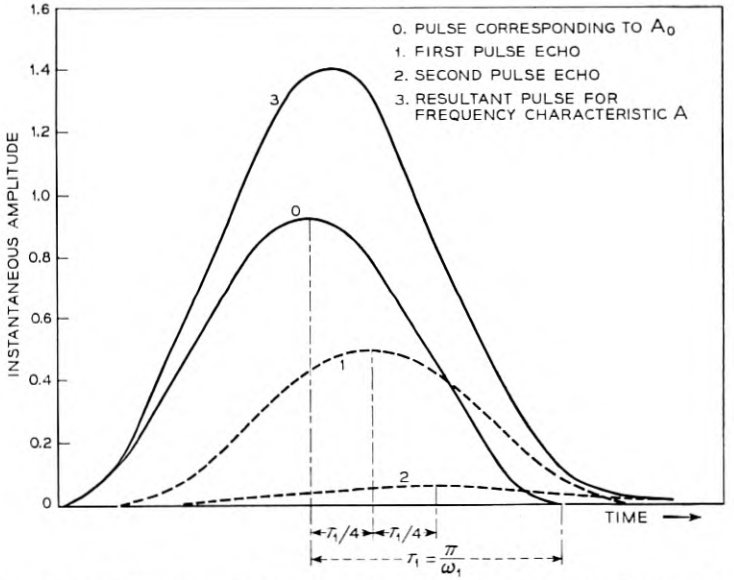


Fig. 28 — Modification of impulses characteristic by slow cosine variation in amplitude characteristic.

echoes in very long lines resulting from impedance mismatches. Fine structure deviations from a specified amplitude characteristic may in principle be represented by a cosine Fourier series, since the amplitude function is an even function of ω. Thus, if the specified amplitude characteristic is $A_0(\omega)$, the actual amplitude characteristic $A(\omega)$ may be represented by an infinite cosine Fourier series as:

$$A(\omega) = A_0(\omega)[1 + a_1 \cos \omega\tau + a_2 \cos 2\omega\tau + \cdots + a_m \cos m\omega\tau + \cdots]. \tag{8.01}$$

The coefficients a_1, $a_2 \cdots a_m \cdots$ are determined in the usual manner by Fourier series analysis to represent the function

$$f(\omega) = \frac{A(\omega)}{A_0(\omega)} = 1 + \alpha(\omega) \tag{8.02}$$

over the frequency band. If $A_0(\omega)$ closely approaches $A(\omega)$ the fine

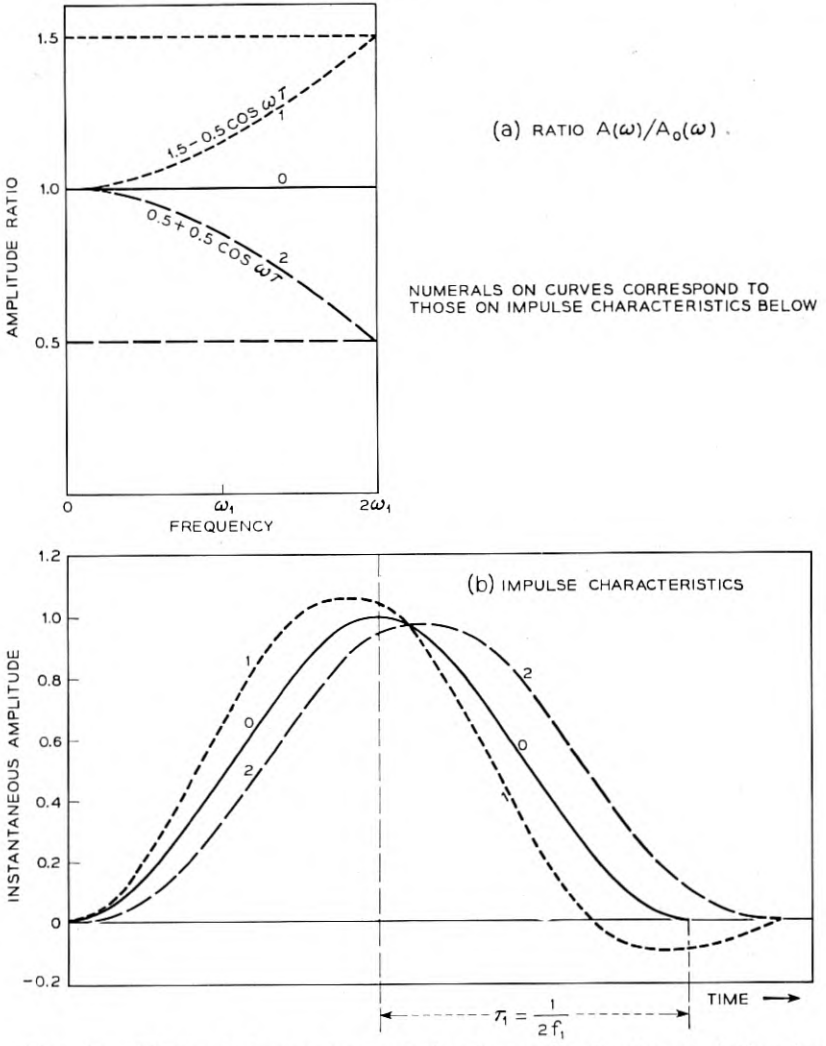


Fig. 29 — Effect of slow cosine variation in amplitude characteristic when amplitude at zero frequency is held constant.

structure departures $\alpha(\omega)$ in the transmission characteristic and hence the coefficients a_1 , $a_2 \cdots a_m \cdots$ will be small.

In the above representation $A_0(\omega)$ can also be regarded as the amplitude characteristic of a terminal network or as the frequency spectrum of a pulse applied to a transmission system with an amplitude characteristic $f(\omega) = 1 + \alpha(\omega)$.

In a Fourier series analysis of the deviation in the amplitude characteristic, the fundamental period of the amplitude variation would be selected so that there is one complete cycle between $-\omega_1$ and ω_1 , the cutoff frequency, in which case $\omega_1\tau = \pi$ or

$$\tau = \frac{\pi}{\omega_1}.$$

This is the interval between pulse echoes when the amplitude characteristic is represented by (8.01). It is identical with the interval τ_1 given by (3.02) at which pulses can be transmitted without mutual interference with a constant amplitude transmission frequency characteristic.

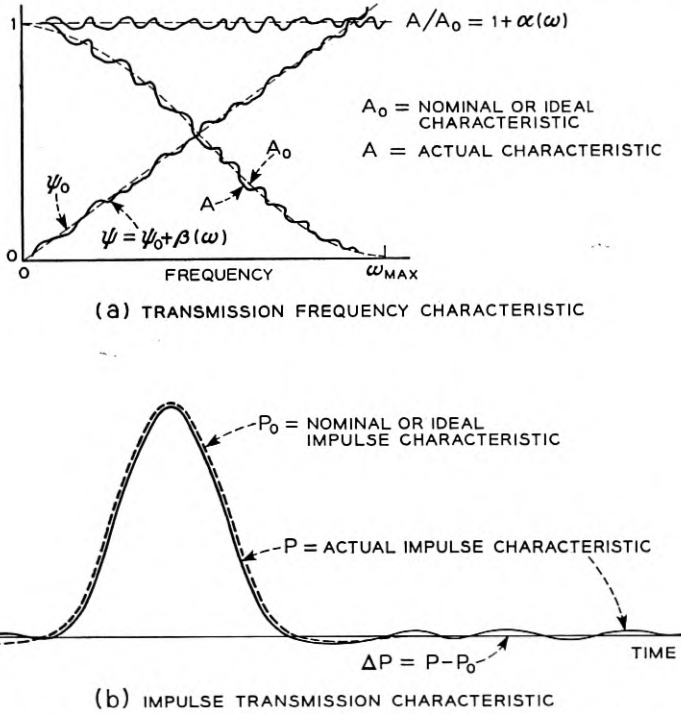


Fig. 30 — Fine structure imperfections in transmission frequency characteristic and resultant prolongation of impulse characteristic.

Assume that pulses of unit peak amplitude but varying polarity are transmitted at intervals $\tau = \tau_1$ and consider the interference with a given pulse from all pulses. As illustrated in Fig. 31, the first preceding and following pulses will in accordance with (7.02) give rise to a pulse echo $\pm a_1/2$ and the second preceding and following pulses to a pulse echo $\pm a_2/2$ etc., where the signs of the echoes depend on the polarity of the pulses and on the signs of the coefficients a_1, a_2. The resultant intersymbol interference $U_a(t)$ will depend on the polarity of the various

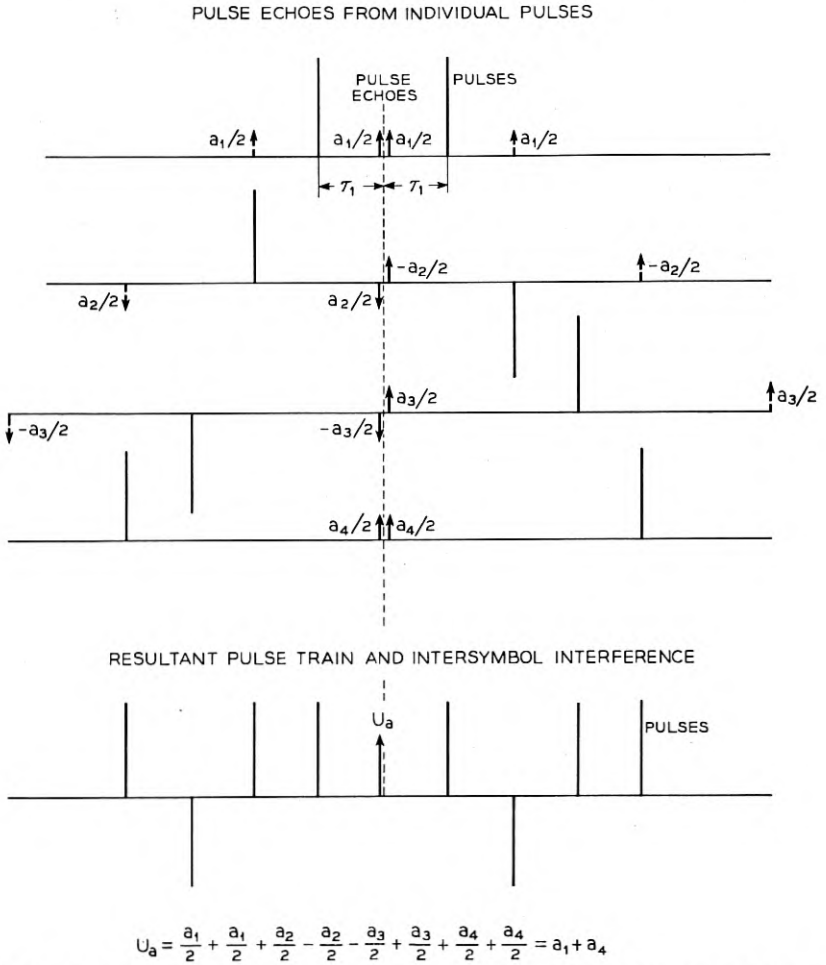


Fig. 31 — Combination of pulse echoes into intersymbol interference for a particular case.

pulses and will thus vary with time. It can have any value assumed by the expression

$$U_a(t) = \pm\frac{a_1}{2} \pm \frac{a_1}{2} \pm \frac{a_2}{2} \pm \frac{a_2}{2} \cdots \pm \frac{a_m}{2} \pm \frac{a_m}{2} + \cdots \qquad (8.03)$$

The maximum possible intersymbol interference will thus be the sum of the absolute values of the coefficients a_m.

$$\hat{U}_a = |a_1| + |a_2| + |a_3| + \cdots + |a_m| + \cdots \qquad (8.04)$$

In certain pulse systems, such as PAM time division systems, rms intersymbol interference is of main importance, while in others, such as PCM or telegraph systems, peak intersymbol interference is of principal interest. If the fine structure imperfections are regarded as of random nature, in the sense that they are not predictable and vary between systems having the same nominal transmission characteristics, peak intersymbol interference can be estimated from rms interference by applying a peak factor of about 4. With random variation in the amplitude of intersymbol interference, the probability of exceeding 4 times the rms value is in accordance with the normal law about 5×10^{-5}. Peaks in excess of 4 times the rms value will thus be so rare that they can for practical purposes be neglected.

The rms intersymbol interference is equal to the root mean square of all the different values which can be assumed by expression (8.03). This turns out to be equal to the root sum square of the amplitudes $a_m/2$ and $-a_m/2$ of the pulse echoes, or

$$\underline{U}_a = \left[\sum_1^\infty \left(\frac{a_m}{2}\right)^2 + \sum_1^\infty \left(\frac{-a_m}{2}\right)^2\right]^{1/2} = \left(\frac{1}{2}\sum_1^\infty a_m{}^2\right)^{1/2}. \qquad (8.05)$$

When a_m are the various coefficients in the Fourier representation of $\alpha(\omega)$ over the frequency band from $-\omega_1$ to ω_1, the following relation holds.

$$\frac{1}{2}\sum_1^\infty a_m{}^2 = \frac{1}{2\omega_1}\int_{-\omega_1}^{\omega_1} \alpha^2(\omega)\, d\omega = \frac{1}{\omega_1}\int_0^{\omega_1} \alpha^2(\omega)\, d\omega \qquad (8.06)$$

where $\alpha(\omega)$ in the present case is given by (8.02) and represents the departure in the ratio $A(\omega)/A_0(\omega)$ from unity.

With (8.06) in (8.05) the following expression is obtained for rms intersymbol interference due to amplitude deviations $\alpha(\omega)$ not accompanied by phase deviations

$$\underline{U}_a = \underline{a} \qquad (8.07)$$

where $\underline{a}$ is the rms deviation in $\alpha(\omega)$ over the transmission band as given by

$$\begin{aligned} \underline{a} &= \left[\frac{1}{\omega_1}\int_0^{\omega_1} \alpha^2(\omega)\, d\omega\right]^{1/2} \\ &= \left[\frac{1}{f_1}\int_0^{f_1} \alpha^2(f)\, df\right]^{1/2} \end{aligned} \tag{8.08}$$

The rms amplitude deviation expressed in db is

$$\begin{aligned} \underline{a}' &= 20 \log_{10}(1 + \underline{a}) \\ &\backsimeq 8.69\, \underline{a} \qquad \text{when } \underline{a} < 0.1 \end{aligned} \tag{8.09}$$

A corresponding analysis can be made for fine structure imperfections in the phase characteristic. The deviation $\beta(\omega) = \psi(\omega) - \psi_0(\omega)$ from a prescribed phase characteristic $\psi_0(\omega)$ may in this case be represented by a sine Fourier series since the phase characteristic is an odd function of ω:

$$\beta(\omega) = b_1 \sin \omega\tau + b_2 \sin 2\omega\tau + \cdots + b_m \sin m\omega\tau + \cdots \tag{8.10}$$

The resultant peak intersymbol interference becomes

$$\hat{U}_b = |\, b_1 \,| + |\, b_2 \,| + \cdots + |\, b_m \,| + \cdots \tag{8.11}$$

and the rms intersymbol interference

$$\underline{U}_b = \left[\frac{1}{2}\sum_1^{\infty} b_m{}^2\right]^{1/2} = \underline{b}, \tag{8.12}$$

where $\underline{b}$ is the rms phase deviation in radians as given by

$$\begin{aligned} \underline{b} &= \left[\frac{1}{\omega_1}\int_0^{\omega_1} \beta^2(\omega)\, d\omega\right]^{1/2}, \\ &= \left[\frac{1}{f_1}\int_0^{f_1} \beta^2(f)\, df\right]^{1/2}. \end{aligned} \tag{8.13}$$

In the above derivation, the amplitude and phase deviations were assumed independent of each other. The resultant rms intersymbol interference from both is in this case

$$\underline{U} = (\underline{U}_a{}^2 + \underline{U}_b{}^2)^{1/2} = (\underline{a}^2 + \underline{b}^2)^{1/2}. \tag{8.14}$$

This relationship, applying to an ideal transmission characteristic, has been established by a different method in a paper by W. R. Bennett.[10]

From (7.05) it will be seen that with minimum phase shift relationships a small cosine deviation of amplitude a_m in the amplitude characteristic will be accompanied by a phase deviation $b_m = a_m$. Hence in this case (8.14) gives

$$\underline{U} = 2^{1/2}\, \underline{a} \tag{8.15}$$

This also follows when it is considered that in this case all the pulse echoes occur after the main pulse, and have amplitudes a_1, $a_2 \cdots a_m$. The root sum square of the amplitudes is in this case $[\sum_1^\infty a_m^2]^{1/2}$, which is greater than $\underline{U}_a$ as given by (8.05) by the factor $2^{1/2}$.

The above analysis was based on an infinite sequence of pulse echoes, which combine to give the proper pulse distortion but may be regarded as fictitious in nature. The assumption of an infinite sequence of pulse echoes can be avoided by a different method of analysis outlined below, which does not involve the assumption that the coefficients are known from a Fourier series analysis, and furthermore, does not assume an ideal amplitude characteristic with a sharp cut-off as above.

Let $Ae^{-i\psi}$ and $A_0e^{-i\Psi_0}$ designate two transmission—frequency characteristics, where A, A_0, ψ and ψ_0 are functions of ω, which for convenience is omitted in the following. The squared absolute value of the difference in the transmission frequency characteristics is then

$$|A\ e^{-i\psi} - A_0e^{-i\psi_0}|^2 = A_0^2[2(1 - \cos\beta)\ (1 + \alpha) + \alpha^2], \tag{8.16}$$

where $\alpha = \alpha(\omega) = (A - A_0)/A_0$ represents the deviation in the ratio of the amplitude characteristics from unity and $\beta = \beta(\omega) = \psi - \psi_0$ the deviation in the phase characteristic.

Let P and P_0 designate the impulse characteristics corresponding to the above transmission frequency characteristics, and let $\Delta P = P - P_0$. Assume that unit impulses of varying polarity are transmitted at uniform intervals τ_1. The rms value of ΔP over the interval τ_1 in relation to the maximum amplitude $P(0)$ of the received pulses, or the rms intersymbol interference $\underline{U}$, is then given by

$$\begin{aligned}\underline{U} &= \frac{1}{P(0)}\left[\frac{1}{\tau_1}\int_{-\infty}^{\infty}(\Delta P)^2\,dt\right]^{1/2},\\ &= \frac{1}{P(0)}\left[\frac{1}{\pi\tau_1}\int_0^\infty A^2[2(1-\cos\beta)(1+\alpha)+\alpha^2]\,d\omega\right]^{1/2}.\end{aligned} \tag{8.17}$$

For small values of α and β, this expression becomes

$$\underline{U} = \frac{1}{P(0)}\left[\frac{1}{\pi\tau_1}\int_0^\infty A_0^2(\alpha^2+\beta^2)\,d\omega\right]^{1/2}. \tag{8.18}$$

If α and β are random variables representing fine structure deviations uniformly distributed over the transmission band, it is permissible to simplify (8.18) to:

$$\underline{U} = \eta \left(\frac{1}{\omega_1 \tau_1}\right)^{1/2} (\underline{a}^2 + \underline{b}^2)^{1/2}, \tag{8.19}$$

where

$$\eta = \frac{1}{\pi P(0)} \left(\omega_1 \int_0^{\omega_{max}} A_0^{\,2}\, d\omega\right)^{1/2}, \tag{8.20}$$

$$\underline{a} = \left(\frac{1}{\omega_{max}} \int_0^{\omega_{max}} \alpha^2\, d\omega\right)^{1/2}, \quad \text{and} \tag{8.21}$$

$$\underline{b} = \left(\frac{1}{\omega_{max}} \int_0^{\omega_{max}} \beta^2\, d\omega\right)^{1/2}, \tag{8.22}$$

where ω_{max} is defined as in Fig. 30 and ω_1 is the bandwidth at the half amplitude point.

For a transmission characteristic with linear phase shift, aside from small random imperfections as considered here:

$$P(0) = \frac{1}{\pi} \int_0^{\omega_{max}} A_0\, d\omega. \tag{8.23}$$

For the particular case of a transmission characteristic with constant amplitude between $\omega = 0$ and $\omega_1 = \omega_{max}$, $\eta = 1$. Pulses would in this case be transmitted at intervals $\tau_1 = \pi/\omega_1$ so that $\pi/\omega_1\tau_1 = 1$ and (8.19) is identical with (8.14).

For a transmission characteristic of the type shown in Fig. 13, pulses would also be transmitted at intervals $\tau_1 = \pi/\omega_1$ so that $\pi/\omega_1\tau_1 = 1$. In this case $\omega_{max} = 2\omega_1$, and evaluation of (8.20) gives $\eta = 3^{1/2}/2 = 0.866$. Rms intersymbol interference is thus reduced by the factor 0.866, for the same values of $\underline{a}$ and $\underline{b}$. However, these are now the rms deviations taken over a band which is twice as great as with a sharp cut-off at ω_1.

Expressions (8.14) and (8.19) can also be applied to localized imperfections in the amplitude and phase characteristics confined to a narrow portion of the transmission band. This follows when it is considered that such deviations can be represented by Fourier series containing a large number of coefficients, so that the resultant intersymbol interference can attain a great number of different values depending on the sequence of transmitted pulses. A particular case of a localized imperfection in the amplitude characteristic in the form of a low-frequency cut-off is considered in the following section.

9. TRANSMISSION DISTORTION BY LOW FREQUENCY CUT-OFF

A low-frequency cut-off in the transmission frequency characteristic of wire systems is unavoidable with transformers as employed for increased transmission efficiency or other reasons. In single sideband frequency division systems, there is a low-frequency cut-off in individual channels caused by elimination of the carrier and part of the desired sideband. The effect of a low-frequency cut-off can be avoided by employing a symmetrical band-pass characteristic as illustrated in Fig. 16, or more generally by double sideband transmission with a two-fold increase in bandwidth as compared to a low-pass system. It can also be overcome by vestigial sideband transmission with inappreciable bandwidth penalty, but with complications in terminal instrumentation. The effect of a low-frequency cut-off can, furthermore, be reduced without frequency translation as involved in double or vestigial sideband transmission, by certain methods of shaping or transmission of pulses, as discussed in the following, and by certain methods of compensation at the receiving end or at points of pulse regeneration not considered here.

The nature of the pulse distortion resulting from a low-frequency cut-off is illustrated in Fig. 32. If the phase characteristic is assumed linear, the amplitude characteristic may be regarded as made up of two components, in accordance with the following identity:

$$A(\omega) = A_0(\omega) + [A(\omega) - A_0(\omega)], \tag{9.01}$$

where $A_0(\omega)$ is the amplitude characteristic without a low-frequency cut-off and $[A(\omega) - A_0(\omega)]$ a supplementary characteristic of negative amplitude, as indicated in Fig. 32.

The impulse characteristic may correspondingly be written

$$P(t) = P_0(t) + [P(t) - P_0(t)]. \tag{9.02}$$

If the cut-off is confined to rather low frequencies, the impulse characteristic $\Delta P(t) = P(t) - P_0(t)$ will extend over time intervals substantially longer than the duration of $P_0(t)$ or the interval at which pulses are transmitted. The total area under the resultant pulse is always zero.

When a sufficiently long sequence of pulses of one polarity is transmitted, the cumulative effect of the pulse overlaps resulting from the modification $P(t) - P_0(t)$ in the impulse characteristic will be a displacement of the received pulse train, as illustrated in Fig. 33 for various intervals between the pulses. This apparent displacement of the zero line, often referred to as "zero wander," will reduce the margin for dis-

tinction between the presence and absence of pulses in a random pulse train. In the particular case when pulses are transmitted at the minimum interval $\tau_1 = 1/2f_1$ possible without intersymbol interference in the absence of a low-frequency cut-off, the pulse train will ultimately vanish when an infinite sequence of pulses of one polarity is transmitted, as illustrated for the last case in Fig. 33.

The number of pulses of one polarity, or nearly all of the same polarity, which can be transmitted before the limiting condition illustrated in Fig. 33 is approached depends on the extent of the low-frequency cut-off. If the low-frequency cut-off is inappreciable, this number may be sufficiently great so that the probability of encountering such a sequence in a random pulse train and resultant errors in reception may be so small that it can be disregarded. The requirement of the low-frequency cut-off which is necessary to this end is evaluated below for pulses transmitted at intervals $\tau_1 = 1/2f_1$.

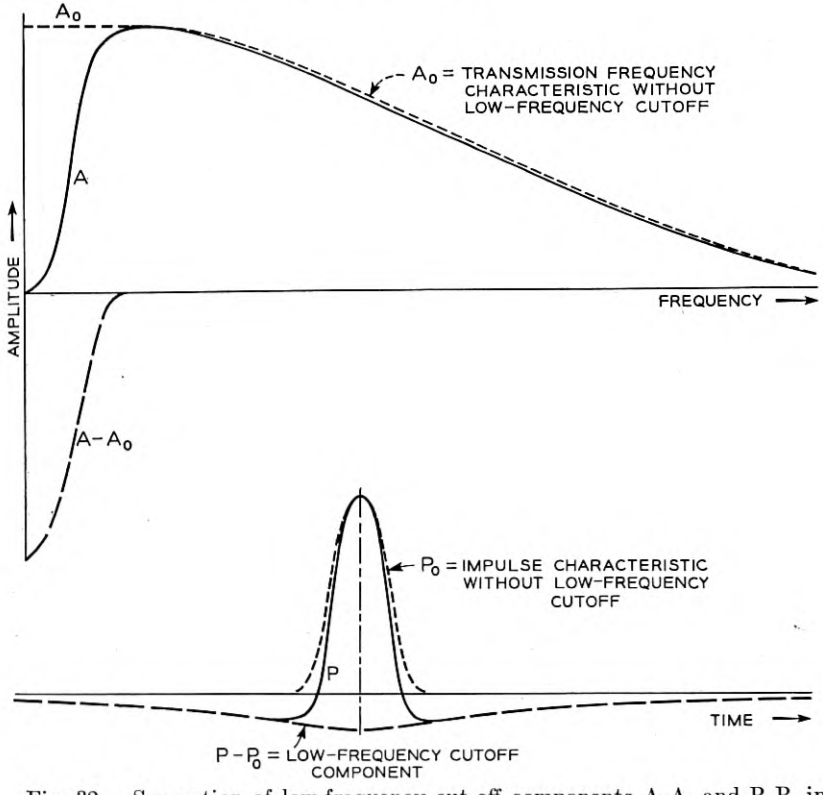


Fig. 32 — Separation of low-frequency cut-off componente $A-A_0$ and $P-P_0$ in transmission frequency and impulse characteristics.

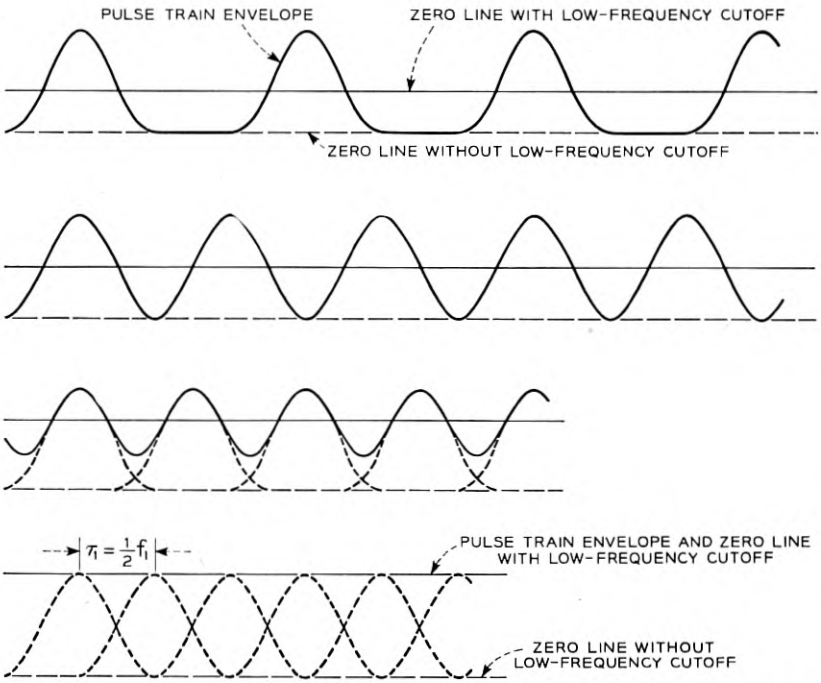


Fig. 33 — Effect of low-frequency cut-off on recurrent pulses as pulse interval is decreased.

If it is assumed that positive and negative impulses are applied at random to the transmission systems at intervals τ_1, the rms intersymbol interference resulting from a low-frequency cut-off can be evaluated by essentially the same method as employed in Section 8 for fine structure imperfections in the transmission characteristic, provided ω_0 is much smaller than ω_1. On this basis, rms intersymbol interference in relation to the peak amplitude $P_0(0)$ of the pulses in the absence of a low-frequency cut-off becomes:

$$\underline{U} = \frac{1}{P_0(0)} \left(\frac{1}{\tau_1} \int_{-\infty}^{\infty} [P(t) - P_0(t)]^2 \, dt \right)^{1/2}, \tag{9.03}$$

$$= \frac{1}{P_0(0)} \left(\frac{1}{\pi \tau_1} \int_0^{\infty} [A(\omega) - A_0(\omega)]^2 \, d\omega \right)^{1/2}. \tag{9.04}$$

For a transmission characteristic with linear phase shift

$$P_0(0) = \frac{1}{\pi} \int_0^{\infty} A_0(\omega) \, d\omega. \tag{9.05}$$

For the particular case of sharp cut-offs at ω_0 and ω_1

$$A_0(\omega) = 1 \qquad 0 < \omega < \omega_1 ,$$

$$A(\omega) - A_0(\omega) = -1 \qquad 0 < \omega < \omega_0 , \text{ and}$$

$$P_0(0) = \omega_1/\pi \qquad \tau_1 = \pi/\omega_1 ,$$

and

$$\underline{U} = \frac{\pi}{\omega_1}\left(\frac{\omega_1\omega_0}{\pi^2}\right)^{1/2} = \left(\frac{\omega_0}{\omega_1}\right)^{1/2}. \tag{9.06}$$

It will be noticed that the same result is obtained from (8.07) with the amplitude deviation $\alpha = [A(\omega) - A_0(\omega)] = -1$ between 0 and ω_0.

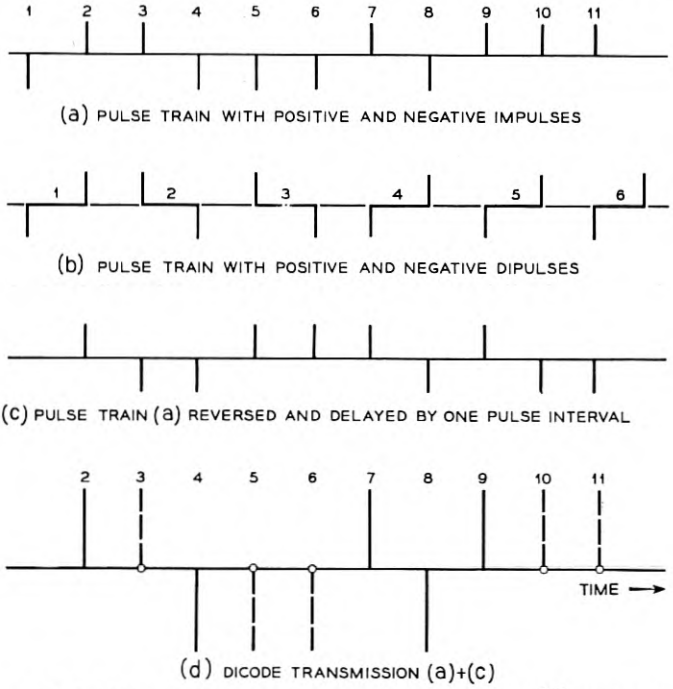
In actual systems, the low-frequency cut-off will be gradual between $\omega = 0$ and ω_0, rather than abrupt as assumed above. With a linear variation in the amplitude characteristic between 0 and ω_0, $A(\omega) - A_0(\omega) = (-1 + \omega/\omega_0)A_0(0)$ and $\underline{U} = (\omega_0/3\omega_1)^{1/2}$.

If a sufficient number of pulses of one polarity is transmitted in succession at intervals $\tau_1 = 1/2f_1$ the received pulses will as noted before in the limit be reduced to zero amplitude by the low-frequency cut-off. The maximum pulse distortion resulting from pulse overlaps when a train of pulses as transmitted is thus equal and opposite to the amplitude $P_0(0)$ of the received pulses in the absence of a low-frequency cut-off, so that peak intersymbol interference $\hat{U} = -1$. If rms intersymbol interference is held at one-quarter the peak value, i.e., $\underline{U} = 0.25$, the probability of encountering the maximum tolerable intersymbol interference and resultant errors in reception is low enough to be disregarded. On this basis the ratio ω_0/ω_1 would in accordance with (9.06) have to be less than 0.0625. Actually a substantially smaller ratio would be required because of intersymbol interference from other imperfections in the transmission characteristic and noise. Furthermore, a low-frequency cut-off will be accompanied by phase distortion at the low end of the transmission band, disregarded in the above evaluation. The requirements imposed on the low-frequency cut-off will thus be rather severe for a pulse system as assumed above in which random sequences of pulses are transmitted at intervals $\tau_1 = 1/2f_1$. Two pulse amplitudes were assumed above, and with a greater number of amplitudes the requirements would be more severe.

From Fig. 33 it is evident that the effect of a low-frequency cut-off on a received pulse train can be reduced by transmitting pulses at longer intervals than $\tau_1 = 1/2f_1$ considered above. For example, with a two-fold

increase in the pulse interval, as represented by the second case in Fig. 33, the maximum displacement of the zero line would be half the peak amplitude of the pulses. There would then be a 50 per cent reduction in the margin for distinction between the presence and absence of a pulse in a random pulse train, rather than a complete elimination of the margin for an infinite train of pulses of the same polarity transmitted at intervals $\tau_1 = 1/2f_1$. This improvement would be achieved at the expense of a two-fold increase in bandwidth for a given pulse transmission rate. A further improvement, for the same two-fold increase in bandwidth, can be achieved by "dipulse" transmission, as discussed below.

In dipulse transmission a positive pulse followed by a negative pulse in the next pulse position would be transmitted to indicate "on," and a negative pulse followed by a positive pulse to indicate "off," as indicated in Fig. 34. There will then be a substantial reduction in the pulse



THE PULSES AND ZEROS IN THE RECEIVED PULSE TRAIN (d) HAVE THE FOLLOWING RELATIONS TO THE ORIGINAL PULSES (a)

1. POSITIVE AND NEGATIVE PULSES IN (d) REPRESENT CORRESPONDING PULSES IN (a)
2. POINTS ON PULSE TRAIN IN (d) REPRESENT A REPETITION OF PREVIOUS PULSE, AS INDICATED BY DASHED LINES

Fig. 34 — Dipulse and dicode pulse transmission methods.

overlaps resulting from a low-frequency cut-off, as illustrated in Fig. 35, and in peak intersymbol interference.

If $\Delta P(t) = P(t) - P_0(t)$ is the modification in the impulse characteristic shown in Fig. 32, the modification in the dipulse transmission characteristic resulting from a low-frequency cut-off becomes

$$\Delta_1 P(t) = \Delta P(t) - \Delta P(t - \tau_1), \tag{9.07}$$

where τ_1 is the interval between the positive and negative dipulse components.

The difference given by (9.07) represents the differential in the curve $P(t) - P_0(t)$ shown in Fig. 32 over an interval τ_1. It can be shown that the maximum cumulative effect or peak intersymbol interference for a long pulse train is represented by the sum of the differentials given by (9.07) and is approximately equal to

$$\hat{U} \leqq \Delta P(\tau_1) = P(\tau_1) - P_0(\tau_1). \tag{9.08}$$

As an example, if the shape of $A - A_0$ in Fig. 32 were about the same as that of A_0, $\Delta P(t)$ would have the same shape as $P_0(t)$ but would be lower in peak amplitude by the factor f_0/f_1 and would have the time scale increased by the factor f_1/f_0. Peak intersymbol interference as

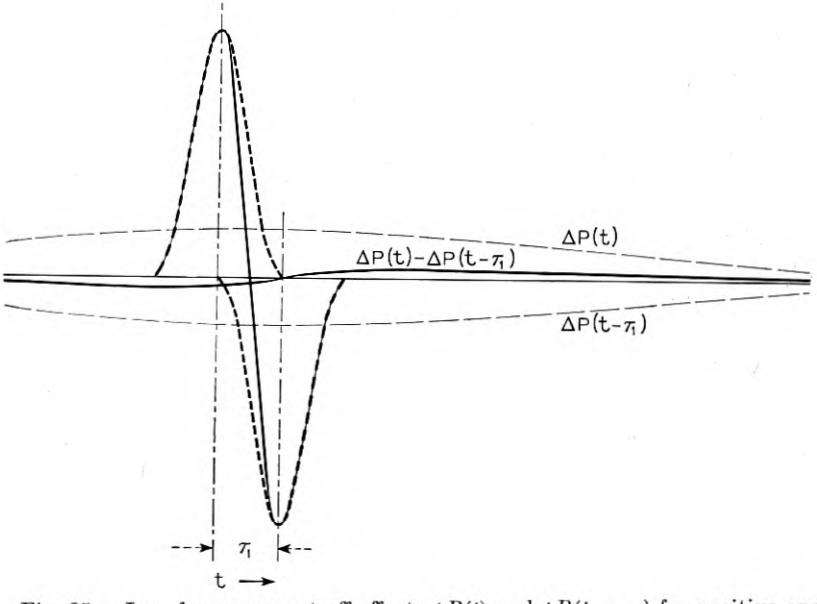


Fig. 35 — Low-frequency cut-off effects $\Delta P(t)$ and $\Delta P(t - \tau_1)$ for positive and negative pulses and resultant effect $\Delta_1 P(t) = \Delta P(t) - \Delta P(t - \tau_1)$ for a dipulse.

obtained from (9.08) would then be about $\hat{U} = f_0/f_1$ and thus in the order of 10 per cent of the peak pulse amplitude for $f_0/f_1 = 0.10$.

The bandwidth penalty incurred in dipulse transmission can be avoided by transmitting two identical pulse trains, one of which is delayed by one pulse interval and reversed in polarity with respect to the other.* The combined pulse train will then be as indicated in Fig. 34, and one or the other of the two original component pulse trains can be restored at the receiving end by suitable conversion equipment. In the combined pulse train, a pulse of one polarity is always followed by a pulse of opposite polarity, but not necessarily in the next pulse position. For this reason the low-frequency cut-off compensation with the above method of "dicode" transmission is not quite as effective as with dipulse transmission. Furthermore, since it is necessary to distinguish between three pulse amplitudes (1, 0, −1), in the received pulse train, the maximum tolerable pulse distortion in relation to the peak pulse amplitude is only half as great as with two pulse amplitudes (1, −1) in an ordinary code.

10. TRANSMISSION DISTORTION FROM BAND-EDGE PHASE DEVIATIONS

In pulse transmission systems where phase equalization is employed, it may be impracticable or unnecessary to equalize over the entire transmission band. There will then be residual phase distortion near the band-edges, as indicated in Fig. 36. This type of phase deviation will give rise to pulse distortion extending over appreciable time intervals if the band-edge phase deviations are large, as indicated in the above figure, for the reason that the frequency components outside the linear phase range will be received with increased transmission delay. Evaluation of the pulse shape is in this case a rather elaborate procedure, but rms pulse distortion or intersymbol interference resulting from such phase distortion can readily be determined as outlined below. In certain pulse modulation systems, such as PAM time division systems, rms intersymbol interference is of principal interest. In other systems where peak intersymbol interference is controlling, it may usually be estimated with engineering accuracy by applying a peak factor.

When the pulse shape is known, peak intersymbol interference may be determined by methods outlined in Section 13. Comparison of peak intersymbol interference evaluated in this manner with rms pulse distortion, for some cases in which the pulse shapes in the presence of phase

* L. A. Meacham originally proposed this method is an unpublished memorandum.

distortion were determined, indicates that the peak factor is about 3 when phase distortion is appreciable and the pulses are substantially prolonged in duration.

Returning to equation (8.17) and assuming $\alpha = 0$, the following relationship is obtained for rms intersymbol interference due to phase deviations

$$\underline{U} = \frac{1}{P(0)}\left(\frac{1}{\pi\tau_1}\right)^{1/2}\left[\int_0^\infty 2A_0^2(1 - \cos\beta)\,d\omega\right]^{1/2}, \qquad (10.01)$$

where $\beta = \beta(\omega)$ is the deviation from a linear phase characteristic.

For transmission systems with a linear phase shift, the peak amplitude of the pulses is given by (8.23) and with this relation in (10.01)

$$\underline{U} = \left(\frac{\pi}{\omega_1\tau_1}\right)^{1/2}\lambda, \qquad (10.02)$$

where

$$\lambda = \left[\omega_1\int_0^{\omega_{max}} 2A_0^2(1 - \cos\beta)\,d\omega\right]^{1/2} \Big/ \left[\int_0^{\omega_{max}} A_0\,d\omega\right]. \qquad (10.03)$$

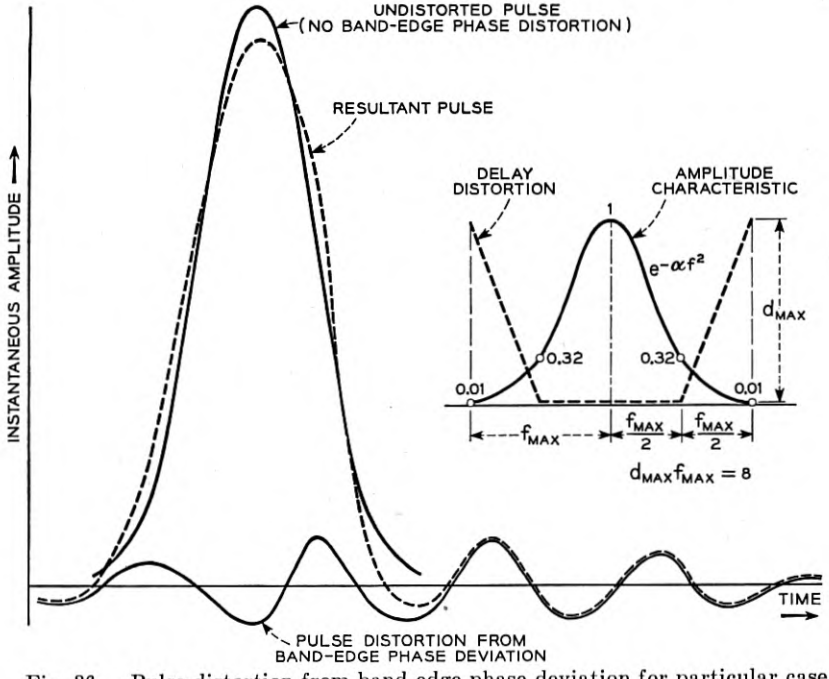


Fig. 36 — Pulse distortion from band-edge phase deviation for particular case of linear band-edge delay distortion.

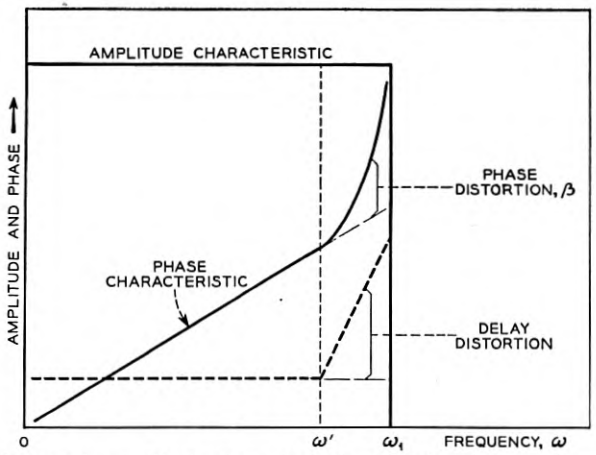


Fig. 37 — Constant amplitude characteristic with band-edge phase distortion.

If there is no phase distortion, i.e., $\beta = 0$, between $\omega = 0$ and ω', equation (10.03) becomes

$$\lambda = \left[\omega_1 \int_{\omega'}^{\omega_{max}} 2A_0^2(1 - \cos \beta\, d\omega\right]^{1/2} \Big/ \left[\int_0^{\omega_{max}} A_0\, d\omega\right]. \quad (10.04)$$

As an example, consider a parabolic deviation from a linear phase characteristic between ω' and ω_{max}, in which case delay distortion would vary linearily in this band, as indicated in Fig. 37 for a constant amplitude characteristic for which $\omega_{max} = \omega_1$. In this case

$$\beta = \beta_1 \left(\frac{\omega - \omega'}{\omega_1 - \omega'}\right)^2, \quad (10.05)$$

where β_1 is the maximum phase deviation, obtained for $\omega = \omega_1$.

Equation (10.04) in the above case becomes:

$$\begin{aligned} \lambda^2 &= \frac{2}{\omega_1} \int_{\omega'}^{\omega_1} \left[1 - \cos \beta_1 \left(\frac{\omega - \omega'}{\omega_1 - \omega'}\right)^2\right] d\omega, \\ &= \frac{2(\omega_1 - \omega')}{\omega_1} \left[1 - \beta_1^{-1/2} \int_0^{\beta_1^{1/2}} \cos u^2\, du\right], \quad (10.06) \\ &= \frac{2(\omega_1 - \omega')}{\omega_1} \left[1 - \frac{1}{2}\left(\frac{\pi}{2\beta_1}\right)^{1/2} (R + X)\right], \end{aligned}$$

where $R + iX = \text{erf}\,(\beta_1^{1/2} e^{i\pi/4})$ in which erf is the error function.

For a constant amplitude transmission characteristic as assumed above, $(\pi/\omega_1\tau_1) = 1$, so that (10.02) becomes $\underline{U} = \lambda$, which may also

be written:

$$\underline{U} = \left(\frac{\omega_1 - \omega'}{\omega_1}\right)^{1/2} \cdot F(\beta_1), \qquad \text{and} \tag{10.07}$$

$$F(\beta_1) = 2^{1/2}\left[1 - \frac{1}{2}\left(\frac{\pi}{2\beta_1}\right)^{1/2}(R + X)\right]^{1/2}. \tag{10.08}$$

For various values of the maximum phase deviation β_1 in radians the function F becomes:

β_1	0	0.25	1	4	∞
F	0	0.14	0.43	1.24	1.42

If, for example, phase distortion were confined to 10 per cent of the transmission band, then $(\omega_1 - \omega')/\omega_1 = 0.1$. For a maximum phase deviation of 1 radian at the edge of the transmission band, $F = 0.43$ and $\underline{U} = 0.135$. For a maximum phase distortion of 4 radians, $F = 1.24$ and $\underline{U} = 0.39$. Since peak intersymbol interference may exceed the above rms values by a factor of about 3, and the maximum tolerable peak intersymbol interference in a system employing two pulse amplitudes would be less than 1, it is evident that band-edge phase deviations must be held at rather small values, less than about 3 radians, in the upper 10 per cent of the transmission band.

The above severe tolerances on band-edge phase distortion can be overcome by employing a transmission frequency characteristic of the type shown in Fig. 38 and previously discussed in Section 5. If the phase characteristic is linear between $\omega = 0$ and ω_1, and phase distortion between ω_1 and $2\omega_1$ varies as

$$\beta = \beta_1\left(\frac{\omega - \omega_1}{2\omega_1 - \omega_1}\right)^2 = \beta_1\left(1 - \frac{\omega}{\omega_1}\right)^2, \tag{10.09}$$

equation (10.04) can be written

$$\begin{aligned}\lambda^2 &= \frac{1}{\omega_1}\int_{\omega_1}^{2\omega_1}\left(1 + \cos\frac{\pi\omega}{2\omega_1}\right)^2\left[1 - \cos\beta_1\left(1 - \frac{\omega}{\omega_1}\right)^2\right]d\omega, \\ &= \int_0^1\left(1 - \sin\frac{\pi}{2}u\right)^2(1 - \cos\beta_1 u^2)\,du.\end{aligned} \tag{10.10}$$

Pulses may also in this case be transmitted at intervals $\tau_1 = \pi/\omega_1$ without intersymbol interference in the absence of phase distortion, so

that (10.02) becomes $\underline{U} = \lambda$ or

$$\underline{U} = \left[\frac{1}{2}\int_0^1 \left(1 - \sin\frac{\pi}{2}u\right)^2 (1 - \cos\beta_1 u^2)\, du\right]^{1/2}. \qquad (10.11)$$

The maximum delay distortion at the edge of the transmission band, i.e., $\omega = 2\omega_1$, is $d_{max} = 2\beta_1/\omega_1$. The product of this delay distortion with the maximum frequency $f_{max} = 2f_1$ is $d_{max}f_{max} = 2\beta_1/\pi$. For various values of maximum phase distortion β_1 and the corresponding product $d_{max}f_{max}$, the following values of rms intersymbol interference are obtained by numerical integration of (10.11). (This integral can be expressed in terms of a number of Fresnel integrals, but numerical integration is simpler and sufficiently accurate for the present purpose.)

β_1	π	2π	4π	∞
$d_{max}\, f_{max}$	2	4	8	∞
$\underline{U}$	0.070	0.120	0.185	0.330

The particular case $d_{max}f_{max} = 8$ is similar to that shown in Fig. 36, except that this figure applies to a Gaussian characteristic, for which the amplitude at $\omega = \omega_1$ has been taken as 0.32 rather than 0.5 in the case considered here. For this reason rms intersymbol interference from phase distortion would be greater in the present case.

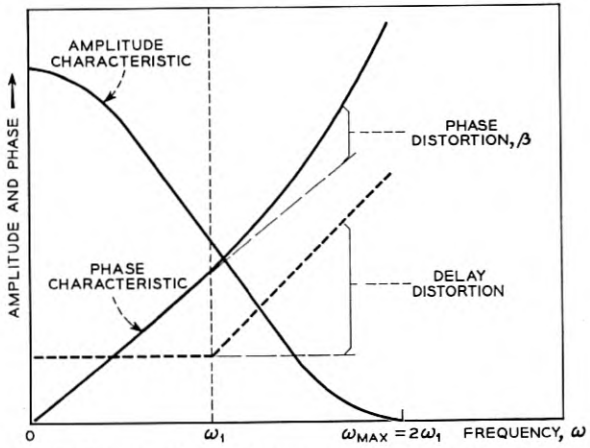


Fig. 38 — Typical transmission frequency characteristic with phase equalization over 50 per cent of transmission band.

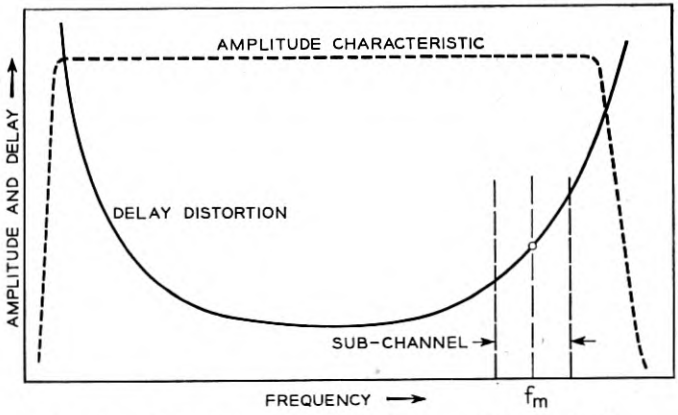


Fig. 39 — Sub-channel with nearly linear delay distortion.

Peak intersymbol interference may exceed the above rms values by a factor of about 3. In a system employing two pulse amplitude (1 and -1), the maximum tolerable intersymbol interference is 1. This value would thus be attained in the above case for $d_{\max} f_{\max} = \infty$. Hence, in a system employing two pulse amplitudes, and in the absence of noise and intersymbol interference from other sources, there would be no limitation on phase distortion for $\omega > \omega_1$, provided the phase characteristic is linear between $\omega = 0$ and ω_1.

11. BAND-PASS CHARACTERISTICS WITH LINEAR DELAY DISTORTION

In Fig. 39 is shown a transmission frequency characteristic together with an assumed delay distortion $d\psi/d\omega$. When a portion of the transmission band is employed for pulse transmission, as for example in pulse signalling, data or telegraph transmission over portion of a voice channel, there may be an appreciable component of substantially linear delay distortion, as indicated in the above figure. The departure from a linear variation can usually be approximated by a cosine variation in delay, and the system can then be regarded as made up of two components in tandem, one with linear the other with cosine variation in delay. The effect of the latter can be evaluated by the methods outlined in Section 6, and the effect of a linear variation by methods established in this section.

In Fig. 40 is shown a symmetrical amplitude characteristic with linear delay distortion over the transmission band. Phase distortion with respect to the midband frequency is in this case

$$\Psi(u) = \beta u^2 \qquad \text{and} \qquad \Psi(-u) = \beta u^2, \tag{11.01}$$

and delay distortion

$$d\Psi(u)/du = 2\beta u, \; d\Psi(-u)/du = -2\beta u. \tag{11.02}$$

(The symbol β, together with α, η, a and b used later in this section do not have the same meaning as in earlier sections.) With (11.01) in (2.10) and (2.11), the in-phase and quadrature components in (2.09) become

$$\begin{aligned} R_- + R_+ &= \frac{2\delta}{\pi} \int_0^\infty \mathfrak{a}(u) \cos ut \cos \beta u^2, \quad \text{and} \\ Q_- + Q_+ &= \frac{2\delta}{\pi} \int_0^\infty \mathfrak{a}(u) \cos ut \sin \beta u^2. \end{aligned} \tag{11.03}$$

The in-phase and quadrature components can accordingly be identified with the real and the negative imaginary component of the integral

$$J = \frac{2\delta}{\pi} \int_0^\infty \mathfrak{a}(u) \cos ute^{-i\beta u^2} \, du. \tag{11.04}$$

The solution of this integral is rather simple for the particular case of a Gaussian transmission characteristic

$$\mathfrak{a}(u) = e^{-\alpha u^2}, \tag{11.05}$$

in which case

$$\begin{aligned} J &= \frac{2\delta}{\pi} \int_0^\infty e^{-(\alpha + i\beta)u^2} \cos ut \, du, \\ &= \frac{\delta}{[\pi(\alpha + i\beta)]^{1/2}} e^{-t^2/4(\alpha + i\beta)}. \end{aligned} \tag{11.06}$$

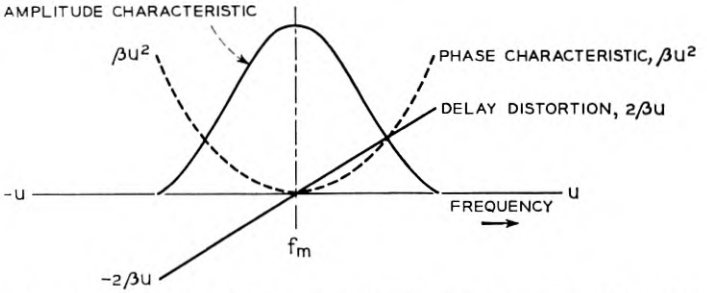


Fig. 40 — Symmetrical band-pass amplitude characteristic with linear delay distortion.

The real and negative imaginary components of this expression are

$$R_- + R_+ = 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} \cos(\Theta - bt^2), \quad \text{and}$$
$$Q_- - Q_+ = 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} \sin(\Theta - bt^2), \tag{11.07}$$

where

$$a = \frac{\alpha}{4(\alpha^2 + \beta^2)} \qquad b = \frac{\beta}{4(\alpha^2 + \beta^2)} \qquad c = (\alpha^2 + \beta^2)^{1/2}$$
$$\tan 2\Theta = \beta/\alpha = b/a$$

The impulse characteristic obtained with (11.07) in (2.09) becomes

$$P(t) = 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} [\cos(\omega_r t - \psi_r) \cos(\Theta - bt^2) + \sin(\omega_r t - \psi r) \sin(\Theta - bt^2)]. \tag{11.08}$$

From (11.08) it is seen that the envelope is

$$\overline{P}(t) = 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2}. \tag{11.09}$$

The peak of the envelope obtained with $t = 0$ is smaller than without delay distortion ($\beta = 0$) by the factor

$$\eta = \frac{1}{[1 + (\beta/\alpha)^2]^{1/2}}. \tag{11.10}$$

The constant a is smaller than without delay distortion by the factor η^4. If t_0 designates the time required for the instantaneous amplitude of a pulse to decay from its peak to a given value without delay distortion, the time t_1 to reach the same amplitude with delay distortion is

$$t_1 = t_0/\eta^2 = t_0[1 + (\beta/\alpha)^2]^{1/2}. \tag{11.11}$$

If ω_{max} indicates the frequency at the 40 db down point on the transmission frequency characteristic, $\alpha\omega_{max}^2 = 4.6$, The corresponding delay distortion is $d_{max} = 2\omega_{max}\beta$. Thus $\beta/\alpha = .68\, d_{max} f_{max}$ so that (11.11) becomes:

$$t_1 = t_0[1 + 0.46\,(d_{max} f_{max})^2]^{1/2}. \tag{11.12}$$

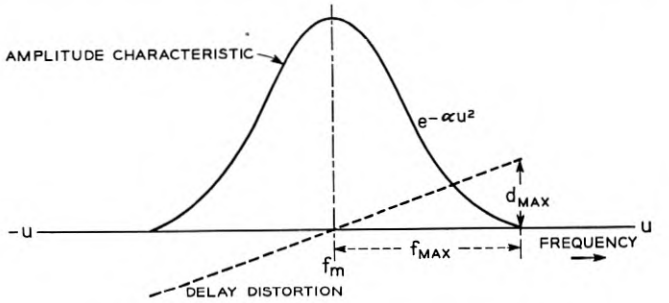
The effect of a linear delay distortion across the transmission band is thus to disperse or broaden the envelope of the received pulses, as illus-

trated in Fig. 41. For a specified pulse overlap or intersymbol interference the pulse spacing must accordingly be increased by the factor t_1/t_0, so that for a given transmission performance the transmission capacity is reduced by the factor t_0/t_1. About the same effect would be expected for other pulse shapes or amplitude characteristics resembling the Gaussian shape assumed in the above derivation.

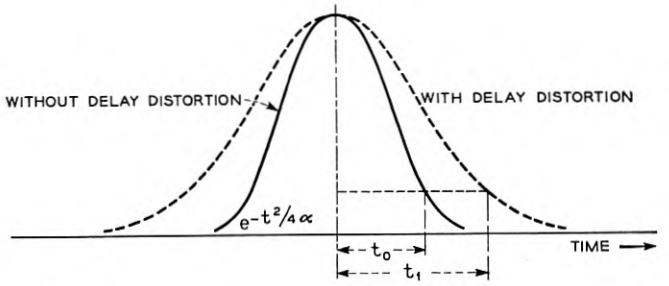
Comparison of (11.08) with (2.13) shows that the function $\varphi(t)$ with respect to the midband frequency is

$$\varphi(t) = \theta - bt^2. \tag{11.13}$$

If the reference or carrier frequency is displaced from the midband by



(a) TRANSMISSION FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTICS WITHOUT AND WITH DELAY DISTORTION

$$t_1 = t_0[1+0.46(d_{MAX}\, f_{MAX})^2]^{\frac{1}{2}}$$

f_{MAX} = FREQUENCY FROM MIDBAND AT WHICH AMPLITUDE OF TRANSMISSION FREQUENCY CHARACTERISTIC IS REDUCED 40 DECIBELS

d_{MAX} = DELAY DISTORTION AT f_{MAX} IN SECONDS

Fig. 41 — Lengthening of impulse envelope by linear delay distortion for Gaussian transmission characteristic.

ω_y, the in-phase and quadrative components are in accordance with (2.18)

$$\begin{aligned} R_-' + R_+' &= \cos(\theta - bt^2 + \omega_y t - \psi_y)\,\overline{P}(t), \quad \text{and} \\ Q_-' - Q_+' &= \sin(\theta - bt^2 + \omega_y t - \psi_y)\,\overline{P}(t), \end{aligned} \tag{11.14}$$

where $\psi_y = \beta\omega_y^2$ and $\overline{P}(t)$ is given by (11.09).

REFERENCES

1. Y. W. Lee, Synthesis of Electric Networks by Means of the Fourier Transforms of Laguere's Functions, J. Math. & Phys., June, 1932.
2. H. W. Bode, *Network Analysis and Negative Feedback Amplifier Design*, D. Van Nostrand Book Company, 1945.
3. H. Nyquist, Certain Topics in Telegraph Transmission Theory, A.I.E.E. Trans., April, 1928.
4. H. Nyquist and K. W. Pfleger, Effect of Quadrature Component in Single Sideband Transmission, B.S.T.J., Jan., 1940.
5. H. A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, Proc. I.R.E., June, 1939.
6. C. R. Burrows, Discussion of 5 above, Proc. I.R.E., June, 1939.
7. G. N. Watson, *Theory of Bessel Functions*, Cambridge University Press, 1944.
8. E. Jahnke and F. Emde, *Funktionentafeln*, 1928, p. 149.
9. E. A. Guillemin, *Communication Networks*, John Wiley & Sons, Inc., 1935.
10. W. R. Bennett, Time Division Multiplex Systems, B.S.T.J., April, 1941.
11. S. Goldman, *Frequency Analysis, Modulation and Noise*, McGraw-Hill Book Co., 1948.
12. C. E. Shannon, A Mathematical Theory of Communication, B.S.T.J., October, 1948.
13. B. M. Oliver, J. R. Pierce and C. E. Shannon, The Philosophy of PCM, Proc. I.R.E., Nov., 1948.

Bell System Technical Papers Not Published in this Journal

AIKENS, A. J.,[1] and C. S. THAELER.[2]

Noise and Crosstalk Control on N1 Carrier Systems, Elec. Eng., **72,** pp. 1075–1080, Dec., 1953.

ALLEY, R. E., JR.,[1] and F. J. SCHNETTLER.[1]

Effect of Cross-Section Area and Compression Upon the Relaxation in Permeability for Toroidal Samples of Ferrites, Letter to the Editor, J. Appl. Phys., **24,** pp. 1524–1525, Dec., 1953.

ALLIS, W. P.,[1] and D. J. ROSE.[1]

The Transition From Free to Ambipolar Diffusion, Phys. Rev., **93,** p. 84–93, Jan. 1, 1954.

ANDERSON, O. L.,[1] and D. A. STUART.[4]

Statistical Theories as Applied to the Glassy State, Ind. Eng. Chem., **46,** pp. 154–160, Jan., 1954.

ARNOLD, S. M., see S. E. KOONCE.

BARSTOW, J. M.,[1] and H. N. CHRISTOPHER.[1]

The Measurement of Random Monochrome Video Interference. A.I.E.E., Commun. and Electronics, pp. 735–741, Jan., 1954.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[4] Cornell University.

BASHKOW, T. R.[1]

Stability Analysis of a Basic Transistor Switching Circuit, Proc. National Electronics Conference, **9,** p. 748, Feb. 15, 1954.

BLECHER, F. H.[1]

Automatic Gain Control of Junction Transistor Amplifiers, Proc. National Electronics Conference, **9,** p. 731, Feb. 15, 1954.

BOGERT, B. P.[1]

Erratum: On the Band Width of Vowel-Formants, [published in J. Acous. Soc. Am., **25,** p. 791, (1953)], J. Acous. Soc. Am., **25,** p. 1203, Nov., 1953. The statement in the abstract which reads "The mean values for bars 1, 2, and 3 were 130, 100, and 185 cps, respectively," should be corrected to read "The median values for bars 1, 2, and 3, were 130, 150, and 185 cps, respectively."

BROWN, W. L.,[1] R. C. FLETCHER,[1] and K. A. WRIGHT.[5]

Annealing of Bombardment Damage in Germanium — Experimental, Phys. Rev., **92,** pp. 591–596, Nov. 1, 1953.

BROWN, W. L., see R. C. FLETCHER.

BURTON, J. A.,[1] G. W. HULL,[1] F. J. MORIN,[1] and J. C. SEVERIENS.[1]

Effect of Nickel and Copper Impurities on the Recombination of Holes and Electrons in Germanium, J. Phys. Chem., **57,** pp. 853–859, Nov., 1953.

BURTON, J. A.,[1] R. C. PRIM,[1] and W. P. SLICHTER.[1]

Distribution of Solute in Crystals Grown from the Melt — Theoretical, J. Chem. Phys., **21,** pp. 1987–1991, Nov., 1953.

BURTON, J. A.,[1] E. D. KOLB,[1] W. P. SLICHTER,[1] and J. D. STRUTHERS.[1]

Distribution of Solute in Crystals Grown from the Melt — Experimental, J. Chem. Phys., **21,** pp. 1991–1996, Nov., 1953.

CAMPBELL, M. E., see C. L. LUKE.

[1] Bell Telephone Laboratories, Inc.
[5] Massachusetts Institute of Technology.

CASE, R. L.,[1] and IDEN KERNEY.[1]

Program Transmission Over Type-N Carrier Telephone, A.I.E.E.. Commun. and Electronics, pp. 791–795, Jan., 1954.

CHRISTOPHER, H. N., see J. M. BARSTOW.

CLARK, M. A.[1]

An Acoustic Lens as a Directional Microphone, J. Acous. Soc. Am., **25,** pp. 1152–1153, Nov., 1953.

CORENZWIT, E., see S. GELLER.

COY, J. A.[1]

Heat Dissipation from Toll Transmission Equipment, A.I.E.E., Commun. and Electronics, pp. 762–768, Jan., 1954.

DUNN, H. K.[1]

Remarks on a Paper Entitled "Multiple Helmholtz Resonators," Letter to the Editor, J. Acous. Soc. Am., **26,** p. 103, Jan., 1954.

FLETCHER, R. C.,[1] and W. L. BROWN.[1]

Annealing of Bombardment Damage in a Diamond-Type Lattice — Theoretical, Phys. Rev., **92,** pp. 585–590, Nov. 1, 1954.

FLETCHER, R. C., see W. L. BROWN.

FRACASSI, R. D.,[1] and H. KAHL.[1]

Type ON Carrier Telephone, A.I.E.E., Commun and Electronics, pp. 713–721, Jan., 1954.

FULLER, C. S., see J. R. SEVERIENS.

GELLER, S.,[1] and E. CORENZWIT.[1]

Hafnium Oxide, HfO_2 (Monoclinic), Anal. Chem., **25,** p. 1774, Nov., 1953.

---

[1] Bell Telephone Laboratories, Inc.

GIBBS, W. B.[3]

Investment Cast Artificial Larynx Eases Fabrication Difficulties, Precision Metal Molding, pp. 34, 35, 75, Dec., 1953.

GREEN, F. O.[3]

Cost Control — Bonus from Centralized Maintenance, Plant Engineering, pp. 88–91, Jan., 1954.

HAGSTRUM, H. D.[1]

Instrumentation and Experimental Procedure for Studies of Electron Ejection by Ions and Ionization by Electron Impact, Rev. Sci. Instr., **24,** pp. 1122–1142, Dec., 1953.

HANSON, A. N.[1]

Automatic Testing of Wired Relay Circuits, A.I.E.E., Commun. and Electronics, pp. 805–857, Jan., 1954.

HULL, G. W., see J. A. BURTON.

KAHL, H., see R. D. FRACASSI.

KERNEY, IDEN, see R. L. CASE.

KOCH, W. E.[1]

Use of the Sound Spectrograph for Appraising the Relative Quality of Musical Instruments, Letter to the Editor, J. Acous. Soc. Am., **26,** p. 105, Jan., 1954.

KOLB, E. G., see J. A. BURTON.

KOONCE, S. E.[1] and S. M. ARNOLD.[1]

Metal Whiskers, Letter to the Editor, J. Appl. Phys., **25,** pp. 134–135, Jan., 1954.

---

[1] Bell Telephone Laboratories, Inc.
[3] Western Electric Company, Inc.

KRETZMER, E. R.[1]

An Amplitude-Stabilized Transistor Oscillator, Proc. National Electronics Conference, p. 756, Feb. 15, 1954.

LEWIS, H. W.[1]

Search for the Hall Effect in a Super-Conductor — Experiment, Phys. Rev., **92,** pp. 1149–1151, Dec., 1953.

LINVILLE, J. G.[1]

A New RC Filter Employing Active Elements, Proc. National Electronics Conference, **9,** p. 342, Feb. 15, 1954.

LUKE, C. L.,[1] and M. E. CAMPBELL.[1]

Determination of Impurities in Germanium and Silicon, Anal. Chem., **25,** pp. 1588–1593, Nov., 1953.

MAHONEY, J. J., see E. H. PERKINS.

MAY, J. E.[1]

Characteristics of Ultrasonic Delay Lines Using Quartz and Barium Titanite Ceramic Transducer, Proc. National Electronics Conference, **9,** p. 264, Feb. 15, 1954.

MORIN, F. J.[1]

Lattice Scattering Mobility in Germanium, Phys. Rev., **93,** pp. 62–63, Jan. 1, 1954.

MORIN, F. J., see J. A. BURTON.

MURPHY, E. J.[1]

Surface Migration of Water Molecules in Ice, J. Chem. Phys., **21,** pp. 1831–1835, Oct., 1953.

PENNELL, E. S.[1]

A Temperature Controlled Ultrasonic Solid Delay Line, Proc. National Electronics Conference, **9,** p. 255, Feb. 15, 1954.

[1] Bell Telephone Laboratories, Inc.

PERKINS, E. H.,[1] and J. J. MAHONEY.[1]

Type-N Carrier Telephone Deviation Regulator, A.I.E.E., Commun. and Electronics, pp. 757–762, Jan., 1954.

PETERSON, G. E.,[6] and GORDON RAISBECK.[1]

The Measurement of Noise with the Sound Spectrograph, J. Acous. Soc. Am., **25,** p. 1157, Nov., 1953.

PRIM, R. C., see J. A. BURTON.

PRINCE, M. B.[1]

Drift Mobilities in Semi-Conductors. I — Germanium, Phys. Rev., **92,** pp. 681–687, Nov. 1, 1953.

RAISBECK, GORDON, see G. E. PETERSON.

REA, W. W.[3]

Oscilloscope Reduces Cost of Jig Grinding Operations, Machinery, pp. 201–203, Dec., 1953.

REMEIKA, J. P.

Method for Growing Barium Titanate Single Crystals. J. Am. Chem. Soc., **76,** pp. 940–941, Feb. 5, 1954.

ROSE, D. J., see W. P. ALLIS.

SCHLAACK, N. F.[1]

Development of the LD Radio System, Trans. I.R.E., Professional Group on Communication Systems, pp. 29–38, Jan., 1954.

SCHNETTLER, F. J., see R. E. ALLEY, JR.

SEVERIENS, J. C., see J. A. BURTON.

---

[1] Bell Telephone Laboratories, Inc.
[3] Western Electric Company, Inc.
[6] University of Michigan.

SEVERIENS, J. R.,[1] and C. S. FULLER.[1]

Mobility of Impurity Ions in Germanium and Silicon, Letter to the Editor, Phys. Rev., **92,** pp. 1322, Dec., 1953.

SHOCKLEY, W.[1]

Transistor Physics, Am. Scientist, **42,** pp. 41–72, Jan., 1954.

SHOCKLEY, W.[1]

Some Predicted Effects of Temperature Gradient on Diffusion in Crystals, Letter to the Editor, Phys. Rev., **93,** pp. 345–346, Jan. 15, 1954.

SLICHTER, E. D., see J. A. BURTON.

SNOREK, F.[3]

Cut Tool Costs with Precision Casting, Iron Age, pp. 136–139, Feb. 11, 1954.

STILES, K. P.[2]

Overseas Radio Telephone Services of A. T. and T. Co., Trans. I.R.E., Professional Group Communications Systems, pp. 39–44, Jan., 1954.

STRUTHERS, J. D., see J. A. BURTON, and C. D. THURMOND.

STUART, D. A., see O. L. ANDERSON.

THAELER, C. S., see A. J. AIKENS.

THURMOND, C. D.[1]

Equilibrium Thermochemistry of Solid and Liquid Alloys of Germanium and Silicon — The Solubility of Ge and Si in Elements of Group III, IV and V, J. Phys. Chem., **57,** pp. 827–830, Nov., 1953.

THURMOND, C. D.,[1] and J. D. STRUTHERS.[1]

Equilibrium Thermochemistry of Solid and Liquid Alloys of Ger-

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[3] Western Electric Company, Inc.

manium and of Silicon — The Retrograde Solid Solubilities of Sb in Ge, Cu in Ge, and Cu in Si, J. Phys. Chem., **57,** pp. 831–834, Nov., 1953.

WALKER, L. R.[1]

Dispersion Formula for Plasma Waves, Letter to the Editor, J. Appl. Phys., **25,** pp. 131–132, Jan., 1954.

WOLFF, P. A.[1]

Theory of Plasma Waves in Metals, Phys. Rev., **92,** pp. 18–23, Oct. 1, 1953.

WRIGHT, K. A., see W. L. BROWN.

[1] Bell Telephone Laboratories, Inc.

Contributors to this Issue

M. M. Atalla, B.S., Cairo University, 1945; M.S., Purdue University, 1947; Ph.D., Purdue University, 1949; Studies at Purdue undertaken as the result of a scholarship from Cairo University for four years of graduate work. Bell Telephone Laboratories, 1950–. For the past three years he has been a member of the Switching Apparatus Development Department, in which he is supervising a group doing fundamental research work on contact physics and engineering. Current projects include fundamental studies of gas discharge phenomena between contacts, their mechanisms, and their physical effects on contact behavior; also fundamental studies of contact opens and resistance. In 1950, an article by him was awarded first prize in the junior member category of the A.S.M.E. He is a member of Sigma Xi, Sigma Pi Sigma, Pi Tau Sigma, the American Physical Society, and an associate member of the A.S.M.E.

James M. Early, B.S., cum laude, New York State College of Forestry, 1943; M.S. and Ph.D. Ohio State University, 1948 and 1951. Bell Telephone Laboratories 1951–. After teaching Electrical Engineering at Ohio State University for five years while studying for his Master's and Ph.D. degrees, Dr. Early joined an electronic apparatus development group, participating in the development of the junction transistor. At present he is doing general theoretical work as well as development work on high frequency junction transistors. Member of the I.R.E. and Eta Kappa Nu. Associate of Sigma Xi.

Walter T. Eppler, B.S., in E.E., Tufts College, 1927; Western Electric Company, purchase of special machinery and development of coil manufacture, 1927–1932; Crystal Golf Ball Company, Superintendent of Manufacture, 1933–1936; New Haven Clock Company, 1936–1937; Western Electric Company, 1937–. In 1941, he engineered a conveyorized dispatch control system for key assembly at the Kearny plant. For the past six years, he has been engaged in cable sheath engineering on Alpeth and Stalpeth cable. Member of New Jersey Society of Professional Engineers.

STEWART E. MILLER, University of Wisconsin, 1936–39; B.S. and M.S., Massachusetts Institute of Technology, 1941. Bell Telephone Laboratories, 1941–. Except for World War II work on airborne radar systems, Mr. Miller's first eight years at the Laboratories were concerned with studies on coaxial carrier transmissions systems. A member of the radio research group, he is currently in charge of research on guided systems and associated millimeter and microwave techniques at Holmdel. Member of the I.R.E., Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

HARRY SUHL, B.Sc., University of Wales, 1943; Ph.D., Oriel College, University of Oxford, 1948. Admiralty Signal Establishment, 1943–46; Bell Telephone Laboratories, 1948–. Dr. Suhl conducted research on the properties of germanium until 1950 when he became concerned with electron dynamics and solid state physics research. His current work is in the applied physics of solids. Member of the American Institute of Physics and Fellow of the American Physical Society.

ERLING D. SUNDE, E.E., Technische Hochschule, Darmstadt, Germany, 1926. Brooklyn Edison Company, 1927; American Telephone and Telegraph Company, 1927–1934; Bell Telephone Laboratories, 1934–. Mr. Sunde's work has been centered on theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series Book. Member of the A.I.E.E., the American Mathematical Society, and the American Association for the Advancement of Science.

LAURENCE R. WALKER, B.Sc. and Ph.D., McGill University, 1935 and 1939; University of California, 1939–41. Radiation Laboratory, Massachusetts Institute of Technology, 1941–1945; Bell Telephone Laboratories, 1945–. Dr. Walker has been primarily engaged in research on microwave oscillators and amplifiers. At present he is a member of the physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.