THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIV JULY 1955 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at $3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIV JULY 1955 NUMBER 4

A New Local Video Transmission System

By STEPHEN DOBA, JR., and A. ROBERT KOLDING

(Manuscript received March 4, 1955)

The design features of the new A2A television transmission system are described. Under the diverse conditions of circuit length and repeater spacing in short-haul television connecting circuits, considerable flexibility is required. Several examples are given of the use of the A2A system to illustrate this flexibility.

INTRODUCTION

The A2A system is a new broadband wire transmission system for providing television connecting circuits over short distances. The system provides video transmission for frequencies up to 4.5 mc over balanced pairs designed for such use. Its design is predicated on meeting high quality performance objectives for a 4,000 mile network, which may comprise a number of A2A circuits along with intercity television systems of other types.

Television connecting circuits are used for a variety of purposes in network operations. Some examples of these uses are illustrated by the diagram of Fig. 1. Large broadcasters usually have their studios, master control, and broadcast transmitter distributed among several buildings in various locations in a city or metropolitan area. The A2A system furnishes broadband tielines to interconnect these facilities. Two-way con-

nections between the master control location and the studio are often required for programming purposes. For example, filmed material from the master control location may be sent to the studio, there to be combined with the live program and returned to the master control over a second circuit. For local broadcast transmission a third circuit to the radio transmitter is required.

For network operation, connecting circuits are required between the master control and the central switching point where connections to the intercity coaxial or microwave radio systems are made. From the Television Operating Center, as this switching point is called, wire circuits are frequently used to send to or receive from a microwave terminal which may be located outside the city. It is interesting to note that three local circuits may appear in tandem when the station is receiving from the network, and either three or four may occur in tandem when the station is feeding the network with material from a remote studio.

Local networks for theater television and for other closed circuit arrangements such as medical demonstrations are other uses for local wire systems.

The first systems used for network operation were the L1 coaxial system[1] for intercity circuits and the A2 video transmission system[2] for local

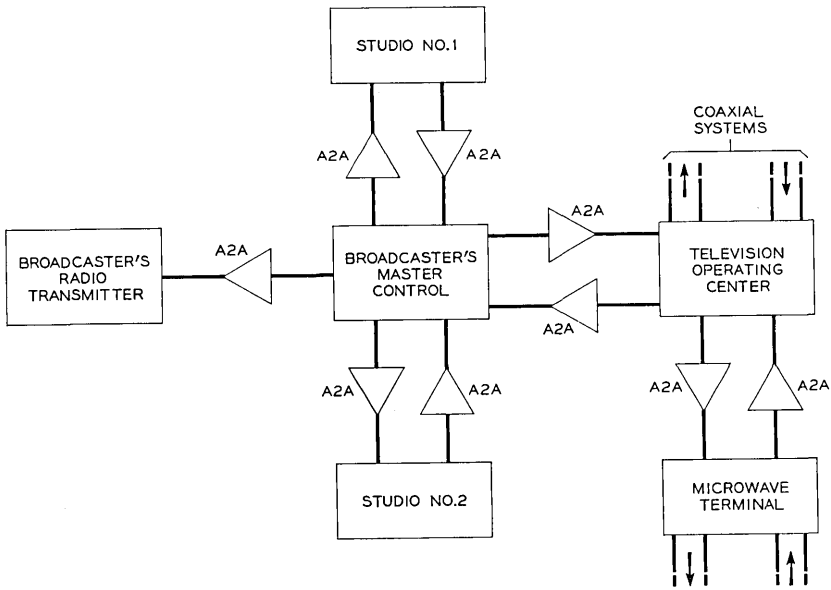


Fig. 1 — Typical A2A uses.

distribution. Subsequent developments produced two improved long distance broadband systems; the TD-2 microwave system[3] and the L3 coaxial system.[4] The A2A system provides the complementary improvement needed for the short-haul connections and should result in reduced maintenance effort and a minimum of special engineering to provide acceptable transmission quality.

The wide range of circuit lengths over which the A2A must operate and the erratic pattern of terminal and repeater locations constitute the outstanding requirements dictating the system design. The circuit lengths range from a fraction of a mile to approximately 10 miles; with the median length being about 1.5 miles. These requirements have been met by providing a series of amplifier and equalizer components which may be assembled in a large variety of combinations. By comparison, in long distance systems such as the L3 coaxial system, the repeaters have almost constant spacing, with the result that all repeaters have the same gain and cable equalization.

Many local circuits are ordered by the television broadcaster for short time use, such as program pickups of sporting events. Since these circuits are frequently requested on short notice, there is usually little time for circuit engineering. For these purposes portable equipment units which are readily convertible from one circuit length to another have been made available.

Another major factor affecting system design lies in the varied and, in general, non-predictable usage of the local links. Frequently the makeup of a complex transmission network varies from hour to hour as demand dictates, and it is not possible to line up or equalize the network on an overall basis. Hence, each link must be capable of a transmission quality such that when all the necessary links are connected in tandem the signal will not be degraded.

The A2A system has been designed to use the extensive existing video cable plant. This plant consists of 16-gauge polyethylene insulated pairs[5] incorporated in standard lead and composite sheathed cables installed since 1947. Ordinary telephone cable paper pairs, which have been used to a limited extent with the existing A2 system, are not suitable for use with the A2A system. In the earliest commercial form of the cable the individual video pairs were covered with two copper wrappings wound spirally in opposite directions. A pair having this construction is referred to as a 16 PSVS video pair. Early in 1950 the design was changed to a form with a longitudinal inner wrapping with the outer tape applied spirally as before. This design is known as 16 PSVL video pair and gives lower crosstalk and a small reduction in attenuation.

In 1953, production began on a new design in which the polyethylene string and strip insulation of the earlier types was supplanted by individual insulation of the conductors with expanded polyethylene together with expanded polyethylene fillers. This type, referred to as 16 PEVL video pair, is a cable whose impedance is held to closer tolerances and has reduced internal echoes. Both 16 PSVL and 16 PEVL are currently in production, and all three types of cable may be referred to collectively as 16 PSV video pair. The A2A system may be used on any of the types. Fig. 2 shows the construction of each of these cables.

The requirement to equalize a wide range of circuit lengths and repeater spacings has a further effect on system design. As is well known, lack of proper terminations produce interaction effects (multiple echoes) which are not simple functions of the cable length. Adequate equalization is exceedingly difficult under these conditions. This difficulty is avoided by terminating the cable at each end in networks which match the image impedance of the cable. This results in an insertion loss characteristic which is the attenuation characteristic of the cable with negligible terminal or reflection effects. The loss of the 16 PSVL or 16 PEVL pairs in db per mile is shown on Fig. 3.

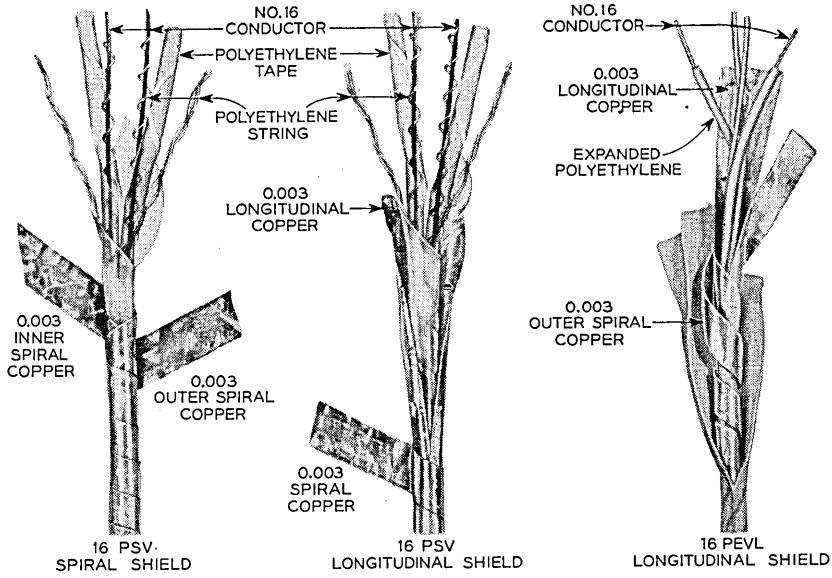


Fig. 2 — Polyethylene insulated cables used in the A2A system.

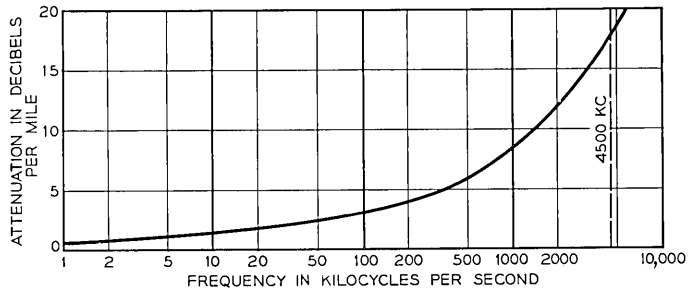


Fig. 3 — Attenuation of PSVL or PEVL cable.

DESCRIPTION OF SYSTEM

In this section the A2A system is described in terms of its application to several typical circuits. A subsequent section describes the transmission performance of a representative circuit. This is followed by a section devoted to a detailed description of the separate components of the system, and the manner of physical arrangement of terminals and repeaters. Since the equalization design of the A2A system is the subject of a campanion article[6] the treatment here is limited to functional description of the several equalizers.

The wide range of circuit lengths used in local circuits is demonstrated by Fig. 4 which shows the distribution of circuit lengths in miles of A2

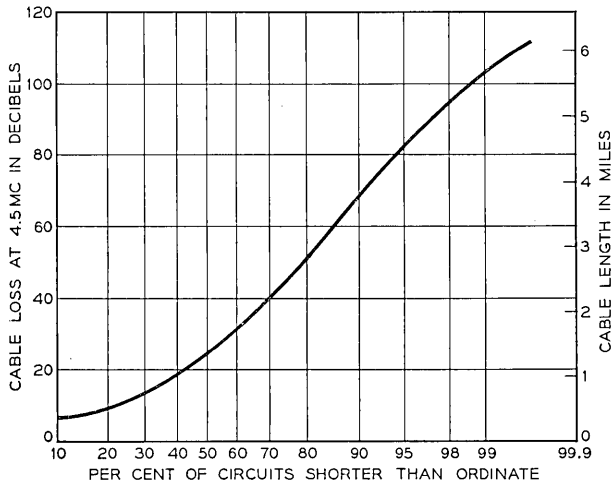


Fig. 4 — Distribution of A2 system circuit lengths in 1952.

circuit installations existing in 1952. Since the A2A is intended for use under similar conditions these data are a useful design guide. This wide range of circuit lengths requires both repeatered circuits and single section, non-repeatered circuits. The A2A solution to this problem is provided by a series of amplifiers, equalizers, and attenuators that can be interconnected in many ways to provide the needed flexibility. In general, the amplifiers have flat gain over the 0 to 4.5 mc band; however, two of the amplifier designs are provided with step-up networks which partially equalize for fixed lengths of cable. Fixed cable equalizers in sizes of 20, 15, 10, 7.5, 5, and 2.5 db are provided. In addition to the fixed equalizers two adjustable equalizers are available which have a total of nine variable equalization shapes.

Because the cable is terminated in essentially its image impedance, the low-frequency insertion loss of the cable is very low, approaching zero. The amplifier gains and equalizer components used are determined by the high-frequency loss of the cable which is proportional to cable length.

The provision of the necessary equalization and gain is achieved by connecting in tandem the proper combinations of amplifiers, equalizers, and attenuators in the proper associated combinations.

The system functions can best be described with reference to several typical circuit layouts. The first of these, shown on Fig. 5, is a 2.2-mile single-link circuit between two broadcaster's locations. The transmission equivalent of this circuit is unity. The normal input is one volt peak-to-peak of video signal. The example chosen is about half as long as the maximum non-repeatered length capability of the system.

The first unit of the transmitting terminal is an equalizer, which pre-equalizes for 15 db of the cable slope. The output amplifier has a voltage gain of 11 db and converts from 75-ohm unbalanced input to a balanced output to drive the video pair. The output impedance matches the cable impedance down to about 200 cps.

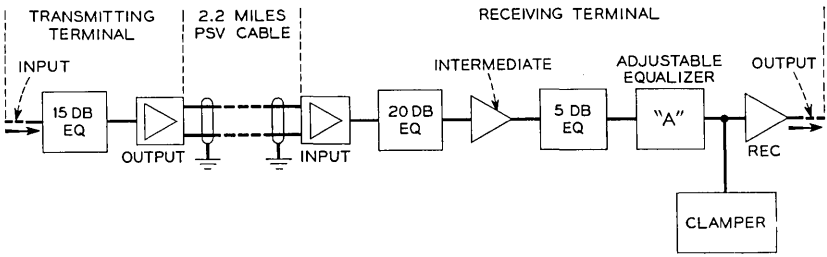


Fig. 5 — Block schematic of a typical single-link circuit.

With impedance matching terminations at both ends of the cable pair, the insertion loss equals the cable attenuation, which is essentially zero at dc, and for the example chosen is 41 db at 4.5 mc. Since the gain of the amplifiers is flat with frequency, the low-frequency line level is made low to prevent excessive modulation in the input amplifier of the receiving terminal. On the other hand the high-frequency sending level is made high to override the cable loss. In the arrangement shown on Fig. 5 the transmitting terminal has a voltage gain of 5 db at 4.5 mc, and due to the shaping of the cable equalizer has introduced a loss of 10 db at very low frequencies.

The input amplifier of the receiving terminal terminates the balanced cable, matching the impedance down to low-frequencies and provides 10 db of gain preceding the first equalizer.

Of the 44 db of slope in the assumed cable length, 15 db is pre-equalized in the transmitting terminal, and 26 db remains to be made up in the receiving terminal. A 20-db block of equalization is provided in the form of the 20-db cable equalizer and intermediate amplifier, whose 24-db flat gain equals the dc loss of the equalizer. The remaining 6-db slope is reduced to 1 db by a 5 db fixed equalizer. Using the fixed equalizers which are available in 2.5-db increments, the cable slope can be reduced to within ± 1.25 db. The remainder, which in the circuit chosen is 1 db, is equalized in the variable "A" equalizer, which provides equalization shape in very small adjustable steps over a range of ± 3 db at 4.5 mc. The meaning of ± 3 db is that at one extreme setting of the dial 3 db of cable slope is equalized; at the mid-position of the dial a flat loss shape is obtained, and at the other extreme of the dial, the loss shape simulates 3 db of cable.

The excess range of the variable equalizer over the ± 1.25 db needed to augment the 2.5 db fixed equalizer is used for periodic equalization adjustment for changes in loss due to change in cable temperature.

In addition to its function as a variable cable length equalizer, the "A" equalizer also provides four other equalization shapes to accommodate differences among the cable types used with this system. For circuit lengths up to about 3.5 miles these five variable shapes and the fixed equalizers suffice to meet transmission objectives.

The remaining components of the block schematic of Fig. 5 are a clamper and a receiving terminal amplifier. The clamper[10] operates as a shunt device across the 75-ohm circuit reducing low-frequency noise and correcting for the signal waveform distortion due to the low frequency cut-off of the amplifiers.

The receiving terminal amplifier has an adjustable amount of flat gain

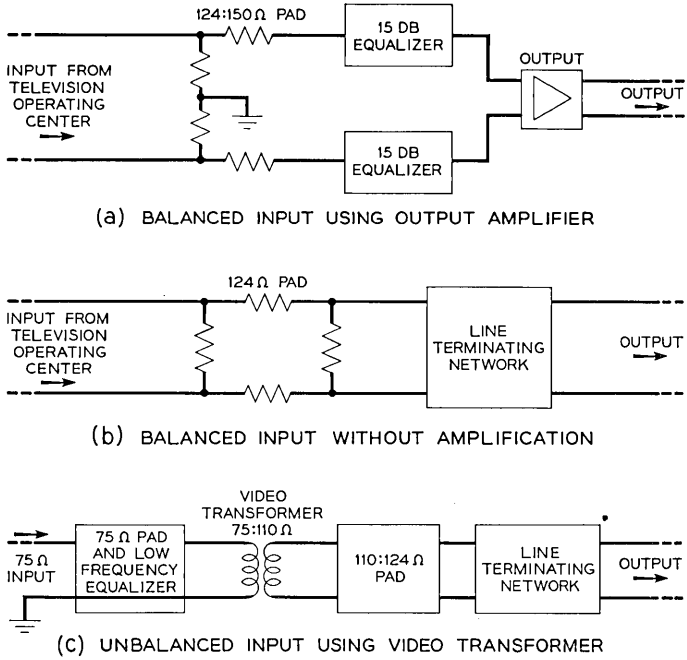


Fig. 6 — Other transmitting terminals. (a) Balanced input using output amplifier; (b) Balanced input without amplification; and (c) Unbalanced input using video transformer.

to make up for the loss of the fixed and variable equalizers following the intermediate amplifier. The gain adjustment is provided to compensate for the variation in flat loss of the "A" equalizer with dial setting, and also to set the end-to-end circuit equivalent.

The receiving terminal amplifier has alternate output arrangements: a balanced output at 124-ohm impedance for use in circuits terminating at television operating centers and a 75-ohm unbalanced output for circuits terminating at customers' locations.

In the example of Fig. 5 the equalization was the maximum for a receiving terminal with one intermediate amplifier. The changes to accommodate shorter circuits will next be described. The 5-db fixed equalization can be reduced in 2.5-db steps down to 0. Plug-in loss pads are then used to make up for the reduced equalizer loss so that the low-frequency gain remains fixed. For still shorter cable lengths the 20-db equalizer can be reduced to 15 db or 10 db, and again loss pads are added as required. Beyond this, the intermediate amplifier may be omitted. When this is done the maximum fixed equalizer range is the sum of the

15 db in the transmitting terminal and 5 db in the receiving terminal or 20 db. Adding the variable equalizer range gives a total of 23 db or about 1.3 miles, which is the median length of the present A2 system installations.

A further reduction in the equalization is obtained by reducing the 15-db fixed cable equalizer in the transmitting terminal to a smaller size or zero and adding loss pads to keep the low-frequency loss constant, as before.

Longer circuits than that of Fig. 5 may be used by adding a second intermediate amplifier and 20-db equalizer to the receiving terminal. The total of the fixed equalizers would then be 60 db, giving about 63 db or 3.5 miles when the range of the variable equalizer is included. This is the maximum single-link circuit length using the flat gain input and output amplifiers which is set by signal-to-noise requirements.

Provision for balanced input and output circuit arrangements have been made for the A2A circuits that originate or terminate in the television operating center. The receiving terminal amplifier has alternate output connections to provide either 124-ohm balanced output or 75-ohm unbalanced output. The output amplifier has a 150-ohm balanced-to-ground input impedance, which was chosen so that the 75-ohm unbalanced cable equalizers could be used in pairs to form a balanced equalizer, instead of duplicating the fixed equalizer designs in 124-ohm balanced form. An impedance transforming pad converts the 124-ohm input to 150 ohms. This arrangement is shown on Fig. 6(a).

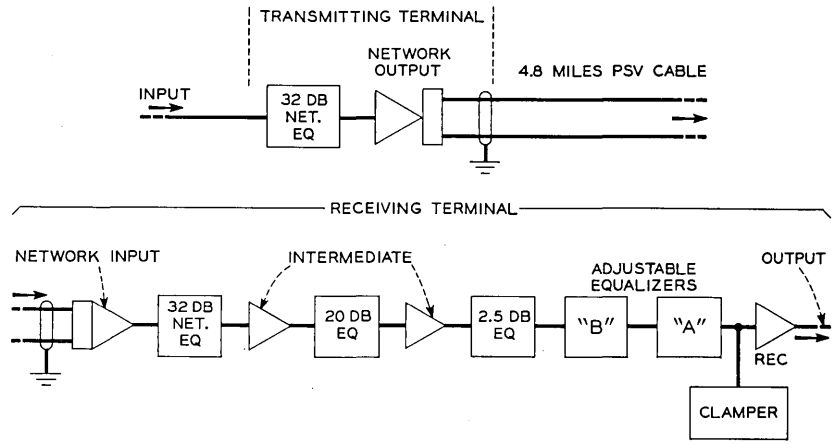
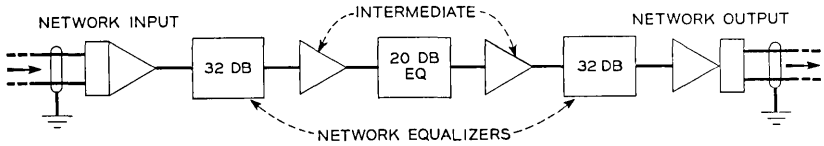
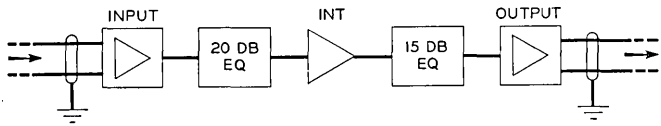


Fig. 7 — Maximum length single-link circuit.



(a) USING NETWORK INPUT & OUTPUT AMPLIFIERS



(b) USING FLAT GAIN INPUT & OUTPUT AMPLIFIERS

Fig. 8 — Typical repeater circuits. (a) Using network input and output amplifiers, and (b) Using flat gain input and output amplifier.

An additional balanced input transmitting terminal is shown on Fig. 6(b). The amplifier and fixed cable equalizers are omitted, so that the output level is flat. This arrangement may be used for short links.

A similar circuit for unbalanced inputs is shown on Fig. 6(c). The amplifier and fixed equalizers are omitted. A video transformer[2] converts the unbalanced input to balanced output to drive the cable. Low-frequency adjustable equalization is provided to compensate for the low-end cut-off characteristic of the transformer.

The use of the network output and input amplifiers permits the maximum single link span to be increased to 86.5 db of loss at 4.5 mc or 4.8 miles of cable. Such a circuit is illustrated in the block schematic of Fig. 7. The transmitting terminal uses an output amplifier in which the output tubes operate into a network embodying transformer action over the upper part of the 4.5-mc band. At low-frequencies the transformer is by-passed and the vacuum tubes operate into a 124-ohm resistive load. The transformer coupling increases the output level capability at high frequencies where it is needed to overcome the cable loss. The circuit of Fig. 7 thus delivers a greater high-frequency output to the cable than the flat-gain output amplifier of Fig. 5.

The gain step-up shape of the output network differs from the required shape for equalizing cable and requires a corrective equalizer to provide the equivalent of the fixed equalizer designs. This corrective function is combined with about 17 db of the fixed equalizer shape in the network equalizer to make a total of 32 db of cable equalization in this transmitting terminal, compared to a maximum of 15 db in the first case. The

amount of pre-equalization in this transmitting terminal cannot be reduced below the 32 db provided by the network equalizer and network amplifier.

The receiving terminal shown on Fig. 7 uses a network input amplifier which contains a step-up network similar to that in the network output amplifier. The impedance step-up in the upper part of the band provides a signal-to-noise improvement over the flat-gain imput amplifier. The combination of both network input and output amplifiers increases the allowable circuit length or repeater spacing about 26.5 db at 4.5 mc.

The input network equalizer serves the same function as the output network equalizer and also provides a total equalization of 32 db, in combination with the input network amplifier.

The receiving terminal of Fig. 7 contains three sections. The first provides the network input amplifier, network equalizer and intermediate amplifier. It equalizes for 32 db of cable slope, with no option for lesser equalization. The second section providing a maximum of 20 db of equalization, contains a fixed equalizer and intermediate amplifier. As before, the equalization may be reduced below 20 db to 15, 10 db, or less; or the amplifier and equalizer may be omitted entirely if the cable loss slope permits.

The third part contains the small size fixed equalizer (2.5 db in the case shown on Fig. 7), variable equalizers, clamper and receiving terminal output amplifier. This group differs from the counterpart of Fig. 5 only in the addition of a second variable equalizer, the "B" unit.

The "B" equalizer[6] contains four manually adjustable loss shapes located in frequency regions between the loss shapes of the "A" equalizer. For circuits shorter than about 3.5 miles, the "B" equalizer is usually not required.

The methods of circuit arrangement for other cable lengths are similar to those already described in connection with the circuit of Fig. 5. One limit is reached at 64 db which is the sum of the irreducible equalization associated with the network input and output amplifiers. A further reduction in slope equalization is available by replacing the network input amplifier and its 32 db of equalization with the flat gain input amplifier. The minimum slope would then be the 32 db of the network output amplifier and its accompanying network equalizer. The use of the flat gain amplifiers at both input and output as in Fig. 5 provides the means for handling circuits shorter than 32 db. However, the upper limit of length is then about 60 db.

The application of the A2A system to circuits longer than 4.5 miles requires one or more intermediate repeaters. In general the repeater

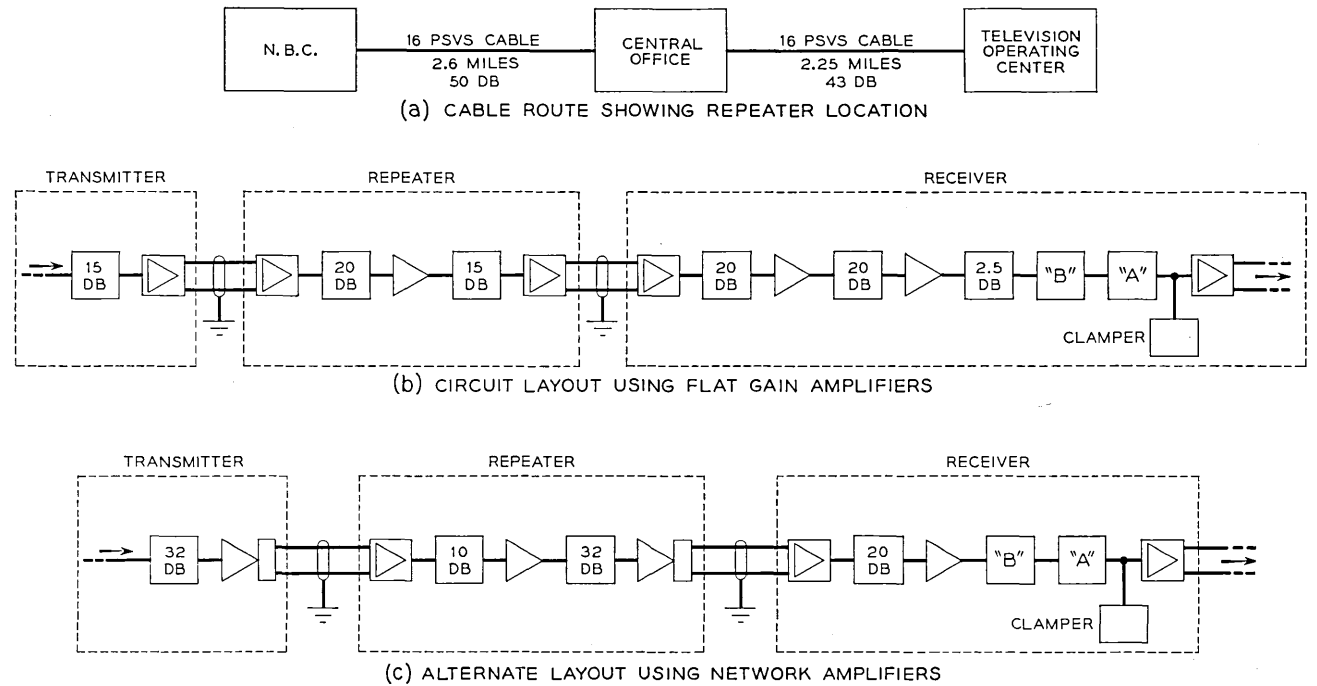


Fig. 9 — Typical two-link circuit. (a) Cable route showing repeater locations; (b) Circuit layout using flat gain amplifiers; and (c) Alternate layout using network amplifiers.

spacing will be variable since it is desirable to place the repeaters in existing Telephone Company buildings. Two typical repeater arrangements are illustrated by the block schematics of Fig. 8. The repeaters use the same amplifier and fixed equalizer components as the transmitting and receiving terminals.

Fig. 8(a) shows a maximum equalization repeater using network amplifiers for both input and output. The input section provides 32 db, the intermediate section 20 db, and the output section 32 db, for a total of 84 db. Reduction of the equalization is accomplished in the same manner as in receiving terminals, the minimum here being 64 db of cable slope.

A repeater using flat gain input and output amplifiers and one intermediate amplifier is illustrated in Fig. 8(b). The arrangement shown provides 35 db of equalization. This can be reduced to 15 db by removing the intermediate amplifier and 20 db equalizer. Further reduction is handled by replacing the 15 db equalizer with smaller units.

The addition of a 20-db equalizer and intermediate amplifier to the circuit of Fig. 8(b) provides a means for extending the equalization to a maximum of 55 db for repeaters using flat gain input and output amplifiers.

The position of a repeater in an A2A circuit may be such that one line section is short and the other long due to the use of existing Telephone Company offices. For these cases repeaters with flat gain input amplifiers and network output amplifiers or vice versa are used. In all, 10 different amplifier arrangements for repeaters are available.

The A2A field trial installation in New York is an illustration of a repeatered circuit. Fig. 9(a) shows the location of the terminals and an available repeater location in a central office. Each of the two cable spans is within the range of the flat gain input and output amplifiers, although the network amplifiers could have been used to some advantage.

The principles of circuit layout for multi-link circuits have been developed to meet the requirement imposed by the wide range of existing repeater location spacings.

By providing only coarse steps of equalization at each repeater, the minimum complement of equalizers and amplifiers is employed. Only two factors need be taken into account in determining the coarseness of equalization. These are the overload limit of the output amplifier and noise level at the succeeding input amplifier.

In the receiving terminal the circuit is equalized to close tolerances using the smaller fixed equalizers and the variable equalizers.

This layout principle is illustrated in the circuits of Fig. 9(b) and (c).

Fig. 9(c) shows an alternate circuit layout for the same cable spans. The transmitter and the repeater use network output amplifiers, but the cable losses are not high enough to permit the use of network input amplifiers. Although the 10-db equalizer and intermediate amplifier could be omitted from the repeater and still provide a high enough output level for the 43 db cable loss, no overall saving would result, since these units would then be added to the receiving terminal. In comparing the two circuit layouts it should be noted that the (c) layout has one less equalizer and intermediate amplifier than the (b) circuit layout.

PERFORMANCE CHARACTERISTICS

Realization of the performance objectives was confirmed in the field trial of a representative two link A2A circuit. The circuit was installed between 30 Rockefeller Plaza and 32 Avenue of the Americas in New York using facilities leased from the New York Telephone Company. Figure 9(a) shows the cable route, and Figure 9(b) shows the circuit layout. This installation was made in June, 1954, and tests were in progress over a period of about 10 months.

Equalization

The measured overall gain-versus-frequency characteristic is given on Figure 10. The residual gain ripples are less than ±0.04 db up to about 4.8 mc. Above this frequency the cut-off shape is gradual; the 6 db loss point occurs at about 7 mc.

The relative ease and speed of achieving this degree of equalization flatness is one of the features of the A2A system contributing to reduced maintenance cost. The process of setting the 9 dials of the "A" and "B" equalizers can be completed in about 10 minutes. Each dial is set by measurements at a particular frequency, making the overall circuit gain

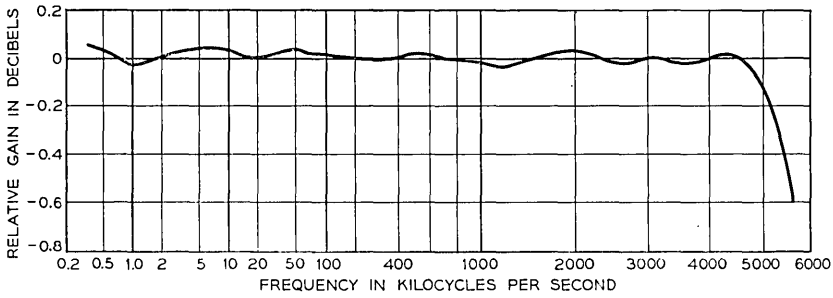


Fig. 10 — Overall gain characteristic of typical 5-mile circuit.

at that frequency equal to the gain at 300 kc. A test set for this purpose alternately sends the test frequency and the 300-kc reference. At the receiving end of the circuit the appropriate equalizer dial is adjusted for equal outputs at the two frequencies. The equalizer shapes[6] and the sequence of their adjustment avoid interaction, making it unnecessary to repeat any dial adjustment. The lineup process yields 10 points on the gain-frequency characteristic which is ususally a sufficient record of the performance, so that a complete gain frequency run need not be taken.

During ten months of the A2A trial the only change in variable equalizers required was correction for temperature change of the underground video cable. This was done periodically by adjustment of only one of the "A" equalizer shapes.

Signal-to-Noise

The presently accepted noise objective for a long television network is a peak-to-peak signal-to-r.m.s.-noise ratio of 45 db for "flat" or "white" noise.[7] For random noise of other spectral characteristics a noise weighting shape is used.[8] The shaping network used has a loss of 9 db to flat noise so that the signal-to-weighted noise objective becomes 54 db. The contribution to this total by all local video systems involved in such a network has been taken as a signal-to-noise ratio of 59 db. Considering the tandem usage of the local circuits, and the distribution of circuit lengths, an objective of 67.5 db for the minimum signal-to-noise performance of one A2A circuit link has been established.

The noise weighting used here includes the weighting versus frequency imposed by the NTSC color system. This is about 5 db more severe than the noise weighting for monochrome television, because the color information which is transmitted as a carrier signal at 3.58 mc over the transmission system is demodulated in the receiver to base-band frequency, where the noise presents a coarser pattern than the noise in the vicinity of 3.58 mc. The weighting also includes the low frequency noise shaping effects of the clamper.

The computed maximum link length to meet a 67.5-db signal-to-noise performance is 3.45 miles of cable for circuits using the flat gain amplifiers. The measured performance of this length of an actual circuit was 67.1 db. In the field trial circuit each of the two circuit links is shorter than this maximum spacing. The measured color weighted signal-to-noise performance of 71 db is in good agreement with the computed performance of 72.6 db.

The use of the network output and input amplifiers permits the maximum spacing to be increased to 4.8 miles for the same signal-to-noise performance as 3.45 miles with the flat gain amplifiers.

Differential Gain and Phase

In the NTSC color system, the color information is transmitted as a double sideband signal at a subcarrier frequency of 3.58 mc. The phase of the subcarrier contains the hue information and the amplitude controls the color saturation. This use of the high-frequency end of the band for color information imposes severe requirements on transmission, noise, and distortion. The change in phase at 3.58 mc with instantaneous signal level variations is known as the differential phase and causes variations in hue. The gain change of the system at 3.5 mc with instantaneous signal level variations is known as differential gain, and results in distortion in the color saturation.

The use of balanced amplifier stages with relatively wide band interstage networks has resulted in low differential gain and phase distortion. The differential gain and phase performance of the A2A system, of course, is dependent on the number of amplifiers used in each circuit. The 5-mile field trial circuit is representative of the performance of a long circuit. Measurements have been taken with the newly developed 47-A transmission measuring set.[9] For the maximum signal voltage excursion the differential phase was 0.2° and the differential gain was 0.3 db.

DETAILED DESCRIPTION OF THE AMPLIFIERS

In this section the principal features of each of the six amplifier types will be described with reference to simplified schematics.

These designs use balanced amplifier stages for all applications including those requiring unbalanced terminations. Several advantages accrue from the use of balanced circuits. The most important of these is the reduction in distortion due to balancing the even order distortion outputs of the electron tubes. Equally important is the fact that the residual even order distortions will add on a random rather than systematic basis, in tandem amplifiers. Since the number of tandem amplifiers is expected to be the order of 50 in a large network the difference between systematic and randon addition is appreciable.

The use of balanced stages also reduces the requirements on impedance of the DC plate supply permitting the use of non-electronically regulated rectifiers. By comparison, the A2 system uses a 10-tube regulated rectifier.

Triodes instead of pentodes are used for the input and output stages

of the flat gain amplifiers in order to obtain improved modulation and signal-to-noise performance. For the network input and output amplifiers pentodes are used to obtain high impedance terminations for the input and output networks.

The interstage video coupling networks use conventional circuits, but the band width at the 0.1 db relative loss point is made approximately 9 mc. The required sacrifice in gain to obtain this wide band yields several advantages. Transmission deviations within the 4.5-mc band of the system due to element deviations and capacitance variations in the tubes are greatly reduced. In addition the differential phase performance of the amplifier is improved.

The principal mechanism causing the differential phase in these amplifiers may be explained as follows. The instantaneous video signal voltage on the control grid results in a modulation of the effective grid-cathode capacitance, in turn producing a differential phase shift in the interstage network. The reduced impedance level of the 9-mc interstage compared to one for a narrower band results in a smaller differential phase shift. Here again, the even order components are suppressed by the balanced circuit arrangement.

Output Amplifier

A simplified schematic of the output amplifier is shown on Fig. 11. Alternate input terminations are provided to handle either a 150-ohm balanced input when used at a television operating center, or a 75-ohm unbalanced input at a broadcaster's location. Parasitic capacitances shunting the resistive termination are built out with small inductors to provide an input impedance with a return loss greater than 40 db up to 5 mc.

A manual gain adjustment is provided to compensate for flat gain deviations caused by variations in electron tubes and in addition, to furnish some range for system purposes. The gain of the amplifier is controlled by a potentiometer between the cathodes of tubes V1 and V2 in the input stage. The range of control is 10 db and the working gain of the amplifier is 11 db. This variable local feedback also causes the plate resistances of the tubes to vary depending upon the setting of the potentiometer. Since the plate resistances of the tubes contribute to the effective load resistance of the first interstage, any change in these resistances results in a change at high frequencies in the interstage gain-versus-frequency characteristic. To compensate for this effect a small positive feedback is introduced by cross-connecting the voltage

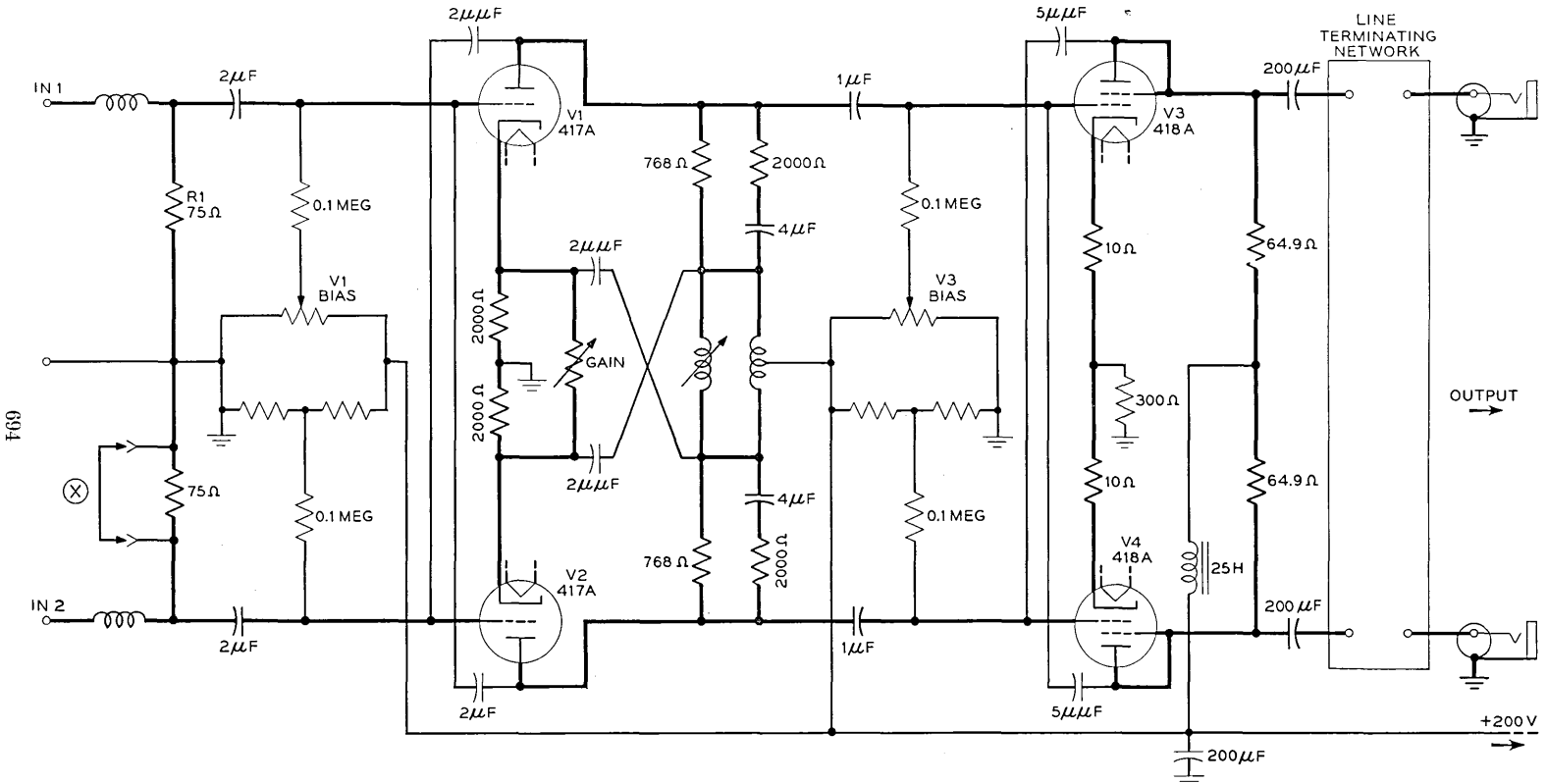


Fig. 11 — Simplified schematic of output amplifier.

drop across the peaking coil of the interstage network to the cathodes of tubes V1 and V2 by means of 2 mmf capacitors. At the higher frequencies, these capacitors couple a small current of phase opposite to that of the cathode current through the potentiometer resistance thereby reducing its voltage drop. Hence the relative gain of the stage is increased at higher frequencies by an amount equal to the gain reduction caused by the increase in plate resistance.

To make the high-frequency input impedance substantially independent of the gain potentiometer setting, cross-neutralization of the grid-plate capacitance is employed. In the absence of neutralization, the grid-to-plate capacitance would be a function of the feedback introduced by the gain potentiometer.

Conversion of unbalanced input signals to a balanced drive for the output stage is provided by the large longitudinal feedback in the cathode circuit of the first stage. Further increase in balance takes place in the output stage which also has longitudinal feedback in the cathode circuit.

The interstage network of the output amplifier uses shunt peaking of a standard type with constants derived for flat amplitude response with a bandwidth of about 9 mc. The small phase distortion in the 4.5-mc band is equalized by the fixed cable equalizers.

A new method for aligning the variable inductor of the interstage network has been developed for the A2A system. Previously the alignment procedure required the measurement of the gain-frequency characteristic at several frequencies for each interstage, isolated from all other frequency shaping components of the circuit. In the new method the measurement is made at one frequency only. No isolation of the stage under adjustment is required, and the test equipment has no requirement on flatness of gain characteristic.

The procedure is illustrated by the simplified schematic of an interstage network given in Fig. 12 where R is the interstage resistance, L is

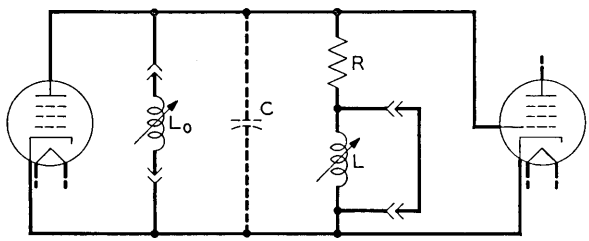


Fig. 12 — Amplifier interstage alignment.

the adjustable peaking coil, and C is the total parasitic capacitance. At low frequencies C and L have negligible effect on the impedance of the network, and the gain of the stage is proportional to R only. The objective is to adjust L for equal gains at 4.5 mc and low frequency. The first step is to short out L, and anti-resonate C at 4.5 mc by the addition of the test shunt inductor Lo. If the test coil has no dissipation, the interstage impedance will be R, and the gain at 4.5 mc will be equal to the low-frequency gain of the network. The 4.5-mc output is then measured at any convenient point in the circuit beyond the interstage. The second step is to remove the test coil Lo and the short circuit around L and adjust L to yield the same output reading obtained above, thereby making the gain at 4.5 mc the same as the low-frequency gain.

In the actual circuit a fixed inductance test coil shunted by a variable condenser is used instead of a variable inductor in order to maintain a constant Q. Because of dissipation in the test coil, and the fact that the parasitic capacitance is distributed, and several other small factors, the interstage resistance at the anti-resonant test condition is reduced below R. This is offset by adding a low resistance in series with the shorting plug used to short out L, making the total resistance again equal to R. Test jacks are accessible from the apparatus side of each panel to facilitate making the alignment.

The series circuit containing a 4-mf capacitor in the interstage network of the output amplifier provides a low-frequency gain step to compensate for the phase shift of several RC coupling circuits. Since the input and output amplifiers are always used in pairs the low-frequency compensation of the output amplifier is proportioned to correct for the time constants of the coupling circuits of both amplifiers.

At all frequencies within the 4.5-mc band, the output impedance of this amplifier approximately matches the characteristic impedance of the transmission line which it terminates. This is accomplished by the line terminating network which, when terminated in 62 ohms on each side to ground, presents an impedance equal to that of the line down to about 200 cycles. The 62-ohm terminations are provided by resistors in parallel with the plate resistance of the output tubes. In order to reduce the effects of parasitic capacitances in shunt with these terminations, the grid-to-plate capacitances of tubes V3 and V4 are cross-neutralized.

The center point of the output termination connects of a 25-henry inductor in series with the 200-volt dc plate supply. Its function is to raise the longitudinal output impedance so as to reduce low frequency longitudinal interference currents, principally at 60 cycles arising in the 16-gauge video cable conductors.

The longitudinal feedback in each stage of the amplifier maintains constant the sum of the cathode currents of the two tubes of each stage. A differential grid bias potentiometer provides a means for making the cathode currents equal. These controls are shown on Fig. 11.

The electron tubes in all of the A2A amplifiers are operated at a heater voltage of 6.1 volts instead of the nominal 6.3 volts in order to obtain increased life. This reduced voltage must be regulated to a close tolerance to insure adequate thermionic emission. The 115-volt 60-cycle voltage applied to each amplifier panel is regulated to about ± 1 per cent. In addition the filament transformers contain thermistor networks to reduce heater voltage changes due to transformer temperature changes. Provision is made in each amplifier to measure the activity of each tube on an in-socket but out-of-service basis. To make the measurement the heater voltage is reduced from 6.1 to 5.7 volts by means of a switch on the amplifier panel. The reduction in heater voltage is accompanied by a change in the grid-cathode bias, the magnitude of which is a measure of the relative age of the tube.

Input Amplifier

The input amplifier provides the first block of gain in A2A receiving terminals and repeaters. The circuit is described with reference to Fig. 13, which shows a simplified schematic.

The amplifier uses two balanced triode stages and provides 10 db of voltage gain into the unbalanced 75-ohm equalizer which follows it. It will be noted that several of the circuit features are similar to those of the output amplifier.

Cable termination is provided by the line terminating network in series with 62-ohm resistors. Small inductors, provide high frequency peaking to compensate for the parasitic capacitance of the input tubes and coupling elements.

Suppression of longitudinal noise voltages arriving over the cable is obtained by means of the 418A tetrode V3 connected in the common cathode circuit of V1 and V2. This constitutes local feedback such as to reduce the gain of the stage to longitudinal signals and minimize their modulation of metallic signals. A tube is used because it presents a high resistance to ac currents and a low resistance to dc currents, whereas a resistor would present high resistance to both ac and dc currents. A low dc drop across this resistance is desirable because it conserves power supply voltage.

The output circuit uses two 417A triodes in a cathode follower circuit

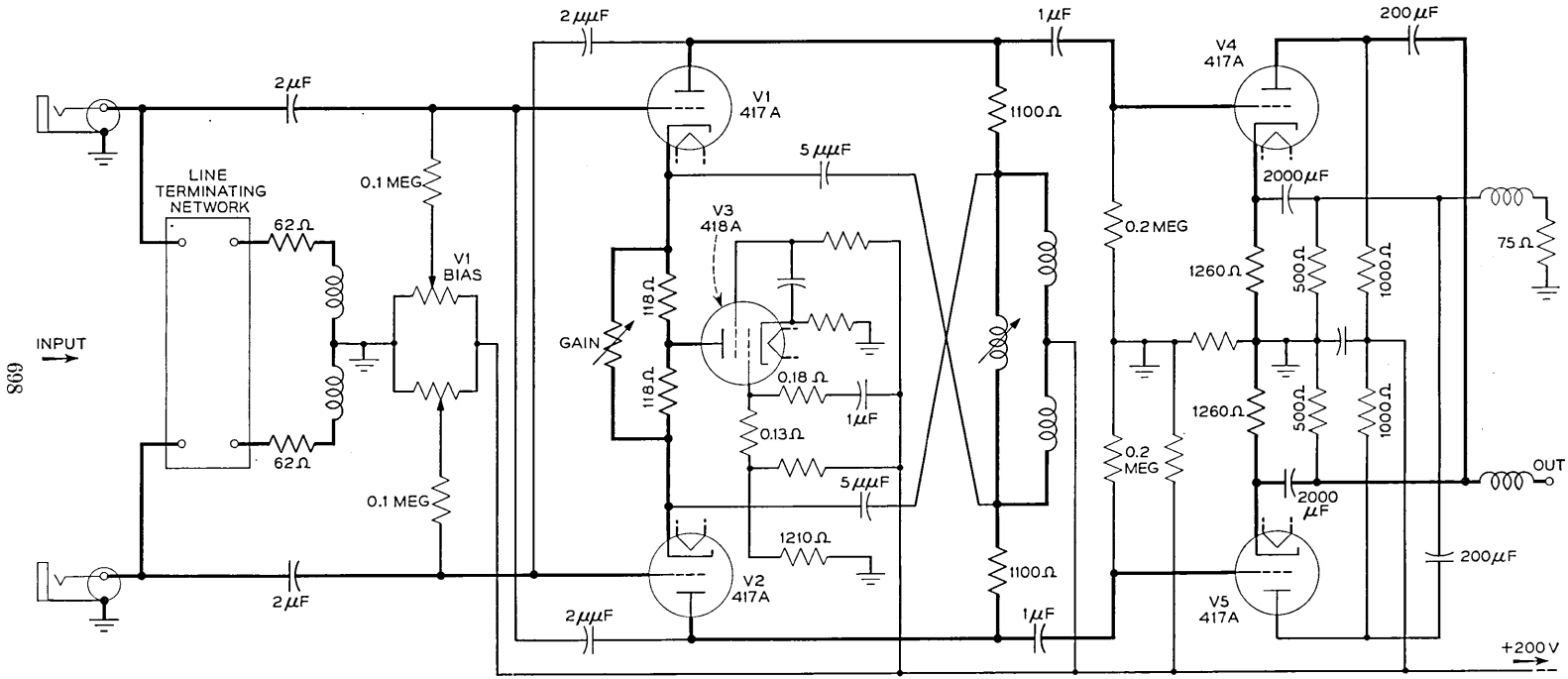


Fig. 13 — Simplified schematic of input amplifier.

which provides low distortion into the 75-ohm unbalanced load. The input of the output stage is a balanced signal. Output from both tubes V4 and V5 is obtained by combining the plate output of one tube with the cathode output of the other, through suitable coupling condensers as shown in Fig. 13. The plate signal current required in the output tube for a given output voltage is 11 db lower in this circuit than it would be in a circuit where the output is taken from the plate circuit of one of the triodes, with a 75-ohm resistive termination. Of the 11 db, 6 db is due to the utilization of the output of both tubes, and 5 db is due to the low cathode impedance which requires only a high load resistance, 500 ohms, to obtain 75 ohms output impedance.

The 11-db signal current reduction which reduces both even and odd order distortion terms provides a modulation performance equivalent to that of the output circuits of the output amplifier which operate with plate currents of 65 milliamperes. By comparison the triodes of the present circuit operate at plate currents of 20 milliamperes.

This output circuit has a limitation that rules out its use in some of the A2A amplifiers. Since the output is taken in part from the cathode circuit, its impedance is a function of the transconductance of the tube, and will therefore vary somewhat with the age of the tubes. In the input amplifier application the output circuit termination is a constant resistance cable equalizer which is well terminated in the 75-ohm input circuit of the succeeding amplifier. Since this connection is made over a short length of cable, the return-loss requirement is low. In the amplifiers that feed long office cables, the return loss requirement is high and this output circuit is not used.

Special means are employed in several of the A2A amplifiers to suppress the effects of power line voltage "bobble". Bobble may be defined as the voltage fluctuations on commercial power lines, occuring at frequencies from 0 to perhaps 25 cps. The magnetic line voltage regulator used in the A2A system attenuates slow changes in line voltage, but has little effect on bobble above above 5 cps. The non-regulated 200-volt rectifier which supplies dc power to the amplifiers has high loss for 120 and 60 cps but low loss below about 20 cps. Power line bobble in the interval of about 5 to 20 cps may therefore be present in the 200-volt supply to the amplifiers.

Two effects due to bobble are produced. In addition to the appearance of the bobble frequencies as an additive component in the output of the amplifier, there are also modulation products of the bobble and the television signal. The modulation is the more serious of the two effects, since it is not attenuated by the clamper as are the low frequency additive com-

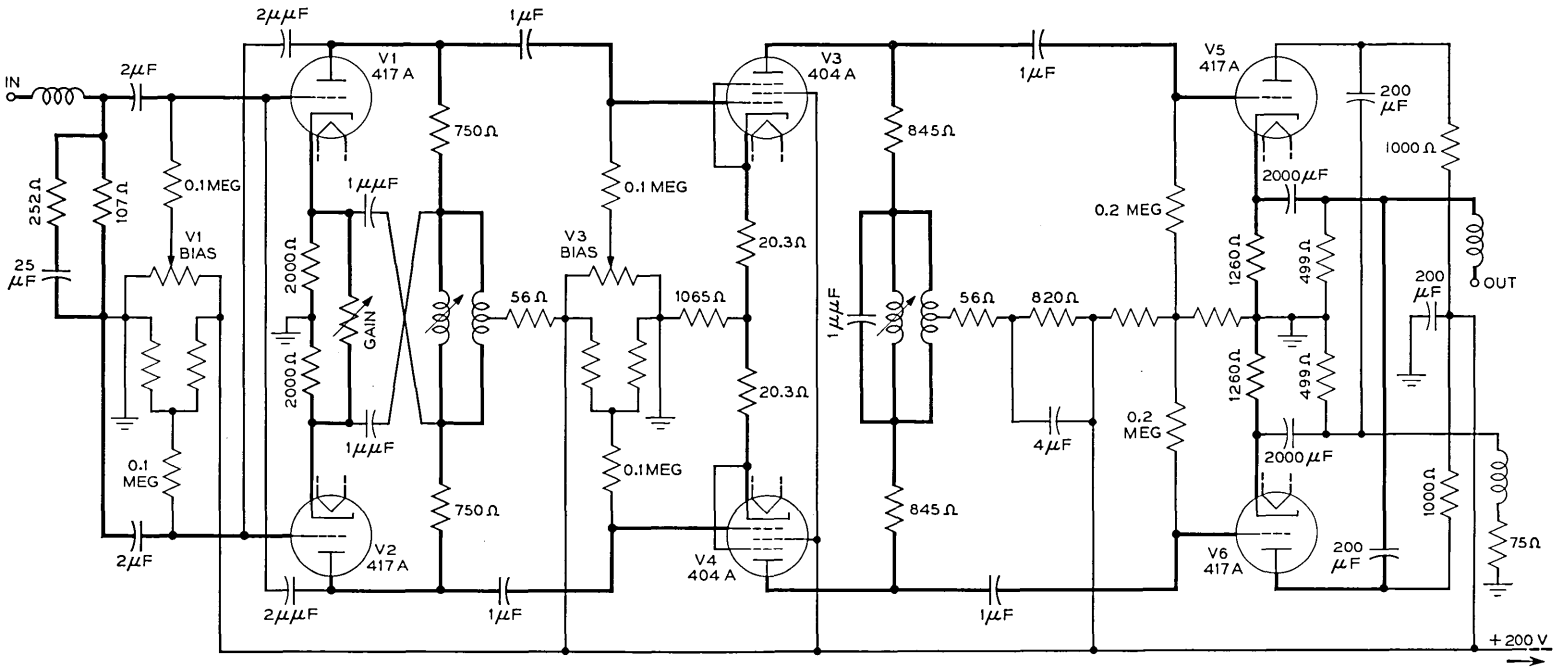


Fig. 14 — Simplified schematic of intermediate amplifier.

ponents. The modulation effects of bobble are due to the high bobble currents in the electron tubes. In most of the amplifier stages high longitudinal resistors used in the common cathode circuits prevents large current changes. However, in the output circuits as in Fig. 13 this condition does not hold and special means to suppress this effect are required.

In the input amplifier this is done by a balancing method. Referring to Fig. 13, a resistance divider couples a fraction of the bobble voltage on the 200-volt supply to the grid of the longitudinal tube V3. The longitudinal current that flows in the load resistors of V1 and V2 due to V3 is equal in amplitude and opposite in phase to the bobble voltage coupled from the plate supply. The resulting balance reduces the effect to a tolerable value.

Intermediate Amplifier

Fig. 14 shows a simplified schematic of the intermediate amplifier. It furnishes 24 db of gain between 75-ohm unbalanced terminations in repeaters and receiving terminals.

Most of the component circuits are similar to those already described. The first stage is comparable to that of the output amplifier. The low-frequency phase compensation circuit is placed in the input circuit instead of in the interstage network. The compensation is for low-frequency coupling elements in the intermediate amplifier only.

The output stage will be recognized as the same circuit used in the input amplifier. The output terminates in a 75-ohm equalizer over a short length of cable so that the output impedance requirements are lenient as for the input amplifier.

The intermediate stage of the amplifier uses 404A pentodes since it operates between high impedances. The interstage network and the provision for one-frequency alignment are the same as those already described.

As in the input amplifier a circuit for suppressing power supply bobble in the output is required. The method employed is a balancing arrangement wherein the bobble gain of the middle stage is so proportioned that the amplified bobble voltage of that stage is made equal to the direct bobble voltage applied from the power supply to the plate circuit. Refering to Fig. 14, the bobble input to the grids of V3 and V4 arrives via the plate feed. The bobble output appearing across the plate coupling resistors just equals the direct bobble voltage from the power supply. Since the two voltages are of opposite phase a cancellation takes place,

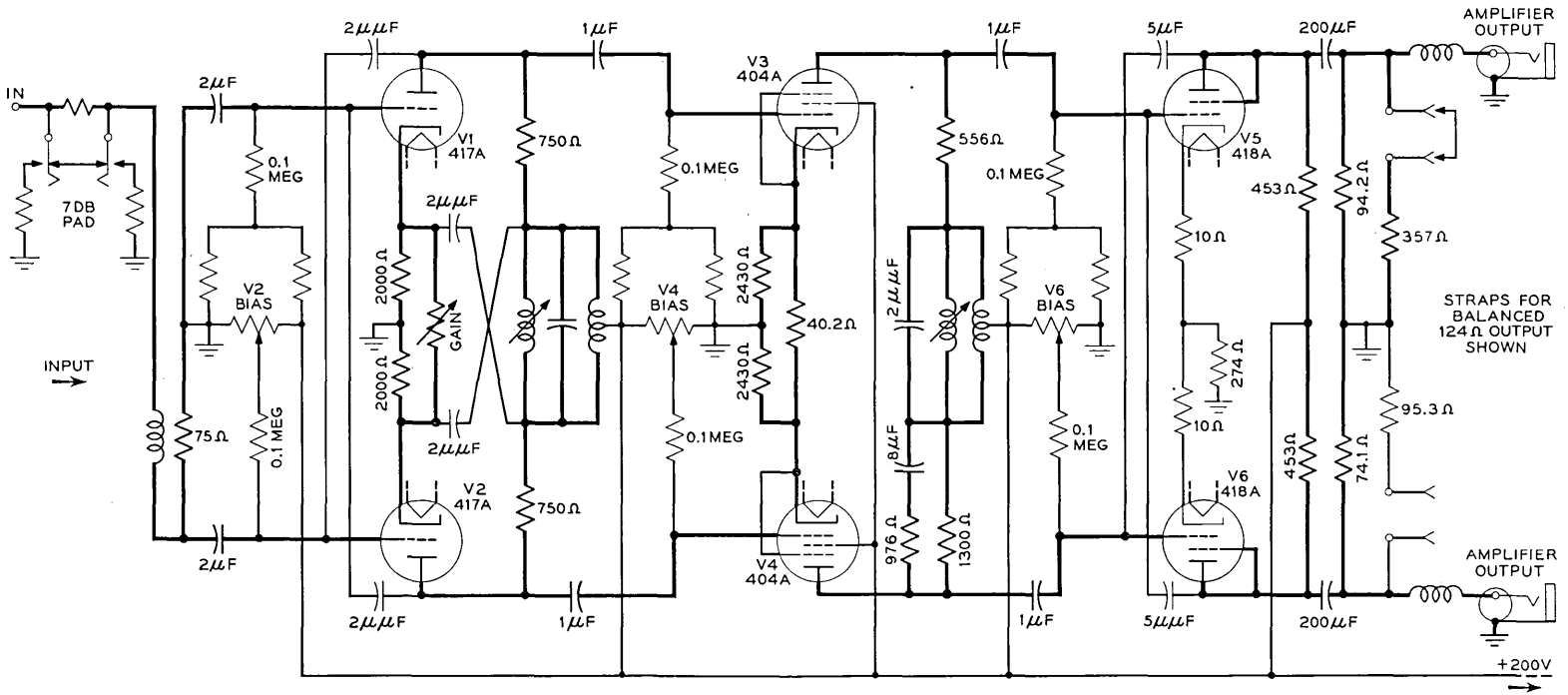


Fig. 15 — Simplified schematic of receiving terminal amplifier.

and the bobble voltage on the output stage grids is reduced with corresponding reduction of the modulation. The increased longitudinal gain of V3 and V4 is restricted to low frequencies by means of a by-pass condenser.

Receiving Terminal Amplifier

The receiving terminal amplifier shown on Fig. 15 supplies the last block of gain in each A2A circuit. Optional 75-ohm unbalanced or 124-ohm balanced outputs are provided to handle terminals at either customer's premises or television operating centers.

Nominally the gain is set to 20 db for 75-ohm output or 23 db for 124-ohm output. Additional gain is available for compensating the variable flat loss of the "A" equalizer, and for setting the overall circuit equivalent from transmitter to receiver. The cathode potentiometer in the first stage has a 10 db range of gain control, and a 75-ohm switchable loss pad provides an additional 7 db of control.

The first and second stages of the amplifier are very similar to the corresponding gain stages of the output and intermediate amplifiers. Compensation for low-frequency phase shift in the coupling circuits is provided by a low-frequency peaking circuit in the second interstage. It will be noted that the compensation is only on one side of the balanced circuit. The large longitudinal suppression of the output stage corrects the low frequency unbalance thus introduced.

The output of the receiving terminal amplifier connects to the broadcaster's equipment or to the television operating center over cables which may be as long as 500 feet. Over this length of cable good termination is required to keep echoes due to reflection down to requirements. For this reason the economical output circuit used in the input and intermediate amplifiers cannot be used here. A conventional plate output is taken from the triode connected 418-A tubes. Wiring options permit either balanced or unbalanced output.

Network Input and Output Amplifiers

These amplifiers are very similar to the flat-gain counterparts which have been described above. The principal difference is the use of networks[6] that provide a rising gain shape in each amplifier.

In the network output amplifier the 418A output tubes are operated as tetrodes to provide a plate impedance high relative to the high side resistance termination of the network. The reduced transconductance of the tetrode compared to the triode is offset by the reduced interstage

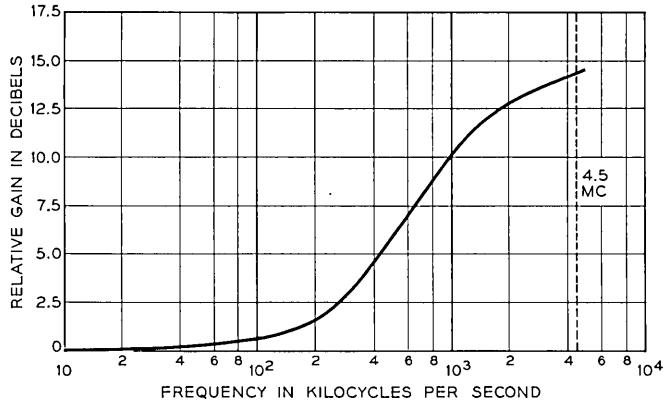


Fig. 16 — Gain characteristic of network input and output amplifier.

capacitance to make the operating low frequency gain of this amplifier substantially the same as the flat-gain output amplifier. Gain margin is provided to handle tube variations.

The overall gain shape of the amplifier is shown on Fig. 16. The gain step-up is 14.5 db at 4.5 mc.

In the network input amplifier the first stage tubes are 404A pentodes, in order to achieve a high impedance shunting the input network termination. The use of pentodes diminishes the benefits of the input network in two ways. Since the transconductance is lower, the amplifier has about 3 db less gain at low frequencies than the flat-gain input amplifier. The higher noise of the pentode compared to an equivalent triode reduces the noise advantage of the input network by about 4 db.

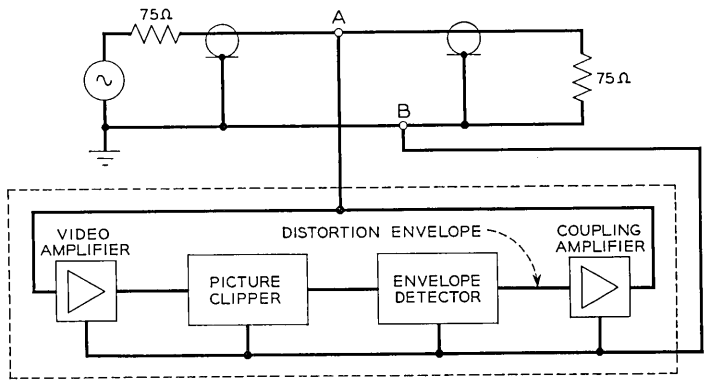


Fig. 17 — Block diagram of clamper.

The input network is similar to the network of the network output amplifier. The line termination elements are an integral part of the network and appear as a shunt circuit across the input, whereas the line termination in the network output amplifier is a series circuit.

The relative transmission is shown on Fig. 16. Variable cathode feedback as a continuous gain control is not used since it would result in a variable input capacitance terminating the high side of the network. Instead, a 3 db switchable pad is provided in the 75-ohm output of the amplifier to be used as a coarse gain adjustment. The continuous gain controls of the subsequent amplifiers in a receiving terminal or repeater are used to set levels accurately.

Clamper

The A2A clamper is an electronic circuit which samples the crest level of each sync pulse of a video signal and subtracts the envelope, derived by the sampling, from the transmitted signal. Since the derived envelope closely approximates any low-frequency interference or distortion which may have been superimposed during transmission, subtraction reduces this interference or distortion to a small residual amount.

This clamper[10] is a feedback device which bridges or operates in shunt with the 75 ohm line as shown in the block diagram of Fig. 17. The composite television signal at the bridging point A-B is amplified; the picture portion is removed by a clipper, and the remaining signal (composed only of sync pulses containing the distortion information) is applied to an envelope detector. The detector output, representing the envelope, is returned to the transmission line at the bridging point by means of a coupling amplifier. Since this feedback envelope is nearly equal in amplitude and opposite in phase to the original distortion envelope, substantial cancellation occurs on the line.

The circuit arrangements for performing these functions are similar to those shown in Reference 10.

Passive Transmitting Terminal

This transmitting terminal uses a wide band video transformer[2] to convert the unbalanced signal from the customer to a balanced signal for transmission over the video cable. It is intended for use in short cable circuit applications where pre-equalization at the sending end is not required.

Fig. 18 shows a schematic of the terminal. The impedance matching

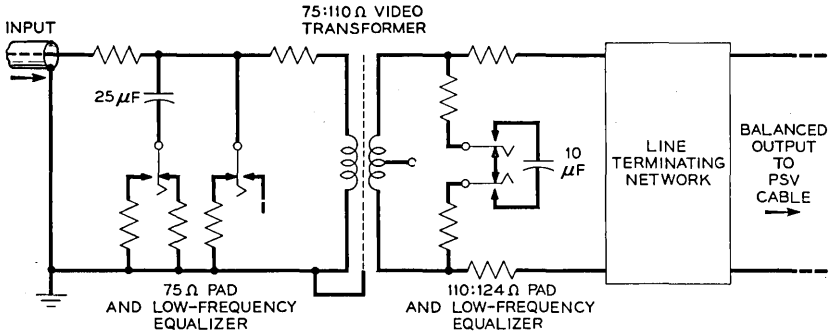


Fig. 18 — Schematic of passive transmitting terminal.

pad in the output side of the circuit converts the 110-ohm impedance of the video transformer to 124 ohms which is the high-frequency asymptotic impedance of the cable pair. The pad at the input serves as impedance correction and as isolation between the broadcaster's circuit feeding the terminal and the outgoing cable.

The loss pads also provide a means for compensating for the transmission effects of the low-frequency cut-off of the transformer. Low-frequency phase correction is accomplished by inserting capacitors in the shunt arms of the pads as shown on Fig. 18. The compensation is made

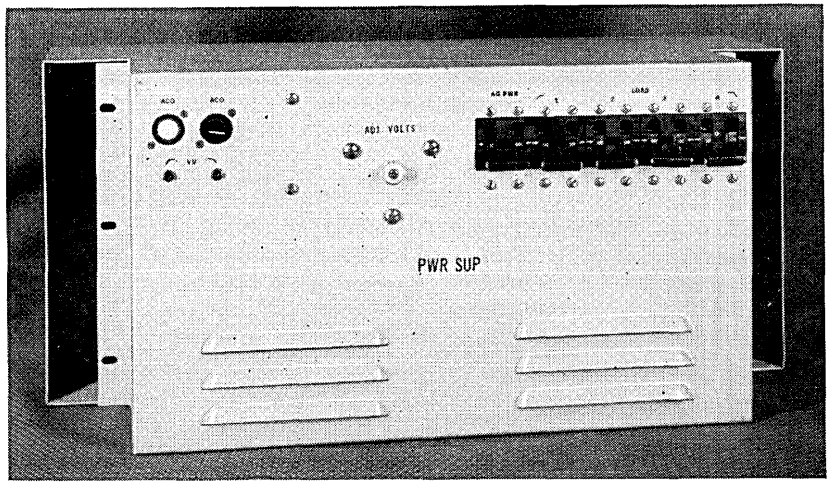


Fig. 19 — Regulator and alarm panel.

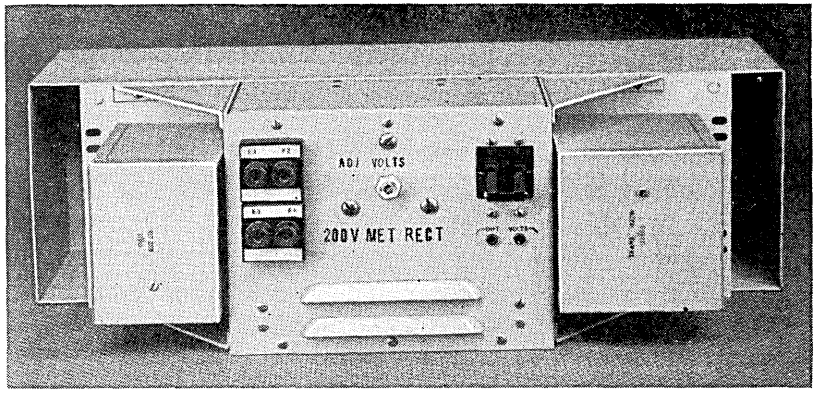


Fig. 20 — Rectifier.

adjustable in four steps by means of two keys to accommodate manufacturing variations in the transformer loss shape at low frequencies.

The line terminating network is the same network used in the output and input amplifiers. The center tap of the balanced winding of the transformer is ungrounded to maintain a high impedance to the flow of longitudinal interference currents.

Power Supply

The A2A system operates entirely from 115-volt 60-cycle commercial ac power. The power supply equipment consists of two units, a regulator and alarm panel, and a 200-volt rectifier. These panels are pictured in Figures 19 and 20.

The regulator and alarm panel uses an ac voltage regulator of the ferro-resonance type to supply closely regulated voltage to the rectifier and to the individual filament transformers of each amplifier. An adjustable auto transformer is provided to set initially the output voltage of the regulator for each amplifier filament load condition. The regulator thereafter will correct for variations in line voltage.

The rectifier is a conventional circuit using a selenium bridge stack and can supply a maximum output of 625 milliamperes at 200 volts dc. One of these rectifiers is required for the largest receiving terminal or repeater, but it can supply more than one of the smaller units. For example, three transmitting terminals may be supplied from one rectifier. An adjustable auto transformer permits setting the output voltage for each amplifier load combination.

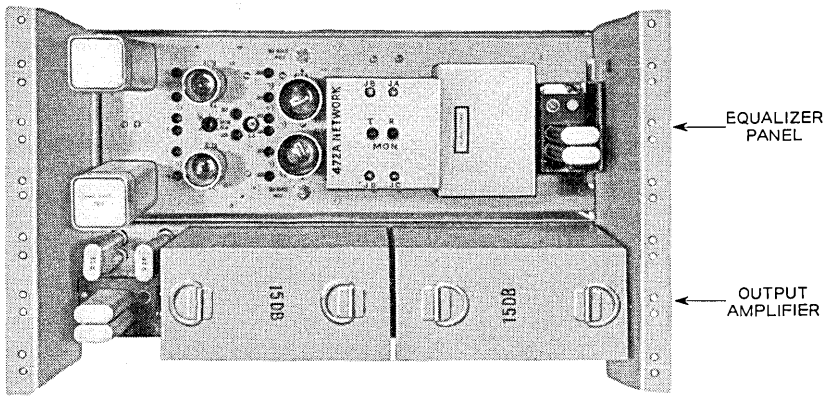


Fig. 21 — Typical transmitting terminal.

Overload protection is provided by fuses in the rectifier and by circuit breakers in the regulator and alarm panel. The alarm feature provides for triggering the central office alarm system for either dc or ac voltage failure or overload current in the circuit breakers or fuses.

A saving of 40 to 50 per cent in power input compared to an equivalent A2 repeater or receiving terminal has been achieved by the use of magnetic regulators and metallic rectifiers instead of electronically regulated rectifiers.

Equipment Mounting Arrangements

All of the equipment panels and assemblies are designed to mount on standard 19″ duct-type bays. These bays are available in 11½-foot height for use in central office installations and in 6- and 7-foot heights for use in "off-premise" locations such as quarters provided in a customer's building. Portable cabinets are also provided for use in temporary circuit arrangements.

Fig. 21 shows the physical arrangement of the panels for a typical transmitting terminal, less the power equipment. The upper panel is the output amplifier which occupies 7″ of vertical space. This panel is mechanically insulated from the supporting framework by means of rubber mountings to reduce microphonic effects due to mechanical shocks or vibrations.

The equalizer panel, occupying 3½ inches of panel space, provides mounting space for two equalizers and two loss pads for gain adjustment. The transmitting terminal pictured is arranged for balanced

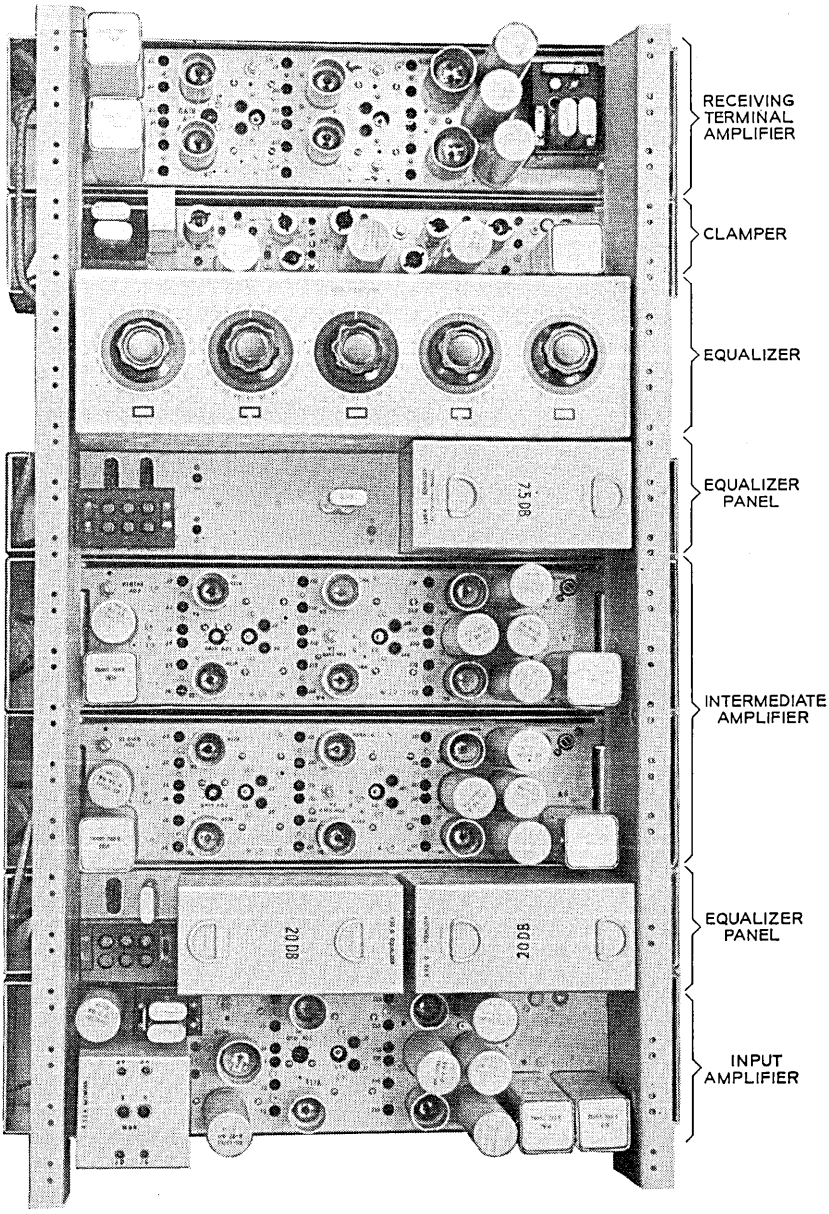


Fig. 22 — Typical receiving terminal.

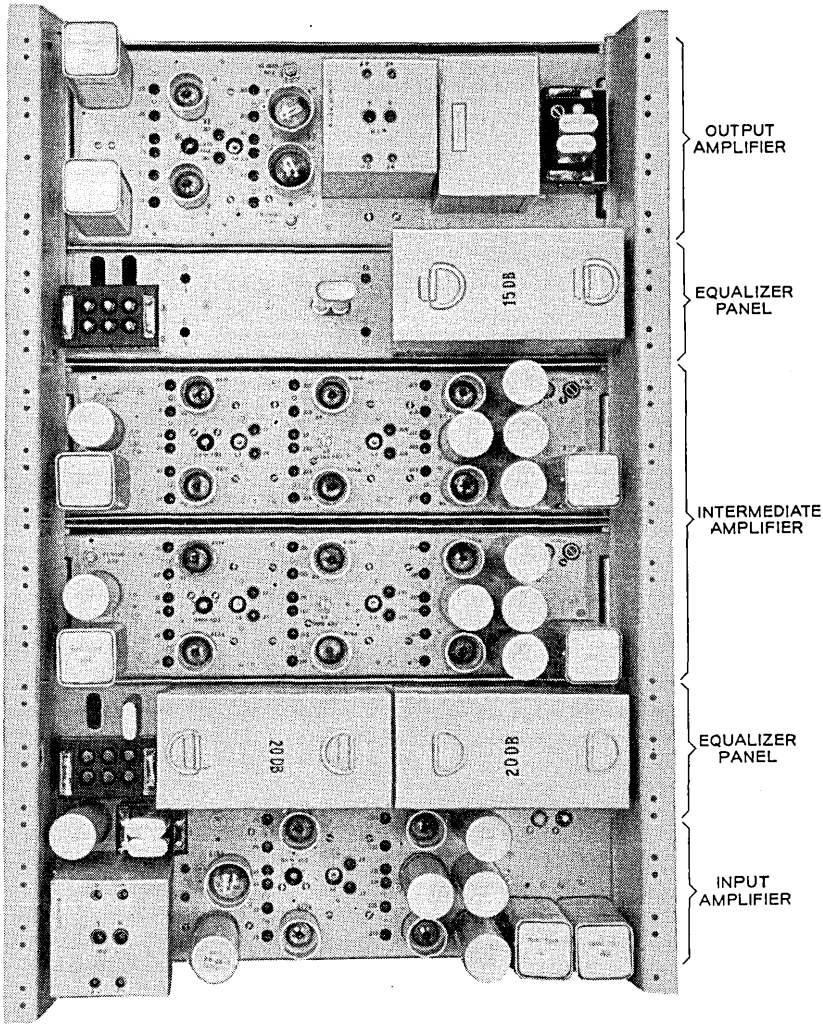


Fig. 23 — Typical repeater.

input, and uses two equalizers in a balanced connection. The manner of attachment of the equalizers to the equalizer panel is illustrated in Fig. 24. Connections are made through coaxial jacks and plugs, and snap fasteners are used to hold the equalizer in place.

The receiving terminal pictured on Fig. 22 contains the maximum complement of amplifiers, and occupies about 3 feet of bay height, exclusive of the power equipment. The progression of the circuit is from the input amplifier at the bottom to the receiving terminal amplifier at the top. Two 20-db fixed cable equalizers associated with the two intermediate amplifiers are mounted on one of the two equalizer panels. The second equalizer panel is shown with one fixed equalizer. The unused space on this panel is available for the variable "B" equalizer when this is required.

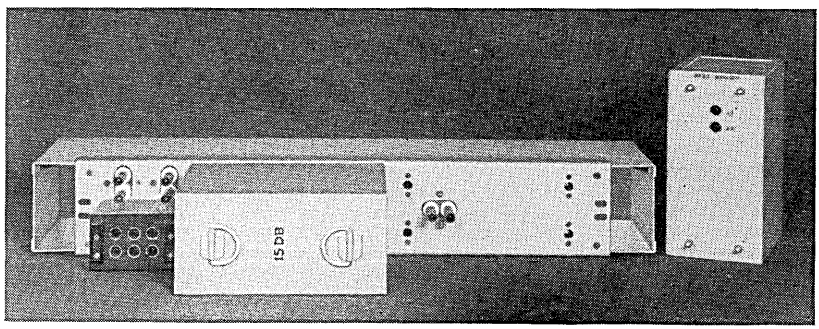


Fig. 24 — Equalizer panel and plug-in equalizers.

The variable "A" equalizer which is required for all circuits is permanently wired in.

Coaxial jack or plug access is provided for the terminal as a whole, and also for each component amplifier and equalizer. This feature facilitates maintenance testing, and permits temporary replacement of a panel by the use of patch cords in event of failure.

The repeater of Fig. 23 contains the maximum number of amplifier panels and occupies about 2½ feet of vertical space on the bay, exclusive of the power equipment. The equipment features are the same as those of the transmitting and receiving terminals.

ACKNOWLEDGMENT

Although this account of the A2A system has been presented by only two persons, the authors wish to stress that the development of the

system represents the combined accomplishment of a large number of people in several areas of the Laboratories, of members of the Pacific Telephone and Telegraph Company, and the American Telephone and Telegraph Company on loan to the Laboratories.

REFERENCES

1. Abraham, L. G., Progress in Coaxial Telephone and Television Systems, A.I.E.E. Trans., **67,** Part II, pp. 1520–1527, 1948.
2. Nebel, C. N., Local Wire Video Television Networks, A.I.E.E. Trans., **69,** Part II, pp. 1451–1460, 1950.
3. Roetken, A. A., Smith, K. D., and Friis, R. W., The TD-2 Microwave Radio Relay System, B.S.T.J., **30,** pp. 1041–1077, Oct., 1951.
4. The L3 Coaxial System, B.S.T.J., **32,** pp. 779–1005, July, 1953.
5. Windeler, A. S., Video-pair Cable, Bell Laboratories Record, **26,** pp. 201–204, May, 1948.
6. Rounds, P. W., and Lakin, G. L., Equalization of Cables for Local Television Transmission, page 713 of this issue.
7. Mertz, P., Data on Random Noise Requirements for Theatre Television, J.S.M.P.T.E., **57,** pp. 89–107, Aug., 1951.
8. Barstow, J. M., and Christopher, H. N., The Measurement of Random Monochrome Video Interference, A.I.E.E. Communication and Electronics, **73,** Part I, pp. 735–741, Jan., 1954.
9. Kelly, H. P., Differential Phase and Gain Measurements in Color Television Systems, A.I.E.E. Trans., **73,** pp. 799–802, Sept., 1954.
10. Doba, S., Jr. and Rieke, J. W., Clampers in Video Transmission, A.I.E.E. Trans., **69,** Part I, pp. 477–487, 1950.

Equalization of Cables for Local Television Transmission

By P. W. ROUNDS and MISS G. L. LAKIN

(Manuscript received March 4, 1955)

An improved equalization system has been developed for video cable. Equalizer design is based on an analysis of cable performance, which shows that an analytic expression can be obtained for the loss and phase of the cable over the video band. Fixed equalizers were designed to handle the loss characteristics of a nominal cable pair and manually-adjustable equalizers to handle expected variations. New wideband approximation methods were used to match arbitrary impedance or loss characteristics to close tolerances. With these equalizers an initial cable distortion of a hundred or more db may be reduced to less than 0.05 db.

INTRODUCTION

Television transmission places stringent requirements on equalization of the transmission medium. This is particularly true on short-haul intracity circuits. A large number of these short-haul links may be connected in tandem for overall long distance television service. Because the arrangement of circuits is subject to frequent change, each link must be independently equalized to close limits. In contrast with the fixed repeater spacings used in long-haul coaxial systems, the lengths of these individual short-haul circuits range from a fraction of a mile to ten or more miles. This requires an equalization system sufficiently flexible to handle all of the circuit lengths encountered in the field. Transmission over these local circuits is handled at video frequencies using special balanced cables. The video band is logarithmically very wide, extending from 30 cps to 4.5 mc — a span of over 17 octaves in frequency. This wide band adds to the complexity of the equalizers and makes necessary the use of special impedance networks for terminating the cable.

In connection with the A2A system described in a companion paper,[1] improved equalizers have been developed for video cable. The design of these equalizers is based on an analysis of the cable performance, which

shows that an analytic expression can be obtained for the loss and phase of the cable over the video band. Fixed equalizers were designed to handle the loss and phase characteristics of a nominal cable pair as derived from this expression, and manually-adjustable equalizers were designed to compensate for expected variations from these nominal values. The equalizers are arranged for plug-in mounting and may be installed as required to fit the desired circuit length.

CABLE CHARACTERISTICS

Loss

The insertion loss and phase of a section of cable between sending and receiving impedances provided by terminals or repeaters may be broken down into two terms; first the image transfer loss and phase, and second the reflection and interaction loss and phase.[2] The first of these two terms is an inherent property of the cable itself. The second involves the ratio of the cable image impedance to the termination. As discussed later, networks are provided in the A2A system for matching the image impedance of the cable at the two ends, thereby eliminating reflection and interaction effects. Consequently it is sufficient to consider the transfer loss and phase alone.

The image transfer constant of the cable may be expressed in the form

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{1}$$

where

α = attenuation

β = phase

ω = radian frequency

R, L, G, and C = total distributed resistance, inductance, conductance, and capacitance in the section of cable

At very low frequencies, where the conductance is negligible and the resistance is constant and large compared with ωL, the attenuation reduces to

$$\alpha = \sqrt{\omega RC/2} = \sqrt{\pi RC}\,\sqrt{f} \tag{2}$$

Thus at low frequencies the attenuation is proportional to the square root of frequency.

At high frequencies the attenuation becomes

$$\alpha = \frac{R/2}{\sqrt{L/C}} + (G/2)\sqrt{L/C} \tag{3}$$

At these frequencies $\sqrt{L/C}$ approaches a constant and the resistance varies as the square root of frequency, due to skin effects. Therefore the first term of (3) also represents a loss increasing as the square root of frequency, but with a different constant multiplier. The second term in (3) is small except at the extreme upper end of the video band and represents a component of loss which varies approximately as the first power of frequency. This component is due in part to the fact that the conductance increases linearly with frequency. Over and above this, there is an increment of loss increasing linearly with frequency, due to eddy current losses in the cable sheath.

The functional form of these losses is illustrated in Fig. 1, which is a plot of the attenuation versus frequency of a typical 1,000-ft length of 16-gauge video cable. In this figure both attenuation and frequency are plotted to a logarithmic scale. At low frequencies the attenuation α is asymptotic to the straight line $\sqrt{\pi RC}\ \sqrt{f}$, and at high frequencies α is asymptotic to a straight line with the same slope but with a different intercept. At the top of the band there is an additional small increment

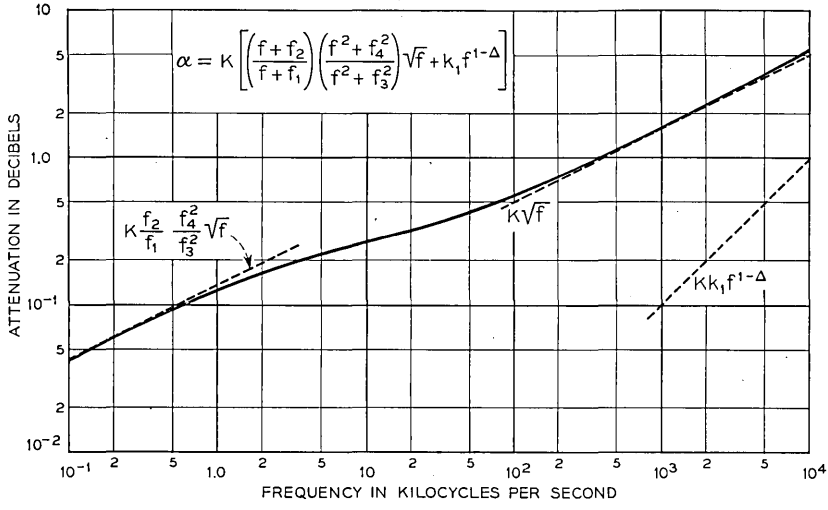


Fig. 1 — Attenuation of video pairs.

of loss given by the second term of (3), which is approximately proportional to frequency.

The curve of Fig. 1 suggests that the attenuation function could be represented by a rather simple analytic expression, provided that this follow the attenuation curve accurately in the transition region between the two straight line asymptotes. Such an expression was found to be

$$\alpha = K\left[\left(\frac{f+f_2}{f+f_1}\right)\left(\frac{f^2+f_4^2}{f^2+f_3^2}\right)\sqrt{f} + k_1 f^{1-\Delta}\right] \tag{4}$$

where $K, f_1, f_2, f_3, f_4, k_1$, and Δ are constants. With the proper choice of constants (4) can be made to match actual measured attenuation of a wide variety of cables. The values of these constants for 1000-ft lengths of several typical cables are given in Table I:

TABLE I — CONSTANTS FOR 1,000-ft LENGTHS OF SEVERAL TYPICAL CABLES

Type of Cable	Temp	K	f_1	f_2	f_3	f_4	k_1	Δ
16-gauge pairs (16 PSVL) longitudinal shield	55°F	1.58	0.00625	0.0164	0.500	0.512	0.012	0
16-gauge pairs (16 PSVS) spiral shield	55°F	1.60	0.00625	0.0150	0.217	0.228	0.043	0
18-gauge pairs (754D) solid dielectric	75°F	2.35	0.0118	0.0400	0.221	0.212	0.019	0
75-ohm coaxial (724) solid dielectric	75°F	2.22	0.0120	0.0480	0.349	0.319	0.041	0

The values of the constants in this table are given for frequency f expressed in megacycles per second and loss α in db.

Phase

As loss and phase equalization are of equal importance in television transmission, the next problem is to determine the non-linear component of cable phase. Direct measurement yields little significant information, because the non-linear component is less than 1 per cent of the total cable phase and becomes masked by normal measurement errors. The existence of an analytic expression for cable loss, however, allows the phase to be determined to an accuracy comparable with the loss accuracy through the use of the minimum phase integral.[3]

By substituting the loss expression of (4) in the minimum phase integral

$$\beta_c = \frac{2\omega_c}{\pi}\int_0^\infty \frac{\alpha - \alpha_c}{\omega^2 - \omega_c^2}\,d\omega \tag{5}$$

expanding the integrand into partial fractions, and integrating over the appropriate contour, the following expression for the cable phase is obtained:

$$\beta = K\left\{\left(\frac{f+f_2}{f+f_1}\right)\left(\frac{f^2+f_4^2}{f^2+f_3^2}\right)\sqrt{f}\left[1-\frac{2\sqrt{f}\,(f_2-f_1)}{(f+f_2)(\sqrt{f}+\sqrt{f_1})}\right]\right.$$
$$+\frac{\sqrt{2}f(f_4^2-f_3^2)}{(f^2+f_3^2)(f_1^2+f_3^2)}\left[(\sqrt{2f_1}-\sqrt{f_3})(f_2-f_1)-\frac{f_3^2+f_1f_2}{\sqrt{f_3}}\right] \quad (6)$$
$$\left.+k_1 f^{1-\Delta}\tan(1-\Delta)\frac{\pi}{2}\right\}$$

Although the phase given by the last term in (6) becomes infinite as Δ approaches zero, the non-linear component remains finite. The character of the non-linearity may be brought out by subtracting a linear phase and writing the limit. Thus

$$\lim_{\Delta\to 0}\left[k_1(f^{1-\Delta}-f)\tan(1-\Delta)\frac{\pi}{2}\right] = -k_1\frac{2}{\pi}f\log f \quad (7)$$

EQUALIZATION PLAN

As mentioned previously, the total insertion loss of a cable section includes reflection and interaction losses which are a function of the impedance presented to the ends of the cable by the terminals or repeaters. Resistive terminations give rise to low-frequency reflection and interaction losses. The reflection losses are independent of cable length and the interaction losses produce ripples in the frequency characteristic having a period inversely proportional to the length. Equalization of these reflection and interaction losses to close limits presents an almost insuperable problem when the cable lengths vary over the wide range expected in A2A systems.

To avoid this problem, the reflection and interaction losses are eliminated by terminating the cables in impedances which match the image impedance of the cable. The impedance match is provided by networks arranged in such a manner that the transmission is flat before insertion of the cable. The method of operation, as shown in Fig. 2, was suggested by C. N. Nebel. With this arrangement the voltage on the grid of the receiving amplifier is exactly one-half the Thevenin open-circuit voltage at the output of the transmitting amplifier and the inserted cable faces its image impedance in both directions. With reflection and interaction losses eliminated by means of these impedance networks, the total loss to be equalized is that given by (4).

This equation provides the key for the entire equalization plan. Fixed equalizers are provided to equalize the loss of the cable as given by the equation with nominal values of the constants. Adjustable equalizers are provided at the ends of the circuits to compensate for changes in loss arising from changes in the constants from their assumed nominal values. By this means it is possible to equalize any cable likely to be encountered in the field. The fixed equalizers are provided in blocks corresponding to cables having losses of 2.5, 5, 7.5, 10, 15 or 20 db at 4.5 mc. As discussed in a companion paper,[1] these fixed equalizers are associated with flat-gain amplifiers to provide flexibility in application.

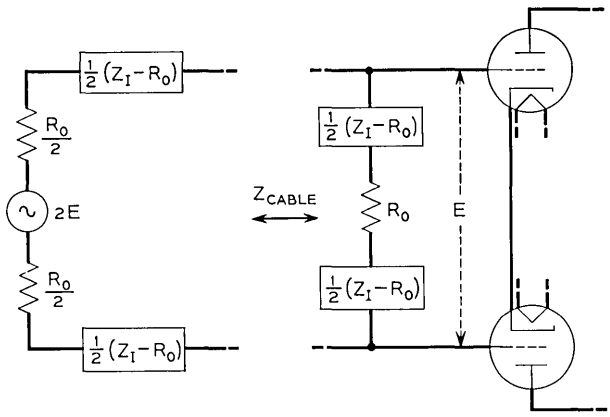


Fig. 2. — Method of terminating cable pairs.

For circuits up to about 3.5 miles in length, it is possible to use flat-gain receiving and transmitting amplifiers with grid and plate directly connected to the line. For longer circuits, however, signal-to-noise considerations necessitate the use of input and output coupling networks which increase the transmission from the line to the grid and from the plate to the line at high frequencies. These networks have simple gain functions and are associated with fixed equalizers which, in combination with the networks, provide equalization for fixed lengths of line.

Since the cable and equalizers are all minimum phase structures, the phase distortion is uniquely related to the loss distortion. If the gain of the system were flat from zero to infinite frequency there would be no phase distortion. The decreasing system gain above the band, however, introduces phase distortion within the band. It is possible to associate the phase equalization directly with the loss equalizers, since (6) relates

the phase distortion of the line directly with the loss distortion. Therefore, each fixed loss equalizer, including those associated with the coupling networks, contains a fixed phase equalizer for handling the phase distortion in the related length of line. This insures proper phase equalization without the necessity of conducting phase measurements in the field. No phase equalizers need to be associated with the adjustable equalizers. Not only is the phase distortion of these equalizers small, but their outband loss performance is such that their phase tends inherently to be complementary to the characteristics being equalized.

DESIGN METHODS

Cable-Impedance Networks

The design of the impedance-matching networks is based upon the fact that the image impedance of the cable is a minimum-reactance function and hence both resistance and reactance are automatically specified when the magnitude of the impedance $|Z|$ is given.[3] Thus, if the impedance magnitude is matched over a sufficiently wide band of frequencies, the resistive and reactive components will also be matched.

In Fig. 3 the logarithm of the image-impedance magnitude $|Z_I|$ for video cable is plotted against the logarithm of frequency. At low frequencies the impedance $|Z_I|$ is inversely proportional to the square root of frequency. It appears in Fig. 3 as a curve approaching a straight-line asymptote having a slope k of $-\frac{1}{2}$. At high frequencies $|Z_I|$ approaches a constant value.

Before proceeding with the design of a network to match this impedance, it is convenient to consider the idealized problem of approximating an impedance $|Z|$ which is proportional to f^{-k}, or which decreases along a straight line of slope $-k$, as shown in Fig. 4. This impedance may be

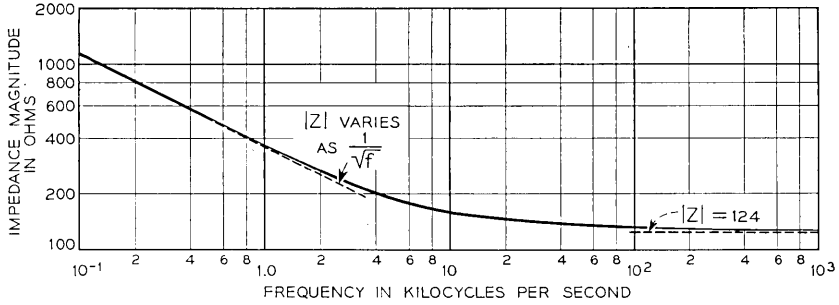


Fig. 3 — Magnitude of image impedance of viedo pairs.

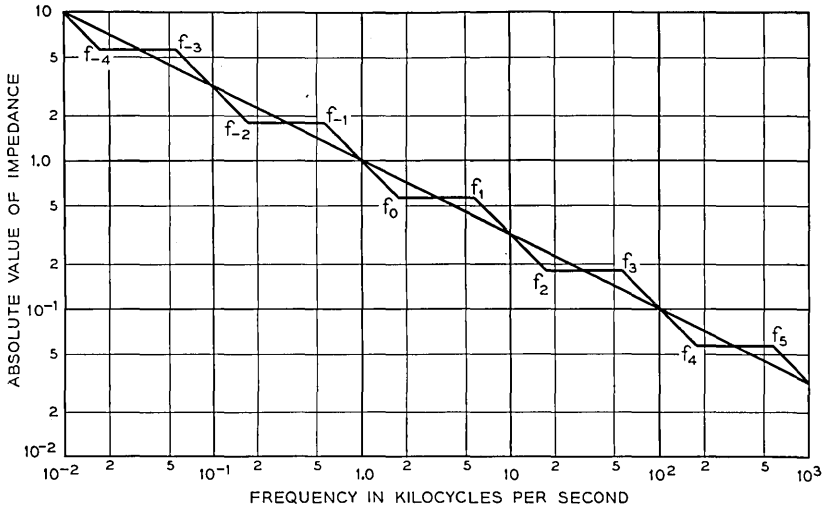


Fig. 4 — Asymptotes for network approximating infinite slope $| Z | = f^{-1/2}$.

approximated with a network which has an infinite number of zero-pole pairs (corresponding to an R-C network). The impedance Z' of such a network is given by the infinite product

$$Z' = H \prod_{-\infty}^{\infty} \frac{p - p_{2n}}{p - p_{2n-1}} \tag{8}$$

where $p = j2\pi f$, and p_{2n} and p_{2n-1} are negative real numbers.

The network behavior may be represented by straight-line asymptotes with slopes of $k = 0$ and $k = -1$ which intersect at frequencies corresponding to $-p_n/2\pi$, as shown in the Fig. 4. When the zero-pole pairs are repetitive on a logarithmic frequency scale, the magnitude $| Z' |$ computed from (8) will ripple about the desired curve with a uniform amplitude and period. The amplitude of this ripple corresponds to the percentage error in the impedance magnitude and may be computed as a function of k and of h, the number of zero-pole pairs per decade, following a method analogous to that described in Reference 4 for computing the error of an infinitely decreasing loss characteristic. Fig. 5 shows the percentage error in $| Z' |$ when k equals $\pm\frac{1}{2}$. For any other value of k the error varies as $\sin \pi k$.

Thus for any impedance magnitude and desired degree of match, the complexity of the network may be determined using Fig. 5 as a guide. The critical frequencies may be found from the asymptotic representa-

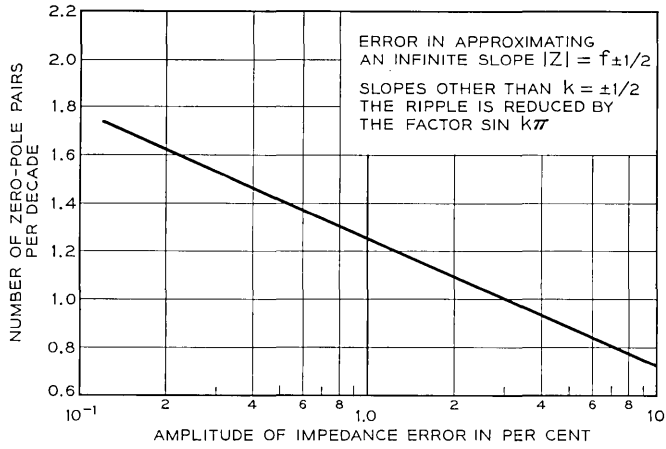


Fig. 5 — Error in approximating $| Z | = f^{\pm k}$.

tion. This is shown in Fig. 6 for the impedance-matching networks for video cable. The figure also gives the expression for the network impedance. The resulting reflection coefficient of the network against the cable image impedance is plotted in Fig. 7. A schematic of the network appears in Fig. 8.

Fixed Cable Equalizers

As mentioned previously, fixed equalizers are available in discrete sizes to handle cables of a variety of lengths. The smallest block equalizes a cable loss of 2.5 db and the largest a cable loss of 20 db at 4.5 mc. The design of each of these fixed equalizers is similar. Fig. 9 shows the desired loss characteristic for the 20-db equalizer. The design of this equal-

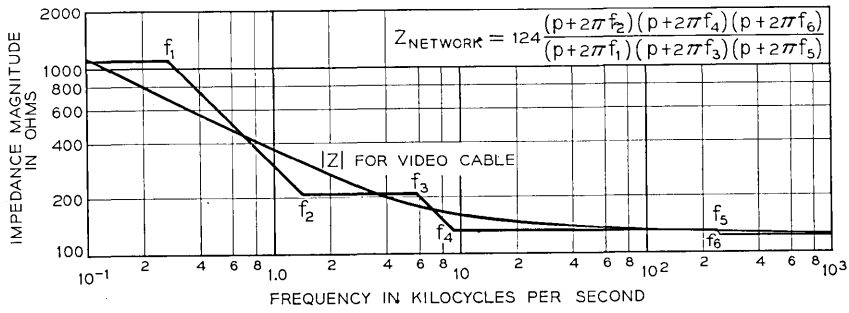


Fig. 6 — Asymptotic representation of impedance network.

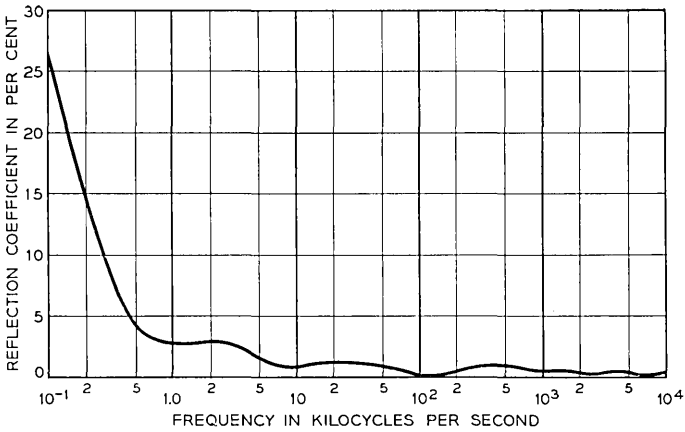


Fig. 7 — Reflection coefficient of network against cable impedance.

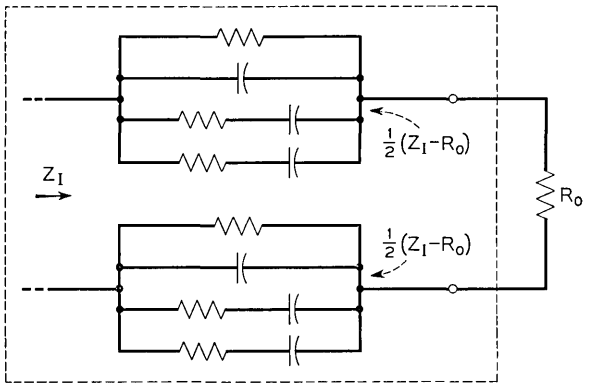


Fig. 8 — Schematic of impedance network.

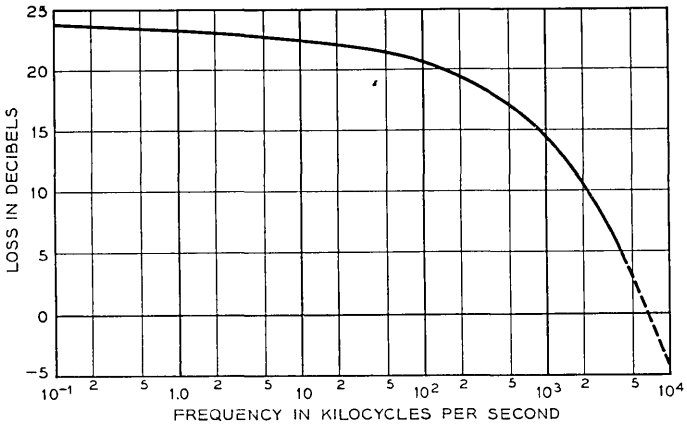


Fig. 9 — Equalizer loss objective.

izer is described in detail in Reference 4 and is analogous to the design of the impedance-matching networks as outlined above. Since the loss is monotonically decreasing over the band, it can be matched with a bridged-T equalizer whose series arm consists of a resistor shunted by resistance-capacitance branches. The complexity of the network can be determined by considering the idealized problem of matching a loss characteristic with an infinite slope of 20k db per decade by an equalizer having an infinite number of uniformly-distributed zero-pole pairs. The loss expression of such an equalizer is given by the infinite product

$$e^{\theta} = \prod_{-\infty}^{\infty} \frac{p - p_{2n}}{p - p_{2n-1}} \tag{9}$$

where $\theta = \alpha + j\beta$ = insertion loss and phase

$p = j2\pi f$

p_{2n} and p_{2n-1} are negative real numbers

As in the case of the impedance networks, the equalizer loss will ripple about the desired straight line characteristic. The amplitude of the ripple is a function of k and of h, the number of zero-pole pairs per decade. The amplitude of the loss error as a function of h is shown in Fig. 10 for a loss characteristic having a slope of 10 db per decade (3 db per octave) corresponding to $k = \pm 1/2$. For any other slope the loss error is sin πk times this amplitude.

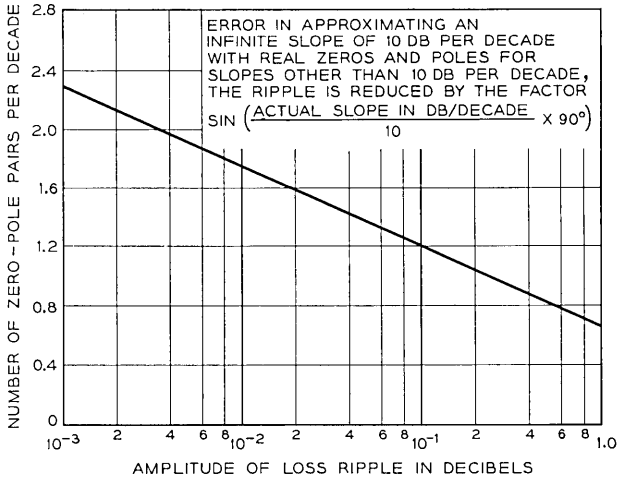


Fig. 10 — Error in approximating infinite loss slope.

This curve was used to determine the complexity necessary for each fixed equalizer. In order to meet tolerances for the requisite number of equalizers in tandem, a design ripple of ±.005 db was selected. The resulting equalizer as shown in Fig. 11 contains nine resistance-capacitance branches in the series arm. A resonant-frequency branch has been added to bring the loss to zero above the band in order to conserve system gain by reducing the total equalizer loss. The configuration of the inverse shunt branch has been modified to allow the inductor dissipation to be absorbed. As shown in the figure, one all-pass section is included to equalize the phase distortion. Fig. 12 shows the sum of the equalizer and cable losses.

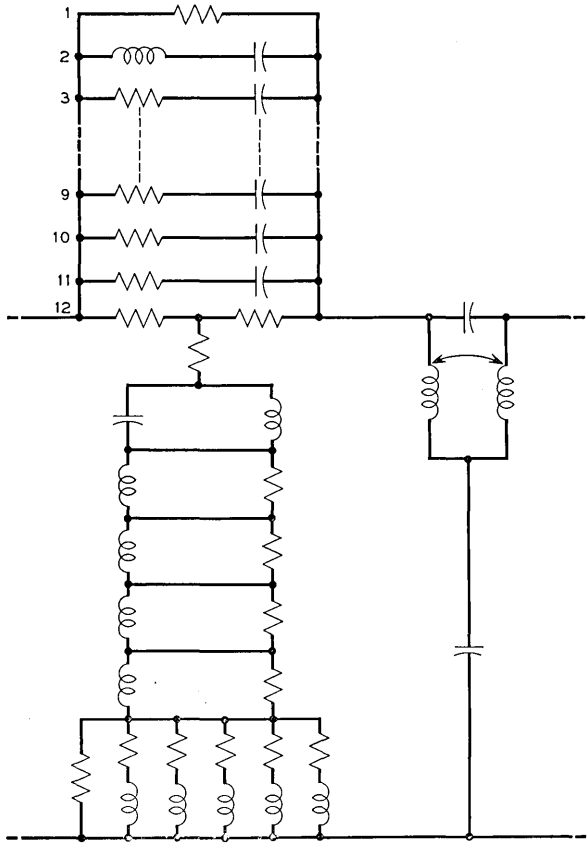


Fig. 11 — Schematic of fixed cable equalizer.

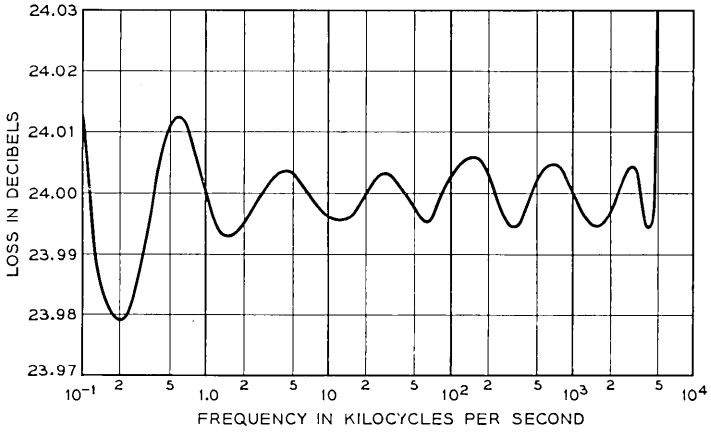


Fig. 12 — Sum of equalizer loss and cable loss.

Input and Output Networks

With flat-gain amplifiers and the fixed equalizers described above, the maximum length of a single A2A link is about 3.5 miles or 65 db.[1] This length can be increased to 4.5 miles through the use of coupling networks which step up the gain at high frequencies. Each network improves the signal-to-noise ratio by approximately 10 db. These networks are balanced-to-ground and use a transformer with a 6 to 1 turns ratio, corresponding to an asymptotic gain of 15.6 db at high frequencies.

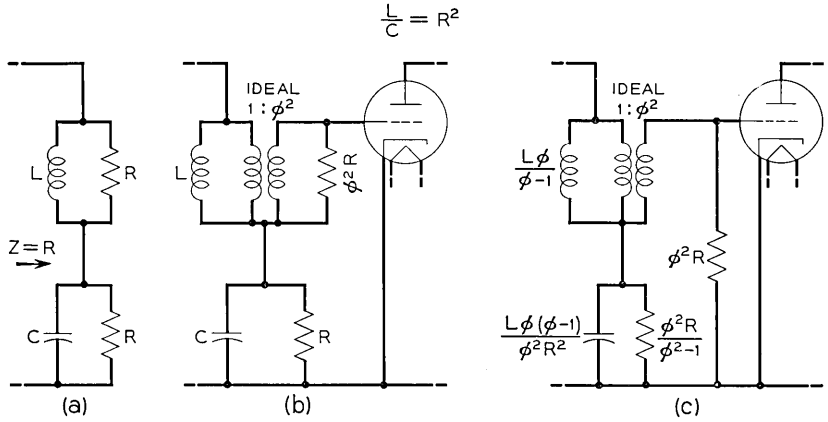


Fig. 13 — Ideal coupling network.

The design of the A2A input and output networks is similar in principle to the design used by E. W. Holman in 1939 for the A2 video system. The basic circuit was derived from the resistive network shown in Fig. 13(a). If the inductor is stepped-up with an ideal transformer of turns ratio $1:\phi$, as shown in Fig. 13(b), the input impedance remains resistive but the ratio of output voltage to input voltage is given by

$$\frac{\text{output voltage}}{\text{input voltage}} = e^{\theta} = \phi \frac{p + R/L\phi}{p + R/L} \tag{10}$$

where $\theta = \alpha + j\beta$ = gain and phase

$p = j2\pi f$

The gain characteristic is given in Fig. 14. At low frequencies the gain is zero, and at high frequencies it is asymptotic to 20 log ϕ. The location of the characteristic in the frequency spectrum is controlled by the ratio of R to L.

In an actual network this ideal gain characteristic is limited by tube and transformer capacitance and by transformer leakage reactance. The parasitic elements may be handled more readily by using the equivalent circuit of Fig. 13(c), which was originally suggested by S. Darlington. In Fig. 13 the voltage from cathode to grid is the sum of the voltages across the R-L and R-C branches. These voltages add in amplitude and phase to produce the simple characteristic of Fig. 14. Any low-pass filter structure used to absorb parasitic capacitances should be designed to shift the phase of both components of the voltage equally in order to avoid ripples in the characteristic. This can be done if the resistor termi-

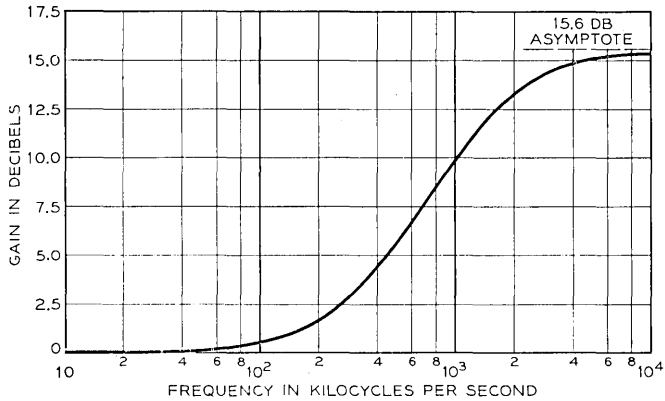


Fig. 14 — Gain characteristic of ideal coupling network.

nating the low-pass filter is connected from grid to ground. Another merit of the configuration of Fig. 13(c) is that, in the case of the output network, the plate resistance can be absorbed by modifying the terminating resistor.

The actual networks are built as balanced-to-ground structures as shown in Figs. 15 and 16. The high-frequency parasites are absorbed into a seven-element low-pass filter having Tchebycheff pass-band characteristics. In the input network the grid is connected to an intermediate point in the filter, as shown. Since the impedance at this point rises with frequency the gain characteristic, shown in Fig. 17, exhibits a rise above the nominal 15.6 db of Fig. 14. In the output network the plate is also connected to an intermediate point in the filter. This filter was designed to have the proper response with a shunt conductance across the capacitor to which the plate is connected in order to allow the plate resistance of the tube to be absorbed. The transmission characteristic is similar to that of the input network. As noted in Figs. 15 and 16, cable impedance networks are added as a shunt element in the input network and as a series element in the output network.

Although the gain characteristic shown in Fig. 17 rises with frequency

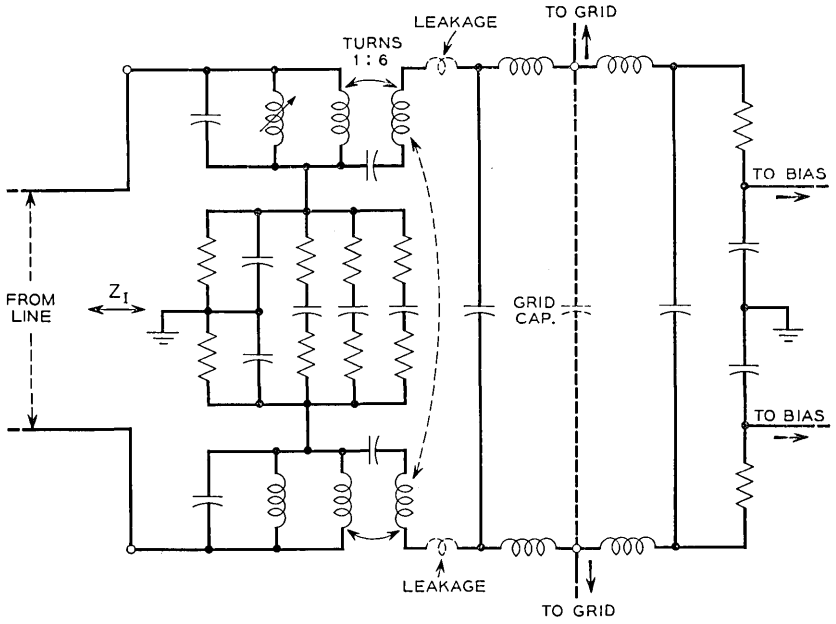


Fig. 15 — Schematic of input network.

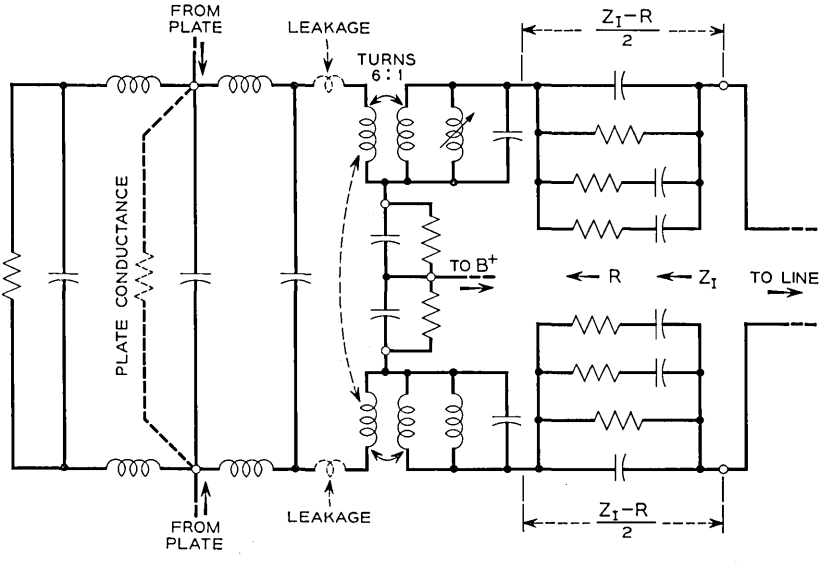


Fig. 16 — Schematic of output network.

the shape is not complementary to the loss of the cable. Fixed equalizers are associated with each of the networks in such a manner that the combination of network and equalizer compensates for 32 db of cable loss at 4.5 mc. The design of these equalizers is similar to that of the fixed cable equalizers. Loss characteristics are shown in Fig. 18 and a schematic in Fig. 19.

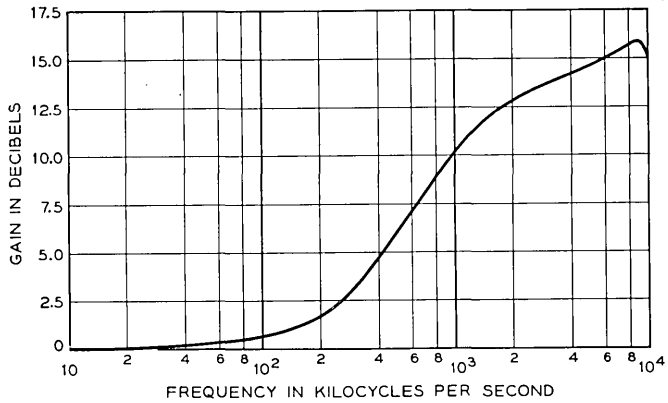


Fig. 17 — Gain characteristic of input network.

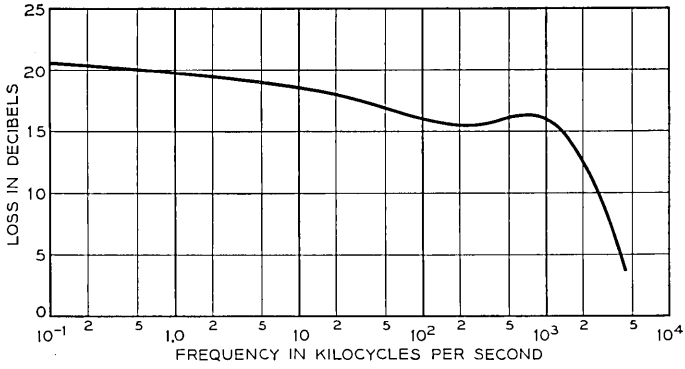


Fig. 18 — Loss characteristics of network equalizers.

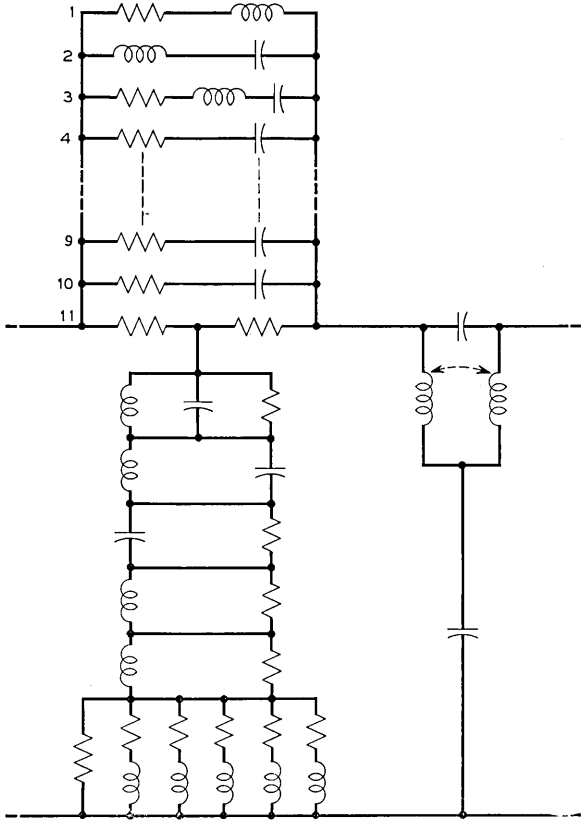


Fig. 19 — Schematic of network equalizer.

ADJUSTABLE EQUALIZERS

The smallest fixed equalizer, which compensates for a cable loss of 2.5 db, allows the circuit to be equalized to the nearest ±1.25 db at 4.5 mc. Equalization for incremental lengths finer than this is handled by manually-adjustable equalizers located at the receiving terminal. These equalizers also allow correction to be made for changes resulting from: (1) manufacturing variations in the cables and changes in cable types; (2) reflection and interaction losses of the resistance-terminated office cables at the ends of the circuits; and (3) variations in amplifiers and equalizers which produce high-frequency loss changes approximately proportional to f^3. The resulting characteristics are shown in Fig. 20. The shapes in the first category were obtained by differentiating (4) with respect to each of the constants. The reflection and interaction shape is that computed for short lengths of coaxial cable terminated in 75 ohms. Provision of the flexibility inherent in these eight shapes would appear to allow any loss variations expected in the system to be successfully equalized.

The optimum choice of loss shapes for a system of adjustable equalizers is dependent on the type of measuring facility available for system line-up. In the A2A system a test set is used which compares the gain at a single test frequency to the gain at 300 kc. The characteristics shown in Fig. 20 overlap in the frequency spectrum. Consequently they are not

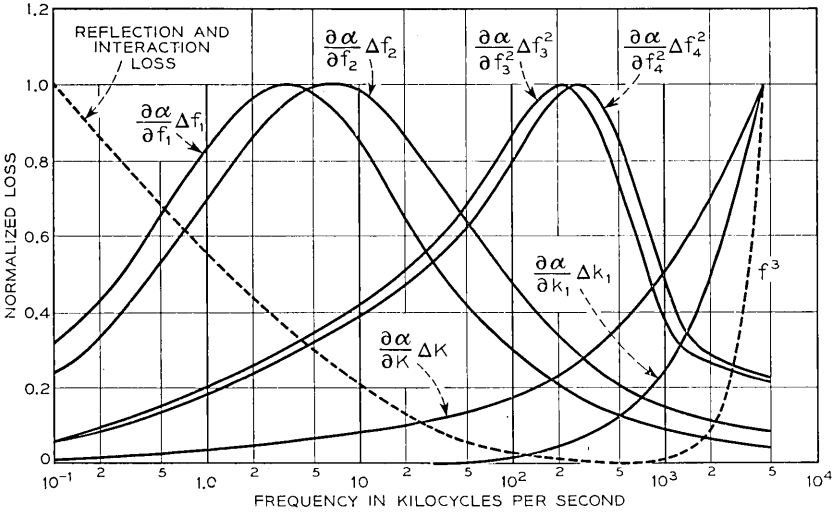


Fig. 20 — Types of loss changes to be equalized.

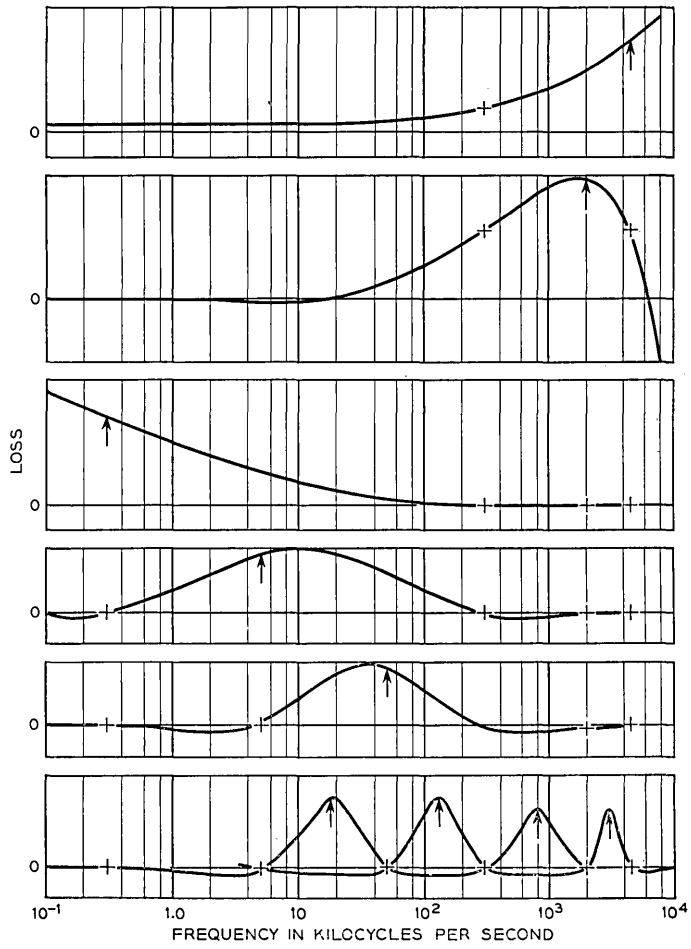


Fig. 21 — Loss characteristics of adjustable equalizers.

amenable to straightforward line-up using this test facility. However, it is possible to resolve this difficulty by taking linear combinations of these characteristics such that each successive shape remains invariant at any of the preceding adjustment frequencies. A ninth shape was added when it became apparent that not all of the changes in the constants of (4) could be treated as differentially small changes. The resulting shapes are illustrated in Fig. 21 with the adjustment frequencies indicated by arrows. Each equalizer is adjusted so that at its adjustment frequency the system gain is made equal to the gain at 300 kc.

The first five shapes, which are the most important, have been packaged in one container known as the "A" Equalizer. The next four shapes comprise the "B" Equalizer, which is normally used only on circuits over 3.5 miles in length.

The structure used to obtain these shapes is a constant-resistance variable equalizer which can readily be adjusted to obtain proportional characteristics varying in either direction from a reference flat condition. This equalizer is a bridged-T structure whose series arm contains a four-terminal constant-resistance network terminated in an adjustable resistor. The overall change in loss and phase from the flat condition may be expressed in terms of the transfer constant of the component four-terminal network. Thus

$$\theta - \alpha_0 = 2 \tanh^{-1} k\rho e^{-2\phi} \approx k\rho e^{-2\phi} \tag{11}$$

where

$\theta - \alpha_0$ = over all change in loss and phase

$k\rho$ is a function of the maximum equalizer range and the setting of the adjustable resistor

ϕ = transfer loss and phase of the four-terminal network

From the approximate relation the transfer constant ϕ may be determined from the desired equalizer loss and phase, $\theta - \alpha_0$. Although Fig. 21 shows only the loss, the phase component of $\theta - \alpha_0$ may be computed

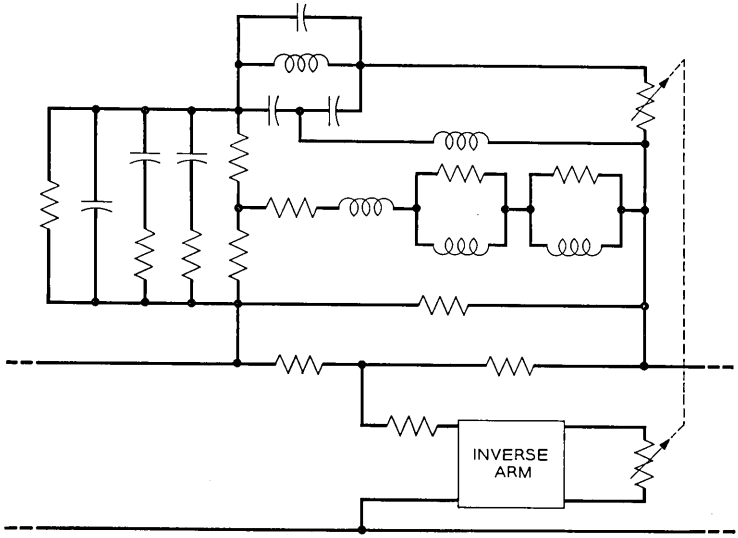


Fig. 22 — Adjustable equalizer with one control.

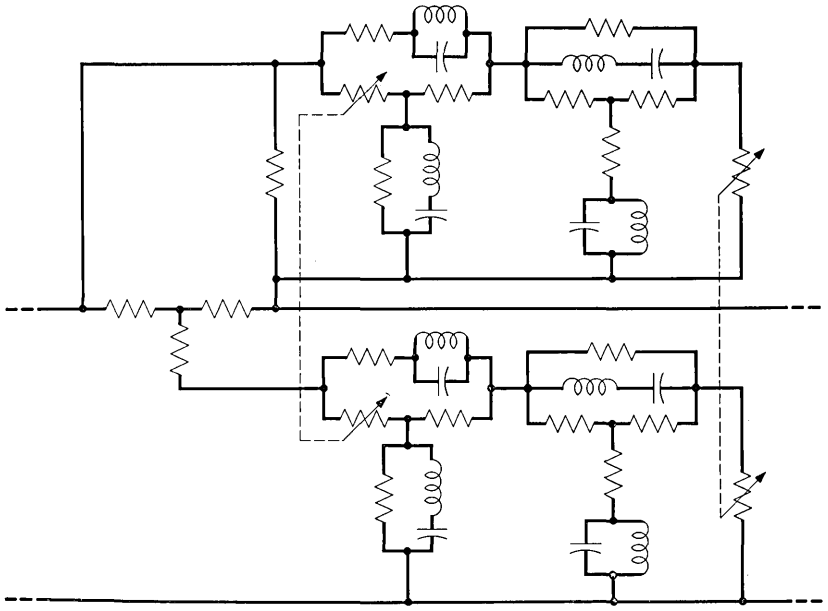


Fig. 23 — Adjustable equalizer with two controls.

using a minimum phase matrix developed by B. A. Kingsbury. This matrix is similar to the one described by W. R. Lundry[5] but is arranged for logarithmic frequency intervals.

With the transfer constant ϕ available, the four-terminal networks can be synthesized by standard methods. Typical schematics of the overall equalizers are shown in Figs. 22 and 23. The configuration of Fig. 23 conserves gain by combining the functions of two controls into one structure.[6]

CONSTRUCTION

The quality of performance of the manufactured product depends on the types of components used, the choice of configuration and the mechanical arrangement of parts. The video band is logarithmically very wide and requires large element values for operation at the low end of the range. The inductors and capacitors should be precise elements at low frequencies. Throughout the rest of the band the inductors should maintain high impedance and the capacitors low impedance. However, low frequency inductors and capacitors characteristically exhibit parasitic secondary resonances and antiresonances within the video band.

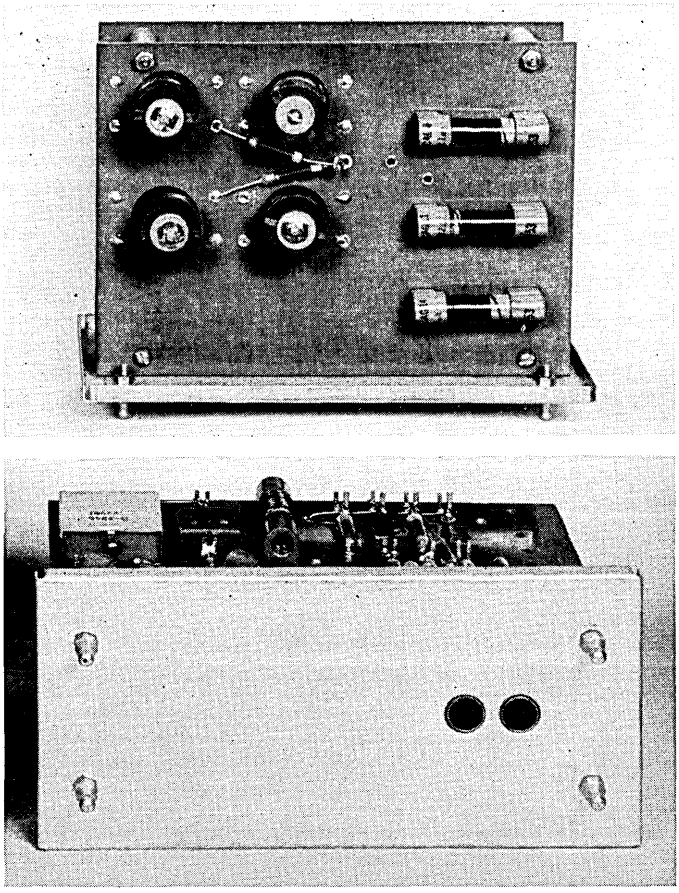


Fig. 24 — Photograph of fixed cable equalizer.

Accordingly it was necessary to develop new types of components which would be free of these spurious characteristics. In the case of the inductors this was done with a ferrite core structure. The magnetic properties of ferrite are superior to those of other magnetic materials over the video range and it is feasible to manufacture ferrite parts in shapes that allow simple uniform windings.[7] The secondary antiresonances in the capacitors were avoided by reducing the lead lengths of the individual units to a minimum and by pairing two units not too disparate in size to achieve the desired precision of adjustment. At high frequencies the

element values are low, and standard molded-mica capacitors and air-core inductors are satisfactory.

Of the many methods available for network synthesis, some allow the network configuration to be prescribed in advance, others place only general restrictions on the functional forms such that the network meets the requirements of physical realizability without specifying in advance what the configuration will be. The advantages of prescribing the configurations in advance are twofold. In the first place, a configuration may be selected in which the effects of element variations are minimized. This is done by arranging the circuit so that the share contributed by each element to the overall characteristic is small and adds to, rather than partially cancels, the portion contributed by other elements. In the second place, the configuration can be chosen in a manner such that the parasites can be absorbed in the structure, or at least such that their effects are minimized. These were the basic considerations in the selection of the configurations for the A2A networks and equalizers.

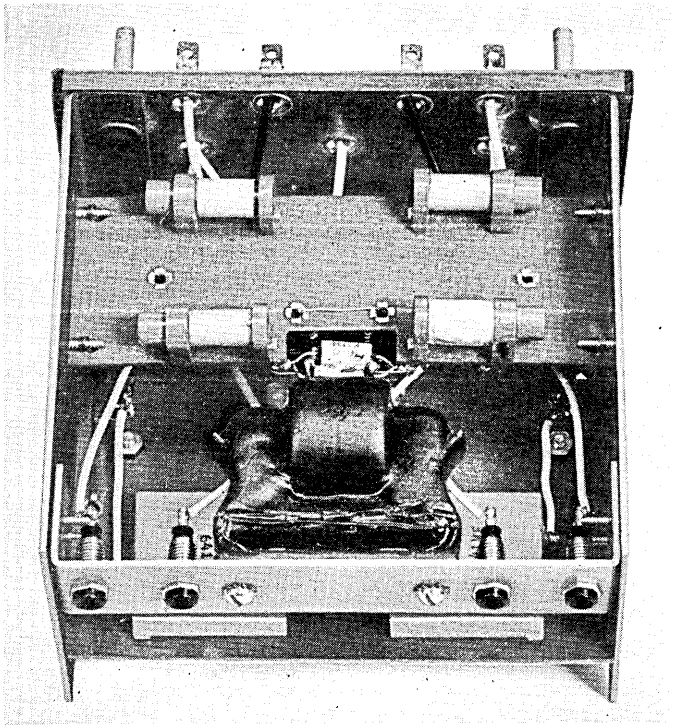


Fig. 25 — Photograph of output network.

The construction of one of the fixed equalizers is illustrated in Fig. 24. A mechanical assembly has been developed to accommodate the diverse types of components used. To permit equipment flexibility the equalizers were designed for plug-in mounting, as shown in the illustration. The locating pins fit into spring catches on the equalizer mounting plate. In the assembly of the equalizer the high-frequency resistance-capacitance branches were placed close to the input-output jacks to keep lead inductance as low as possible. With these precautions the insertion loss of the manufactured product matches the computed performance to high precision. At the same time the maximum reflection coefficient at input and output terminals is held to 2 per cent.

Fig. 25 shows a view of the output network. The ferrite-core transformer appears at the bottom of the figure. The pin jacks adjacent to the transformer are used to short out the cable impedance networks in testing the amplifier. Since this is a balanced network, care was taken to maintain constructional symmetry in order to avoid introduction of longitudinal unbalances.

The adjustable "A" equalizer is made up of five component equalizers assembled in one can as shown in Fig. 26. The adjustment sequence calls for the setting of the knobs from left to right. A typical component equalizer is illustrated in Fig. 27. Control of the characteristic is obtained with a step-type dual rheostat wired with deposited carbon resistors as shown in the figure. Where necessary, the phase shift of the rheostat and its connecting leads was absorbed in the design of the four-terminal network which the rheostat terminates. Each equalizer maintains a maximum reflection of 2 per cent for any setting of the control knob.

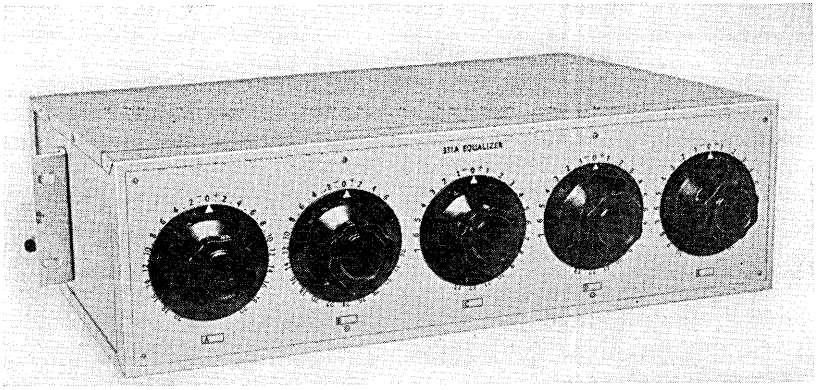


Fig. 26 — Adjustable "A" Equalizer.

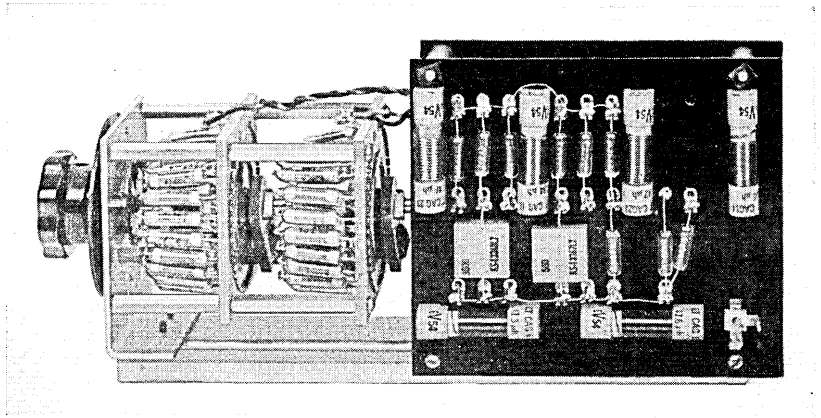


Fig. 27 — Component of "A" Equalizer.

The adjustable "B" equalizer is assembled in a container similar to that used for the fixed equalizer, and hence may be plugged in as required on long circuits. Controls on this equalizer utilize small dual composition-type rheostats.

CONCLUSION

Experience with these equalizers in commercial service indicates that the wide variety of circuit lengths and cable types encountered in the field can be successfully equalized. Fixed blocks of equalization are installed only as required for the circuit length involved. Use of a new design technique allows sufficient precision in the fixed equalizers so that no mop-up shapes are required for their design deviations; the only adjustable equalizers used are those which compensate for manufacturing variations in the cable and changes in cable types, as predicted from the analytic expression for cable losses. These factors aid in reducing costs on the preponderant number of short circuits which exist in the video plant. Circuit alignment is straightforward and rapid, using simple portable test equipment. With this equalization system an initial cable distortion of a hundred or more db may be reduced to less than 0.05 db.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the part played by many of their colleagues in carrying out this development. H. M. Thomson was responsible for the mechanical design of the equalizers and networks and G. B. Thomas for the design of the adjustable equalizers under the

direction of C. L. Semmelman. Dorothea M. Bohling, Frances C. Dunbar, and Kathleen M. Greenfield did much of the detailed design work, including programming the extensive computations performed on the Mark VI computer.[8]

REFERENCES

1. Doba, S. and Kolding, A. R., A New Local Video Transmission System, page 677 of this issue.
2. Johnson, K. S., Transmission Circuits for Telephonic Communication, D. Van Nostrand Co.
3. Bode, H. W., Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co.
4. Rounds, P. W., Equalization of Video Cables, Convention Record of the I.R.E., Part 2, Circuit Theory, March, 1954.
5. Lundry, W. R., Application of a Minimum Phase Matrix to Adjustable Equalizer Design, Convention Record of the I.R.E., Part 2, Circuit Theory, March, 1954.
6. Lundry, W. R., Attenuation and Delay Equalizers for Coaxial Lines, A.I.E.E. Trans., **68,** Part 2, pp. 1174–1179, 1949.
7. Stone, H. A., Jr., Ferrite Core Inductors, B.S.T.J., **32,** 265, March, 1953.
8. Graham, R. S., Relay Computer for Network Analysis, Bell Lab. Record., **31,** p. 152, April, 1953.

Relaxation Phenomena in Ferrites

By A. M. CLOGSTON

(Manuscript received March 2, 1955)

J. K. Galt has recently suggested a mechanism for the losses observed in ferromagnetic resonance and domain wall motion in single crystals of nickel ferrite containing small amounts of divalent iron substituted for divalent nickel. Each such substitution provides one electron able to move between the various octahedral sites in the crystal. Galt's suggestion is that the losses arise from a relaxation associated with the motion of these electrons.

This paper develops a theory of this mechanism based on a thermodynamic model. Expressions are found for the velocity of domain wall motion, for the line width in ferromagnetic resonance and for the displacement of the field for resonance.

INTRODUCTION

In a recent paper, Galt[1, 2] and Wijn and van der Heide[3] have suggested a mechanism for the losses observed during ferromagnetic resonance and domain wall motion in single crystals of nickel ferrite having a small amount of divalent iron substituted for divalent nickel in the octahedral sites. It is supposed that each substitution of a divalent iron atom for a divalent nickel atom provides one 3d electron able to move from one to another of the various iron atoms on the octahedral sites. For convenience, we shall term such electrons the "free electrons". With the crystal magnetized in a particular direction, these free 3d electrons will always so arrange themselves as to minimize the free energy of the crystal. The free energy, therefore, is going to depend upon the direction of magnetization because of this mechanism, in addition to its usual dependence upon the crystalline and shape anisotropy. It is clear that there will be an additional torque acting on the magnetization.

In what follows, we shall assume that the free electrons cannot instantly assume the equilibrium arrangements specified by the motion of the

[1] Galt, Yager and Merritt, Phys. Rev. **93**, No. 5, p. 1119, 1954.
[2] J. K. Galt, B.S.T.J., **33**, No. 5, p. 1023, 1954.
[3] H. P. J. Wijn and H. van der Heide. Rev. Mod. Phys. **80**, p. 744, 1950.

magnetization but tend toward equilibrium with a relaxation time τ. We shall show that this assumption results in energy being extracted from the motion of the magnetization and appearing as heat in the crystal.

The above mechanism has been analyzed by Galt in reference 2 in connection with the velocity of domain wall motion in nickel-iron ferrite. A similar question has been analyzed by L. Néel[4] in discussing the motion of domain walls in iron. In Néel's case the relaxation is associated with the migration of carbon atoms between various sites in the crystal. The theory used here is essentially similar to that applied by Néel. We shall try, however, to emphasize more than does Néel the connections with thermodynamics, and shall apply the theory not only to domain wall motion but also to ferromagnetic resonance. The theory is not in accord with that proposed by Galt and the points of disagreement will be mentioned below.

1. Kinetic Model

We shall consider the assembly of N free electrons per unit volume to constitute a thermodynamic system in heat contact with the crystal lattice. We suppose that there are m possible sites per unit volume of crystal for the free electrons to occupy; m is, of course, just the number of iron atoms per unit volume lying in octahedral sites. We shall let $\varepsilon_i(\vec{M})$ be the energy levels available to the electrons on the octahedral sites where $\vec{M}$ is the magnetization vector. We shall suppose that there are d_i levels per unit volume of energy ε_i and have therefore $\sum d_i = m$. We shall let N_i be the average number of electrons per unit volume in the level ε_i. If the magnetization is held steady in some direction, N_i will approach an equilibrium value that will be denoted by $N_{i\infty}(\vec{M})$. We shall assume a universal relaxation time τ such that N_i approaches equilibrium according to the equation

$$\frac{dN_i}{dt} = \frac{N_{i\infty} - N_i}{\tau} \tag{1-1}$$

and this equation is to hold whether or not $N_{i\infty}$ is a function of time. We observe that $\sum N_i = N$ and that therefore $\sum dN_i/dt = 0$ as must be the case.

$N_{i\infty}$ is obtained in the usual way by minimizing the free energy at constant temperature, and is given by

$$N_{i\infty} = \frac{d_i}{e^{\varepsilon_i - \varepsilon_f/kT} + 1} \tag{1-2}$$

[4] L. Néel, Théorie du Traînage Magnétique de Diffusion, Journal de Physique et Radium, **13**, p. 249, 1952.

In what follows, we shall assume that the percentage of substituted iron is small enough so that $d_i \gg N$ for each level. We have for the total number of electrons per unit volume

$$N = \sum \frac{d_i}{e^{\varepsilon_i - \varepsilon_f/kT} + 1} \tag{1-3}$$

It is clear that $e^{\varepsilon_i - \varepsilon_f/kT}$ must be much greater than unity for all i. Consequently, we may write

$$N_{i\infty} = N \frac{d_i e^{-\varepsilon_i/kT}}{\sum d_i e^{-\varepsilon_i/kT}} \tag{1-4}$$

We now define the internal energy, the entropy, and the free energy per unit volume as

$$U = \sum_i N_i \varepsilon_i \tag{1-5}$$

$$S = k[N \ln N - \sum_i N_i \ln N_i] \tag{1-6}$$

$$F = U - TS \tag{1-7}$$

The equilibrium value of these quantities are obviously given by

$$U_\infty = \sum_i N_{i\infty} \varepsilon_i \tag{1-8}$$

$$S_\infty = k[N \ln N - \sum_i N_{i\infty} \ln N_{i\infty}] \tag{1-9}$$

$$F_\infty = U_\infty - TS_\infty \tag{1-10}$$

If we take a time derivative of equation (1-5) we obtain

$$\frac{dU}{dt} = \sum_i N_i \frac{d\varepsilon_i}{dt} + \sum_i \varepsilon_i \frac{dN_i}{dt} \tag{1-11}$$

and in accordance with the first law, identify the rate of doing work on the system as

$$\frac{dW}{dt} = \sum_i N_i \frac{d\varepsilon_i}{dt} \tag{1-12}$$

while the rate at which heat flows into the system is

$$\frac{dQ}{dt} = \sum_i \varepsilon_i \frac{dN_i}{dt} \tag{1-13}$$

2. *Domain Wall Motion*

We shall first apply the kinetic model to a discussion of the velocity of a 180° domain wall moving perpendicularly to the (110) plane. In this

case the magnetization vector remains essentially in the (110) plane and its direction can be specified by the angle θ as in Fig. 1. We shall take the y-direction into the plane of the paper. Let us suppose that the wall is moving in the positive y direction, and that the magnetization turns through 180° from θ_1 to θ_2 as the wall sweeps by a given point.

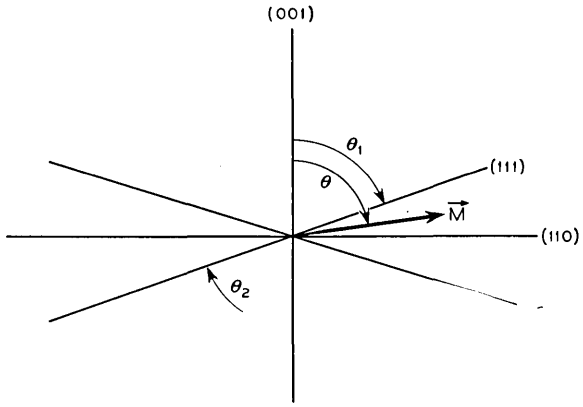


Fig. 1 — Angles used in describing motion of domain wall in (110) plane.

If we follow the procedure outlined by Kittel,[5] it is easy to show that the shape of a domain wall in the (110) plane for a crystal whose easy direction of magnetization is (111), is given by

$$A\left(\frac{d\theta}{dy}\right)^2 = g(\theta) - g(\theta_1) \tag{2-1}$$

where

$$g(\theta) = K[\tfrac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta] \tag{2-2}$$

K is the first order anisotropy constant (a negative number) and A is the usual exchange constant. We have $\sin\theta_1 = \sqrt{\tfrac{2}{3}}$, $\cos\theta_1 = \sqrt{\tfrac{1}{3}}$ so that $g(\theta_1) = \tfrac{1}{3}K$ and $g(\theta) - g(\theta_1) = (-K/12)(2 - 3\sin^2\theta)^2$. Equation (2-1) determines the wall except for its position along y. We have ignored in (2-2) a term depending on the magnetostriction which removes an ambiguity in the shape of the wall at the easy direction intermediate between θ_1 and θ_2. This term is unimportant in what follows.

[5] C. Kittel, Physical Theory of Ferromagnetic Domains, Rev. Mod. Physics **21**, p. 556, Oct., 1949.

We shall now assume that the wall moves along y with a constant velocity v and maintains the static shape given by (2-1). At each point in the wall we can then write,

$$\frac{d\theta}{dt} = -v\frac{d\theta}{dy} \tag{2-3}$$

The domain wall is, of course, moving as the magnetization turns to line up with an applied external field H_o. The rate at which this field does work on the wall per unit area is given by $2H_oMv$. This quantity must be set equal to the net rate at which work is done on our thermodynamic system, or, what is the same thing, the net rate at which heat flows into the crystal lattice. We may write therefore,

$$2H_oMv = \int_{-\infty}^{\infty} \frac{dW}{dt}\, dy \tag{2-4}$$

In order to compute dW/dt from (1-12) we suppose that the domain wall is moving slowly and that $N_{i\infty}$ is therefore a slowly changing function of time. Consequently we obtain from (1-1) an approximate expression for N_i

$$N_i = N_{i\infty} - \tau\frac{dN_{i\infty}}{dt} + \tau^2\frac{d^2N_{i\infty}}{dt^2} + - - -. \tag{2-5}$$

We obtain immediately from (1-12) to first order in τ,

$$\frac{dW}{dt} = \sum_i N_{i\infty}\frac{d\varepsilon_i}{dt} - \tau\sum_i \frac{dN_{i\infty}}{dt}\frac{d\varepsilon_i}{dt} \tag{2-6}$$

The first term in this expression can be written as dF_∞/dt, and contributes to the integral over y a term

$$-v\int_{\theta_1}^{\theta_2}\frac{dF_\infty}{d\theta}\, d\theta$$

We assume that F_∞ has cubic symmetry, and there is therefore no net contribution from this term.

The second term may be treated as follows. Proceeding from (1-4) we find by differentiation

$$\frac{dN_{i\infty}}{dt} = -\frac{1}{kT}N_{i\infty}\frac{d\varepsilon_i}{dt} + \frac{1}{kT}N_{i\infty}\frac{1}{N}\sum_i N_{i\infty}\frac{d\varepsilon_i}{dt} \tag{2-7}$$

Then we have

$$\sum \frac{dN_{i\infty}}{dt}\frac{d\varepsilon_i}{dt} = -\frac{1}{kT}\sum_i N_{i\infty}\left(\frac{d\varepsilon_i}{dt}\right)^2 + \frac{1}{NkT}\left(\sum_i N_{i\infty}\frac{d\varepsilon_i}{dt}\right)^2$$

$$= -\frac{N}{kT}\left[\sum_i \frac{N_{i\infty}}{N}\left(\frac{d\varepsilon_i}{dt}\right)^2 - \left(\sum_i \frac{N_{i\infty}}{N}\frac{d\varepsilon_i}{dt}\right)^2\right] \tag{2-8}$$

$$= -\frac{N}{kT}\sum_i \frac{N_{i\infty}}{N}\left[\frac{d\varepsilon_i}{dt} - \sum_i \frac{N_{i\infty}}{N}\frac{d\varepsilon_i}{dt}\right]^2$$

$$\sum_i \frac{dN_{i\infty}}{dt}\frac{d\varepsilon_i}{dt} = -\frac{N}{kT}\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2\right\rangle\left(\frac{d\theta}{dt}\right)^2 \tag{2-9}$$

where the angular brackets indicate a mean value over the equilibrium system. Upon substitution of (2-9) in (2-6) we find,

$$\frac{dW}{dt} = \frac{dF_\infty}{dt} + \tau\frac{N}{kT}\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2\right\rangle\left(\frac{d\theta}{dt}\right)^2 \tag{2-10}$$

By using the first law, we can write for the heat flow into the system[6]

$$\frac{dQ}{dt} = \frac{dU}{dt} - \frac{dF_\infty}{dt} - \tau\frac{N}{kT}\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2\right\rangle\left(\frac{d\theta}{dt}\right)^2 \tag{2-11}$$

It is clear that, in any process that takes the system between two identical thermodynamic states, the terms dU/dt and dF_∞/dt will integrate to zero. We must therefore regard the positive definite expression

$$\tau\frac{N}{kT}\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2\right\rangle$$

as an irreversible heat loss from the system, and it is just this heat loss that makes dW/dt greater than dF_∞/dt.

Before proceeding we wish here to make certain comparisons with the theory used by Galt in discussing domain wall motion in Reference 2.

[6] It is possible to formulate this problem in terms of a temperature T_κ such that

$$N_i = Ne^{-\varepsilon_i/kT_\kappa}/\Sigma\, e^{-\varepsilon_i/kT_\kappa}$$

One then obtains

$$\frac{dQ}{dt} = \frac{C_V}{\tau}(T_\kappa - T)$$

where

$$C_V = \frac{N}{kT^2}[\langle\varepsilon_i^2\rangle - \langle\varepsilon_i\rangle^2].$$

From (15) to (24) in Galt's paper one finds, to first order in τ,

$$\frac{dW}{dt} = \tau \frac{d^2 g_{i\infty}}{d\theta^2} \left(\frac{d\theta}{dt}\right)^2$$

which is to be compared with (2-10). This formulation seems somehow to neglect the equilibrium free energy of the free electrons, but since this term is indistinguishable from the ordinary crystalline anisotropy energy the omission is unimportant. The expression above, however must be interpreted as the irreversible heat loss and should therefore be a positive definite quantity as in (2-10). The quantity $(d^2 g_{i\infty}/d\theta^2)$, however, is not positive definite, and this would seem to be a serious objection to Galt's formulation of the theory. In Reference 2, $g_{i\infty}$ is set proportional to $g(\theta)$ defined in (2-2) so that $(d^2 g_{i\infty}/d\theta^2)$ in that case specifically has regions of positive and negative value, which should not be the case.

Using equation (2-10) we find

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{dW}{dt}\, dy &= \tau \frac{N}{kT} \int_{-\infty}^{\infty} \left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle \left(\frac{d\theta}{dt}\right)^2 dy \\ &= v^2 \tau \frac{N}{kT} \int_{\theta_1}^{\theta_2} \left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle \left(\frac{d\theta}{dy}\right) d\theta \end{aligned} \tag{2-12}$$

where $(d\theta/dy)$ is taken as positive by convention. This integral is the net rate at which work is done on the wall and is positive definite. The corresponding expression developed by Galt (30) is

$$\int_{-\infty}^{\infty} \frac{dW}{dt}\, dy = v^2 \tau \int_{\theta_1}^{\theta_2} \frac{d^2 g_{1\infty}}{d\theta^2} \left(\frac{d\theta}{dy}\right) d\theta$$

We can only regard a positive value for this integral as accidental and to depend upon the shape of the domain wall.

Combining (2-4), (2-12) and (2-1), we find for the velocity of domain wall motion

$$v = \frac{2H_0 M}{\tau \left(\frac{N}{kT}\right) \sqrt{\frac{-K}{12A}} \int_{\theta_1}^{\theta_2} \left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle |\, 2 - 3 \sin^2 \theta \,|\, d\theta} \tag{2-13}$$

This expression exhibits the characteristic dependence of v on $1/\tau$ discussed by Galt in reference 2 and shown there to account for the very marked dependence of v on temperature found in his experiments. It also contains the obvious result that v depends inversely on the number of free electrons.

It would be very desirable now to proceed to calculate the coefficient of $1/\tau$. This can only be done by introducing some physical ideas to arrive at an explicit dependence of ε_i on θ. This may be done according to two simple schemes that we shall consider later. First, however, we shall discuss the phenomenon of ferromagnetic resonance. One point that will soon become clear is that the present theory allows no direct comparison between domain wall motion experiments and ferromagnetic resonance experiments unless a particular model is adopted.

3. *Ferromagnetic Resonance*

We shall now apply the kinetic model to a calculation of the line width for ferromagnetic resonance and the corresponding displacement of the field for resonance. We suppose that the applied field and equilibrium direction of magnetization lie in the (110) plane at angles η and θ_0 as shown in Fig. 2. The direction z lies along $\vec{M}_0$. The direction x lies in the 110 plane perpendicular to z, while the y direction is perpendicular to the (110) plane in such a sense that x, y and z form a right-handed system.

We begin by writing down an expression for the total free energy of the system.

$$F_T = -MH \cos(\theta - \eta) \cos\Phi + G(\theta, \Phi) + \sum_i N_i \varepsilon_i - kT(N \ln N - \sum_i N_i \ln N_i) \quad (3\text{-}1)$$

where θ and Φ are the instantaneous directions of the magnetization.

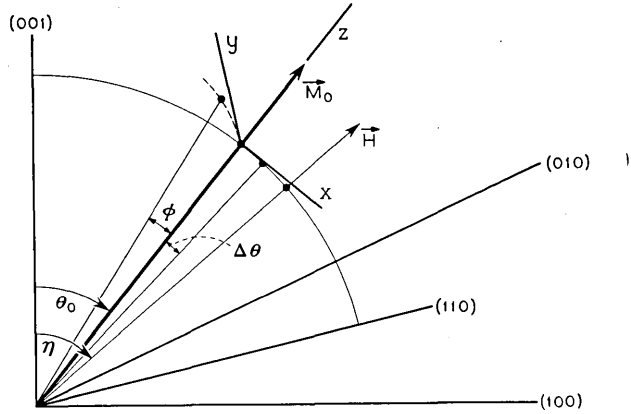


Fig. 2 — Coordinate systems used in discussing ferromagnetic resonance in the (110) plane.

It will be observed that θ and Φ are not the usual polar coördinates but are angles especially suitable to describing events occurring in the (110) plane. The first term in this expression is the energy of the magnetization in the external field. The second term includes the crystalline anisotropy energy, the shape anisotropy of the specimen and any energy due to magnetostriction. The third term is the free energy of the free electrons. We now wish to obtain the x and y components of the torque acting on the magnetization. These torques will be given by

$$T_y = -\frac{\partial F}{\partial \theta} \qquad T_x = \frac{\partial F}{\partial \Phi} \tag{3-2}$$

where the derivatives are taken with N_i held constant. This procedure is consistent with (1-12) for dW/dt.

We obtain for T_y, setting $\Phi = 0$

$$T_y = -MH \sin(\theta - \eta) - \frac{\partial G}{\partial \theta} - \sum_i N_i \frac{\partial \varepsilon_i}{\partial \theta} \tag{3-3}$$

Returning to (1-1), we can write generally for N_i, ignoring a transient term,

$$N_i = N_{i\infty} - e^{-t/\tau} \int^t e^{t/\tau} \frac{dN_{i\infty}}{dt}\, dt \tag{3-4}$$

If we insert (3-4) into (3-3) we obtain,

$$\begin{aligned} T_y = -MH \sin(\theta - \eta) - \frac{\partial G}{\partial \theta} - \sum_i N_{i\infty} \frac{\partial \varepsilon_i}{\partial \theta} \\ + \sum_i \frac{\partial \varepsilon_i}{\partial \theta} e^{-t/\tau} \int^t e^{t/\tau} \left(\frac{\partial N_{i\infty}}{\partial \theta}\right) \frac{d\theta}{dt}\, dt \end{aligned} \tag{3-5}$$

Let us expand this expression around the equilibrium angle θ_0 to obtain

$$\begin{aligned} T_y = -MH \sin(\theta_0 - \eta) - \left(\frac{\partial G}{\partial \theta}\right)_{\theta_0} - \sum_i N_{i\infty} \left(\frac{\partial \varepsilon_i}{\partial \theta}\right)_{\theta_0} \\ - \left(MH \cos(\theta_0 - \eta) + \left(\frac{\partial^2 G}{\partial \theta^2}\right)_{\theta_0} + \sum_i \left[\frac{\partial}{\partial \theta}\left(N_{i\infty} \frac{\partial \varepsilon_i}{\partial \theta}\right)\right]_{\theta_0}\right) \Delta\theta \\ + \sum_i \left(\frac{\partial \varepsilon_i}{\partial \theta}\right)_{\theta_0} \left(\frac{\partial N_{i\infty}}{\partial \theta}\right)_{\theta_0} e^{-t/\tau} \int^t e^{t/\tau} \left(\frac{d\theta}{dt}\right) dt \end{aligned} \tag{3-6}$$

which can be expressed in terms of the equilibrium free energy as

$$T_y = -MH \sin(\theta_0 - \eta) - \left(\frac{\partial G}{\partial \theta}\right)_{\theta_0} - \left(\frac{\partial F_\infty}{\partial \theta}\right)_{\theta_0} - \left(MH \cos(\theta_0 - \eta) + \left(\frac{\partial^2 G}{\partial \theta^2}\right)_{\theta_0} + \left(\frac{\partial^2 F_\infty}{\partial \theta^2}\right)_{\theta_0}\right)\Delta\theta + \sum_i \left(\frac{\partial \varepsilon_i}{\partial \theta}\right)_{\theta_0}\left(\frac{\partial N_{i\infty}}{\partial \theta_0}\right) e^{-t/\tau} \int^t e^{t/\tau}\left(\frac{d\theta}{dt}\right) dt \tag{3-7}$$

We choose θ_0 so that $MH \sin(\theta_0 - \eta) + (\partial G/\partial\theta)_{\theta_0} + (\partial F_\infty/\partial\theta)_{\theta_0} = 0$ and suppose that H is large enough so that θ_0 is substantially equal to η. We shall now drop subscripts and suppose that the field and equilibrium magnetization are in the direction θ. We further recognize that $\Delta\theta = M_x/M$. The expression for T_y then becomes,

$$T_y = -\left(MH + \frac{\partial^2 G}{\partial \theta^2} + \frac{\partial^2 F_\infty}{\partial \theta^2}\right)\frac{M_x}{M} + \sum_i \left(\frac{\partial \varepsilon_i}{\partial \theta}\right)\left(\frac{\partial N_{i\infty}}{\partial \theta}\right) e^{-t/\tau} \int^t e^{t/\tau} \frac{1}{M}\frac{dM_x}{dt} dt \tag{3-8}$$

An exactly similar expression is obtained for T_x

$$T_x = \left(MH + \frac{\partial^2 G}{\partial \Phi^2} + \frac{\partial^2 F_\infty}{\partial \Phi^2}\right)\frac{M_y}{M} - \sum_i \left(\frac{\partial \varepsilon_i}{\partial \Phi}\right)\left(\frac{\partial N_{i\infty}}{\partial \Phi}\right) e^{-t/\tau} \int^t e^{t/\tau} \frac{1}{M}\frac{dM_y}{dt} dt \tag{3-9}$$

The torques T_x and T_y may now be substituted into the equation of motion of the magnetization

$$\frac{d\vec{M}}{dt} = -\gamma \vec{T} - \gamma \vec{M} \times \vec{h} \tag{3-10}$$

where $\gamma = ge/2mc$ and $\vec{h}$ is the driving field. It will be observed that M_x and M_y will vary sinusoidally with frequency ω of the driving field for small amplitude motion. In that case, we can write

$$e^{-t/\tau} \int^t e^{t/\tau} \frac{dM_x}{dt} dt = \frac{\tau}{1 + (\omega\tau)^2}\frac{dM_x}{dt} + \frac{(\omega\tau)^2}{1 + (\omega\tau)^2} M_x \tag{3-11}$$

and similarly for M_y. It now becomes an elementary matter to compute the resonance line width and the field shift brought about by the relaxation process.

The total line width is given by,

$$\Delta H = -\frac{1}{M}\sum_i \left(\left(\frac{\partial \varepsilon_i}{\partial \theta}\right)\left(\frac{\partial N_{i\infty}}{\partial \theta}\right) + \left(\frac{\partial \varepsilon_i}{\partial \Phi}\right)\left(\frac{\partial N_{i\infty}}{\partial \Phi}\right)\right)\frac{\omega\tau}{1 + (\omega\tau)^2} \tag{3-12}$$

while the field required for resonance is increased by,

$$\begin{aligned}\delta H = &-\frac{1}{2M}\left(\frac{\partial^2 F_\infty}{\partial \theta^2} + \frac{\partial^2 F_\infty}{\partial \Phi^2}\right) \\ &+ \frac{1}{2M}\sum_i \left(\left(\frac{\partial \varepsilon_i}{\partial \theta}\right)\left(\frac{\partial N_{i\infty}}{\partial \theta}\right) + \left(\frac{\partial \varepsilon_i}{\partial \Phi}\right)\left(\frac{\partial N_{i\infty}}{\partial \Phi}\right)\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2}\end{aligned} \tag{3-13}$$

The expression

$$\sum_i \left(\left(\frac{\partial \varepsilon_i}{\partial \theta}\right)\left(\frac{\partial N_{i\infty}}{\partial \theta}\right) + \left(\frac{\partial \varepsilon_i}{\partial \Phi}\right)\left(\frac{\partial N_{i\infty}}{\partial \Phi}\right)\right)$$

is very similar to an expression that arose in discussing domain wall motion. We can treat this term exactly as was done in arriving at (2-9). We shall find that

$$\begin{aligned}&\sum_i \left(\left(\frac{\partial \varepsilon_i}{\partial \theta}\right)\left(\frac{\partial N_{i\infty}}{\partial \theta}\right) + \left(\frac{\partial \varepsilon_i}{\partial \Phi}\right)\left(\frac{\partial N_{i\infty}}{\partial \Phi}\right)\right) \\ &\qquad = -\frac{N}{kT}\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2 + \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle\frac{d\varepsilon_i}{d\Phi}\right\rangle\right]^2\right\rangle\end{aligned} \tag{3-14}$$

Let us first notice that ΔH and δH depend upon the quantity

$$\left\langle\left[\frac{d\varepsilon_i}{\partial \Phi} - \left\langle\frac{d\varepsilon_i}{d\Phi}\right\rangle\right]^2\right\rangle$$

which does not appear at all in (2-13) for domain wall velocity. Without making more explicit assumptions about ε_i, therefore, we cannot make any connection between the two experiments. The essential point here of course is that $\vec{M}$ is not constrained in the resonance experiment to move in the (110) plane as it is in the domain wall experiment.

Let us now make the reasonable assumption that the quantity

$$\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2 + \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle\frac{d\varepsilon_i}{d\Phi}\right\rangle\right]^2\right\rangle$$

varies in the (110) plane with cubic symmetry. We write therefore that

$$\begin{aligned}&\left\langle\left[\frac{d\varepsilon_i}{d\theta} - \left\langle\frac{d\varepsilon_i}{d\theta}\right\rangle\right]^2 + \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle\frac{d\varepsilon_i}{d\Phi}\right\rangle\right]^2\right\rangle \\ &\qquad = p + q[\tfrac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta]\end{aligned} \tag{3-15}$$

When we come to discuss specific models, we shall give particular expressions for p and q in terms of one unknown constant. We see that $p > 0$ and $p + q/3 > 0$ to maintain the right hand side of (3-15) positive. We observe that it is not possible to claim cubic symmetry for the terms

$$\left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle \right]^2 \right\rangle \quad \text{and} \quad \left\langle \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle \right]^2 \right\rangle$$

separately.

Using (3-15), the expressions for line width and line shift become,

$$\Delta H = \frac{1}{M} \frac{N}{kT} [p + q(\tfrac{1}{4} \sin^4 \theta + \sin^2 \theta \cos^2 \theta)] \frac{\omega\tau}{1 + (\omega\tau)^2} \tag{3-16}$$

$$\begin{aligned} \delta H = & -\frac{1}{2M} \left(\frac{\partial^2 F_\infty}{\partial \theta^2} + \frac{\partial^2 F_\infty}{\partial \Phi^2} \right) \\ & - \frac{1}{2M} \frac{N}{kT} [p + q(\tfrac{1}{4} \sin^4 \theta + \sin^2 \theta \cos^2 \theta)] \frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \end{aligned} \tag{3-17}$$

Equation (3-16) for the line width shows a dependence on $(\omega\tau)$ that is characteristic of relaxation processes. If a measurement is made of ΔH versus temperature at a given frequency and in a given crystallographic direction, a peak will be observed in the line width at very nearly the temperature where $\tau = 1/\omega$. The peak will not fall accurately at $\tau = 1/\omega$ because of the temperature dependence of the coefficient of $\omega\tau/1 + (\omega\tau)^2$ but comes very close to $1/\omega$ because of the exponential dependence of τ on temperature. Observations of this kind have been made by Galt, Yager and Merritt on nickel-iron ferrite and reported by them in Reference 1.

If we may now extend the measurements over a range of frequencies, and observe how the line shape varies as a function of frequency at constant temperature, ΔH will have its maximum accurately at $\omega = 1/\tau$ In principle, then, τ can be determined as a function of temperature.

If τ is known at a particular temperature, a measurement of ΔH versus θ at constant temperature and frequency will serve to determine $1/M\ N/kT\ p$ and $1/M\ N/kT\ q$ at the temperature in question.

Let us now consider the field shift δH given by (3-17). In the first place, this shift is not the ordinary displacement that accompanies the introduction of loss into any resonant system since such a field shift would be proportional to $(\Delta H)^2$. To the contrary, the field shift is of the same order of magnitude as the line width and we must consider seriously what effect this will have on the measurement of g-value. We have

assumed previously that F_∞ has cubic symmetry, and in that case, the term in (3-17) depending on F_∞ will be indistinguishable from the usual crystal anisotropy and need not be considered further. The second term we must consider more closely.

The following expression is an identity.

$$p + q[\tfrac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta] = \left(p + \frac{q}{5}\right) - \frac{q}{10}(2 - \tfrac{5}{2}\sin^2\theta - \tfrac{15}{8}\sin^2 2\theta) \tag{3-18}$$

Now, the second term on the right has the angular dependence in terms of which one usually expresses the first order anisotropy constant K_1. We may expect then that the measured anisotropy constant will have the form

$$K_1 = K_0 - \frac{N}{kT}\left(\frac{q}{20}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{3-19}$$

where K_0 is the static first order anisotropy constant.

We may look, therefore, for a dependence of K_1 on frequency and temperature like that given in (3-19) and make comparisons with the behavior of the line width.

If the total anisotropy field has been thus accounted for, we will be left with a line shift

$$-\frac{1}{2M}\frac{N}{kT}\left(p + \frac{q}{5}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{3-20}$$

which will appear as a change in the g-value. If a measurement of the field H for resonance is made for a given ω, T and θ, and if this field is corrected for the anisotropy field to give an effective field H_e, we will have

$$H_e = \frac{\omega}{\gamma} - \frac{1}{2M}\frac{N}{kT}\left(p + \frac{q}{5}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{3-21}$$

We shall measure an effective γ, however, given by

$$H_e = \frac{\omega}{\gamma_e} \tag{3-22}$$

Equating (3-21) and (3-22), we obtain

$$\frac{\gamma_e - \gamma}{\gamma} = \frac{1}{H_e}\frac{1}{2M}\frac{N}{kT}\left(p + \frac{q}{5}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{3-23}$$

or, if $\gamma_e = g_e e/2mc$ and $\Delta g = g_e - g$,

$$\frac{\Delta g}{g} = \frac{1}{2H_e M}\frac{N}{kT}\left(p + \frac{q}{5}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{3-24}$$

We, therefore, see that the g-factor will also have a dispersion in frequency and temperature, and the measurement of g-factor may indeed be the most sensitive way of observing the relaxation phenomenon.

4. *First Model*

There are two simple models we may adopt to obtain an explicit dependence of ε_i on θ and Φ. In the first case, let us suppose that there are four distinguishable energy levels in the crystal ε_1, ε_2, ε_3 and ε_4 and that each is associated with one of the body diagonals of the cubic crystal according to the relation

$$\varepsilon_i = w\cos^2\theta_i \tag{4-1}$$

where θ_i is the angle between the direction of magnetization and the i^{th} body diagonal and w is some constant depending on temperature.

This model is particularly tempting because there are four non-equivalent octahedral sites in the ferrite crystal, each associated with a given body diagonal. Expressed in terms of the direction cosines of the magnetization $(\alpha_1\alpha_2\alpha_3)$, we find

$$\begin{aligned}
\varepsilon_1 &= \frac{w}{3}[1 + 2(\alpha_1\alpha_2 + \alpha_2\alpha_3 + \alpha_3\alpha_1)] \\
\varepsilon_2 &= \frac{w}{3}[1 + 2(\alpha_1\alpha_2 - \alpha_2\alpha_3 - \alpha_3\alpha_1)] \\
\varepsilon_3 &= \frac{w}{3}[1 + 2(-\alpha_1\alpha_2 + \alpha_2\alpha_3 - \alpha_3\alpha_1)] \\
\varepsilon_4 &= \frac{w}{3}[1 + 2(-\alpha_1\alpha_2 - \alpha_2\alpha_3 + \alpha_3\alpha_1)]
\end{aligned} \tag{4-2}$$

while expressed in terms of the angles θ and Φ, the energies become,

$$\begin{aligned}
\varepsilon_1 &= \frac{w}{3}\left\{1 + [\sin^2\theta\cos^2\Phi - \sin^2\Phi + 2\sqrt{2}\sin\theta\cos\theta\cos^2\Phi]\right\} \\
\varepsilon_2 &= \frac{w}{3}\left\{1 + [\sin^2\theta\cos^2\Phi - \sin^2\Phi - 2\sqrt{2}\sin\theta\cos\theta\cos^2\Phi]\right\} \\
\varepsilon_3 &= \frac{w}{3}\left\{1 + [-\sin^2\theta\cos^2\Phi + \sin^2\Phi + 2\sqrt{2}\cos\theta\sin\Phi\cos\Phi]\right\} \\
\varepsilon_4 &= \frac{w}{3}\left\{1 + [-\sin^2\theta\cos^2\Phi + \sin^2\Phi - 2\sqrt{2}\cos\theta\sin\Phi\cos\Phi]\right\}
\end{aligned} \tag{4-3}$$

Let us consider now how we shall calculate the various mean values that will be required. If we consider a set of values g_i we have

$$\langle g_i \rangle = \frac{\sum_i g_i\, d_i e^{-\epsilon_i/kT}}{\sum_i d_i e^{-\epsilon_i/kT}} = \frac{\sum_i g_i\, d_i e^{-(\epsilon_i - \langle \epsilon_i \rangle)/kT}}{\sum_i d_i e^{-(\epsilon_i - \langle \epsilon_i \rangle)/kT}} \tag{4-4}$$

We shall suppose that T is large enough so that $(\varepsilon_i - \langle \varepsilon_i \rangle) \ll kT$, and assume that d_i is the same for each level. In that case, approximately

$$\langle g_i \rangle = \frac{1}{k} \sum_{n=1}^{k} g_n \tag{4-5}$$

if there are k levels. We shall show shortly that this assumption is consistent with the experimental data.

Using (4-5) we can calculate for our first model the following mean values,

$$\left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle = \left(\frac{w}{3}\right) \tag{4-6}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle = \left(\frac{w}{3}\right)^2 [1 + 3\cos^2 2\theta] \tag{4-7}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle\right]^2 \right\rangle = \left(\frac{w}{3}\right)^2 4\cos^2\theta \tag{4-8}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 + \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle\right]^2 \right\rangle = \left(\frac{w}{3}\right)^2 [1 + 3\cos^2 2\theta + 4\cos^2\theta] \tag{4-9}$$

$$= \tfrac{8}{9} w^2 [1 - 2(\tfrac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta)] \tag{4-10}$$

Referring back to (2-13), the velocity of domain wall motion becomes

$$v = \frac{2H_0 M}{\tau \left(\frac{N}{kT}\right) \sqrt{\frac{-K}{12A}} \int_{\theta_1}^{\theta_2} \left(\frac{w}{3}\right)^2 [1 + 3\cos^2 2\theta]\,|\,2 - 3\sin^2\theta\,|\,d\theta} = \frac{2H_0 M}{\tau \sqrt{\frac{-K}{12A}} \left(\frac{N}{kT}\right) w^2 (1.02)} \tag{4-11}$$

Comparing with (3-15) we find $p = \frac{8}{9} w^2$ and $q = -\frac{16}{9} w^2$. The line width for resonance becomes

$$\Delta H = \frac{1}{M} \frac{8}{9} \left(\frac{N}{kT}\right) w^2 \left[1 - 2\left(\frac{1}{4} \sin^4 \theta + \sin^2 \theta \cos^2 \theta\right)\right] \frac{\omega\tau}{1 + (\omega\tau)^2} \tag{4-12}$$

The anisotropy constant is given by

$$K_1 = K_0 + \frac{4}{45} \left(\frac{N}{kT}\right) w^2 \frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{4-13}$$

and the shift in g-factor by

$$\frac{\Delta g}{g} = \frac{1}{H_e M} \frac{4}{15} \left(\frac{N}{kT}\right) w^2 \frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{4-14}$$

We thus have four different measurements which we may try to fit with the same value of $N/kT\ w^2$. It should be noted that (4-12) predicts a maximum line width in the (100) direction and a minimum in the (111) direction.

5. *Second Model*

For the second model, we shall assume that there are three distinguishable energy levels ε_1 ε_2 and ε_3 each associated with one of the cube edges of the crystal according to the relation

$$\varepsilon_i = w \cos^2 \theta_i \tag{5-1}$$

where θ_i is the angle between the direction of magnetization and the ith cube edge. Expressed in terms of the direction cosines of $\vec{M}$ we have

$$\begin{aligned} \varepsilon_1 &= w\, \alpha_1^2 \\ \varepsilon_2 &= w\, \alpha_2^2 \\ \varepsilon_3 &= w\, \alpha_3^2 \end{aligned} \tag{5-2}$$

while expressed in terms of the angles θ and Φ the energies become,

$$\begin{aligned} \varepsilon_1 &= \frac{w}{2} (\sin\theta \cos\Phi - \sin\Phi)^2 \\ \varepsilon_2 &= \frac{w}{2} (\sin\theta \cos\Phi + \sin\Phi)^2 \\ \varepsilon_3 &= w (\cos\theta \cos\Phi)^2 \end{aligned} \tag{5-3}$$

This model we now consider is the same as the one used by Néel in Reference 3 to discuss the relaxation of interstitial atoms in iron. Proceeding as before we find,

$$\left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle = \left(\frac{w}{3}\right) \tag{5-4}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle = 2w^2 \sin^2\theta \cos^2\theta \tag{5-5}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle\right]^2 \right\rangle = \frac{2}{3} w^2 \sin^2\theta \tag{5-6}$$

$$\left\langle \left[\frac{d\varepsilon_i}{d\theta} - \left\langle \frac{d\varepsilon_i}{d\theta} \right\rangle\right]^2 \right\rangle + \left\langle \left[\frac{d\varepsilon_i}{d\Phi} - \left\langle \frac{d\varepsilon_i}{d\Phi} \right\rangle\right]^2 \right\rangle = \frac{8}{3} w^2 \left(\frac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta\right) \tag{5-7}$$

In this case we obtain,

$$v = \frac{2H_0M}{\tau\left(\frac{N}{kT}\right)\sqrt{\frac{-K}{12A}}\int_{\theta_1}^{\theta_2} 2w^2 \sin^2\theta\cos^2\theta \mid 2 - 3\sin^2\theta \mid d\theta} = \frac{2H_0M}{\tau\sqrt{\frac{-K}{12A}}\left(\frac{N}{kT}\right)w^2\,(.583)} \tag{5-8}$$

and by comparison with (3-15) $p = 0$ and $q = \frac{8}{3} w^2$. The line width for resonance is

$$\Delta H = \frac{1}{M}\frac{8}{3}\left(\frac{N}{kT}\right)w^2\left(\frac{1}{4}\sin^4\theta + \sin^2\theta\cos^2\theta\right)\frac{\omega\tau}{1 + (\omega\tau)^2} \tag{5-9}$$

The anisotropy constant is

$$K_1 = K_0 - \frac{2}{15}\left(\frac{N}{kT}\right)w^2\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{5-10}$$

The effective field at resonance is

$$H_e = \frac{\omega}{\gamma} - \frac{4}{15}\left(\frac{NW^2}{MkT}\right)\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{5-11}$$

and the g-factor shift is

$$\frac{\Delta g}{g} = \frac{1}{H_eM}\left(\frac{4}{15}\right)\left(\frac{N}{kT}\right)w^2\frac{(\omega\tau)^2}{1 + (\omega\tau)^2} \tag{5-12}$$

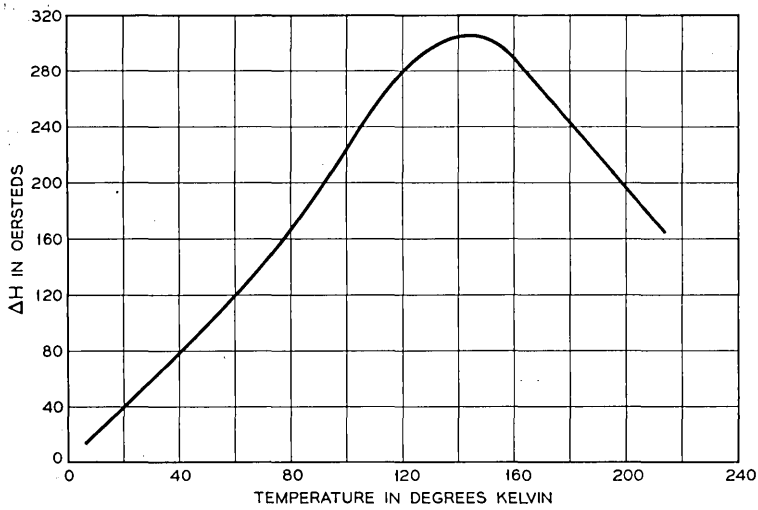


Fig. 3 — Corrected line width in the (111) direction versus temperature for $Ni_{0.75}Fe_{2.25}O_4$ after Yager, Galt and Merritt.

In this case the line width due to relaxation is zero in the (100) direction and a maximum in the (111) direction.

6. *Comparison with Experiment*

We wish now to make some comparisons between the theory developed in the preceding sections and measurements upon nickel-iron ferrite of composition $Ni_{0.75}Fe_{2.25}O_4$ made by Yager, Galt and Merritt.[7] The data we shall consider were made upon a .009″ sphere and are presented in Fig. 4 and Table II of Reference 7. These data include a measurement of line width versus temperature for the (111) direction and (100) direction in the crystal; a measurement of the first order anisotropy constant (as K_1/M) versus temperature; and a measurement of the effective field at resonance, H_e, versus temperature. It is observed from these data that the line width is maximum in the (111) direction and we shall therefore try to fit the data with the second model as discussed in Section 5. This model requires the line width to be zero in the (100) direction. In fitting the data, we shall assume that the line width in the (100) direction represents non-angle-dependent contributions from other mechanisms and shall subtract this width at each temperature from the width in the (111) direction. The resulting curve is shown in

[7] Yager, Galt and Merritt, Phys. Rev., to be published.

Fig. 3. The data for K_1/M and H_e are presented in Table I. The effective fields H_e have been corrected in each case to a common frequency 24,388 mc/sec.

The equations we need are contained in Section 5 and are repeated here for convenience,

$$\Delta H = \frac{8}{9}\left(\frac{Nw^2}{MkT}\right)\frac{\omega\tau}{1+(\omega\tau)^2} \tag{6-1}$$

$$H_e = \frac{\omega}{\gamma} - \frac{4}{15}\left(\frac{Nw^2}{MkT}\right)\frac{(\omega\tau)^2}{1+(\omega\tau)^2} \tag{6-2}$$

$$\frac{-K_1}{M} = \left(\frac{-K_0}{M}\right) + \frac{2}{15}\left(\frac{Nw^2}{MkT}\right)\frac{(\omega\tau)^2}{1+(\omega\tau)^2} \tag{6-3}$$

To proceed, we must make some assumptions about how w depends upon temperature, and shall therefore suppose that (Nw^2/MkT) is a constant independent of temperature. If we assumed instead, for instance, that w was a constant, the curves to be calculated would differ considerably at low temperatures but not enough to affect the conclusions to be drawn.

Proceeding on this assumption, we conclude from Fig. 3 that $\omega\tau = 1$ at $T = 145°$ and calculate that $(Nw^2/MkT) = 690$. We may now use Fig. 3 and (6-1) to calculate $\omega\tau$ at each temperature. With this we may calculate H_e at each temperature, and the resulting curve is shown in Fig. 4. Since we have no independent value for γ, the curve has been fitted to the data at 160°K. It is seen that the predicted changes in H_e are in the correct direction and are of approximately the right magnitude. The experimental points are not fitted in detail, however. The fit above 85°K could be somewhat improved by assuming a different dependence of w on T. The theory as presently developed cannot explain the in-

TABLE I — ANISOTROPY FIELD $-K_1/M$ AND EFFECTIVE FIELD H_e FOR VARIOUS VALUES OF ABSOLUTE TEMPERATURE AFTER YAGER, GALT AND MERRITT.

$T°K$	$-K_1/M$ (oe)	H_e (oe)
300	152	8195
159	218	8070
135	238	
85	259	7953
4.2	421	8047

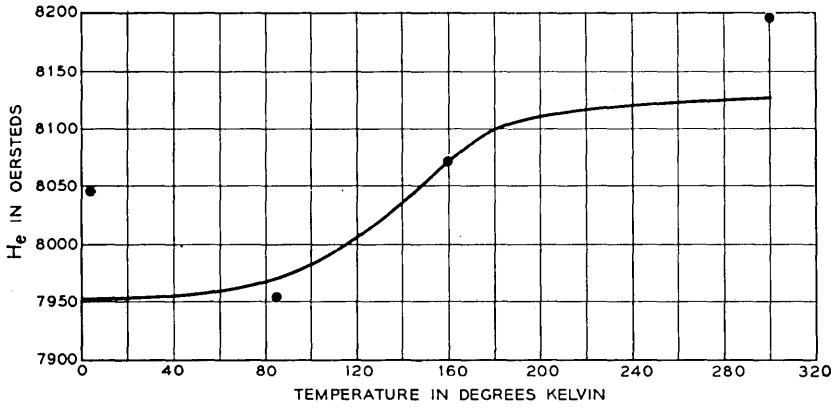


Fig. 4 — Measured and calculated values of effective field H_e as a function of absolute temperature.

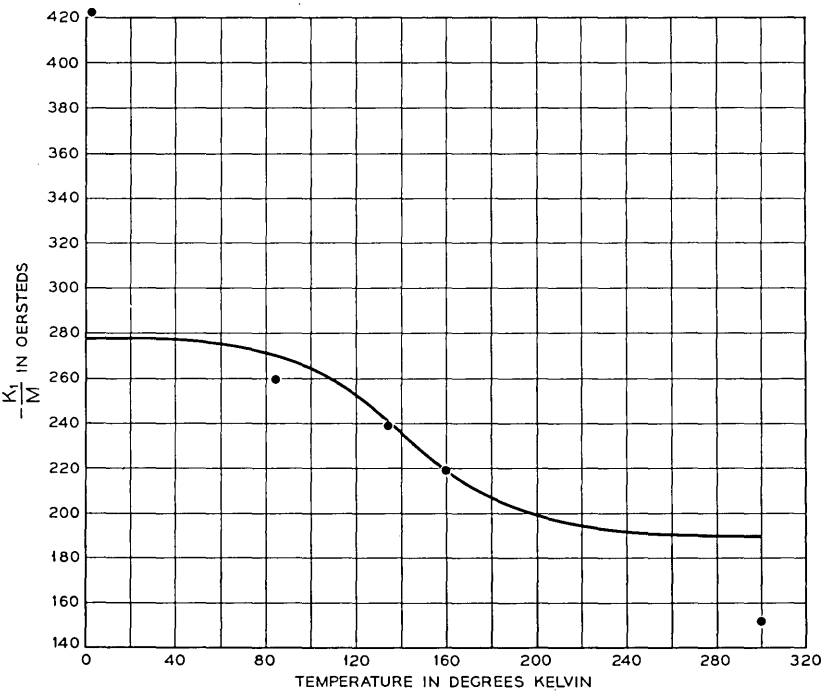


Fig. 5 — Measured and calculated values of anisotropy field $-K_1/M$ as a function of absolute temperature.

crease of H_e at 4.2°K. Yager, Galt, and Merritt believe that the changes observed between 85°K and 4.2°K must be ascribed to some other mechanism.

In arriving at (6-1) to (6-3) we have assumed as in Section 4 that $w/kT \ll 1$. At a temperature of 150°K, kT equal 2.07×10^{-14} ergs. The molecular weight M of nickel ferrite is 234.39 and the density ρ is 5.37 gms/cc. There are then $\rho L/M = 1.38 \times 10^{22}$ molecules per cubic cm, where $L = 6.025 \times 10^{23}$ is Avogadro's number. In nickel ferrite of composition $Ni_{0.75}Fe_{2.25}O_4$ one divalent iron is introduced for each four molecules, and the number of free electrons N is therefore 3.45×10^{21} per cc. The magnetization of this ferrite is about 340 cgs units. We have determined above that $(Nw^2/MkT) = 690$ and can therefore calculate that $w/kT = 0.057$ at $T = 150$°K. We have thus proceeded consistently in assuming $w \ll kT$. If w is constant, our calculation would not however be correct below about 50°K or if w^2/kT is constant, below about 10°K. It would be entirely possible to make accurate calculations for the entire temperature range and, by assuming special variations of w with temperature, perhaps account for some of the rise in H_e at 4.2°K.

We must also recognize that the ratio m/N of levels to number of free electrons in this composition is 5. Since we have assumed the levels

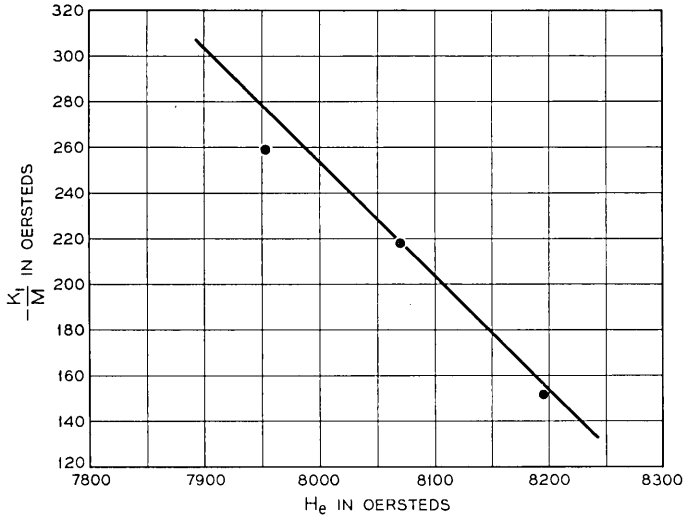


Fig. 6 — The theoretical relation between anisotropy field $-K_1/M$ and effective field H_e compared with experimental points.

divided into three equal groups, $d_i = \frac{5}{3}N$ so that we are hardly justified in considering the electrons to be non-degenerate. We should then really use (1-2) instead of (1-4) in computing our averages. Furthermore, when degeneracy effects are important, the relaxation equation (1-1) should be replaced by a more sophisticated expression. Since the general effect of degeneracy is to reduce the number of electrons taking part in the relaxation process, particularly at low temperatures, a calculation considering degeneracy has a good chance of accounting for the rise in H_e at liquid helium temperatures.

In a ferrite of composition similar to that considered here, Bozorth, Cetlin, Galt, Merritt and Yager[8] report measurements of the static anisotropy which show K_1/M to be nearly constant between 300°K and 196°K. If we make the assumption that the static anisotropy varies very little down to 85°K, we may use (6-3) to draw the curve in Fig. 5. It will be observed that the predicted changes in K_1/M are again in the right direction and of approximately the right size although the detailed fit is poor.

We may make one further comparison with the data. By eliminating between (6-2) and (6-3) one obtains

$$\left(-\frac{K_1}{M}\right) = \left(-\frac{K_0}{M}\right) + \frac{1}{2}\left(\frac{\omega}{\gamma} - H_e\right) \tag{6-4}$$

With the same assumption about K_0, a plot of $(-K_1/M)$ versus H_e should be a straight line with slope $-\frac{1}{2}$. The data has been plotted in Fig. 6 and a straight line of the appropriate slope drawn through the point at 160°K. Although only three experimental points are available, the fit is very satisfactory.

The author would like to acknowledge many helpful discussions with J. K. Galt, P. W. Anderson, H. Suhl and L. R. Walker.

[8] Bozorth, Cetlin, Galt, Merritt and Yager, Physical Review, to be published.

Design of Alloyed Junction Germanium Transistors for High-Speed Switching

By J. J. EBERS and S. L. MILLER

(Manuscript received March 25, 1955)

Alloyed-junction transistors have been critically evaluated with respect to their potential use as high-speed switching devices. Limitations are found to result from avalanche multiplication and space charge layer widening which necessitate design compromises between high frequency and high voltage circuit requirements. It is shown in this article that n-p-n germanium transistors have an advantage over p-n-p type beyond that expected from mobility considerations. The additional advantage derives from differences in avalanche multiplication.

1.0 INTRODUCTION

The present trend in switching systems and computers is toward systems that operate at megacycle rates with pulse rise times in the vicinity of a tenth of a microsecond. In the attendant design of pulse-amplifiers, pulse-generators, and flip-flops there has arisen the need for high-speed junction transistors, particularly in cases where the device must switch high currents for extended periods of time. Roughly, devices are needed with alphas greater than 0.95 and frequency cut-offs of alpha in the range of 10 to 50 mc per second. In considering transistor structures for high pulse rate systems there are several contenders for further development, for example, grown-junction triodes and tetrodes, alloyed-junction triodes, and intrinsic-barrier transistors.[1] A comparison has to be made of the design possibilities and the relative merits of these devices when used as switches. The objective of this paper is to determine the applicability and limitations of alloyed-junction transistors in high-speed switching applications. Many of the considerations in the design of alloyed-junction transistors for switching are also important in large-signal transmission applications.

Alloyed junctions transistors have been extensively used in switching systems having a pulse rate below a megacycle per second with excellent

results which can be attributed to their low impedance in the ON condition when used as a switch, to their high OFF impedance and to the relatively small storage of minority carriers compared to grown junction transistors. In extending the operating range toward higher speeds, limitations are found to result from:

1. punch-through, the spreading of the collector junction space-charge-layer over to the emitter junction,[2]
2. avalanche multiplication of alpha which results in alphas greater than unity at collector voltages well below the theoretical breakdown voltage,[3, 4, 5, 6]
3. collector capacitance which in combination with the load resistance may severely limit pulse rise time, and
4. the practical limitation of reproducibly fabricating transistors with extremely narrow base layers. With increased understanding of the properties of junctions and transistors, it has become possible to determine whether a given set of circuit requirements can be met, even theoretically, by a proposed design. For example, the requirements may be that the device operate up to a certain voltage, that at this voltage the alpha remain below unity and the collector space charge region not punch through to the emitter, and that the frequency response be high enough for the application.

The maximum permissible base layer width is essentially fixed by the specified frequency cutoff. With this base width, the requirement that there be no punch-through at the highest operating voltage puts an upper limit on base resistivity, since the encroachment into the base layer of the collector space charge region is greater the higher the base resistivity. On the other hand, the requirement that the low-voltage alpha times the avalanche multiplication of the collector junction be less than unity, sets a lower limit on permissible collector body breakdown voltage. This limit, in turn, sets a lower limit on permissible base resistivity, since for an alloyed structure the collector breakdown voltage is nearly a linear function of base resistivity. It is only for base resistivities between these two limits, if such a region exists, that designs are possible.

Since in germanium holes produce greater avalanche multiplication than electrons in a given electric field,[5] there exists under the above condition a greater difference in the maximum obtainable frequency cutoff of alpha between npn and pnp transistors than would be expected merely on a basis of the difference in mobility between holes and electrons.

The most general transistor design considerations for a switch involve

the OFF impedance, the ON impedance, the switching time, the current carrying ability, and the maximum open circuit voltage. As implied above, important interrelations between these switch parameters exist which limit the range of possible design. Furthermore, in order to carry through a transistor switch design it is necessary to know how the switch parameters are related to the terminal behavior of the transistor. The large-signal behavior of junction transistors has been discussed in two previous papers, and it has been found possible to relate the large-signal behavior to easily measurable transistor parameters which are in turn intimately related to the physics and structure of the transistor.[7]

The plan of this paper will be: (1) to review the switch requirements and to specify them in terms of transistor device parameters, (2) to show how the device parameters are related to the structure, and (3) to indicate the method of obtaining a design to meet a given set of switch requirements or to determine whether or not such a design is possible. The discussion will be primarily in terms of a common-emitter pulse amplifier since this circuit has proven to be the one which is most widely used and since it exemplifies the problems encountered in transistor design for switching applications.

2.0 SWITCH REQUIREMENTS

2.1 *OFF Impedance*

In the common-emitter pulse amplifier circuit (see Fig. 1) means are usually supplied to reverse bias both the collector and emitter junctions in the absence of a driving pulse. Under these conditions the collector current is given by[7]

$$I_C = \frac{I_{CO}(1 - \alpha_I)}{1 - \alpha_N \alpha_I} \tag{1}$$

where α_N and α_I are the normal and inverted alphas, respectively. In order to make this current as small as possible it may be desirable to have an alloyed-junction transistor with an emitter junction which is larger in diameter than the collector, thus having α_I greater than α_N. In most cases this is not feasible because α_N must be large for other reasons. About the best that can be done is to use a symmetrical structure with $\alpha_N = \alpha_I$. The collector current in the OFF condition is then given by

$$I_C = \frac{I_{CO}}{1 + \alpha_N} \tag{2}$$

which for high alpha transistors approaches $\frac{1}{2}I_{co}$, corresponding to an OFF impedance of $2Vc/I_{co}$. The value of I_{co} increases to some extent with collector voltage due to avalanche multiplication. In many cases there is also a surface leakage current which is not subject to analysis.

In some applications the ac impedance of the switch in the OFF condition is of great importance. It has been found possible to fabricate junction devices in which the slope of the reverse junction characteristic is consistantly greater than 25 megohms. In this event the junction capacitance places an upper limit on the OFF impedance.

2.2 *ON Impedance*

The voltage drop across the switch in the closed condition (common-emitter connection) is given by[7]

$$V_{CE} = (\pm)\frac{kT}{q}\ell n\frac{\alpha_I\left(1-\frac{I_{C1}}{I_{B1}}\left(\frac{1-\alpha_N}{\alpha_N}\right)\right)}{1+\frac{I_{C1}}{I_{B1}}(1-\alpha_I)} \tag{3}$$

where $kT/q = 0.026$ volts at room temperature and the $(+)$ sign applies to pnp transistors and the $(-)$ sign to npn transistors.* (For current definitions see Fig. 1.) The extremely low voltages which result when the device is driven into the current saturation region are attributed to the fact that the collector junction attains a forward bias almost equal to that of the emitter junction. In order to obtain a low value of V_{CE} it is necessary to have both α_N and α_I large, approaching unity, and to maintain these values at high currents. Experience has shown that the best way to accomplish this is to use a symmetrical structure with a thin base layer.

An expression for the ac impedance of a common-emitter switch has also been derived.[7] The expression shows that if this is the switch parameter of primary importance in a given application, it may be necessary to drive the switch well into the current saturation region in addition to using a device with high α_N and α_I.

2.3. *Current Carrying Ability*

The power dissipation in a transistor switch is easily calculable provided the collector current and voltage waveforms are known. In many cases most of the power will be dissipated during the transition from

* This expression is not applicable to grown-junction transistors which may have 10 to 200 ohms of collector body resistance.

OFF to ON or vice versa. Frequently, however, the switch is maintained in the closed condition for extended periods of time. During these intervals the power dissipation is given approximately by

$$P_{\text{dis}} = I_C V_{CE} + I_B^2 r_{B3} \tag{4}$$

where V_{CE} has a value given by (3) and r_{B3} is the base resistance in the current saturation region. For large collector currents, r_{B3} may be only a fraction of the base resistance in the active region. It is apparent that the transistor should be driven well into the saturation region to obtain a low value of V_{CE}. The current carrying ability is determined primarily by the allowable power dissipation.

2.4 *Maximum Open Circuit Voltage*

The maximum open circuit switch voltage exerts a controlling influence on the choice of geometry and material constants. If sufficient collector to base voltage is applied to extend the space charge layer of the collector junction through the entire thickness of the base layer to the emitter junction "punch-through"[2] results and the emitter and collector junctions behave as if shorted together through a battery. In alloyed devices the punch through voltage is determined by the resistivity of the base material and the shortest base distance between emitter and collector.

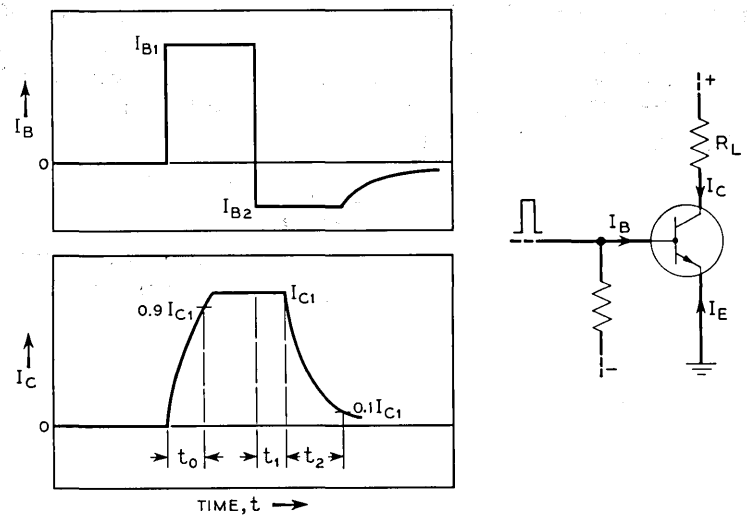


Fig. 1 — Transient behavior of common-emitter pulse amplifier.

Avalanche multiplication[3, 4] of the emitter current, which is also a function of the collector voltage, places an upper limit on the low voltage value of alpha, since at some collector voltage the value of alpha must necessarily become equal to unity[5, 6] (whether due to body or surface avalanche breakdown). This may result in improper operation* in the case of common-emitter circuits and other circuits not designed for $\alpha > 1$. The voltage at which alpha becomes one is also the maximum voltage which can be sustained from collector to emitter with the base open circuited. Avalanche multiplication is related to the body breakdown voltage which in turn is directly related to the base layer resistivity in the case of alloyed-junction transistors.

2.5 *Switching Speed*

Rise time in transistor switching circuits has been analyzed by J. L. Moll.[8] This analysis has been found to give sufficiently accurate results provided the time constant associated with the collector capacitance and load resistance is small compared with the time constant associated with the frequency cut-off of alpha, i.e., $R_L C_C < 1/\omega_N$. Rise time in the common-emitter circuit is then given by

$$t_0 = \frac{1}{(1 - \alpha_N)\omega_N} \ell n \frac{1}{1 - 0.9 \dfrac{1 - \alpha_N}{\alpha_N} \dfrac{I_{C1}}{I_{B1}}} \tag{5}$$

where t_0 is the time required for the collector current to rise to 0.9 of its final value and ω_N is the angular cut-off frequency of the normal alpha. The quantity $t_0\omega_N$ is plotted in Fig. 2 as a function of α_N for various values of I_B/I_C. For sufficiently high values of alpha and sufficiently high values of I_B/I_C, the rise time is given by

$$t_0 = 0.9 \frac{1}{\omega_N \alpha_N} \frac{I_C}{I_B} \tag{6}$$

As an example, if an amplifier is to have a current gain of twenty and a rise time of a tenth of a microsecond, $f_N = \omega_N/2\pi$ must be about 30 mc./sec.

If a common-emitter pulse amplifier is driven by essentially a current source, the base resistance in the active region does not limit the response time.

* It is possible to operate transistors with alpha greater than unity in common-emitter pulse amplifier circuits. However under some driving conditions the circuit may be bistable and difficulty will be encountered in turn-off. In general, large currents may be needed for turn-off and the pulse source should have low impedance.

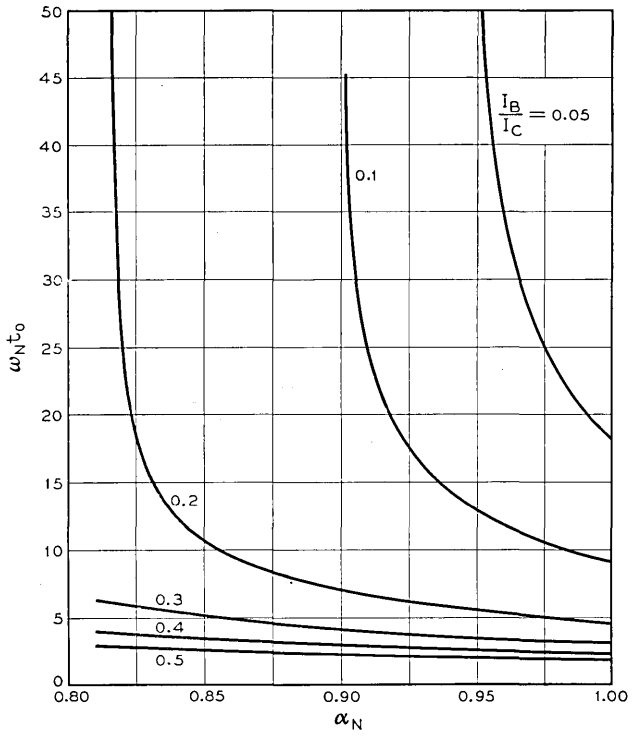


Fig. 2 — Normalized rise time as a function of alpha for common-emitter pulse amplifier.

At the termination of the input pulse the output current remains relatively constant for a certain period of time because of minority carrier storage. During this interval (t_1 in Fig. 1) the stored minority charge in the base layer is flowing out at a rate limited by the circuit resistance. Not until the minority carrier density at the base side of the collector junction has been reduced to zero by this outflow does the decay of circuit current commence (interval t_2 in Fig. 1). The storage time has been shown by Moll to be[8]

$$t_1 = \frac{\omega_N + \omega_I}{\omega_N \omega_I (1 - \alpha_N \alpha_I)} \ell n \frac{I_{B1} - I_{B2}}{\frac{1 - \alpha_N}{\alpha_N} I_{C1} - I_{B2}} \tag{7}$$

This time is small if the values of normal and inverted alpha frequency cut-off are both large. The expeditious way of accomplishing this in an

alloyed structure is to make the structure symmetrical, in which case the expression for storage time becomes

$$t_1 = \frac{2}{\omega_N (1 - \alpha_N^2)} \ln \frac{I_{B1} - I_{B2}}{\frac{1 - \alpha_N}{\alpha_N} I_{C1} - I_{B2}} \tag{8}$$

Storage time is plotted in Fig. 3 as a function of alpha for a given set of current conditions for a symmetrical transistor.

Decay time is the interval of time between the end of the storage interval and the point when the load current reaches 10 per cent of its value before turn-off. It is given by

$$t_2 = \frac{1}{(1 - \alpha_N)\omega_N} \ln \frac{I_{C1} - \frac{\alpha_N}{1 - \alpha_N} I_{B2}}{0.1 I_{C1} - \frac{\alpha_N}{1 - \alpha_N} I_{B2}} \tag{9}$$

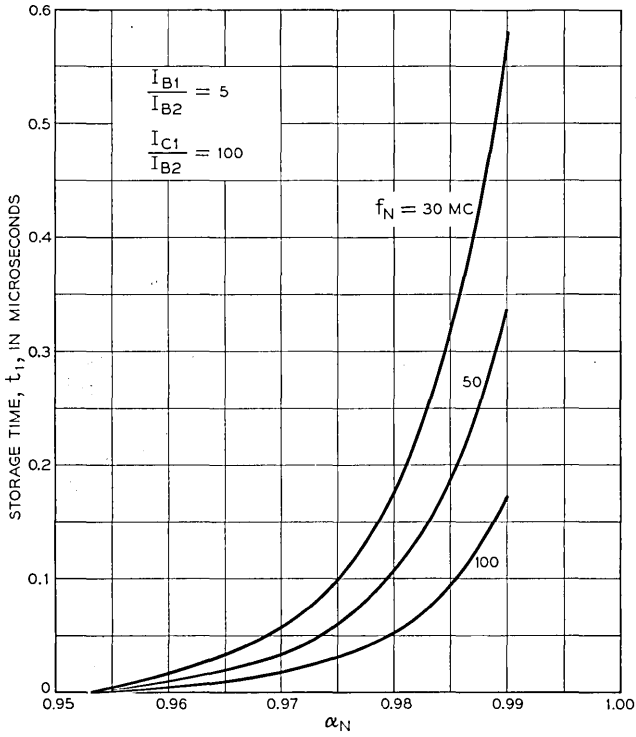


Fig. 3 — Storage time as a function of alpha for common-emitter pulse amplifier.

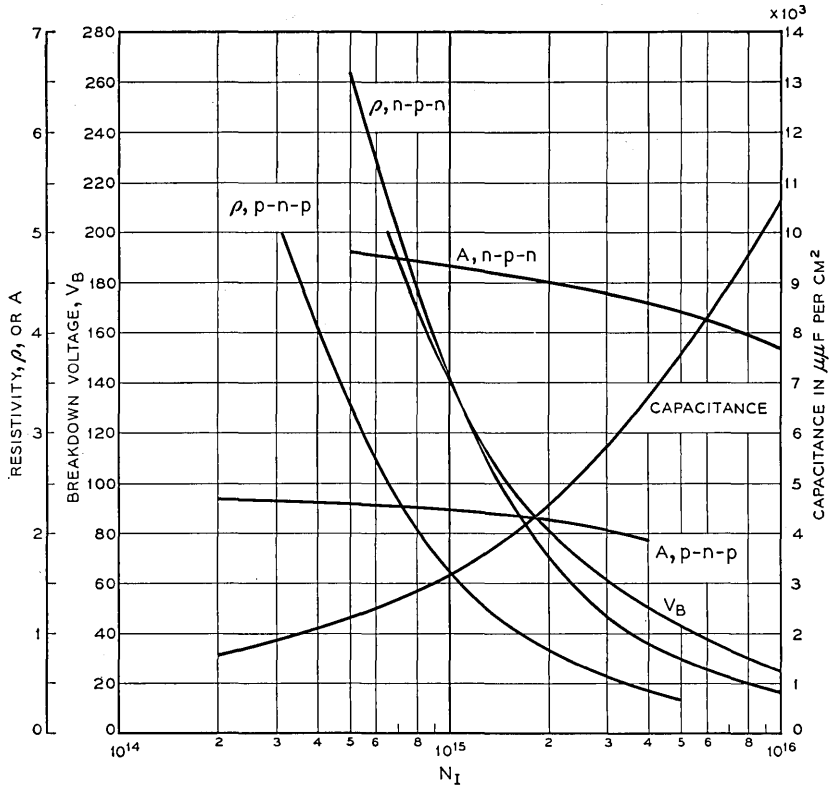


Fig. 4 — Resistivity, capacitance (at 10 *V.*) per unit area, breakdown voltage, and frequency cut-off factors versus density of impurity centers for germanium.

The time constant associated with the decay time is seen to be the same as that associated with the turn-on time. The decay time is usually somewhat larger than the rise time unless relatively large values of base current are used during turn-off.

3.0 DEVICE DESIGN

3.1 *Junction Capacitance*

Alloyed junctions are usually abrupt or step junctions, even though the actual junction may be formed by the diffusion of donors or acceptors ahead of the surface of deepest penetration of the alloying material. For

step junctions of this type the junction capacitance is given by[9]

$$C = \left(\frac{\kappa\varepsilon_0 q}{2}\right)^{1/2}\left(\frac{N_I}{V}\right)^{1/2} \tag{10}$$

or, for germanium,

$$C = 3.4 \times 10^{-4}\left(\frac{N_I}{V}\right)^{1/2} \quad \mu\mu f/cm^2 \tag{11}$$

where V is the total reverse junction bias in volts and $N_I = |N_D - N_A|$ is the net density of donors or acceptors in the base region. Capacitance per unit area is plotted in Fig. 4 as a function of N_I assuming $V = 10$ volts.

3.2 *Space Charge Layer Width*

Since the emitter and collector regions of alloyed transistors are heavily doped with donor or acceptor atoms, their conductivities are high and the space charge layer associated with the junction capacitance extends primarily into the base layer. The thickness of this space charge layer is given by[9]

$$W_S = \left(\frac{2\kappa\varepsilon_0}{q}\right)^{1/2}\left(\frac{V}{N_I}\right)^{1/2} \tag{12}$$

or, for germanium,

$$W_S = 4.2 \times 10^{3}\left(\frac{V}{N_I}\right)^{1/2} \quad \text{cm} \tag{13}$$

The widening of the space charge layer with voltage limits the resistivity and width of the base region assuming the operating collector voltage is specified. If the base layer is sufficiently thin and of high resistivity, the space charge layer may extend over to the emitter at collector voltages below the collector breakdown voltage. The voltage at which this occurs, the punch-through voltage, is the same whether it be applied collector to base, collector to emitter, or emitter to collector. In Fig. 5 the solid curves are contours of constant space charge layer width plotted with junction voltage and base layer resistivity as the independent variables for n-p-n transistors. Similar solid curves in Fig. 6 are for p-n-p transistors. For a given maximum operating collector to base or collector to emitter voltage and a given resistivity of the base layer material these curves indicate the minimum base layer thickness that may be used if punch through is not to occur at a voltage below the given value.

3.3 *Breakdown Voltage*

The breakdown voltage of germanium step junctions when the breakdown voltage is due to avalanche multiplication within the body of the semiconductor has been investigated by S. L. Miller[5] using both alloyed junctions and junctions grown to simulate alloyed junctions. The body

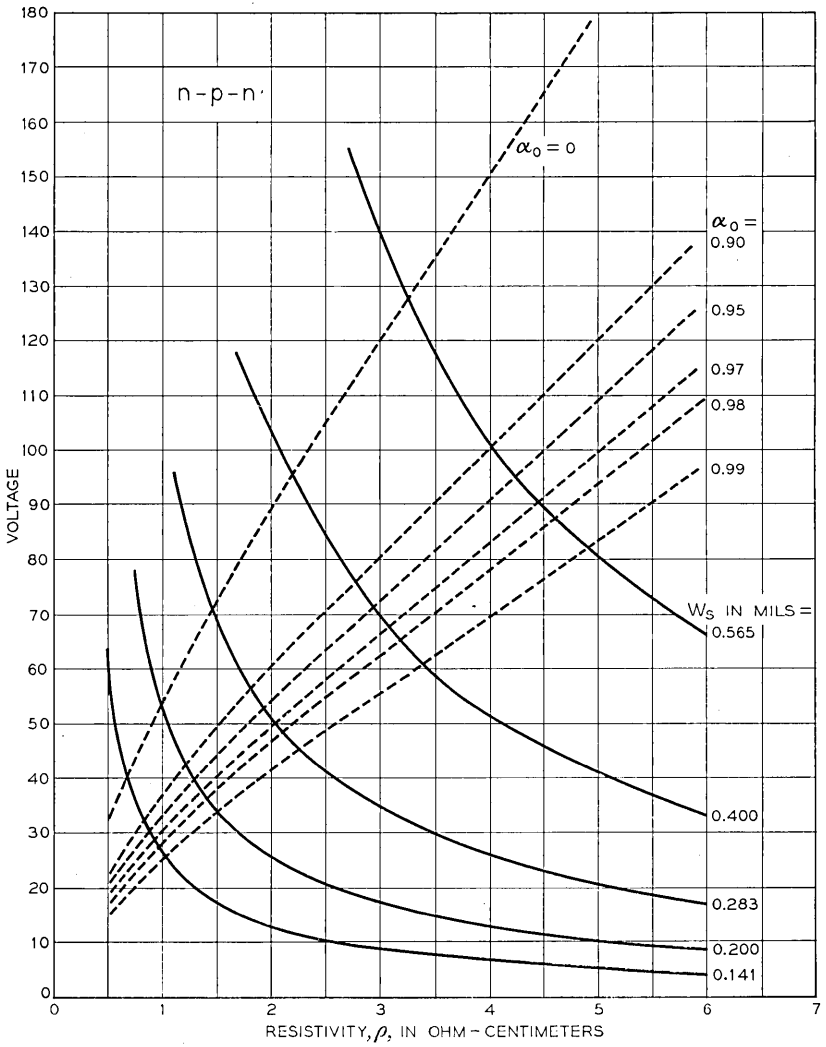


Fig. 5 — Contours of unity total alpha and constant space-charge width for n-p-n alloyed-junction, germanium transistors.

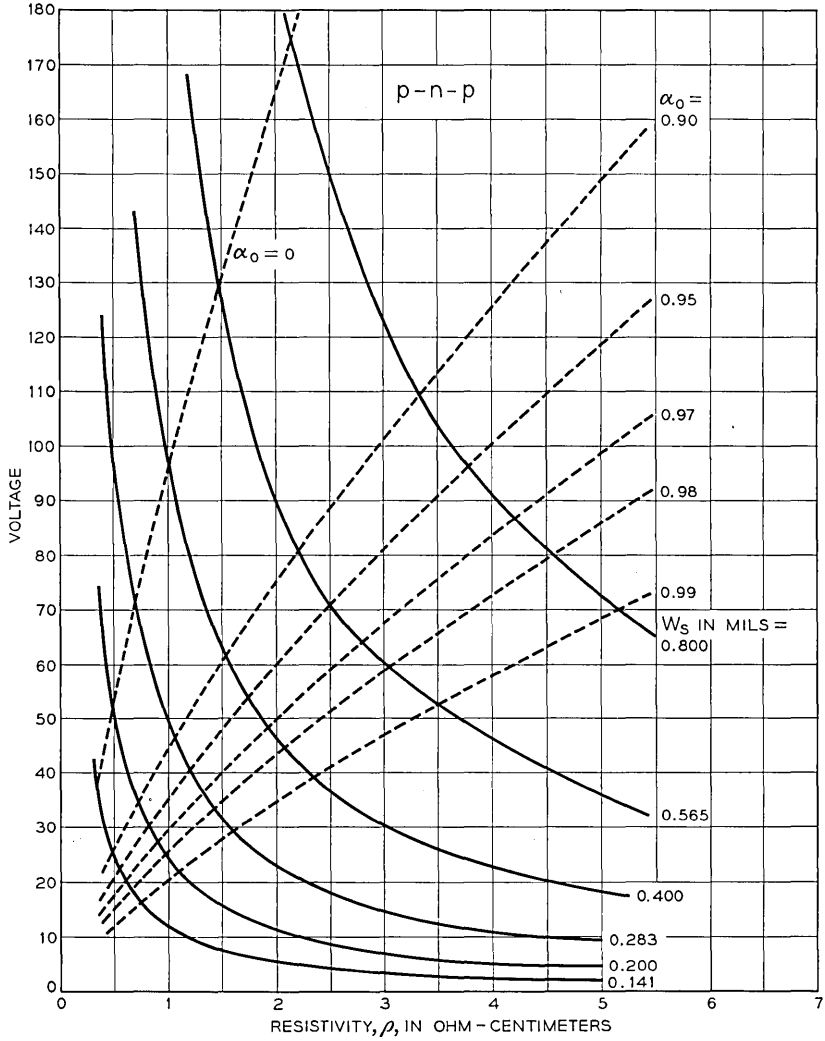


Fig. 6 — Contours of unity total alpha and constant space-charge width for p-n-p, alloyed-junction, germanium transistors.

breakdown voltage thus obtained is plotted in Fig. 4 as a function of N_I. Also plotted in this figure are curves of resistivity versus N_I^{10} for both n-type (labeled p-n-p) and p-type (labeled n-p-n) material. Body breakdown voltage for a given resistivity of base material can thus be read from these curves.

3.4 *Avalanche Multiplication*

The variation of avalanche multiplication with junction voltage for step junctions on germanium is given by the empirical equation[5]

$$M = \frac{1}{1 - (V/V_B)^n} \tag{14}$$

Since current injected by the emitter junction is multiplied by this factor upon being collected, the variation of alpha with collector voltage is given by the equation

$$\alpha = \alpha_0 M = \frac{\alpha_0}{1 - (V/V_B)^n} \tag{15}$$

when space-charge-layer widening effects[11] are neglected.

For alloyed transistors made on base material in the one to five ohm-centimeter range, values of $n = 3$ for p-n-p transistors and $n = 6$ for n-p-n transistors give multiplications which agree well with experiment.[5] Since values of alpha greater than unity may result in failure of junction transistor circuits to operate properly, it is necessary to limit the value of alpha at zero collector voltage to such a value that alpha will not become greater than unity at the maximum operating collector voltage. Alpha is known to vary with temperature and emitter current, hence it is necessary to study α_0 over the expected range of these variables. In Figs. 7 and 8 are shown typical variations of $(1 - \alpha_0)$ with current and temperature for p-n-p and n-p-n transistors of approximately the same geometry. Fortunately, most transistors of a given type made by the same processes show the same type of variation. Therefore, it should be necessary to measure α_0 only at a particular current and temperature to determine its maximum value.

It is apparent that the alpha of n-p-n transistors shows less variation with emitter current than the alpha of p-n-p transistors. Furthermore it has been found that the variation is decreased in both types as the base width is decreased.

The dotted curves in Figs. 5 and 6 are contours of constant α_0 plotted with collector voltage and base resistivity as independent variables, using (15) and appropriate values of n. For a given maximum collector operating voltage and base layer resistivity these curves show the maximum values of α_0 which can be permitted if total alpha is not to exceed unity at the chosen conditions.

The curves shown in Figs. 5 and 6 can be used in several different ways to design transistors or to determine whether or not a given set of requirements are compatible.

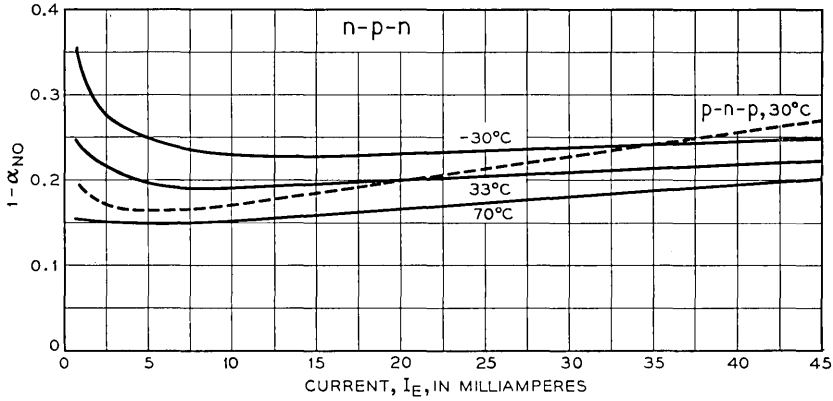


Fig. 7 — Variation of alpha with emitter current and temperature for n-p-n transistors.

3.5 *Base Resistance*

The base resistance for the somewhat idealized transistor structure shown in Fig. 9 has been calculated by J. M. Early.[11] It is given by

$$r_B = \frac{1}{8\pi}\frac{\rho}{W_1} + \frac{1}{2\pi}\frac{\rho}{W_2}\,\ell n\left(\frac{r_2}{r_1}\right) \tag{16}$$

where the first term represents the contribution due to the base layer between the emitter and collector junctions and the second term is the contribution of the base wafer between the edge of the emitter and col-

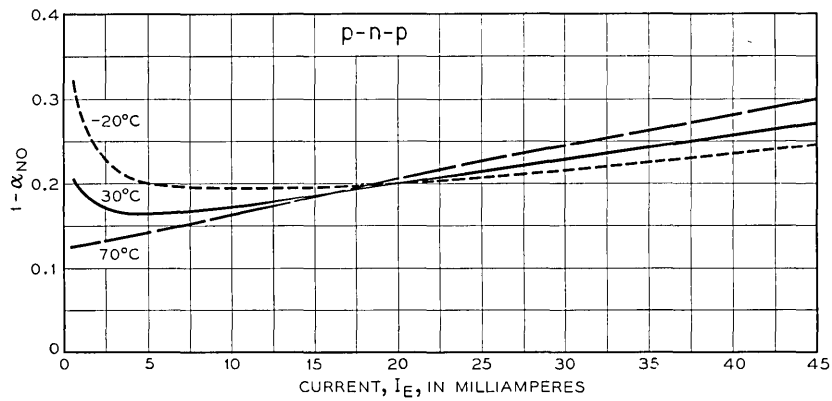


Fig. 8 — Variation of alpha with emitter current and temperature for p-n-p transistors.

lector contacts and the ring base contact. This expression is probably a good approximation to the base resistance of the transistor in the active region, particularly at low currents. As the current increases, the outer edge of the emitter becomes more forward biased than the center, thus effectively decreasing the base resistance. In the current saturation region the base resistance is given approximately by the second term, or

$$r_{B3} = \frac{1}{2\pi} \frac{\rho}{W_2} \ell n \left(\frac{r_2}{r_1} \right) \tag{17}$$

If a common-emitter amplifier is not driven from a high impedance current source, the active region base resistance may seriously limit the pulse rise time.

3.6 *Frequency Cut-Off of Alpha*

Emitted carriers cross the base layer by a diffusion process when current densities are small. Since a dispersion in transit time is inherent in such a process, changes in both magnitude and phase of alpha with frequency are to be expected.[12] The frequency at which the magnitude of alpha falls to $1/\sqrt{2}$ of its low frequency value is given by

$$f_N = \frac{D}{\pi W^2} \tag{18}$$

where D is the diffusion constant for injected carriers. This can be expressed as

$$f_N = \frac{A}{W^2} \qquad \text{mc/sec, where } W \text{ is in mils} \tag{19}$$

The variation of A with the resistivity of the base layer for alloyed-junc-

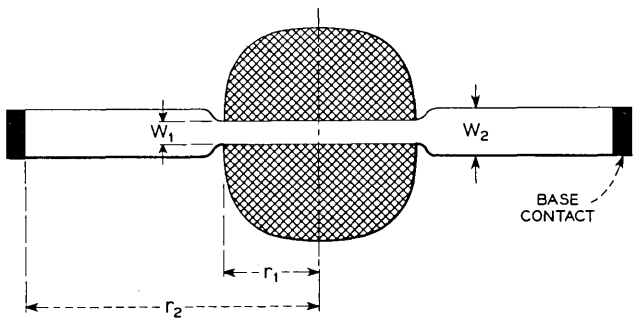


Fig. 9 — Alloyed transistor geometry.

tion germanium transistors is shown in Fig. 4 for both n-p-n and p-n-p transistors. The variation is due to mobility changes.[10]

It is of course true that at high current densities there is a departure from pure diffusion of minority carriers because of induced drift fields.[13, 14] Furthermore, at high currents the effective base layer width changes as a result of a re-distribution of emitter current. There is also a large body of evidence for anomalous changes of frequency cut-off of alpha with surface changes. However, it will be assumed that (18) is a valid criterion for frequency cut-off at low voltages and currents.

Because of space charge layer widening, the cut-off frequency of alpha should vary with collector voltage. Using (13) and (19) this variation can be shown to be given by

$$f_N = \frac{f_{N_0}}{\left[1 - \left(\frac{V_c f_{N_0}}{3.65 \times 10^{-13} N_I A}\right)^{1/2}\right]^2} \tag{20}$$

where f_{N_0} is the cut-off frequency at zero voltage. For n-p-n transistors made on 1.5 ohm-cm material

$$f_N = \frac{f_{N_0}}{[1 - 0.016(V_c f_{N_0})^{1/2}]^2} \tag{21}$$

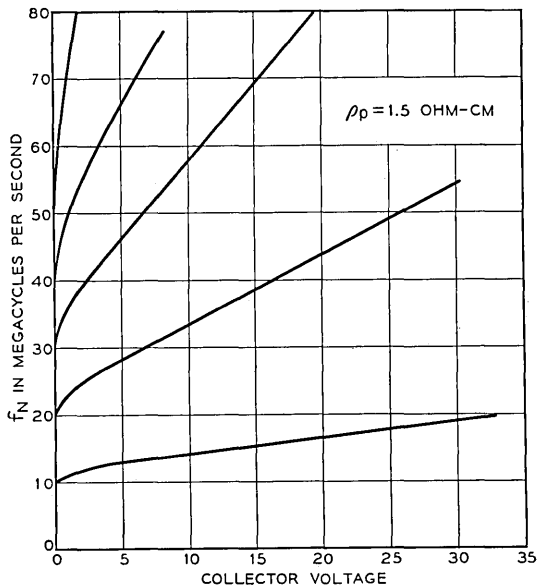


Fig. 10 — Variation of frequency cut-off of alpha with collector voltage for n-p-n transistors made on 1.5 Ω-cm Ge.

The variation of alpha cut-off frequency for alloyed junction n-p-n transistors made on 1.5 ohm-cm germanium is plotted in Fig. 10 for several values of f_{N_0}. Fig. 11 is a more general plot for other resistivities and f_{N_0}'s. It is apparent that a problem exists in alpha cut-off frequency measurement as a result of the large variation of f_N with voltage. Furthermore an exact analysis of rise time and decay time would have to take this variation into account.

It is possible that this increase of f_N with voltage may allow some relaxation of low-voltage requirements in this respect.

3.7 DISCUSSION

Let it be assumed that a transistor is desired which will operate at a collector voltage of 30 volts and which, when used as a common-emitter pulse amplifier will give a current gain of twenty with a pulse rise time of a tenth of a microsecond or less. The maximum current will be assumed to be 100 milliamperes, corresponding to a load impedance of 300 ohms.

A reasonable range for α_0 would be from 0.97 to 0.99, the maximum limit being imposed because of avalanche multiplication. From Fig. 2 it

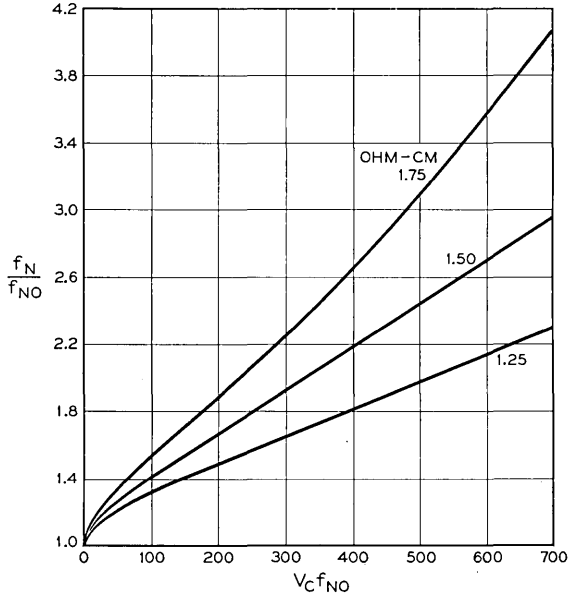


Fig. 11 — Variation of frequency cut-off of alpha with collector voltage for n-p-n transistors.

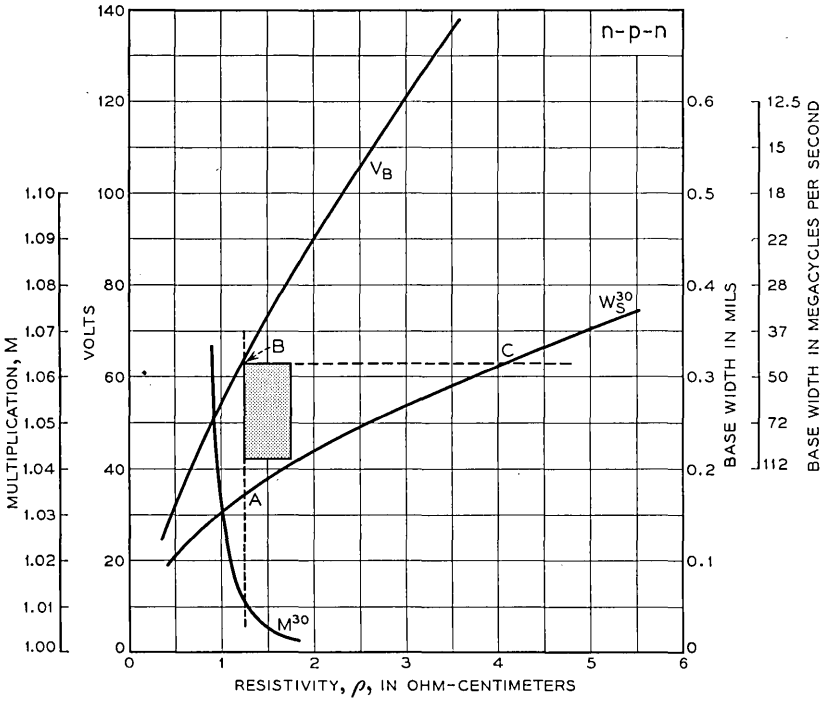


Fig. 12 — Design summary for n-p-n transistors.

is calculated that the minimum allowable value of f_N is 43 mc/sec. From Fig. 5 (for n-p-n transistors) assuming a maximum collector voltage of 30 volts and a maximum permissible α_0 of 0.99 it is seen that the resistivity of the base material must be at least 1.25 ohm-cm and that the minimum base layer thickness is about 0.175 mils, corresponding to a f_N of 144 mc/sec from (19) and Fig. 4. Unfortunately, it is difficult with the design data plotted as they are in Fig. 5 to obtain the limits on the base resistivity and the lower limit on base width which are necessary to insure freedom from punch through and alpha greater than unity for a specified collector voltage. For this reason the design data have been replotted in a different manner in Figs. 12 and 13. Here the avalanche multiplication factor M and space charge layer width W_s, both at 30 volts, have been plotted as a function of base layer resistivity. Again assuming a maximum α_0 of 0.99 this places an upper limit of 1.01 on the multiplication and therefore, a lower limit of 1.25 ohm-cm on the base layer resistivity. Since the base layer width must be greater than the

space charge layer width the range of base layer width chosen must lie entirely above the curve for space charge layer width at 30 volts. In addition, the minimum frequency cut-off requirement of 43 mc/sec means the design must lie below the horizontal dotted line. Thus triangle ABC encloses the region of possible designs. The shaded rectangle seems to be the most appropriate design. The frequency range can be read directly from the base width scale since this scale has been calibrated in mc/sec. Assuming a 0.015″ diameter collector junction, this design can be summarized as follows:

Base layer resistivity — 1.25 to 1.75 ohm-cm
Base layer width — 0.21 to 0.32 mils
Collector capacitance (10V) — 5.5 to 6.4 $\mu\mu$f
Alpha — 0.97 to 0.99
Frequency cut-off of alpha — 43 to 100 mc/sec

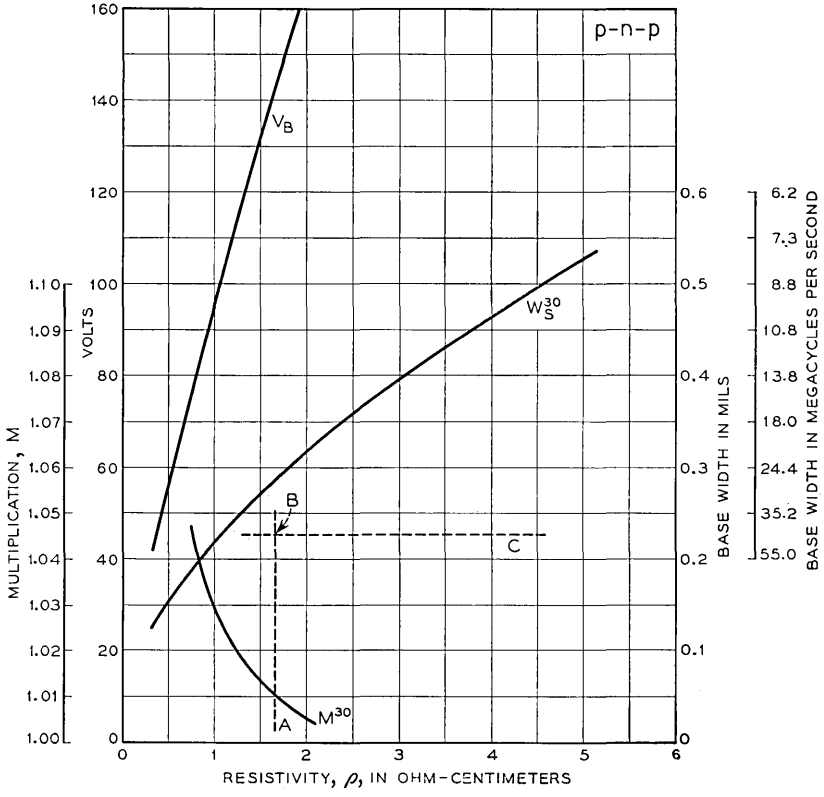


Fig. 13 — Design summary for p-n-p transistors.

By studying the data in Figs. 6 and 13 for p-n-p transistors it is seen that a design to meet the stated requirements is not possible. Assuming a maximum collector voltage of thirty volts, the absolute maximum value of cut-off frequency would be about 27 mc/sec for p-n-p transistors.

Returning to the n-p-n design, assume that the device has the following geometry:

Emitter dia — 0.015″
Collector dia — 0.015″
Base layer width — 0.00025″
Base wafer width — 0.0015″
Base contact dia — 0.030″
Base resistivity — 1.5 ohm-cm

then from (16) and (17), $r_B = 138$ ohms and $r_{B3} = 44$ ohms. The collector capacitance at 10 volts is 5.8 $\mu\mu$f. For a load resistance of 300 ohms, $R_L C_C = 1.7 \times 10^{-9}$ sec which is not very small compared to $1/\omega_N = 3.7 \times 10^{-9}$ sec. If the base is driven by a current source whose internal impedance is large compared with 138 ohms, the rise time would probably be controlled by the alpha cut-off frequency, however.

4.0 CONCLUSIONS

(1) The interdependence of transistor parameters has been shown to set theoretical limits on designs which are possible. In essence, high frequency and high voltage requirements militate against each other. A design theory which takes these factors into account has been described.

(2) For high-speed and high-voltage switching (and transmission) applications, it is shown that the n-p-n enjoys a superiority beyond that expected from only mobility considerations. The additional advantage stems from the different multiplication laws for n-p-n and p-n-p transistors. For example, it is theoretically feasible to make alloyed junction n-p-n transistors with alpha cut-off frequencies in the range of 50 to 100 mc/sec which will sustain collector voltages as high as 30 volts, and be capable of operating as pulse amplifiers in circuits with a current gain of 20 and a pulse rise time of less than a tenth of a microsecond. A design meeting these requirements is not possible with p-n-p's. In the example chosen, there is a ratio of 5.3 between the maximum frequency cut-off of alpha obtainable with n-p-n transistors as compared with p-n-p transistors for the same maximum value of low voltage alpha.

The design theory developed above gives upper limits on what can be achieved. It has been assumed that there are no fabrication problems and that geometries are ideal, namely, that the junctions are perfectly

planar and parallel. Furthermore, it has been assumed that the breakdown and multiplication properties of collector junctions are determined by the bulk properties of the junction. In many cases there are breakdowns on the surface at lower voltages. These are frequently multiplicative but they multiply a very small fraction of the emitter current.

ACKNOWLEDGMENTS

The graphical method of presenting the design information evolved after numerous discussions with R. M. LeLacheur. The assistance of P. A. Harding, who took the data shown in Figs. 7 and 8, is gratefully acknowledged. The authors are indebted to M. A. Clark for information on the operation of common-emitter amplifiers using transistors with alpha greater than unity.

REFERENCES

1. Early, J. M., p-n-i-p and n-p-i-n Junction Transistor Triodes, B.S.T.J., **33,** pp. 517–533, May, 1954.
2. Shockley, W., Transistor Electronics — Imperfections, Unipolar and Analog Transistors, Proc. I.R.E., **40,** pp. 1289–1313, Nov., 1952.
3. McKay, K. G., and McAfee, K. B., Electron Multiplication in Silicon and Germanium, Phys. Rev., **91,** pp. 1079–1084, Sept. 1, 1953.
4. McKay, K. G., Avalanche Breakdown in Silicon, Phys Rev, **94,** pp. 877–884, May, 15, 1954.
5. Miller, S. L., Avalanche Breakdown in Germanium, Phys Rev, to be published.
6. Miller, S. L., and Ebers, J. J., Alloyed Junction Avalanche Transistors, B.S.T.J., to be published.
7. Ebers, J. J., and Moll, J. L., Large-Signal Behavior of Junction Transistors, Proc. I.R.E., **42,** pp. 1761–1772, Dec., 1954.
8. Moll, J. L., Large-Signal Transient Response of Junction Transistors, Proc. I.R.E., **42,** pp. 1773–1784, Dec., 1954.
9. Shockley, W., The Theory of p-n Junctions in Semiconductors and p-n Junction Transistors, B.S.T.J., **28,** pp. 435–489; July, 1949.
10. Prince, M. B., Drift Mobilities in Semiconductors, I. Germanium, Phys Rev, **92,** pp. 681–687, Nov., 1, 1953.
11. Early, J. M., Design Theory of Junction Transistors, B.S.T.J., **32,** pp. 1271–1312, Nov., 1953.
12. Shockley, W., Sparks, M., and Teal, G. K., p-n Junction Transistors, Phys Rev, **83,** pp. 151–162, July 1, 1951.
13. Rittner, E. S., Extension of the Theory of the Junction Transistor, Phys Rev, **94,** pp. 1161–1171, June 1, 1954.
14. Webster, W. M., On the Variation of Junction Transistor Current Amplification Factor with Emitter Current, Proc. I.R.E., **42,** pp. 914–920, June, 1954.

An Experimental Polytonic Signaling System

By C. A. LOVELL, J. H. McGUIGAN and O. J. MURPHY

An experimental high-speed signaling system capable of signaling at the rate of 100 decimal digits per second over a variety of existing Bell System transmission facilities is described. This system uses several frequencies in combinations as its code elements. Because the manner of discriminating between different frequencies is significantly different from present-day practice, the system has been called polytonic *to distinguish it from existing multifrequency systems. Signal separation in this system depends not on the steady-state response of filters, but on the transient response of simple reactive networks. Discrimination is enhanced by the use of rather precise timing in the signal generating mechanism and somewhat less precise timing in the receiver. A second system, capable of signaling at the rate of 300 decimal digits per second over a more restricted range of transmission facilities, is also briefly discussed.*

1.0 INTRODUCTION

The need in future telephone systems for a signaling method of significantly higher speed that that provided by the familiar dial or even by the intertoll multifrequency system[1] has led to a research program in high speed signaling. The program has resulted in laboratory models of several experimental systems based on different principles. This paper describes one of these systems, based on the concept of orthogonal functions. Several different frequencies are used, in combinations, to transmit signal information. The means of generating and recognizing these frequencies represent a departure from present practice; in order to distinguish this experimental system from the multifrequency system, the term *polytonic* was coined. It is hoped that this paper provides an interesting illustration of the way a mathematical conception may be transformed, through successive stages, to a physical system.

[1] C. A. Dahlbom, A. W. Horton and D. L. Moody, Application of Multifrequency Pulsing in Switching, Trans. A.I.E.E., **68**, pp. 392–396, 1949.

Telephone circuits fall naturally into two classes, local and toll. The bandwidths of these two classes of facilities and the types of distortion which they exhibit are so different that it was found expedient to build two separate embodiments of the polytonic system. One was designed to work in the presence of the characteristic forms of distortion found in the toll plant, while the other was built to take advantage of the higher speed permitted by the nature of the local plant.

Most of the following discussion applies equally well to both signaling systems. Throughout the body of the paper the main emphasis will be placed on the toll system; except for occasional mention, the local system will be treated in a separate section. The discussion will begin with the theoretical background for the signaling method under consideration and proceed to a detailed examination of the signals representing the code elements. Problems encountered in the design of practical systems will then be taken up, including certain arbitrary requirements of coding and signal content. This will be followed by a description of some of the circuitry evolved to meet the requirements and an operational description of the toll-signaling system which was built to assess the capabilities of this approach to the signaling problem. A brief account of tests of this system over certain Bell System toll facilities will be given. The discussion will conclude with some remarks about circuitry devised for the local-signaling embodiment which was capable of achieving a substantially higher signaling speed.

2.0 THEORETICAL BACKGROUND

The basis for the signaling systems about to be described lies in the concept of orthogonality as applied to certain classes of mathematical functions. A set of normalized time-functions, $\Phi_1(t)$, $\Phi_2(t) \cdots \Phi_n(t)$ are said to be orthogonal over the time-interval T if

$$\begin{aligned} \int_0^T \Phi_i(t) \cdot \Phi_j(t)\, dt &= 0 \qquad \text{for } i \neq j \\ &= 1 \qquad \text{for } i = j \end{aligned} \tag{1}$$

Sinusoids of various frequencies provide a familiar example of one such class of functions; the period of orthogonality, T, may be any value which embraces an integral number of periods of each of the frequencies under consideration. It was recognized that direct use of this property could be made in separating, at the receiving point, signals made up of combinations of such functions, without employing the steady-state selective properties of reactive filter networks.

As an illustration of how such a signaling system might be mechanized, consider a group of oscillators whose frequencies are multiples of some common base frequency. The period of this frequency then defines one possible value of the period of orthogonality, T, which will satisfy the requirements for all frequencies in the group. Now suppose that some coding procedure is set up which involves sending a tone-spurt of one or more of these frequencies over some common transmission medium to a distant point to represent each of a number of desired signal elements or characters. If the tone-spurts have a duration equal to or greater than the period T, it will be possible to distinguish, at the receiving end, which frequencies are present, by application of the above principle. This may be accomplished by having a group of oscillators at the receiving end whose frequencies are identical with those at the sending end and providing a number of parallel channels for the received signal, equal to the number of oscillators. In each of these channels, the received signal is multiplied, in the mathematical sense, with the aid of suitable circuitry, by the output of the corresponding oscillator and the resulting product integrated for a period of duration T. An examination of the several channel integrators, at the expiration of the period T, reveals a "unit" output from those whose local frequency corresponds to a component of the received signal, while those corresponding to frequency components not present in the received signal show "zero" output. The terms "unit" and "zero" are employed here to represent the ideal case; in practice, there is found, in a properly designed circuit, a substantial value of output representing correspondence and little or no output for a lack of correspondence. After the integrator outputs have been examined, which is done virtually instantaneously on a sampling basis, the electrical integrators are reset to zero, thus clearing the receiver for the reception and recognition of the next tone-spurt from the sending end.

The system just outlined represents a mechanization of the classical concept of orthogonality for signaling purposes. It has been suggested[2] that an extension of the concept of orthogonality can be used to define a new set of properties which seems, in some respects, more useful than those defined by (1). Consider a set of normalized functions $\theta_1(t)$, $\theta_2(t)$ $\cdots$ $\theta_n(t)$ which, over the time interval T, satisfy the following conditions:

$$\begin{aligned} \int_0^T \theta_i(\lambda)\cdot\theta_j(T-\lambda)\,d\lambda &= 0 \qquad \text{for } i \neq j \\ &= 1 \qquad \text{for } i = j \end{aligned} \tag{2}$$

[2] This suggestion was made by R. B. Blackman of Bell Telephone Laboratories.

The practical importance, in the generation and reception of electrical signals, of functions which satisfy (2) is that the defining integral describes the transient response of a network; that is to say, if $\theta_i(t)$ is the response of a network to an impulse time-function, then the integral represents the response of the network when it is acted upon by an applied signal of the form $\theta_j(t)$. Thus the inherent factors which govern the instantaneous transient response of passive electrical networks may be utilized to perform the required mathematical processes of multiplication and integration without recourse to special functional circuits for the purpose.

The following description will deal in some detail with signaling arrangements exploiting these properties of reactive networks, in which the signal generation and recognition is performed on the basis of the instantaneous responses at particular times rather than on the basis of a steady-state frequency-response, as in more conventional systems.

3.0 POLYTONIC SIGNALS

The elements of the signals used in the polytonic signaling system are damped sine waves, such as that shown in Fig. 1(a), which are characterized by the expression:

$$E_m = E_0 e^{-at/T} \sin m\pi t/T$$

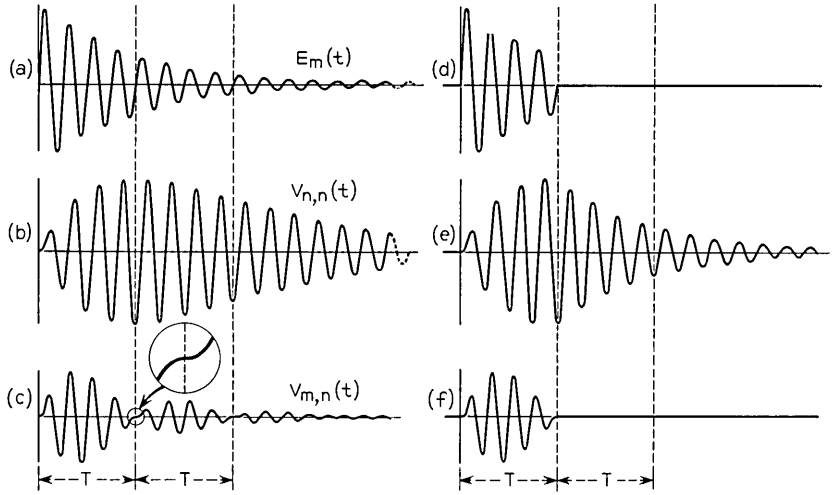


Fig. 1 — Polytonic signal wave forms and network responses.

The frequency variable is expressed in the form shown to emphasize the fact that the frequencies in the set are integral multiples of a common base-frequency, $1/2T$. If the multiplier m is restricted to successive odd integers or successive even integers, as is actually done in the system to be described, the frequency separation between adjacent members of the set is then $1/T$. For any integral value of m, this expression satisfies the conditions of both (1) and (2).

This type of response can be obtained, for example, by applying a step-function of voltage to a series circuit of R, L, and C elements and observing the voltage across the resistive element, which is proportional to the current. Circuits of this kind are, in fact, used to generate the polytonic signals at the transmitting end of the system.

The discriminating devices at the receiving end may be similar simple circuits of R, L, and C elements, matched to the respective generator circuits in both frequency and damping. It is desired that the impulsive voltage transfer-function of the receiving networks match the signals generated by the corresponding transmitting networks. This correspondence may be obtained by taking the voltage across the condenser element, which is proportional to the integral of the current, as the output of the network.

We shall now proceed to determine the response of one of these receiving networks both to a transmitted signal which it matches and to one which it does not match. If we consider the signal

$$E_m(t) = E_0 e^{-at/T} \sin m\pi t/T$$

applied to a network whose voltage transfer function is

$$G_n(t) = Ke^{-at/T} \sin n\pi t/T$$

we have, for the response

$$V_{mn}(t) = KE_0 e^{-at/T} \int_0^t \sin (m\pi\lambda/T) \sin [n\pi(t - \lambda)/T]\, d\lambda \qquad (3)$$

When evaluated, this yields

$$V_{mn}(t) = KE_0 \frac{Te^{-at/T}}{\pi(m^2 - n^2)} (m \sin n\pi t/T - n \sin m\pi t/T) \text{ for } m \neq n \qquad (4)$$

and

$$V_{nn}(t) = KE_0 \frac{Te^{-at/T}}{2n\pi} [\sin n\pi t/T - (n\pi t/T) \cos n\pi t/T] \text{ for } m = n \qquad (5)$$

When equations (4) and (5) are evaluated for $t = T$ we find

$$\begin{aligned} V_{mn}(t) &= 0 \\ V_{nn}(t) &= KE_0(-1)^{n+1}(T/2)e^{-a} \end{aligned} \tag{6}$$

If a suitable normalizing factor is applied to the second expression, this pair is seen to fulfill the conditions of (2).

Thus, if a receiving network is excited only by the signal element, or damped sine wave like that shown in Fig. 1(a), to which it is matched, the voltage across the condenser element of the network is given by (5) and will have the form shown in Fig. 1(b); this will hereafter be referred to as a "desired" response. A desired response is seen to consist of an oscillatory voltage which starts with zero amplitude and zero slope and whose successive peak values are contained within an envelope which builds up to a broad maximum and then decays slowly. One of the peaks is centered at time T; by suitable choice of damping constant, ($a = 1$) the maximum value of the envelope may be made to coincide with this peak, and this is the condition illustrated in the figure.

If a receiving network is excited only by a signal element to which it is not matched, its response, in terms of condenser voltage, is given by (4) and depends on the relationship, in frequency spacing, between the frequency-of-match and that at which it is excited. This will be regarded as an "undesired" response; Fig. 1(c) shows the undesired response when the exciting frequency is the next higher frequency in the set. This response will be seen to consist of an oscillating voltage which starts with zero amplitude and zero slope, builds up to a maximum of amplitude, and then returns to zero amplitude and zero slope at time T. Between successive multiples of T, the envelope of the undesired response exhibits maxima of successively smaller amplitude. More remote signal elements produce undesired responses with zeroes at submultiples of T, but integral multiples of T are the only times at which all undesired responses are simultaneously zero in both slope and amplitude. For stimuli of equal strength, the maximum amplitude of the undesired response is always considerably less than that of the desired response.

When the signal applied to a receiving network contains both desired and undesired components, the response is much more complex in appearance than those shown in Figs. 1(b) and 1(c) which represent excitation by single signal elements. The response is, in fact, a linear combination of components of the two classes described, each of which retains its individual properties. Thus if the voltage outputs of the receiving networks are sampled instantaneously, at time T, all networks which have

found their counterpart signal element in the complex received signal will exhibit an output representing the peak value of an oscillatory wave of corresponding frequency, while those which have not found their counterpart will show, ideally, no output at this instant even though they may have shown a considerable oscillatory voltage before this time and may show an appreciable voltage at a later time.

4.0 CONSIDERATIONS IN THE DESIGN OF A PRACTICAL SYSTEM

The design of a practical signaling system involves considerations not covered by the theory of the signaling method used. Some of these considerations are entirely under the designer's control, while others are influenced by the external conditions under which the system must work. The external influences will be discussed first, and then the design parameters which the designer can choose freely.

4.1 Timing

At the outset the mathematical theory considered the response of receiving networks to signals sent out by transmitting networks and found that at a particular instant the desired response was large and the undesired response was zero. Mathematically it is easy to examine a function at a particular instant; from a circuit designer's viewpoint it is usually easy to sample a network response at a particular instant. However, when there is interposed between the transmitter and the receiver a transmission facility having an unknown delay and a restricted transmission band, it is not easy to know the particular instant at which to take the sample. The starting time, from which the interval must be measured, may not be very sharply defined at the receiving end, and yet a fairly high degree of precision in timing is obviously necessary. If, for example, an error of one-quarter period of the oscillation frequency is made in taking the instantaneous sample of a desired response, a zero output will be obtained instead of the legitimate peak value. Extending the sampling period, to avoid this difficulty, would bring in appreciable energy from the undesired responses, which are zero only at the critical instant. There are still other difficulties, related to timing, which will be taken up later in connection with the design of a specific system, but enough has already been said to point up the desirability of some artifice which would retain the advantages of the orthogonal approach to the signaling problem while either relaxing the stringency of the timing requirements or shifting them from the receiver to some other part of the system, such as the transmitter, where they might be easier to meet.

Several proposals were considered which effected the desired easing of the timing requirement in greater or lesser degree, but the one finally adopted did, in effect, shift the timing mechanism to the transmitter. This arrangement produced a generated signal which no longer perfectly fulfilled the postulates of orthogonal theory, but whose deviations from perfection were found to be small and relatively unimportant in their effect on the ability of the signal receiver to discriminate between signal elements. The justification for, and utility of, this modification will become apparent upon further examination of the behavior of the networks of the receiver.

At the time T, or any multiple thereof, it may be seen that there is no energy stored in the receiving networks exhibiting undesired responses: the zero of voltage means that there is no energy in the condensers (neglecting dielectric polarization) while the zero of slope means that there is no current flowing and hence no energy in the coils. If, then, instead of persisting indefinitely, as was previously assumed, the energizing signal is removed at this time, no further energy will be introduced into the networks. Hence no voltage will appear across the condensers and the undesired response will remain zero thereafter. A network exhibiting a desired response, on the other hand, has considerable stored energy; its condenser voltage, indeed, is at a peak at this critical time. If the source of energy is removed at time T, an oscillating voltage will continue to appear across the condenser, decaying in amplitude somewhat more rapidly than was formerly the case, as the stored energy is dissipated and not replaced. Thus, by removing the source of energy at time T, it is theoretically possible to obtain infinite discrimination, in the absence of noise, between desired and undesired responses *at any time thereafter*. The effect of removing the source of energy may conveniently be achieved by abruptly stopping the decaying oscillations of the transmitting network at time T after their initiation, as shown in Fig. 1(d), instead of allowing them to persist indefinitely. The behavior of the receiving circuits in the cases of wanted and unwanted responses are shown in Figs. 1(e) and 1(f). It is relatively easy to accomplish this timing job since the initiating pulse for the transmitting networks has a steep wave front which defines the starting time precisely.

The timing requirement at the receiver is then merely that the sampling of the network condenser-voltage occur *after* the expiration of a time T measured from the starting time of the signal, and a fairly generous margin for time uncertainty can be provided. The sampling period itself may now be extended, since the undesired responses will remain zero instead of building up again after the null point, and can ad-

vantageously be chosen to be of such length that one or more peaks of desired response are embraced, thus avoiding the possibility of missing a single peak through mis-timing an instantaneous sample.

The means employed for arresting the oscillations of the transmitting networks at the critical time, and thus producing a signal of controlled length, will be taken up in detail in the section devoted to the design of the signal transmitter for a practical system.

4.2 Transmission Facilities

Any signaling system intended for telephone use must be capable of working over the wide variety of transmission facilities existing in the telephone plant. These facilities all fall short of perfect transmission in a variety of ways which do not seriously impair the intelligibility of speech but which do distort signals of all kinds. Non-human signal receivers are limited in complexity both by cost considerations and by their designer's limited skill, so they are less tolerant of distortion than is the ear.

Telephone circuits, as indicated in the introductory section, fall naturally into two classes, local and toll. Subscriber lines and local trunks usually have moderate attenuation and low noise, and their bandwidth restriction appears as a gradual roll-off accompanied by little phase distortion in the useful transmission band. Toll trunks, on the other hand, may have a wide range of both attenuation and noise. Bandwidth is usually less than on local trunks, although carrier circuits may show greater bandwidth than some small-gauge local trunks. Both voice-frequency and carrier toll trunks have sharper cutoffs than local trunks, with attendant phase distortion. In long loaded and repeatered circuits the phase distortion near the edges of the band accumulates to such an extent that it is manifest as an appreciable difference in transmission time, spoken of as delay distortion. Carrier circuits exhibit a characteristic kind of distortion caused by lack of synchronism between the carrier oscillators at the two ends of the circuit. This type of distortion is held within limits which make it unnoticeable to the ear but which may still render an arbitrary signal shape unrecognizable. The consequences of these various types of distortion seriously complicate the receiver design problem.

4.3 Codes

On a two-amplitude or on-off basis, at least four frequencies are required to provide a distinct combination for each of the ten decimal

digits. If a fifth frequency is provided, the familiar two-out-of-five code with its self-checking feature can be used. In a polytonic signaling system operating over transmission facilities of limited bandwidth, the use of the five-frequency code entails a slight loss of signaling speed: since the bandwidth must be divided into five channels rather than four, the frequencies are closer together and therefore the interval T is longer. In addition, the fifth frequency requires extra equipment for its implementation. These costs were felt to be small compared to the advantage of the self-checking code, so the two-out-of-five code was chosen.

4.4 Frequency Choice

One of the first decisions to be made in designing a system of this kind is to choose the frequencies which will be used. The frequencies chosen for the experimental toll signaling system were 500, 700, 900, 1,100 and 1,300 cps. These frequencies lie comfortably within the pass-band of all transmission facilities over which the system might be required to operate.

4.5 Signaling Speed

Since T is the reciprocal of the frequency separation, the frequencies chosen give a value for T of 5 milliseconds. It was decided to take another 5 milliseconds, after the time T for each digit, to allow for detection, network deenergization, and other system operations which were required between digits. Each digit thus occupied a time of 10 milliseconds, and the signaling rate was 100 digits per second.

4.6 Mode of Operation

The signaling speed of one hundred decimal digits per second is much higher than the rate at which a human operator can supply the information to be transmitted. To make efficient use of this speed, therefore, the input information must be accumulated in some kind of temporary storage mechanism, or register, from which it can be spilled at the rate for which the transmitter is designed. One signal transmitter of a high-speed system would send out completed input information derived from any of a number of registers, so that the time required to set up the information, in individual instances, would not constitute a limiting factor. Similarly, one receiver might accept messages from any one of a number of transmitters, thus further exploiting whatever inherent speed capabilities the system might possess.

If the self-checking features of the code, discussed in an earlier section, are to show their full advantage, the system must be arranged to send the same signal information again, upon demand, whenever an implausible received combination causes rejection of the information received. The easiest way to do this, and the one adopted for this system, is to send the complete signal information or "message" repetitively, whether repetition is needed or not, until an acknowledgment signal is obtained from the receiver that a complete message, plausible in all digits, has been received. Reception of this signal, at the transmitting end, would then release the transmitter.

Because of the repetitive character of the signal transmissions, and the possibility of a signal receiver "coming in" on a line from a transmitter in the middle of the message, some internal indication is required in the message itself to mark the end of one iteration and the beginning of the next; this subject is discussed further in Section 5.22.

For the system to be described, the message is defined to consist of any arbitrary succession of eight decimal digits, representing a typical subscriber's number with office code, line number, and party designation. The desired message is completely set up before any part of it is transmitted; and after transmission is begun it is sent, digit by digit, repetitively, with internal indication of sequence beginning, until the receiver acknowledges reception or until, in the case of laboratory tests, it is desired to send some other message.

5.0 DESIGN FEATURES OF A PRACTICAL SYSTEM

5.1 The Signal Transmitter

A block diagram of a signal transmitter conforming to the mode of operation outlined in Section 4.6 is shown in Fig. 2. It begins with a source of frequency which determines the rate at which the digit signals are produced, and which is shown, for convenience, as a multivibrator. The onset of one part of the cycle of this multivibrator is used for control of the stepping process, as shown. The other part, which must be of length T, times the action of the keying circuit to shock the tuned circuits and thus produce the damped sine wave signal elements. The stepper-distributor conveys the pulses originating in the keying circuit to the digit selector switches in succession. It is represented schematically as a rotating switch and might indeed be such a switch in slow-speed systems. In the system herein described, however, it was made up of a gas-tube stepping chain actuating relatively high-speed mercury-contact relays.

The digit selection switches are also shown schematically as rotary switches and they were thus constituted in the experimental system, two-deck switches serving to convey the distributed pulses to any selected combination of two out of the five tuned circuits. In a working system these switches would probably be replaced by aggregations of relays under control of push-button keysets, or some other source of signal information.

The tuned circuits are five in number and are used in pairs to provide the two-out-of-five code elements required. A small resistance was placed in the common lead of all five circuits and the voltage drop across it used as the signal output.

The keying circuit is perhaps the item of greatest interest in the signal transmitter. Figs. 3 and 4 show circuits implementing two methods, which will be described, for generating the controlled-length polytonic signals previously discussed. Both methods have been used and neither has a clear-cut advantage over the other.

The keying circuit has two duties to perform: it must initiate shock-oscillations in selected tuned circuits and it must arrest these oscillations after an integral number of half-cycles occupying the period T. It is also

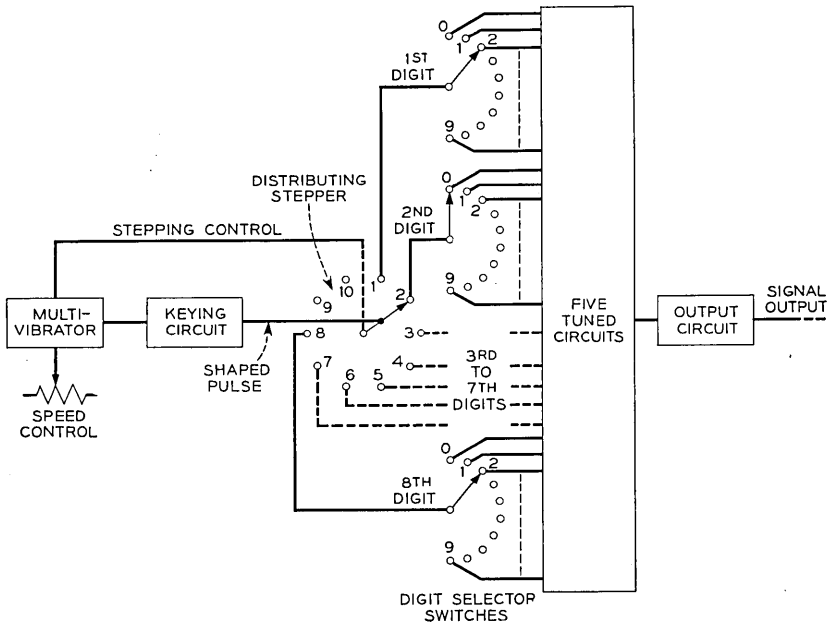


Fig. 2 — Schematic diagram of signal transmitter.

highly desirable, to facilitate the generation of a succession of signals, that the shock excitation voltage supplied by the keying circuit have the same value near the end of a digit period as it had just before the steep wave front occurred, so that a complete cycle is traversed for each signal generated.

One method of arresting the transmitter-network oscillations in an almost transient-free manner at the end of an integral number of half-cycles is to introduce, in series with the circuit, a resistor having a value equal to or somewhat greater than the critical damping resistance. This resistor must be removed, of course, before the next signal-producing shock is applied. If such a damping resistor is introduced at time T, it is then possible to return the exciting voltage to the value from whence it started, by means of an oppositely directed step, without occasioning any further oscillations. An exponentially-damped transient will, however, ensue.

A more nearly transient-free signal may be obtained by employing a shocking pulse having a steeply rising wave front, followed by a smooth decay which will reduce its amplitude to a negligible value at time T. Use of a pulse such as this, instead of the ideal step-function, hitherto contemplated as the exciting voltage, will introduce some anomalous frequency components into the spectrum of the transmitted signal, but these have been found to have little influence on the behavior of the receiving networks.

A keying circuit providing the features of a shaped pulse for excitation

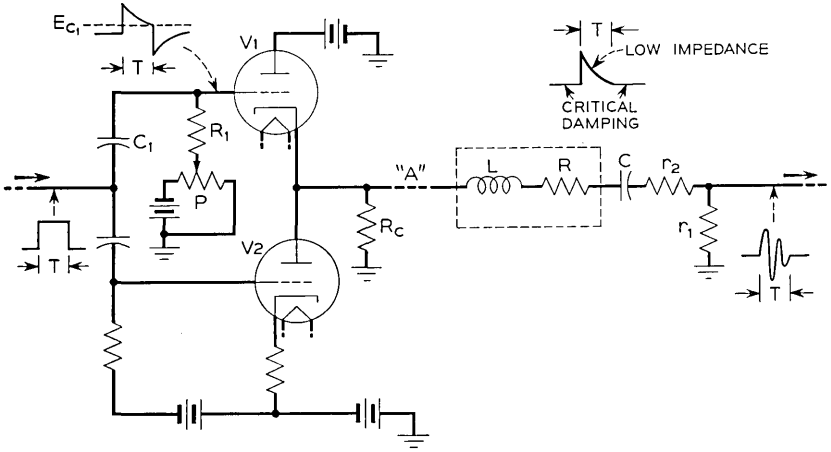


Fig. 3 — A method for generating a controlled-length polytonic signal.

and critical-resistance damping for arresting the oscillations is shown in Fig. 3.

In this arrangement a rectangular voltage pulse of the critical length is applied to tubes V1 and V2. Both tubes are normally cut off. Elements R_1 and C_1 partially differentiate the applied signal at the grid of V1, which is a cathode follower, causing the cathode voltage to rise suddenly and then subside toward ground; the output impedance of the tube, at the cathode, is low during this time, owing to the ordinary dynamic feedback action of the tube. Tube $V2$ is also brought into conduction for the period T through coupling elements of long time-constant; this tube supplies the current required to provide vigorous dynamic action in tube $V1$ as its cathode approaches ground potential and also provides a con-

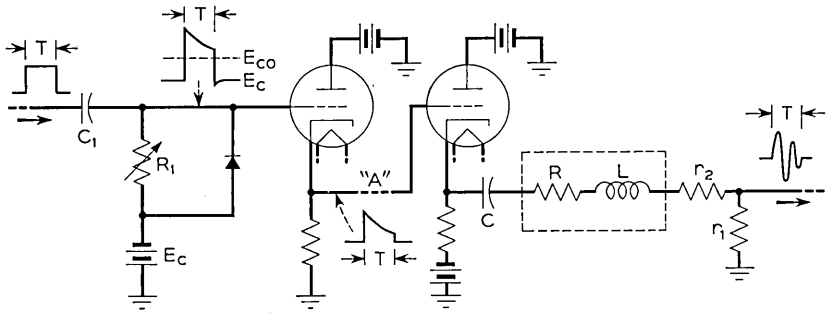


Fig. 4 — Another method for generating a controlled-length polytonic signal.

ductive path for the negative lobes of the oscillatory current in the tuned circuit. The resistor R_c, having a value approximately that of the critical damping resistance for the tuned circuit, is connected from the cathode of $V1$ to ground but plays little part in the action just described, since it is shunted by the low dynamic impedance of $V1$. By properly proportioning the circuit elements and adjusting the grid voltage of $V1$ by means of control P, the cathode of $V1$ may be made to arrive at a voltage of approximately zero with respect to ground at the expiration of the period T. The trailing edge of the input pulse then restores both tubes to the cutoff condition, the impedance from cathode to ground of $V1$ is that of R_c, and the tuned circuit is critically damped by this high impedance. During the period T the total resistance in series with the tuned circuit is the sum of that of the cathode-follower driver, that of the physical coil itself and that of the two resistors r_1 and r_2. These two latter resistors are ordinarily both small; r_1 serves as a means of obtaining an output voltage proportional to the oscillatory current while r_2 serves as a padding re-

sistor to adjust the total to the value required for the prescribed damping constant. The distributing stepper and digit selector switches are interposed at the point marked A.

Another way of terminating the transmitter-network oscillations in a transient-free manner is to apply a second step-function of correct magnitude and polarity to the tuned circuit at the exact completion of any integral number of half-cycles of the oscillation, so that a new oscillatory wave, due to the second shock, will exactly cancel the unwanted remaining portion of the original wave. The number of half-cycles to be chosen is, in this instance, the number occurring within the time interval T. If an odd number of half-cycles have occurred, the terminating shock must have the same polarity as the initial shock, whereas if an even number have occurred a terminating shock of opposite polarity is required. In any case, the magnitude of the terminating shock must be less than that of the initiating shock in exact proportion to the decay of the oscillations. A circuit working on this basis, for an even number of half-cycles within the period T, is shown in Fig. 4.

Here the rectangular input pulse is applied to a cathode-follower through the coupling circuit C_1R_1 whose time-constant is adjusted to match the decay rate of the tuned circuit oscillations and which thereby produces the desired ratio between leading-edge step and trailing-edge step. The diode shunting R_1 serves to discharge C_1 quickly at the termination of the input pulse, which is supplied from a low impedance source. The output of the cathode-follower stage is, then, a voltage pulse of the correct shape which begins and ends at essentially ground potential. This pulse is conveyed to another cathode follower which is arranged to draw a substantial current, and therefore to exhibit a low output impedance, at all times; this tube drives the tuned circuits.

The distributing stepper and digit-selector switches are interposed at the point marked A, as before.

This type of signal generation was employed for the second embodiment of the polytonic system, which was built to attain still higher signaling speeds over local-type facilities. A different set of frequencies from those listed in section 4.4 were chosen for this case. These frequencies, 800, 1,200, 1,600, 2,000 and 2,400 cps, provided an even number of half-waves within the period T.

5.2 The Signal Receiver

5.21 Envelope Detector

It was desired to build the system around a detector which did not have severe timing requirements and which would be capable of with-

standing some distortion of the received signals. The detector devised to meet these requirements based its discrimination essentially on the envelope of the tuned circuit response. In the case of undistorted received signals, after time T the envelope of the desired response is a decaying exponential of appreciable magnitude, while the envelope of the undesired response is zero. It was found experimentally that when the received signals were subjected to the kind of distortion produced by transmission over a carrier system with non-synchronous carrier oscillators, the envelope of the desired response was substantially unchanged while the envelopes of undesired responses, though no longer zero, were still much smaller than those of desired responses. When the signals were transmitted over lines with delay distortion, so that one component arrived a time-interval Δ later than the other, the envelopes of undesired responses were observed to be substantially zero after time $(T + \Delta)$, measured from the arrival of the earlier component, while again the envelopes of the desired responses were little changed.

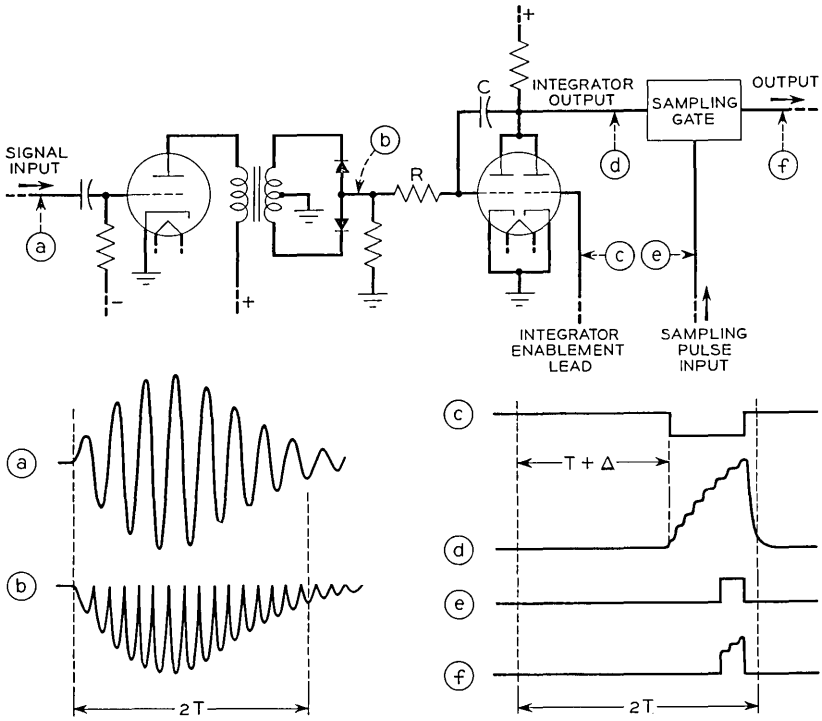


Fig. 5 — Circuit and wave-shapes of receiver channel integrator for envelope detection.

A common method of deriving the envelope of an oscillating signal is to rectify the signal and filter out the carrier and higher frequency components. This method is not feasible here because these envelopes contain significant frequency components at frequencies comparable to the carrier. The approach which was actually used is illustrated in Fig. 5, and consisted in full-wave rectifying the network responses to eliminate their oscillatory character and then integrating over a short period to eliminate the fluctuating nature. This integration should not be confused with that mentioned in the discussion of orthogonal functions. Its only purpose is to make the signal channel outputs monotonic, and it need not, therefore, be mathematically exact. The output voltage from a receiving network which, for a desired response, will have the form shown in line A, is applied to the input of the detector circuit. Full-wave rectification produces the wave shape as shown at B. Elements R and C, together with the left-hand half of the double-triode constitute an integrator of the Blumlein type. The plate of this tube, however, is connected to the plate of the right-hand triode which acts as a clamp during the early part of the digit interval, thereby preventing any substantial integrating action. At time $T + \Delta$ after the beginning of the input signal, the clamp tube is cut off by a negative pulse, derived from a timer circuit and applied to the Integrator Enablement lead; integration then proceeds as indicated in line D of the figure. The integrator output is sent to a sampling gate which blocks it except during a short period when a sampling pulse, also derived from the timer circuit, occurs. The sampling procedure is employed to present the signals from all five channels to a two-out-of-five code-checking circuit at the same instant; without this synchronizing feature the finite slopes of the integrator outputs might cause them to reach threshold value at different times and thereby confuse the checking circuit.

The wave-shapes shown in Fig. 5 refer to a desired response; in the case of an undesired response the input signal is substantially zero during the integration period and no output is produced.

With the signal parameters chosen, the total digit interval was $2T$ or 10 milliseconds. Of this time, 5 milliseconds was allotted for the action of the receiving networks, 1.3 milliseconds for Δ, the margin for timing error and transmission delay-distortion, and 3.2 milliseconds for the integration period. This latter includes the one millisecond occupied by the sampling pulse. Thus 0.5 millisecond remains for resetting the integrators to their initial condition, through the action of the clamp tubes, before the expiration of the 10 millisecond period.

5.22 Start Circuit

It was decided, as a matter of system design, that the transmitter should begin transmitting the desired number as soon as it was set up, without determining whether the receiver was ready to accept the information. The receiver, on the other hand, would begin to consider information received over the line at an unspecified time depending on the completion of its handling of the previous number. This design philosophy leads to the need for a device to which the name *start circuit* has been given. The receiver may find itself beginning to consider an incoming number at a time when some other digit than the first is arriving; it must be able to recognize in some way which digit in the train is the first digit of the number, and to wait until that digit arrives before beginning its operations.

The first digit is marked at the transmitter by the artifice of preceding it by a blank interval: the digits follow each other at regular intervals except that between the last digit of one iteration of the number and the first digit of the next iteration there is left an interval two digit-periods in length during which no signal is sent out. The blank interval is recognized in the receiver by means of an *R-C* timing circuit: a condenser is connected through a resistor to a high voltage and tends to charge up to the supply voltage. As each digit arrives, the condenser is discharged to zero and begins to charge again. There thus exists across the condenser a sawtooth voltage wave, which is monitored by a threshold circuit. As long as the digits arrive at regular intervals, the condenser voltage never rises beyond a certain maximum value which is less than the firing point of the threshold circuit. However, when the blank interval occurs, the condenser voltage is permitted to rise to about three times its normal maximum value. In so doing, it passes the firing point of the threshold circuit, which delivers a signal used to indicate that the next digit will be the first digit of the transmitted number. Whenever the receiver considers a new number, it waits until a blank interval occurs in the train of digits before beginning its operations.

In addition to recognizing the beginning of the number, it is necessary to recognize the arrival of each digit in turn. The sudden drop in voltage across the start circuit condenser is used to mark the beginning of each digit. This drop marks a well defined instant from which various necessary time intervals previously mentioned, can be measured. It will be recognized that the relationship of this instant to the beginning of the digit is not highly precise. The start circuit is necessarily an amplitude-sensitive device, since it must distinguish between the presence and

absence of signals. The sensitivity must be high enough so that the start circuit recognizes weak signals, but not so high as to cause false responses to noise on the line. Signals of the wave form shown in Fig. 1(a) or 1(d) have a definite starting point defined by a discontinuity in the derivative, but this discontinuity will tend to become obscured by transmission over any practical medium. The initial value of the derivative is different for different signaling frequencies, so that the time required to build up to any arbitrarily defined threshold level is not exactly known. It is these considerations, as discussed above in Section 4.1, that makes it impossible to use a detector based on the instantaneous value of the oscillating response of the receiving network. The envelope detector just described, however, is much less time-sensitive and can easily tolerate the lack of precision of the start circuit.

5.23 Receiver Block Diagram

A block diagram of the signaling receiver is given in Fig. 6. It is not proposed to describe the contents of each block in detail, but the sequence of operations in the handling of a received number will be traced out, and the functions of each block will be discussed.

Before describing the block diagram, it is well to point out that the system under discussion was built for purposes of laboratory test rather than practical use and that therefore it contains some features not necessary for plant service and lacks others that would be required. In particular, since in the laboratory it was always known what number was being transmitted, it was possible to include a comparator — a relay circuit which at the end of each receiver operation made a comparison between the transmitted number and the received number and indicated whether the number was correctly registered. Again, since received signals always arrived over the same pair of wires, no circuit was built to

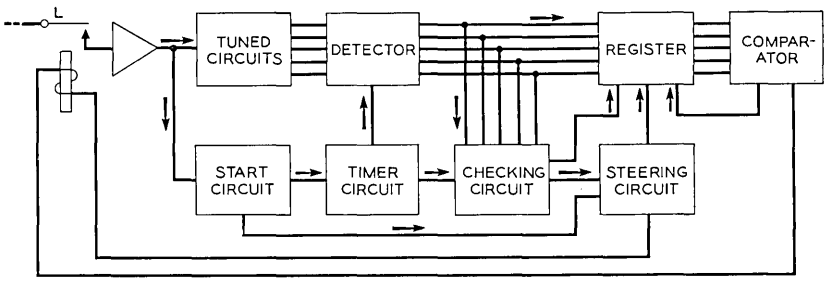


Fig. 6 — Block diagram of signal receiver.

perform the function of connecting the receiver to any one of a number of pairs to receive signals from a variety of sources. In a practical system it would be necessary, as previously mentioned, to send a "dismiss" signal back over the line to cause the transmitter to stop sending when the number was successfully received; this feature was not included here because it was felt to constitute a fairly simple design problem and one that had no special bearing on the particular research problem of interpreting and handling polytonic signals.

The block diagram of the system falls readily into two major divisions. The main detecting and registering functions are performed in the boxes across the top of the diagram, while the lower boxes perform auxiliary control functions.

The sequence of operations begins with the closing of the line relay (L), which connects the line to the input of the receiving amplifier. The L relay, in this laboratory version of the receiver, takes the place of the more elaborate connector circuit which would be required in a practical system with a large number of incoming lines.

When the L relay closes, the received signals, after amplification by the receiving amplifier, are fed to the receiving tuned circuits and to the start circuit. The L relay will close at a random point in the train of signals; one function of the start circuit is to make the register unresponsive to received signals until the blank interval in the train has occurred, thus ensuring that registration of a received number will begin with the first digit of a transmitted sequence.

Another function of the start circuit is to provide a timing reference to mark the beginning of each digit in the received sequence. Starting from this reference, the timer circuit measures off the interval $T + \Delta$ mentioned above and then measures off the integrate period which begins at time $T + \Delta$ and lasts just over three milliseconds.

When a received digit composed of two code elements is applied to all the five receiving tuned circuits, there appear on the five tuned-circuit output leads two desired responses and three undesired responses. These responses are considered, during the integrate period determined by the timer, by the five detectors. At the end of the integrate period there appear pulses on the output leads of the two detectors which were considering desired responses while, ideally, no pulses appear on the other three leads. These pulses are fed to the register circuit.

Until the blank interval occurs in the train of received signals, the register circuit does not respond to the pulses from the detectors. When a

blank interval does occur, the start circuit primes the steering circuit, which enables the first section of the register. The first digit of the transmitted number is then stored.

The outputs of the five detectors are also fed to a checking circuit, which determines whether a desired signal was recognized in exactly two of the five channels. If the digit is plausible, as indicated by two-out-of-five response, the checking circuit advances the steering circuit, so that the next digit can be stored in the next section of the register. If any digit is implausible, as indicated by response in more or less than two channels, the checking circuit does not advance the steering circuit but instead "recycles." That is, it erases all information stored in the register before the arrival of the questionable digit, and it returns the steering circuit to the state that existed before the recognition of the blank interval. The receiver then attempts on a succeeding repetition of the transmitted number to recognize a complete sequence of plausible digits.

For the purpose of simplifying the laboratory set-up, a skeletonized register capable of registering only the first three digits of the number was built. All eight digits, however, are checked. If all digits are plausible, the steering chain delivers a "number complete" signal, which releases the line relay, on the next appearance of the blank interval. The operation of the receiver is then complete.

The storage of the received number in the register represents the final output of the receiver. For laboratory purposes the information was used to operate various data-accumulating message registers as well as the comparator mentioned above. Upon completion of their operations, the register is cleared, and the receiver is then ready to handle another number.

6.0 TESTS

Conditions affecting the successful operation of a signaling system may arise in the transmitter, in the receiver, or in the transmission medium connecting the two. Laboratory tests, in which one condition at a time could be introduced and varied in a controlled manner, were first undertaken to evaluate the capabilities of the system. The results of these tests were so encouraging that arrangements were then made to try the system on a representative variety of toll circuits available on a loop basis out of New York. Good operation was obtained over two-wire voice-frequency circuits up to 175 miles in length and over four-wire

voice-frequency circuits up to 350 miles in length. Seven hundred and fifty miles of voice-frequency four-wire circuit caused signaling failure owing to excessive delay-distortion. Two links of K carrier, aggregating 1900 miles, gave satisfactory performance with some restriction in the allowable range of amplitude of signals.

7.0 HIGHER SPEED SIGNALING — LOCAL LINES

The system just described was designed with the objective of tolerating large amounts of certain kinds of distortion. After the construction and testing of this system was completed, investigation was concentrated on the design of a system which would be capable of still higher signaling speed, even though it might be more vulnerable to distortion. Such a system would be useful in the local plant where the most severe types of distortion — carrier shift and delay distortion — do not occur.

In designing a system to signal at higher speed, it is obvious that the time interval T must be shortened, by spacing the signal frequencies further apart. In addition, a further increase of speed can be obtained by shortening the "handling" interval between digits. In the previously described system, it will be remembered, digits were transmitted at intervals of $2T$. This timing allowed an interval T for signal discrimination to appear in the network responses, and another interval T for detection and manipulation of the digit information. By spacing the signal frequencies at intervals of 400 cycles, which shortened T from 5 milliseconds to 2.5 milliseconds, and by reducing the handling time allowance from T to $\frac{1}{3}$ T, it was possible to raise the signaling speed from 100 digits per second to 300 digits per second and to signal successfully at this speed over local loops. A new type of detector was required for this embodiment. The rest of the system was unchanged in principle from the previous design, so it will not be described.

7.1 Energy Detection

The previously described detector merely observed a voltage manifestation of the energy in the receiving network without making direct use of that energy. After detection of a digit signal, networks exhibiting desired responses still contained substantial amounts of energy, which decayed at a slow rate. With the altered signal parameters used in the higher-speed system, there was a likelihood that the decaying oscillations of a desired response might last into the next digit interval and obscure the discrimination for that digit. For this reason, and also be-

cause the available time for signal handling was much shorter, the new detector was devised with the dual objective of recognizing the presence of a desired response quickly and also of terminating as quickly as possible the slowly-decaying oscillations characteristic of a desired response. To accomplish this objective, the energy is used directly: it is made to advertise its own presence and simultaneously dissipates itself as quickly as the network permits. Because of the direct use of the energy, the detector is called an energy detector.

The scheme consists merely of inserting in series with the receiving resonant circuit, at time T, a resistance of such value as to make the circuit approximately critically damped. Then practically all the energy in the network is dissipated in the resistance within about one-half cycle of the resonant frequency. A simple circuit to perform this operation is shown in Fig. 7. The critical damping resistance is in the circuit at all times, shunted by a varistor through which steady current flows. As long as the steady current is larger than the peak amplitude of the oscillatory current in the tuned circuit, the varistor remains a low impedance and the circuit behaves as if the damping resistance were not there. When the time for detection arrives, the steady current is removed by driving to cut-off the vacuum tube through which it is supplied. If, at this time, the oscillatory current happens to be flowing in the direction for which the varistor is a low impedance (counter clockwise in the figure) nothing happens; but when the current reverses direction, the energy is dissipated in the damping resistance, producing a voltage pulse which can be used to actuate a storage register. The varistor serves two purposes: not only does it constitute a convenient, fast-acting switch for introducing the damping resistance, but also it ensures that the output pulse will always have the same polarity when it occurs. If the detection is initiated

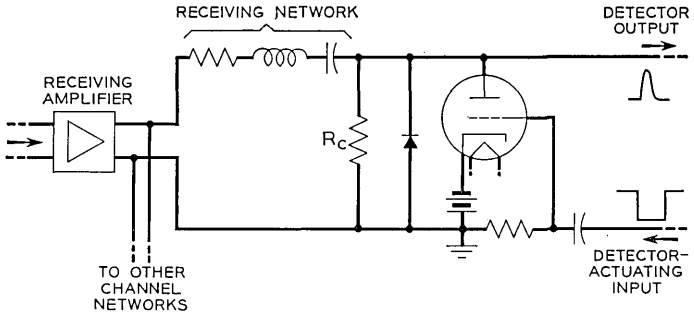


Fig. 7 — Circuit for polytonic signal energy detector.

at a time when the oscillating current is flowing in such a direction as to tend to give an output pulse of the wrong polarity, the output is simply delayed by an amount not greater than one-half period of the oscillation frequency.

8.0 CONCLUSION

A research program in high-speed signaling has resulted in laboratory models of several experimental systems based on different principles. This paper traces one of these principles from its origin in a mathematical concept to its application in two versions of workable physical circuitry. Both versions were found to give satisfactory results in laboratory tests and in limited trials with actual plant circuits.

Tapered Velocity Couplers

By J. S. COOK

(Manuscript received March 15, 1955)

The advent of very broad band microwave amplifiers and oscillators necessitates microwave circuitry of comparable frequency range. A scheme is presented which makes it possible to build microwave couplers of various kinds having bandwidths of well over two octaves. A preliminary investigation of mechanical and mathematical analogs is described; and some particular devices using the new principle are mentioned.

INTRODUCTION

In waveguide and transmission line networks, it is often desired to transfer power from one line to another. To this end, various hybrid junctions and ordinary directional couplers were developed some years ago.

More recently, S. E. Miller[1] has made use of the fact that when two transmission lines with equal phase velocities are continuously coupled over some length, a signal introduced into one line will be completely transferred periodically back and forth between the lines. (Fig. 1.) This principle is employed in power-"splitting" devices where power is, for instance, equally divided between two lines. Half of the signal power is permitted to transfer from one line to the other and at this point the coupling is discontinued. (Fig. 2(a).) Another obvious application of this coupling principle is a means to effect complete transfer of a signal from one line to another. (Fig. 2(b).)

Now the trend is toward systems which are capable of handling even greater bandwidths. Miller's approach to the problem of coupled waves has led to waveguide couplers which are useful over bandwidths of 25 per cent and better. However, with the advent of traveling-wave tube amplifiers and similar devices, useful bandwidths of two-to-one and more have become desirable. The bandwidth of the Miller scheme is limited because the strength of coupling and the electrical length of the coupling

[1] S. E. Miller, Coupled Wave Theory and Waveguide Applications, B.S.T.J., **33**, pp. 677–692, May, 1954.

section vary with frequency. This directly affects the power division between the lines. In the following two paragraphs are the background and thought which led to the new broadband coupling scheme.

Miller[1] has shown that if two coupled lines have a fixed difference in phase velocities the signal power transfer also is periodic. In this case, however, only a part of the power is transferred, the amount depending on the ratio of velocity difference to coupling factor. (Fig. 3.)

Assume, now, that two coupled lines have slightly different phase velocities and a signal is introduced into one line. Let the velocity difference increase slightly at the point where maximum energy has been transferred to the adjoining line. Repeat this process until the amount of power exchange becomes vanishingly small. It is conceivable that if the velocity

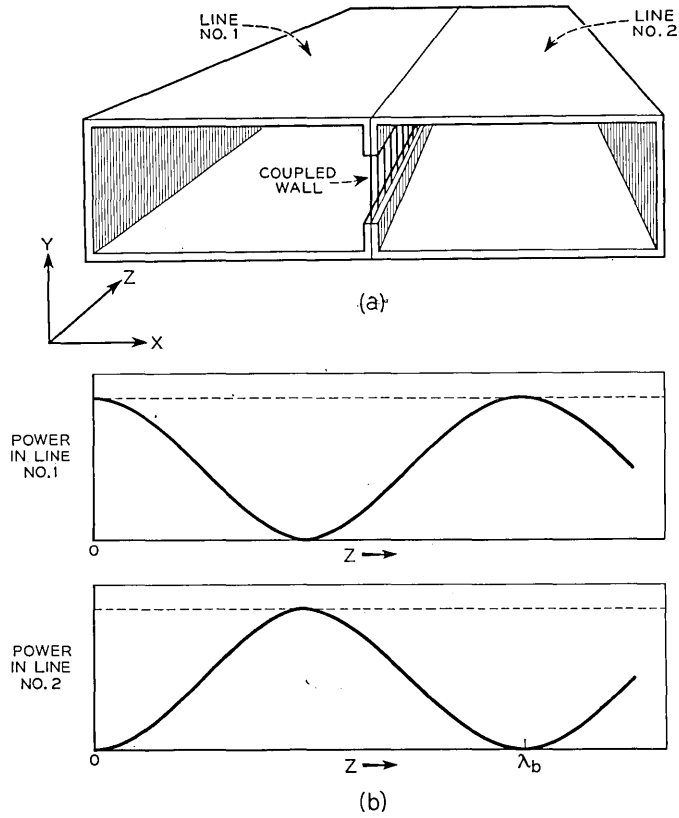


Fig. 1 — (a) Coupled waveguide transmission lines. (b) Wave power distribution along the lines.

steps are small the signal will finally be equally divided between the lines. (Fig. 4.) If this is so, it follows that the same effect must result from a continuous gradual change in the relative phase velocities of the two lines. Furthermore, if the coupling section is made long enough, this device should be frequency independent.

MECHANICAL ANALOG

It has long been recognized that coupled pendulums are analogous to coupled transmission lines. This analogy is demonstrated in the Appendix. To test the above conclusions about lines with changing phase velocities two pendulums were coupled together in such a way that their relative lengths could be adjusted while in motion. (Fig. 5.) It was ob-

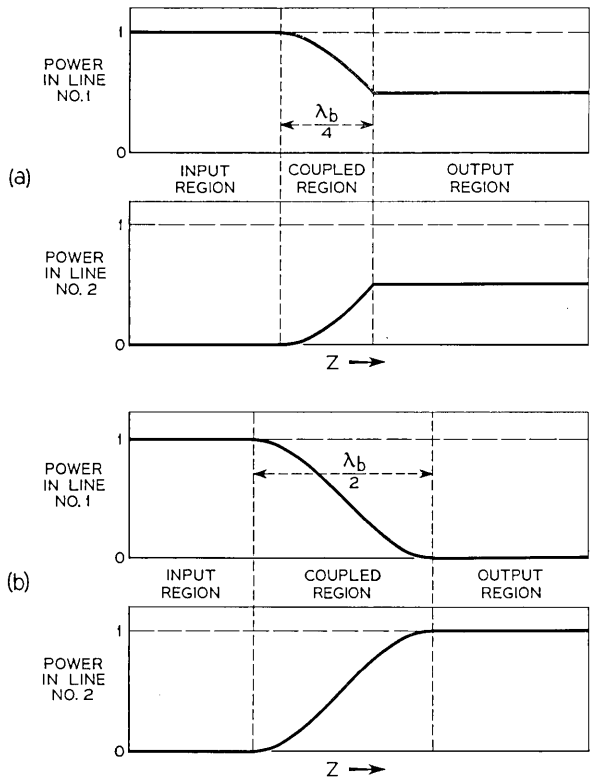


Fig. 2 — (a) Power distribution through a half-power coupler. (b) Power distribution through a complete-power coupler.

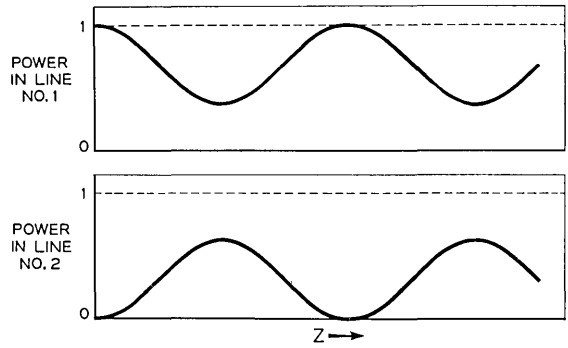


Fig. 3 — Incomplete power transfer resulting from coupling two lines having different phase velocities.

served that when the pendulum lengths were fixed and equal, and one pendulum was set swinging, it would swing less and less until it stopped altogether, while the other pendulum would swing higher and higher until it reached a maximum. Then the process reversed until the first pendulum reached a maximum and the second pendulum came to rest once more. This cycle repeated until friction finally brought both pen-

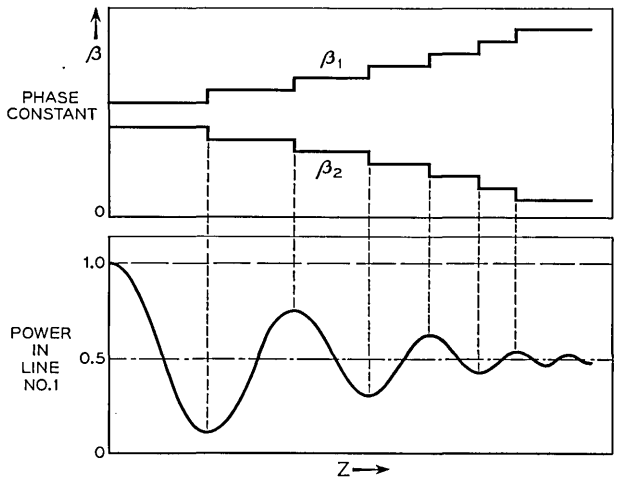


Fig. 4 — Power distribution in the driving line where the difference in phase velocity between two coupled lines is increased each half beat wavelength.

β_1 = phase constant of line 1

β_2 = phase constant of line 2

dulums to rest. This is a classical experiment performed in most early physics courses.

Now one pendulum was made quite long and the other relatively short. The longer pendulum was set swinging. It was then gradually shortened and the other gradually lengthened until they were finally of equal length. At that point they were swinging nearly together and with about the same amplitude. It was found, also, that if the short pendulum instead of the long was set swinging, and then the pendulum lengths were gradually equalized, the pendulums ended up swinging opposite to each other and again, of course, with about equal amplitudes. The initial and final conditions of the pendulums in these two experiments are indicated in Fig. 6. Although these experiments are the inverse of the original proposition reciprocity must hold, and the proposition, is thus valid.

It is known that a fixed system of two coupled pendulums will support two independent or "normal" modes of oscillation. These two normal modes are characterized by the pendulums swinging exactly "in phase"

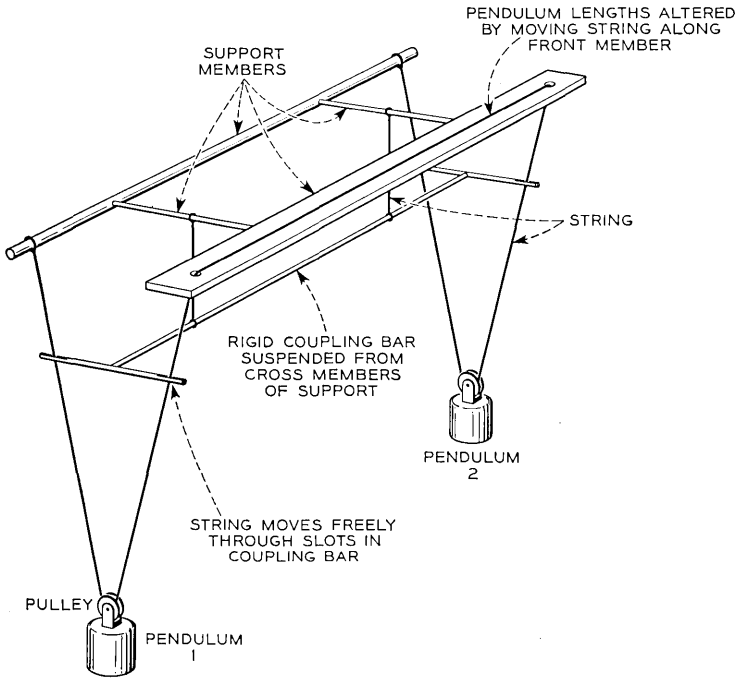


Fig. 5 — A system of two coupled pendulums whose lengths may be altered while in motion.

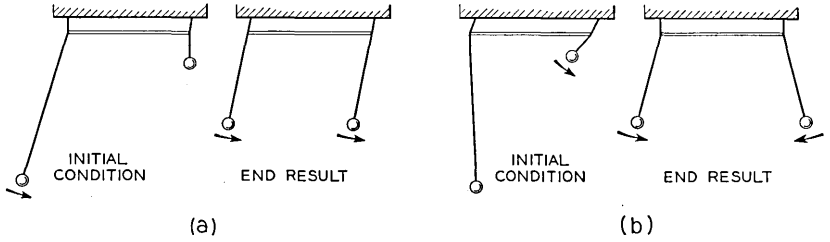


Fig. 6 — Initial and final states of oscillation of a system of two coupled pendulums where, (a), only the longer one is initially excited, and (b), only the shorter one is initially excited. In each case the lengths have been slowly equalized to produce the end results.

or exactly "out of phase" with each other. The end results of the two excitations in the experiment, then, were the two normal modes.

MATHEMATICAL ANALOG COMPUTER

As is shown in the Appendix, the pendulum analogy could not provide true quantitative solutions in terms of the transmission line constants. For this reason, and because of the mathematical complexity encountered in trying to solve the transmission line equations directly, we resorted to an analog computer. The coupled transmission line equations as proposed by Miller and given in the Appendix were appropriately rationalized and adapted to permit the phase constants of the lines to vary linearly with distance. Symbolically:

$$\beta_1 = \beta + \mu z, \qquad \beta_2 = \beta - \mu z \tag{1}$$

where β_1 = phase constant of line 1

β_2 = phase constant of line 2

β = initial phase constant of both lines

z = distance along coupled system

and μ = constant which determines rate of phase change.

In the process of rationalizing it was found convenient to measure z in terms of coupling wavelengths, λ_{b0}. The latter quantity is defined as the hypothetical distance required to transfer all the power from one line to the other and back again where $\beta_1 = \beta_2 = \beta =$ constant. If we let

$$\zeta = 2\pi z/\lambda_{b0}$$

we can write (1) as

$$\beta_1 = \beta\left(1 + \frac{\zeta}{2\pi n}\right), \qquad \beta_2 = \beta\left(1 - \frac{\zeta}{2\pi n}\right) \tag{2}$$

where n = number of coupling wavelengths the problem is allowed to run, i.e., until $\beta_2 = 0$, or $0 \leqq \zeta \leqq 2\pi n$.

The quantity E_1^2 was plotted as a function of ζ where E_1 was the wave amplitude in line 1 and $E_1(0) = 1$, $E_2(0) = 0$.

These plots are shown in Fig. 7 for $n = 2$ and $n = 6$. The dotted lines roughly indicate the centers of oscillation of these curves.

Inspection of the curves shows that:

1. The accuracy with which the power was ultimately divided between the two guides seemed to be determined within the first two or three exchange cycles.

2. The ultimate phase constant difference here chosen was not great enough since there was too much power still being exchanged at the end of the plot.

3. The exchange amplitude was much too constant after the first few cycles. This indicates that a linear phase constant change is inefficient if minimum coupler length is to be realized.

4. Although, initially, a slow rate of change of phase velocity is necessary, subsequently, a progressively faster rate should reduce the final exchange amplitude.

From the dotted lines indicating the centers of oscillation of the curves it may be seen that the initial rate of phase change for $n = 2$ is too great while that for $n = 6$ is slow enough to keep the oscillation fairly well centered.

It can be shown from Miller's equations for coupled lines of different fixed phase velocities[1] that the peak-to-peak amplitude of the power exchange between the lines, P_x, when the distance-averaged powers in the lines are equal is given by the following equation.

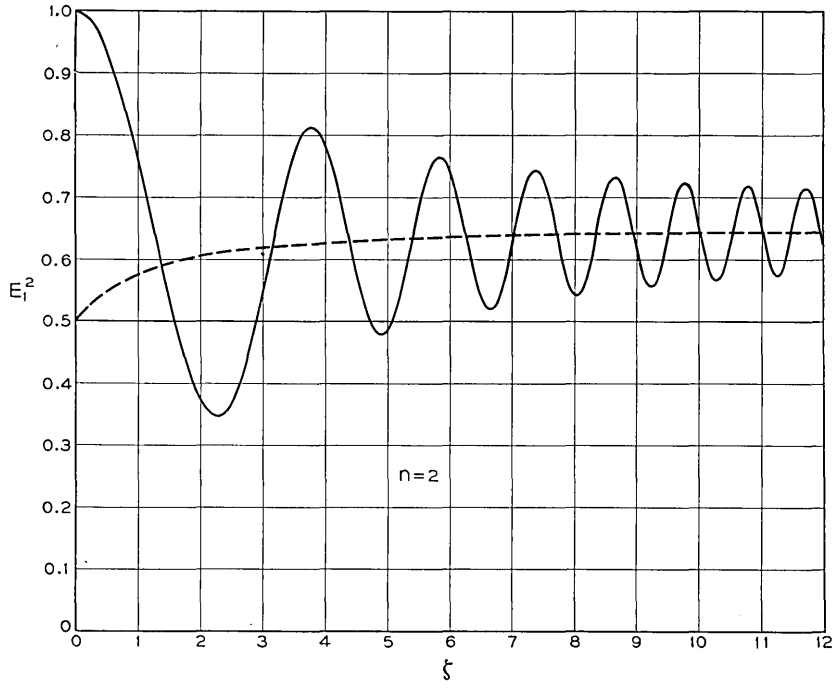
$$P_x = \frac{P}{\sqrt{\left[\frac{\Delta\beta}{2c}\right]^2 + 1}} \tag{3}$$

where P = total power in the coupled system,

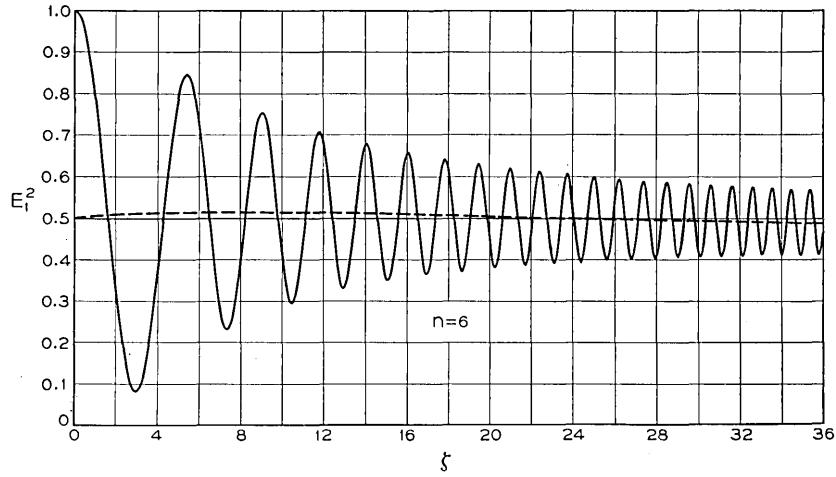
$$\Delta\beta = \beta_1 - \beta_2$$

and c = coefficient of coupling between the lines.

In spite of the fact that this calculation assumes a constant velocity



(a)



(b)

Fig. 7 — Power distribution in the driving line where the difference in phase constant between the two lines is increased linearly. Curve (a) shows the result of increasing too rapidly, (b) the result of an acceptable rate of increase. These curves are reproductions of those recorded by a mathematical analogue computer. Curve (b) discloses that there was a small amount of computer "drift."

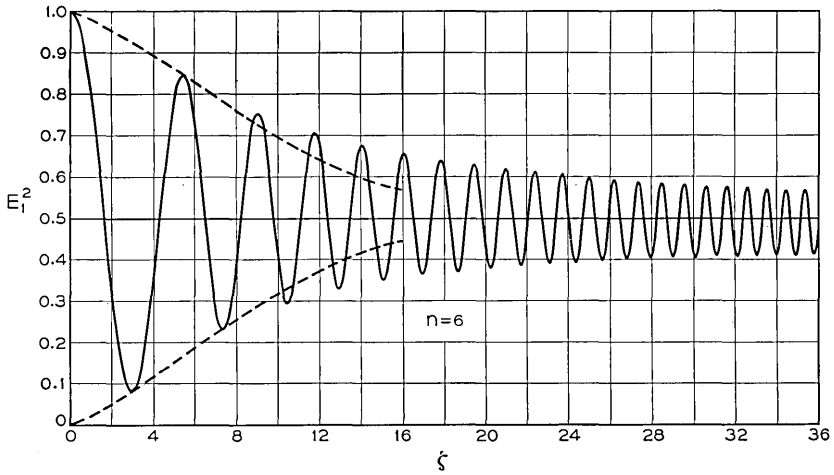


Fig. 8 — A reproduction of the curve of Fig. 7(b) showing the proposed envelope which might result from a more efficient coupler.

difference it can be shown from Fig. 7(b) that it provides a good approximation for the power exchange at any point along a coupled system where the phase velocities are varying slowly. Equation (3) indicates that the final power exchange amplitude may be reduced either by increasing the final velocity difference or by decreasing the coupling.

The next question is how to choose a variation of phase velocities in such a way that the power will be well divided in the shortest possible distance. Assume for the present that the velocity difference between the lines is constant throughout each power exchange shown in Fig. 7(b). An envelope of oscillation amplitude was arbitrarily superimposed on the $n = 6$ curve as shown in Fig. 8. It was proposed to find a law of velocity change which would produce such an envelope. From the equations for fixed velocity difference the value of $\Delta\beta/2c$ was calculated which would produce the power amplitude predicted by this envelope after each power exchange. The value thus calculated was considered as the value half way through the exchange distance, this distance also being determined by the particular value of $\Delta\beta/2c$. These points were then plotted as a function of ζ as shown in Fig. 9. The solid line represents the linear variation actually used for the computed curve. Note that the first three points lie very close to this line proving the validity of the approximation.

W. H. Louisell has made a thorough mathematical examination of this sort of coupler.[2] It is interesting to note that Louisell's analysis

[2] B.S.T.J., page 853 of this issue.

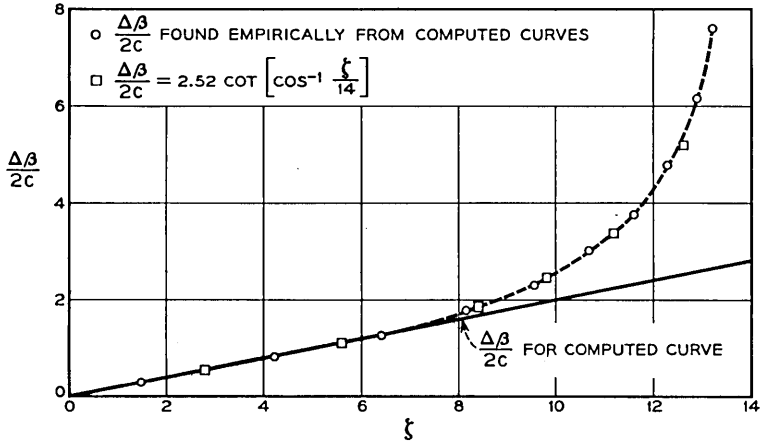


Fig. 9 — The solid line shows the linear variation of $\Delta\beta/2c$ used on the analogue computer to produce the curve of Fig. 7(b). The broken line is the variation of $\Delta\beta/2c$ calculated to produce the envelope superimposed on that curve in Fig. 8.

indicates that when the coupling is held constant the phase velocity difference, $\Delta\beta$, should vary as follows for greatest efficiency:

$$\Delta\beta \sim \cot[\cos^{-1} \zeta/2\pi n]$$

where n = the length of the coupler measured in coupling wave lengths, λ_{b0}. A suitable multiplier and a value for n were chosen for the above formula to make it coincide with the empirical curve of Fig. 9 as nearly as possible. Points found from the resulting equation appear as boxes on that figure.

WAVEGUIDE COUPLER

Using the function of $\Delta\beta/2c$ thus calculated, an experimental coupler was made. Two rectangular waveguides (1.145″ x 2.290″) were placed side by side so that they had a common narrow wall. They were electrically coupled through a slot like that shown in Fig. 1. One guide was loaded with a gradually increasing amount of polystyrene such as to taper the phase velocity to match the curve of Fig. 9. Since the phase constant difference $\Delta\beta$ could not be made great enough by this method, the coupling near the end was decreased by tapering the slot width linearly to zero through the section where $\Delta\beta$ was maximum, i.e., where the loaded guide was completely filled with polystyrene. To reduce reflections the other end of the slot was also tapered to zero, though over a much shorter distance. Finally, the polystyrene at the fully

loaded end was tapered off outside the coupled section slowly enough to provide a low reflection transition to unloaded guide. The effective coupling length was about 88″.

When a signal was fed into either guide at one end of the coupler the signal output at the other end was equally divided between the two guides within 0.2 db over a frequency range of 2,700 to 3,100 mc/s. Since guide cutoff is about 2,600 mc/s this frequency range represented a large range of λ_{b0}. The coupling wavelength, indeed, varied from about 18″ at 2,700 mc/s to 40″ at 3,100 mc/s. The coupler accuracy degenerated rapidly for frequencies above 3,100 mc/s both because the coupling wavelength got too large compared with the length of the coupler and because of large reflections due, possibly, to the TE_{20} mode which can exist above about 3,200 mc/s in the fully loaded guide.

APPLICATIONS

One rather serious shortcoming besets this kind of coupler. It is inherently many wavelengths long. This presents a physical size problem for fast-wave structures at frequencies of 5000 mc/s and lower, and a loss problem at all frequencies. Thus, some of the most practical applications are found in connection with traveling-wave tubes where the associated helices provide an ideal medium for varying phase and coupling constants on a slow wave structure. If two helices are wound in opposite sense and mounted coaxially, strong coupling occurs between them.[3] If the axial phase constants of the two helices are equal, and an out-of-phase normal mode is launched on the system, the electric field in the region between the helices will be oriented transverse to the axis. It has long been proposed that an axially directed electron beam in such a transverse field region would provide low noise amplification.[4] C. F. Quate has suggested tapering the relative pitch of the two helices (Fig. 10) at the beginning such as to launch the transverse normal mode over a wide frequency range by introducing the signal to the outer helix only. By reversing the relative pitches at the input in Fig. 10, an in-phase or or longitudinal, normal mode could be excited.

Coupling power into and out from a traveling-wave tube over a wide frequency range has long been a problem. R. Kompfner[5, 6] has sugested using a short section of helix after the principle Miller used, i.e., an outer

[3] J. R. Pierce, Travelling Wave Tubes, pp. 46–47, D. Van Nostrand, 1950.
[4] J. R. Pierce, Travelling Wave Tubes, p. 156, D. Van Nostrand, 1950.
[5] Kompfner, Experiments on Coupled Helices, A.E.R.E. Report No. G/M98, Sept., 1951.
[6] R. Kompfner, Coupled Helices, Paper presented at I.R.E. Electron Tube Conference, 1953, Stanford, Cal.

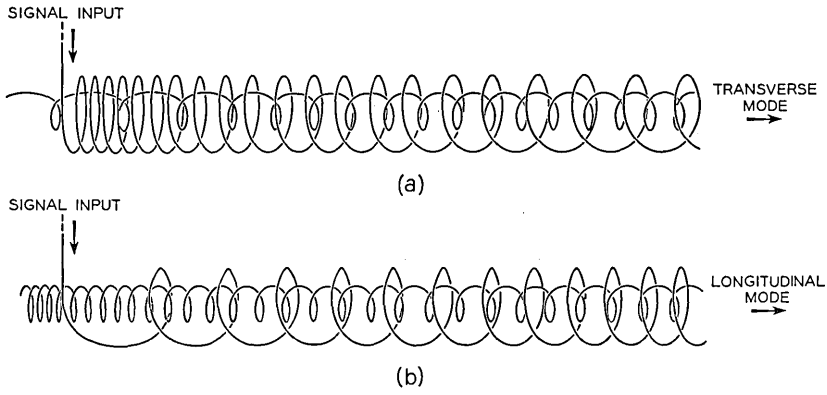


Fig. 10 — Coupled helices arranged so that essentially pure, (a), transverse or, (b), longitudinal normal modes may be excited over a wide frequency range by introducing the signal to the outer helix only.

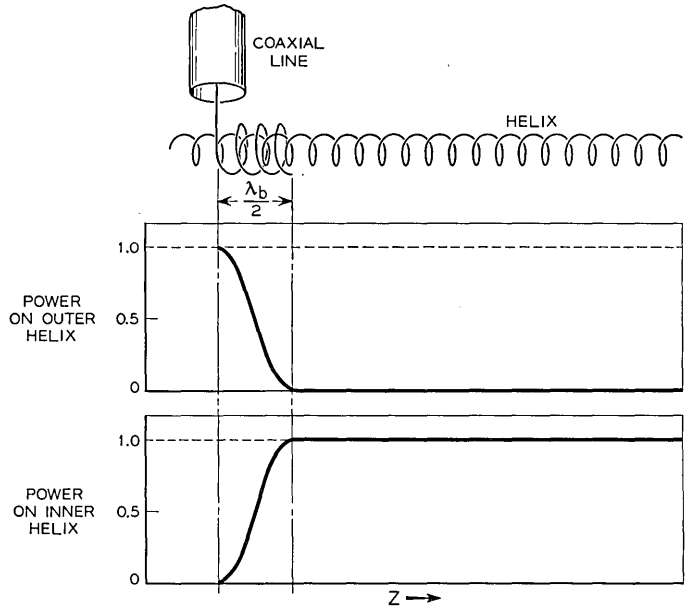


Fig. 11 — Helix coupler proposed by R. Kompfner and currently in common use for coupling to traveling-wave tubes.

helix concentric with the traveling-wave tube helix (Fig. 11) and just long enough for all the power launched on it from a matched coaxial line to be transferred to the inner helix. Under certain conditions of coupling and geometry this kind of coupler may have a useful frequency range of two to one. There are times, however, when even wider bandwidths are desirable. The new coupling scheme again may be applied. Let the inner helix turns be wound very close at point "A" (Fig. 12) and then the pitch increased gradually, until it is a very loose or "fast" helix at point "B". By constrast let the outer helix be very fast at A and slow at B. At some point between A and B the helix phase velocities will be equal. Now if a wave is launched on the outer helix at A it will appear as the in-phase normal mode at the point where the helix velocities are equal. As it proceeds along the coupler the wave energy will again be

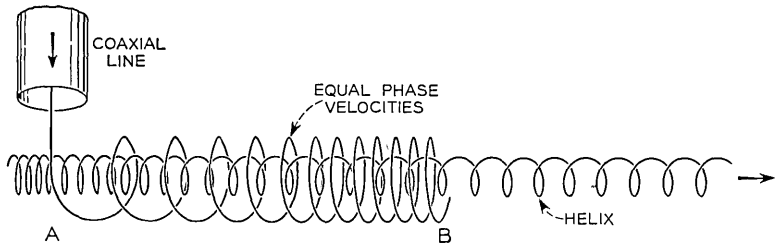


Fig. 12 — A tapered-velocity coupler to a traveling-wave tube helix. Though the inner helix is here shown tapered, it might, in practice, be made uniform if the outer helix is sufficiently tapered.

transferred to the faster helix which now is the inner one. At point B, then, the wave energy will be entirely on the inner helix. In other words, if the coupling is not too great and the taper sufficiently long this device is a 100 per cent, or complete, coupler. It turns out that such a coupler on a 4,000 mc/s traveling-wave tube may be less than 2″ long and operate effectively over a bandwidth of three or more octaves.

Since a helix may be matched to a coaxial transmission line over a considerable bandwidth, and since a 3-db coupler may be made with concentric helices to cover a similar frequency range, it should be possible to build a compact hybrid junction with coaxial connectors which will operate usefully throughout a bandwidth of three octaves.

CONCLUSION

Using distributed coupling between transmission lines with tapered phase velocities has made possible extremely broad band directional couplers. This principle can be used to give an equal and accurate division

of power delivered to two conjugate lines (a hybrid coupler) or to give essentially complete power transfer from one line to the other.

Probably the most useful applications are to traveling-wave tubes. A very broad band coupling can be made through the envelope to a standard traveling-wave tube helix. The principle also provides a means to couple into and out from either normal mode of a double, concentric traveling-wave tube helix.

Finally, it should be possible to make a broad band (3 or more octaves) hybrid junction by coupling to tapered helices from coaxial lines.

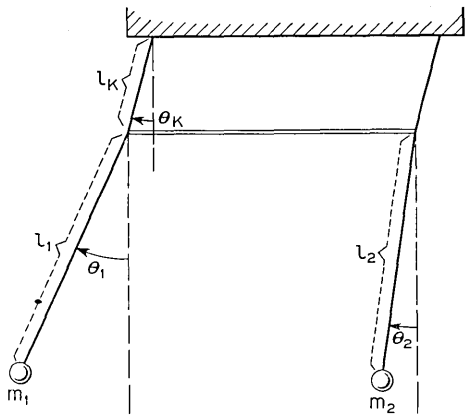


Fig. 13 — A system of two coupled pendulums with the symbols used in the appendix.

APPENDIX

Fig. 13 shows the coupled pendulums and the symbols used in the following analysis. From the energy equations for the system the Lagrangian and Hamiltonian equations were found. All third order and greater terms were dropped and since no second order terms appeared the following linear equations resulted.

$$m_1\ell_1\ddot{\theta}_1 + m_1\ell_k\ddot{\theta}_k = -m_1g\theta_1 \tag{A1}$$

$$m_2\ell_2\ddot{\theta}_2 + m_2\ell_k\ddot{\theta}_k = -m_2g\theta_2 \tag{A2}$$

$$(m_1 + m_2)\ell_k\ddot{\theta}_k + m_1\ell_1\ddot{\theta}_1 + m_2\ell_2\ddot{\theta}_2 = -(m_1 + m_2)g\theta_k \tag{A3}$$

Substituting equations (A1) and (A2) in (A3) we find

$$m_1g\theta_1 + m_2g\theta_2 = (m_1 + m_2)g\theta_k \tag{A4}$$

In our experimental pendulums the masses were equal so

$$m_1 = m_2$$

and equations (A1), (A2), and (A4) become

$$\ell_1\ddot{\theta}_1 + \ell_k\ddot{\theta}_k = -g\theta_1 \tag{A5}$$

$$\ell_2\ddot{\theta}_2 + \ell_k\ddot{\theta}_k = -g\theta_2 \tag{A6}$$

$$\theta_1 + \theta_2 = 2\theta_k \tag{A7}$$

Let us choose S_1 and S_2 as a measure of pendulum displacement from rest; and define it as

$$\begin{aligned} S_1 &= \ell_1\theta_1 + \ell_k\theta_k\,, \\ S_2 &= \ell_2\theta_2 + \ell_k\theta_k \end{aligned} \tag{A8}$$

Then

$$\begin{aligned} \ddot{S}_1 &= \ell_1\ddot{\theta}_1 + \ell_k\ddot{\theta}_k \\ \ddot{S}_2 &= \ell_2\ddot{\theta}_2 + \ell_k\ddot{\theta}_k \end{aligned} \tag{A9}$$

Substitute (A9) in (A5) and (A6)

$$\begin{aligned} \ddot{S}_1 &= -g\theta_1 \\ \ddot{S}_2 &= -g\theta_2 \end{aligned} \tag{A10}$$

From (A7) and (A8) find

$$\begin{aligned} \theta_1 &= \frac{\ell_k S_1 + 2\ell_2 S_1 - \ell_k S_2}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \\ \theta_2 &= \frac{\ell_k S_2 + 2\ell_1 S_2 - \ell_k S_1}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \end{aligned} \tag{A11}$$

Substitute (A11) in (A10)

$$\begin{aligned} \ddot{S}_1 &= -g\,\frac{[\ell_k S_1 + 2\ell_2 S_1 - \ell_k S_2]}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \\ \ddot{S}_2 &= -g\,\frac{[\ell_k S_2 + 2\ell_1 S_2 - \ell_k S_1]}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \end{aligned} \tag{A12}$$

Miller has expressed the general wave equations for two coupled

lines as

$$\begin{aligned} \frac{dE_1}{dz} &= -j(c + \beta_1)E_1 + jcE_2 \\ \frac{dE_2}{dz} &= -j(c + \beta_2)E_2 + jcE_1 \end{aligned} \tag{A13}$$

Differentiate (A13) and make the proper substitutions to find

$$\frac{d^2E_1}{dz^2} = -c(2c + \beta_1 + \beta_2)E_1 - [c(\beta_1 - \beta_2) + \beta_1^2]E_1 + c(2c + \beta_1 + \beta_2)E_2$$

$$\frac{d^2E_2}{dz^2} = -c(2c + \beta_1 + \beta_2)E_2 - [c(\beta_2 - \beta_1) + \beta_2^2]E_2 + c(2c + \beta_1 + \beta_2)E_1$$

where E_1 = wave amplitude in line 1

E_2 = wave amplitude in line 2

β_1 = wave phase constant in line 1 before coupling

β_2 = wave phase constant in line 2 before coupling

c = coupling constant

Now equations (A12) and (A14) are of exactly the same form, namely:

$$\begin{aligned} \ddot{S}_1 &= -AS_1 - BS_1 + AS_2 \\ \ddot{S}_2 &= -AS_2 - BS_2 + AS_1 \end{aligned} \tag{A15}$$

Thus the solutions for the two sets of equations differ only in their constants. The constants have the following correspondence:

$$\frac{g\ell_k}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \doteq c(2c + \beta_1 + \beta_2) \tag{A16}$$

$$\frac{2g\ell_2}{\ell_k(\ell_1 + \ell_2) + 2\ell_1\ell_2} \doteq c(\beta_1 - \beta_2) + \beta_1^2 \tag{A17}$$

The complexity of the constants seems to defy separating them out. Suffice it to say that ℓ_k roughly corresponds to c and ℓ_n to reciprocal β_n. Note, however, that if $\beta_1 = \beta_2 = \beta$ and $\ell_1 = \ell_2 = \ell$, (A17) becomes

$$\beta^2 \doteq \frac{g}{\ell + \ell_k} \tag{A18}$$

It may be recognized from the equations for a simple pendulum that $\sqrt{g/l + l_k}$ is, indeed, the phase constant of the uncoupled pendulums, just as β is, by definition, the phase constant for the uncoupled transmission lines.

Wave Coupling by Warped Normal Modes

By A. G. FOX

(Manuscript received February 23, 1955)

It has been shown by J. S. Cook that wave power may be transferred from one to another of two coupled waveguides through a variation of their phase constants. It is now clear that this is but one example of a new principle of coupling which is here called "normal mode warping." Wave power inserted at one end of a coupled waveguide system may be made to appear at the other end with any desired power distribution by gradual warping of the normal mode field patterns along the coupler. In general, variation of both the coupling coefficient and phase constants are required. Much wider bands are theoretically possible than with any other distributed type of coupler. This principle may be applied to dielectric waveguides, birefringent media, and waveguides containing ferrite, to obtain both reciprocal and non-reciprocal couplers.

INTRODUCTION

It is now well known that complete transfer of power can be effected from one to another of two waveguides provided there is distributed coupling between the waveguides and provided the phase velocities are equal.[1, 2] A good illustrative analog is a pair of coupled pendulums having the same period. The periods correspond to the wavelengths in the waveguides; passage of time for the pendulums corresponds to distance along the waveguides; the energies in the pendulums at a particular time correspond to the wave powers in the two waveguides at a particularly point along the waveguides. As energy is interchanged between pendulums with increasing time, power is interchanged between waveguides with distance.

To obtain a complete interchange of power for the waveguides requires a particular length which is determined by the coupling. As long as the

[1] S. E. Miller, Coupled Wave Theory and Waveguide Applications, B.S.T.J., **33**, p. 661, May, 1954. See this paper for additional bibliography on wave coupling.

[2] B. M. Oliver, Directional Electromagnetic Couplers, Proc. I.R.E., **42**, p. 1686, Nov., 1954.

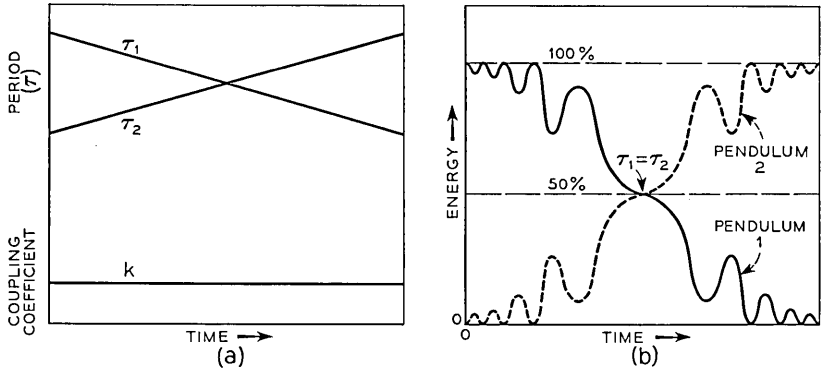


Fig. 1 — Transfer of energy from one to another of two coupled pendulums whose periods are varying with time.

coupling is constant, the power transfer should be independent of frequency. However, in practice, the coupling does vary with frequency, and hence the power division is a slowly varying function of frequency.

Recently a new and rather surprising method of transferring power has been described by J. S. Cook. He pointed out that if the phase constants of two waveguides are grossly *unequal* at one end, but are continuously varied so that they become *equal* in the middle of the coupling region and again grossly *unequal* in the opposite sense at the other end, complete power transfer should take place. Moreover this transfer is independent of the size of the coupling coefficient so that a coupler built in this way should be very broadband indeed. The conclusion is an unexpected one since we know that a uniform coupler having unequal phase constants can never give complete power transfer.

Nevertheless, the principle may be demonstrated using a pair of coupled pendulums whose lengths are continuously varied so that the one which is initially shorter finally becomes the longer of the two. In Fig. 1 is shown a typical result when the longer pendulum is initially excited. Fig. 1(a) shows how the periods (τ) and the coupling (k) vary with time. There is a fluctuation of energy which is quite small at first, but which increases until at the time the periods become equal, the energy is equally divided and the pendulums are in phase. (When the shorter pendulum is initially excited, they are 180 degrees out of phase.) With increasing time the energy is finally transferred to the other pendulum with small residual fluctuations which gradually diminish. It appears then that while the transfer of energy is almost complete, it will not be

complete unless the difference between the periods approaches infinity at the beginning and end of the process.

In looking for applications for Cook's coupling scheme, the writer discovered that by varying both the coupling coefficient and the phase constants of the waveguides simultaneously, the residual power fluctuations may be substantially eliminated. The design of such a coupler may appear to be complicated, but it turns out that the requirements are very simply expressed by a new principle of broad-band coupling which is here called "normal mode warping". Using this principle it should be possible to build wave couplers providing any desired degree of power division over very large bandwidths, limited only by the bandwidth capacity of the waveguides themselves. It may be applied equally well to nonreciprocal and reciprocal structures, and in a wide variety of ways. In this paper the principle of normal mode warping will be developed in terms of physical concepts and with the aid of some rather straightforward analysis.

NORMAL MODES VERSUS COUPLED MODES

Before attempting a discussion of normal mode warping it will be necessary to clarify somewhat the meaning of "normal modes." In a uniform metal waveguide, the usual understanding is quite adequate. Normal modes are characterized by unique distributions of transverse electric and magnetic field components, which distributions are independent of frequency or position along the waveguide. Also for a given frequency, the normal modes have unique phase velocities (except for certain pairs which are said to be degenerate because they have the same phase velocity). This description automatically requires that the normal modes be orthogonal, which means that the flow of energy in one mode does not contribute to energy flow in any other mode. If this were not so, the excitation of one normal mode would result in a transverse field pattern which would change with distance along the waveguide. These statements are equally true whether we are talking about modes in one waveguide or in a system of several waveguides.

When two identical dominant mode waveguides are uniformly coupled throughout a certain portion of their length, a field excitation corresponding to a dominant wave in one of the waveguides is no longer a normal mode. Such modes are now coupled, and power will cycle back and forth between them as stated earlier. However, there will be a new pair of normal modes which are orthogonal to each other, and for which the power is equally divided between the waveguides. One of these may be

called the even mode because the electric field will have the same direction in both waveguides. The other may be called the odd mode because the field will be oppositely directed in the two waveguides. While the two coupled modes will have the same phase velocities, the normal modes will have different phase velocities. The behavior of such a coupler can be explained either in terms of coupled modes, or in terms of normal modes, and these two concepts are completely equivalent.[3] According to the normal mode explanation, if all of the power is initially introduced into one of the waveguides, the subsequent transfer of power to the other waveguide is due to the excitation of both normal modes and to the interference between them which is a result of their unequal phase velocities. On the other hand, if all of the power were introduced into both waveguides with the proper phase and amplitude to correspond with one of the normal modes, it would travel through the coupler with the phase velocity of that mode and would emerge at the far end in that mode.

When the two coupled waveguides have different phase velocities, we will again have a pair of normal modes which are orthogonal to one another, but they will no longer have equal amounts of energy in the two waveguides. Instead, one will have more of its energy in one waveguide, and the other will have more of its energy in the other waveguide, the unbalance depending upon the difference in the phase velocities of the coupled modes and upon the magnitude of the coupling coefficient.

Finally, if the coupled waveguides are not uniform, but are "warped" so that their phase constants and coupling coefficient vary along the coupling region, the normal mode concept clearly requires re-examination. It is no longer possible to define a normal mode in terms of a characteristic and invariant field distribution in any cross sectional plane. The field distribution will change along the coupler. However, we will show that if the warping is sufficiently gradual, the normal modes will have transverse field patterns at any cross section which are approximately the same as the normal mode patterns for a uniform coupler having the same cross sectional properties. We will call the normal modes for the equivalent uniform coupler the "local normal modes" to distinguish them from the true normal modes for the warped coupler.

TWISTED WAVEGUIDE AS PROTOTYPE BROADBAND COUPLER

Perhaps the most easily understood example of normal mode warping is a twisted birefringent waveguide. Fig. 2 shows in cross-section a long

[3] A. G., Fox, Miller, S. E., Weiss, M. T. Behavior and Applications of Ferrites in the Microwave Range, B.S.T.J., **34**, p. 16, Jan., 1955.

dielectric strip which is twisting in a clockwise direction with propagation into the paper. We may think of this as being either an unshielded dielectric waveguide, or as a dielectric fin inside of a circular waveguide sheath. In either case we know from experiment that if we launch a linearly polarized wave with its electric polarization either parallel or perpendicular to the dielectric fin, it will propagate along the twist with its polarization remaining, to a first approximation, either parallel or perpendicular to the fin at all points. Thus, the polarization will rotate with rotation of the fin. We know experimentally that if the twist is performed too rapidly, some depolarization will result. On the other hand, if the twist is long and gradual, the polarization will remain quite linear at all points.

Now if the twist in the waveguide is just 90°, as suggested in Fig. 2, a vertically polarized wave introduced at one end will emerge as a horizontally polarized wave at the other end. Thus, we find that the twist section, by some coupling mechanism not yet defined, can transfer 100 per cent of the power in a vertically polarized wave to a horizontally polarized wave. Moreover, we know that this transfer is not frequency sensitive. We suspect, therefore, that the twist represents a preferred way of effecting a 100 per cent power transfer between the two modes. Let us arbitrarily identify the vertically and horizontally polarized modes as the modes between which coupling takes place, and ask how the coupling coefficient and phase constants for these modes vary along the twist.

We will define the phase constants and coupling coefficient at any point along the twist as being the same as those of a uniform (non-twisted) dielectric strip having the same cross sectional geometry. Let us now determine these parameters for the uniform birefringent waveguide indicated schematically in Fig. 3. If a vertically polarized wave

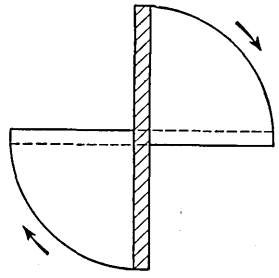


Fig. 2 — End view of a twisted dielectric waveguide.

with amplitude e_0 is launched on this waveguide, there will be a partial transfer of power from the vertical to the horizontal polarization and back again, just as though these two polarizations were coupled waveguides with unequal phase velocities. We can analyze the situation by breaking the input polarization into components along the cross sectional axes (A and B) of the waveguide, allowing for the difference in the phase constants β_a and β_b, and determining the vertically and horizontally polarized components which result at some other cross section. The result is:

$$e_v = e_0 \varepsilon^{-j\beta_0 z} [(\cos^2 \theta) \varepsilon^{-j\Delta z} + (\sin^2 \theta) \varepsilon^{+j\Delta z}] \tag{1}$$

$$e_h = -je_0 \varepsilon^{-j\beta_0 z} [\sin 2\theta \sin \Delta z] \tag{2}$$

where

$$\beta_0 = \frac{\beta_a + \beta_b}{2} \tag{3}$$

$$\Delta = \frac{\beta_a - \beta_b}{2} \tag{4}$$

(1) and (2) must now be compared with the equations for a pair of coupled waveguides.

The basic equations governing the amplitudes of waves in coupled waveguides are:

$$\frac{de_1}{dz} = -j\beta_1 e_1 + k_{21} e_2 \tag{5}$$

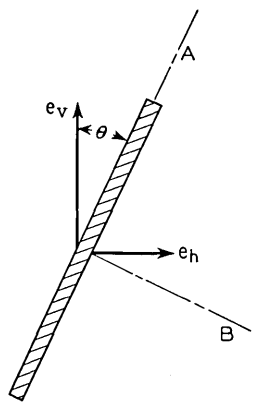


Fig. 3 — A uniform birefringent medium producing coupling between vertical and horizontal polarizations.

$$\frac{de_2}{dz} = -j\beta_2 e_2 + k_{12}e_1 \tag{6}$$

This has already been treated by S. E. Miller[1] using slightly different notation. For simplicity we are assuming that there is no attenuation. e_1 and e_2 are the amplitudes of the forward travelling coupled waves. These waves have phase constants β_1 and β_2 and field distributions which are perturbed by the presence of the coupling. Hence they are truly the coupled modes and not the modes which would exist in the absence of coupling. There must also be a pair of backward traveling waves in the two waveguides since power leaking through the coupling aperture will set up both forward and backward traveling components. However, if the coupling per unit length is small so that the coupler would have to be many wavelengths long to produce a complete power transfer, then the backward traveling components will interfere with one another, so that the backward waves may be safely neglected. This greatly simplifies the analysis and the weak coupling assumption will be made here and throughout the rest of this paper. Characteristic impedances are normalized so that the power in either mode is equal to the square of the amplitude (e). Propagation takes place in the $+ z$ direction. Power conservation requires that

$$|k_{12}| = k = |k_{21}| \tag{7}$$

$$k_{12}k_{21} = -k^2 \tag{8}$$

The solution of (5) and (6) then gives,

$$e_1 = \varepsilon^{-j\beta_0 z}\left[A_1\varepsilon^{-j(\sqrt{\delta^2+k^2})z} + B_1\varepsilon^{+j(\sqrt{\delta^2+k^2})z}\right] \tag{9}$$

$$e_2 = \varepsilon^{-j\beta_0 z}\left[A_2\varepsilon^{-j(\sqrt{\delta^2+k^2})z} + B_2\varepsilon^{+j(\sqrt{\delta^2+k^2})z}\right] \tag{10}$$

where

$$\beta_0 = \frac{\beta_1 + \beta_2}{2} \tag{11}$$

$$\delta = \frac{\beta_1 - \beta_2}{2} \tag{12}$$

and the coefficients A and B have the following relation:

$$\frac{A_2}{A_1} = j\,\frac{\delta - \sqrt{\delta^2 + k^2}}{k_{21}} \tag{13}$$

$$\frac{B_2}{B_1} = j\,\frac{k_{12}}{\delta - \sqrt{\delta^2 + k^2}} \tag{14}$$

The amplitudes of these two waves will oscillate with distance along the structure, and a beat wavelength may be defined as

$$\lambda_b = \frac{\pi}{\sqrt{\delta^2 + k^2}}$$

in which interval there will occur one complete power transfer cycle. If all of the power is initially in mode 1, then

$$A_1 = \left(\frac{1}{2} + \frac{\delta}{2\sqrt{\delta^2 + k^2}}\right) e_0 \tag{15}$$

$$B_1 = \left(\frac{1}{2} - \frac{\delta}{2\sqrt{\delta^2 + k^2}}\right) e_0 \tag{16}$$

By comparing (1) and (2), which give the amplitudes of vertically and horizontally polarized modes on the birefringent waveguide, with (9) and (10) for a pair of coupled modes, we can see that they are of the same form. They may be made identical if we make the following substitutions:

$$\beta_1 = \beta_a \cos^2 \theta + \beta_b \sin^2 \theta \tag{17}$$

$$\beta_2 = \beta_a \sin^2 \theta + \beta_b \cos^2 \theta \tag{18}$$

$$k = \frac{\beta_a - \beta_b}{2} \sin 2\theta \tag{19}$$

It follows also that

$$\frac{\beta_1 + \beta_2}{2} = \beta_0 = \frac{\beta_a + \beta_b}{2} \tag{20}$$

$$\delta = \frac{\beta_a - \beta_b}{2} \cos 2\theta \tag{21}$$

Thus, the behavior of the uniform birefringent medium can be described in terms of (9) and (10) for coupled transmission modes. Conversely, any pair of coupled waveguides can be described in terms of parameters for an equivalent birefringent medium.

We shall find it convenient to make use of this analogy since warped mode structures can be easily visualized in terms of an equivalent birefringent twist.

We are now in position to interpret the birefringent twist of Fig. 2 in terms of coupling between vertical and horizontal polarizations. Although θ is now variable, we use it to define at each point along the twist section a set of equivalent coupled line parameters corresponding to a

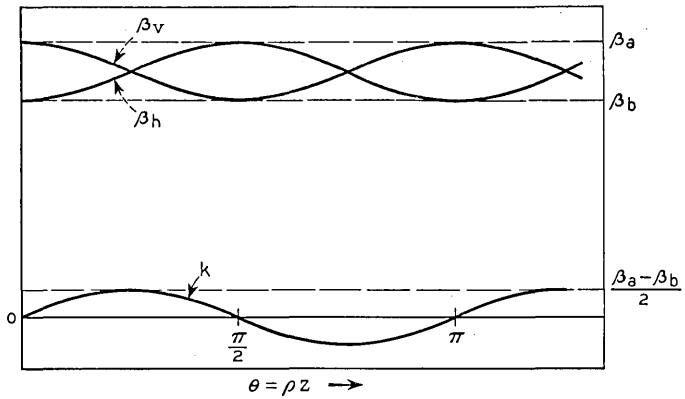


Fig. 4 — Coupling coefficient and phase constants for vertical and horizontal polarizations in a twisted birefringent medium.

non-twisted section having the angle θ. Fig. 4 shows how the local coupling coefficients and phase constants [(17), (18), (19)] vary with distance along a uniform birefringent twist section (θ is directly proportional to z). We conclude that if we should build a pair of coupled waveguides [1 and 2] where the phase constants and coupling coefficient vary with z as shown in Fig. 4, the power division between the two waveguides will be as shown in Fig. 5, since this is the way the twist prototype behaves. θ is now simply a parameter relating k, β_1 and β_2 as a function of length, but it is still convenient to think of it as the angle of the equivalent twist section. Fig. 5 shows that if the coupler is made of length $\theta = \pi/2$, a complete power transfer will take place, and this will occur smoothly and without the fluctuations of Fig. 1(b). A 3-db coupler will be pro-

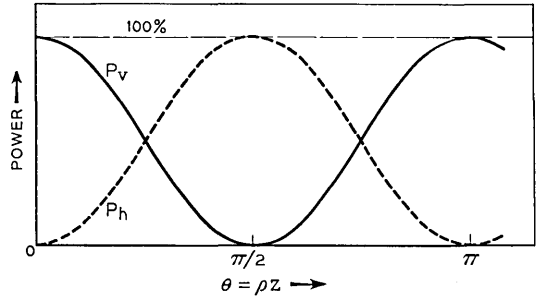


Fig. 5 — Power transfer between vertical and horizontal polarizations in a twisted birefringent medium.

vided by length $\theta = \pi/4$; and in fact any desired division is obtainable by using the proper length. All of these should be broadband because we know that the twist prototype is broadband.

It might appear that since the desired power transfer is obtained for only a particular length, it would therefore be as frequency sensitive as matched velocity couplers. Actually, the electrical length is not important. It is simpler to think in terms of the twist prototype where we see that the only requirement for the desired transfer is that the total twist angle θ be chosen correctly.

At this point we have shown that we can think about coupled waveguides in terms of a twist medium, and when we do so, we discover that this medium holds the secret of how to make a broad-band power transfer from one waveguide to another. Specifically, we conclude that both the coupling coefficient and the phase constants should be varied.

One way in which these design requirements may be met is suggested in Fig. 6 for a pair of coupled rectangular waveguides providing complete power transfer ($\theta = \pi/2$). The top wall has been removed to show a divided aperture so tapered as to give a coupling coefficient which varies as the sine of the distance from one end. At the same time the dividing partition is warped so as to produce the cosinusoidal cross-over of phase

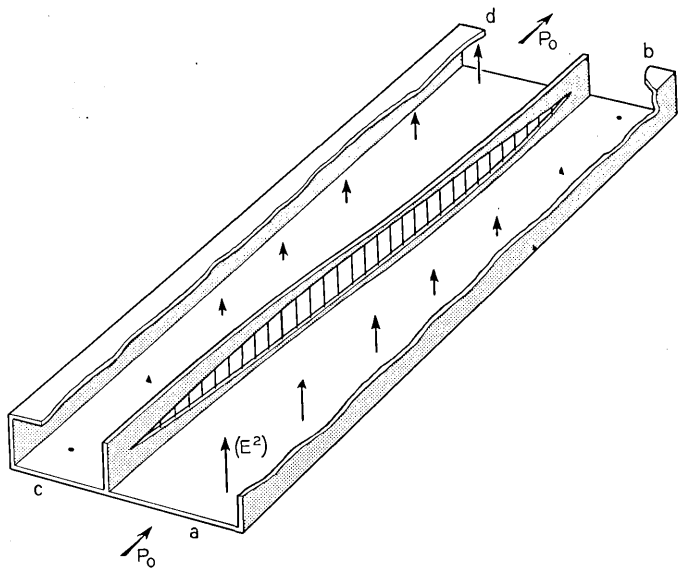


Fig. 6 — A broadband 100 per cent power transfer coupler using mode warping.

constants. The phase constants could also be adjusted by the insertion of a variable amount of dielectric loading. The vertical vectors represent the square of the electric field present in the two waveguides at various cross sections along the structure when all of the power is initially inserted at a. A complete transfer takes place with all of the field appearing at d and none at b.

The phase relations between the field vectors on the two sides of the coupling aperture are of interest. We know that in the case of a coupler employing uniform waveguides, the induced wave in one waveguide is always 90 degrees out of phase with the driving wave in the other waveguide ($k_{12} = k_{21} = -jk$). This is also true for the two coupled modes (e_v and e_h) in a uniform birefringent medium [see (1) and (2)] and hence $k_{vh} = k_{hv} = -jk$. However, we know that if we launch a linearly polarized wave on the twist medium with polarization paralled to one of the principal cross sectional axes, then the wave will remain linearly polarized. Consequently for this medium the vertically and horizontally polarized modes (e_v and e_h) will have a zero or 180-degree phase relation at all points along the medium. It follows that the coupler of Fig. 6 which was derived from the twist medium must also have a zero or 180-degree phase relation between the field components on opposite sides of the coupling partition at every cross section. This situation is illustrated in Fig. 7 where the transverse electric field is plotted for a series of cross sections of the coupler. The input end is shown at the top, and the output end at the bottom. The left hand column represents the field configurations when the wave is initially launched in the smaller of the two waveguides. The right hand column is for a wave launched in the larger of the two waveguides.

It may be seen that when the wave is launched in the larger waveguide terminal, it emerges at the other end of the coupler from the larger waveguide terminal. Moreover, at the center cross section where the two waveguides have the same phase constant, the energy is equally divided and we recognize this as the even symmetric normal mode for the local cross section. In fact, the wave travels throughout the length of the coupler in the local normal mode which has the higher phase constant.

Conversely, if the wave is launched in one of the smaller waveguide terminals it appears at the center cross section in the odd symmetric mode and emerges from the smaller waveguide at the far end. Thus it travels throughout the coupler in the local normal mode having the lower phase constant.

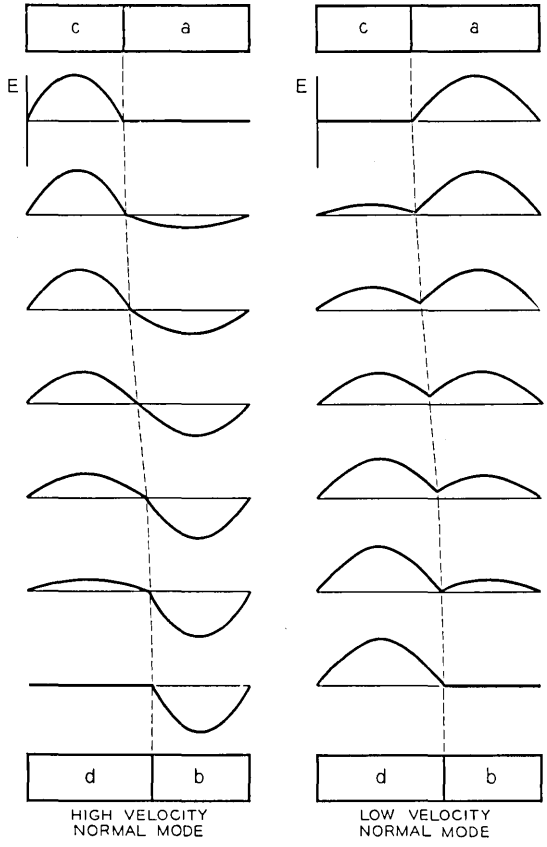


Fig. 7 — Warping of odd and even symmetric modes in the coupler of Fig. 6.

PRINCIPLE OF BROADBAND COUPLING BY NORMAL MODE WARPING

We have discussed a twisted birefringent medium and a rectangular waveguide coupler, both of which are examples of normal mode warping. We will now attempt a statement of the principle which is basic to all such couplers. We assume a waveguide system having two modes of propagation which are to be coupled so as to effect transfer of power. There are then two normal modes for this system, which are the dual of the coupled modes, and which may vary in field pattern from point to point along the structure depending upon the phase constants and coupling coefficients of the coupled modes. Provided these parameters vary slowly and smoothly along the structure, then if all of the power is in-

jected in one of the normal modes at one end, it will remain in one of the normal modes at successive cross sections and will emerge in one of the normal modes at the far end. This situation will be independent of frequency.

On the other hand, if both normal modes are excited at one end, power will be transmitted through the structure in both normal modes and emerge in both modes. This is what happens in conventional matched β couplers, and interference between the modes can vary as frequency changes.

The objective of a broadband design should then be: (1) to adjust parameters at the ends of the structure so that the normal modes at those points are identical with the desired input and output field excitations, and (2) to smoothly vary the parameters along the structure so that the normal modes are transmuted or "warped" from the one to the other set of field excitations. In this way the power will at all points be in only one of the normal modes, and interference between modes is avoided.

EXAMPLES OF MODE WARPING

We have already seen how this objective was achieved in the birefringent twist. Power was injected in one of the normal modes. This mode was smoothly warped from vertical polarization at one end to horizontal polarization at the other end by twisting. If we had chosen to excite this medium with a wave polarized at 45° to the birefringent axes, both of the normal modes of the medium would have been excited equally. As a result, equal amounts of power would have propagated down the twist in the two normal modes and would have arrived at the far end with relative phases which would depend upon the phase velocities of the two modes and the total distance travelled along the twist. The output polarization would, in general, be elliptical and would be frequency dependent.

In the case of the coupled waveguides of Figure 6, the normal modes at the ends corresponded to dominant wave excitation of the separate waveguide terminals, and they are smoothly warped from one set of terminals to the other. Excitation of one of the waveguide terminals would cause power to flow through the system in only one of the normal modes. If the waveguide structure had been cut at the center, either half of it would constitute a 3 db coupler. With excitation of one of the end terminals, the power at the center cross-section will exist entirely in one of the normal modes, which requires equal voltages in the two waveguides. Thus, the three db power division should be very broadband. On the other hand, if both of the waveguide terminals at one end were ex-

cited power would flow in both of the normal modes. The relative phases of these modes arriving at the center would vary with frequency, and hence the power division would be frequency dependent.

In Fig. 8 is shown another example of mode warping which is interesting because it illustrates some departures from the type of warping used in the previous examples. This device is a broadband converter from linear to circular polarization. It comprises a long section of round waveguide containing a slender axially magnetized rod of ferrite. Throughout a certain portion of its length the waveguide is gradually flattened to produce an elliptical cross-section. The rod of ferrite is also tapered to pointed ends. Transverse cross-section views are shown below the longitudinal cross section, and at the bottom is shown the way in which the magnitude of the coupling coefficient and the phase constants for vertically polarized and horizontally polarized waves vary along the length. We may identify three different parts of the transducer. At either end is a region in which the normal modes undergo no warping, and which functions solely as a taper section for matching the normal modes into the central region. Between these two matching regions is the region where mode warping occurs. At the left of this central region the elliptical cross section produces a linear birefringence where vertically polarized waves have a larger phase constant than horizontally polarized waves. The absence of any ferrite at this point means that the coupling coefficient between the two polarizations is zero. These relations are shown by the β

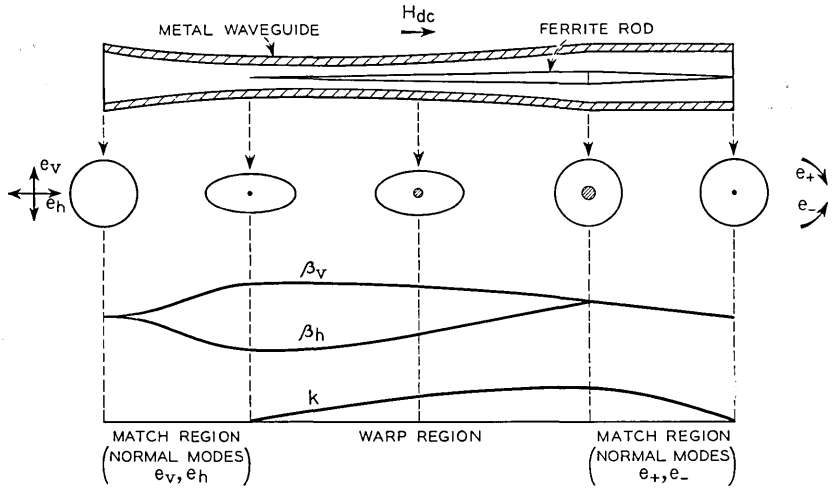


Fig. 8 — Broadband linear-to-circular polarization converter.

and k curves at the bottom of the figure. Proceeding toward the right, the gradual insertion of ferrite and decrease in ellipticity of the waveguide sheath cause the difference between the phase constants to decrease and the coupling coefficient to increase. At the right hand end of the central region, the sheath is round and the linear birefringence is zero. At the same time the presence of a maximum amount of ferrite causes the coupling to be a maximum. We see that the programming of the coupling coefficient and phase constant difference corresponds to 45 degrees of the prototype twist medium, and a linearly polarized wave at the left should have half its power transferred to horizontal polarization at the right. Unlike the birefringent twist, however, where the coupling coefficient is imaginary, the coupling coefficient due to the ferrite is real. As a result there will be a 90-degree phase relation between the horizontal and vertical polarizations, yielding circular polarization, which we already know is the normal mode for the longitudinally magnetized ferrite. Because a vertically polarized wave at the left will have the higher phase constant, it will be warped into the negative circularly polarized wave at the right (counterclockwise) since this wave sees the higher permeability, and hence phase constant, for the usual case where the applied magnetic field is less than that required for gyromagnetic resonance.

Unlike the couplers described previously, this coupler does not keep β_0 constant. The phase constant difference and the coupling coefficient still vary as the cosine and sine of distance respectively from the left hand end of the central section. But the increase in dielectric loading due to the ferrite rod causes both β_v and β_h to be larger than they would be if the dielectric constant of the ferrite were unity. However, examination of (9) and (10) show that within the bracketed factor which controls the amplitude of the wave, only δ and k appear. Provided these parameters vary in the proper way, the power transfer should take place regardless of β_0. Change in β_0 will cause the velocity of a normal mode to change from point to point along the structure, but will not change the field distribution for the normal mode. What this means physically can be visualized in the case of the birefringent twist by assuming that the dielectric strip either changes in dielectric constant or in total cross section from point to point without changing birefringence or rate of twist. Clearly, the rate at which the polarization of the wave is altered as a function of distance is not affected.

Another peculiarity of this coupler is that because of the non-reciprocal nature of the ferrite, the signs of the coupling coefficients will be opposite for opposite directions of propagation. Thus, for propagation

from left to right, $k_{vh} = +k$ and $k_{hv} = -k$, while for propagation from right to left, $k_{vh} = -k$ and $k_{hv} = +k$. It is simpler to see what this means, however, by considering that a counterclockwise polarized wave sent in from the right will have positive polarization relative to the ferrite magnetization, and will therefore have a lower phase constant than a clockwise polarized wave. As a result, this wave will be warped into the horizontally polarized wave at the left hand end, because this wave has the lower phase constant in the squashed region. Thus the circular to linear polarization conversion properties of this section are nonreciprocal and it is possible to combine it with reciprocally birefringent elements to obtain a circulator.*

In Fig. 9 is shown another example of mode warping where ferrite is employed in rectangular waveguide to produce a broadband circulator.

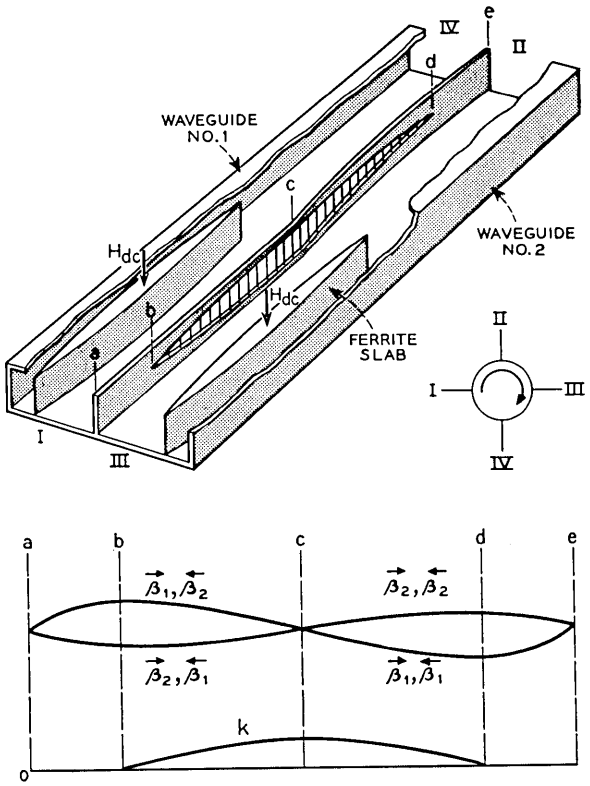


Fig. 9 — Rectangular waveguide circulator employing mode warping.

* See page 86 of Reference 3.

Two rectangular waveguides are coupled by a long divided aperture. At the left hand end, both waveguides are occupied by thin tapered slabs of ferrite located off center in the waveguide cross section. These terminate in knife edges at the center where the coupling aperture is largest. From here on the dividing partition is deflected so as to alter the phase constants of the two waveguides. We may identify four distinct regions in this coupler. The regions between *a* and *b*, and between *d* and *e* are for the purpose of matching to the waveguide terminals. In the regions between *b* and *c*, and between *c* and *d* mode warping occurs. Both of these regions are equivalent to 45-degree twists of the birefringent prototype, and power in either of their input terminals will be equally divided between the two waveguides at the center cross section *c*. Thus they operate as broadband hybrids. Section *cd* is like one half of the structure of Fig. 6, and it operates in the same manner. Section *bc*, on the other hand, operates like a non-reciprocal hybrid, and by virtue of its transversely magnetized ferrite slabs, it has non-reciprocal phase constants as shown at the bottom of the illustration. For propagation from left to right, the ferrite in waveguide 1 exhibits a permeability greater than one, while the ferrite in waveguide 2 exhibits a permeability less than one. For propagation from right to left, the situation is reversed. We may therefore analyze the behavior as follows:

A wave entering at terminal I will, by the time it arrives at cross section b, be travelling in the normal mode having the higher phase constant. As it passes through section *bc* the mode will be warped so that at *c* the power will be equally divided between the two waveguides in the even symmetric mode, which has the higher phase constant. Mode warping will continue through section *cd*, and all of the power will be transferred to waveguide 2 which has the higher phase constant at section *d*. Thus, power entering at I will be delivered to II.

A wave entering at II will return to cross section c with power equally divided in the even symmetric mode. From this point on, however, the situation is changed because of the non-reciprocal behavior of the ferrite. Now, waveguide 2 will have a phase constant which is higher than waveguide 1, and the wave which is travelling in the higher phase constant mode will be delivered to terminal III. The circulation order of the terminals is therefore as shown by the circulator symbol at the right of the waveguide.

In this structure, as in the one which preceded it, we note that the average phase constant is not necessarily constant. Nevertheless, the coupling coefficient and phase constant difference are varied sinusoidally as shown in Fig. 4.

NORMAL MODES ON A RAPID TWIST

So far we have shown that a slowly twisting birefringent medium can be used as a prototype for the design of broadband directional couplers having any desired power division ratio. We have not yet said anything about the length such a coupler must have, nor about how the twist may vary with distance. While a constant rate of twist was assumed, this is not necessary. The rate of twist can be varied arbitrarily along the coupler provided only that the total twist angle is such as to give the desired power division. However, the twist rate must not exceed some maximum value or the assumption that the polarization of the wave will twist with the medium will no longer be satisfied. Practically speaking, it would appear that the rate of twist should be constant and equal to the maximum permissible value, since to use a smaller rate in some portions of the coupler would only make the coupler longer than necessary.

It is evident at this point that we need to know more about the maximum twist rate and what happens if it is exceeded in the interest of making the coupler short. Our previous assumption of a gradual twist amounted to the assumption that the normal modes at any cross-section of the medium were linearly polarized along the principal axes of birefringence for the cross section, i.e., were the same as the local normal modes. We will now show that if the twist is rapid the normal modes are perturbed and are no longer linearly polarized.

Figs. 10(a) and 10(b) show symbolically two successive cross sectional views of a rapid twist having principal axes of birefringence A and B.

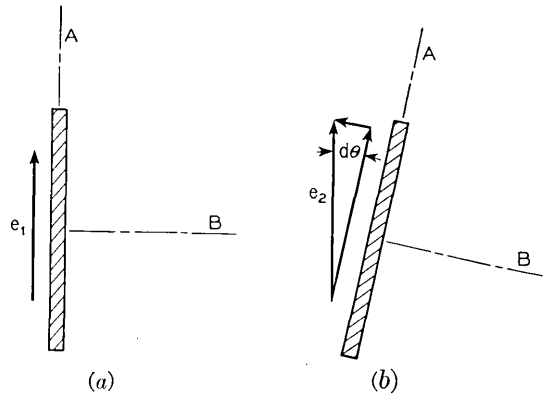


Fig. 10 — Coupling produced between A and B polarizations in a rapidly twisting birefringent medium.

We assume that the two cross sections are an infinitesimal distance apart, and hence that the angle $d\theta$ is infinitesimal. Between these cross sections we assume the axes of birefringence are fixed as shown in Fig. 10(a). Thus, a wave with polarization e_1 along the A axis at the input will emerge from the first infinitesimal step with unchanged polarization as shown by e_2. On entering the second step we see that both the A and B polarizations will be excited. Now if the fin twists at a constant rate $d\theta/dz = \rho$, then an electric field polarized along A is coupled to an electric field polarized along B as given by

$$de_b = -e_a\, d\theta; \qquad de_a = +e_b\, d\theta. \tag{22}$$

As before, we have assumed that there is no scattering or attenuation and that backward traveling components may be neglected. We can now write the basic equations for waves of polarization A and B.

$$\frac{de_a}{dz} = -j\beta_a e_a + e_b \frac{d\theta}{dz} \tag{23}$$

$$\frac{de_b}{dz} = -j\beta_b e_b - e_a \frac{d\theta}{dz} \tag{24}$$

These equations when solved simultaneously give

$$e_a = \varepsilon^{-j\beta_0 z}\left[C_a\, \varepsilon^{-j(\sqrt{\Delta^2+\rho^2})z} + D_a\, \varepsilon^{+j(\sqrt{\Delta^2+\rho^2})z}\right] \tag{25}$$

$$e_b = \varepsilon^{-j\beta_0 z}\left[C_b\, \varepsilon^{-j(\sqrt{\Delta^2+\rho^2})z} + D_b\, \varepsilon^{+j(\sqrt{\Delta^2+\rho^2})z}\right] \tag{26}$$

where β_0 and Δ have the meanings given in (3) and (4), and the coefficients C and D are related by

$$\frac{C_b}{C_a} = +j\left[\frac{\Delta - \sqrt{\Delta^2+\rho^2}}{\rho}\right] = \frac{D_a}{D_b} \tag{27}$$

These equations say that in such a twist medium there are two normal modes of propagation. The mode with the larger phase constant is obtained when $D_b = 0$ and $D_a = 0$. For this mode

$$\frac{e_b}{e_a} = +j\left[\frac{\Delta - \sqrt{\Delta^2+\rho^2}}{\rho}\right] \tag{28}$$

If the twist rate is low ($\rho \to 0$), e_b is very small relative to e_a. Therefore, this mode is characterized by having most of its energy polarized along the A axis. It is elliptically polarized because e_b lags e_a by 90°.

If C_a and C_b equal zero, we have the other normal mode which is also elliptically polarized, but has most of its energy polarized along the B axis and therefore has the lower phase constant.

If we launch all of the wave power into such a twist section in one of the normal modes, (25) and (26) tell us that the wave will continue in this mode until the twist is terminated.

These conclusions may now be translated back into our fixed frame of reference (vertical and horizontal axes) and we can determine how power will be transferred from one polarization to the other. If ρ is small enough, the conclusion will be the same as given in Fig. 5. But if ρ is large, a vertically polarized input wave (which is a local normal mode) no longer corresponds to one of the normal modes for the medium. Consequently, some of both modes will be excited, and since they travel at different velocities, the output polarization will depend upon the electrical length of the coupler and hence upon the frequency. The power transfer curves will have ripples as indicated in Fig. 11. This picture is similar to that of the pendulums shown in Fig. 1, and we can now see that the reason is much the same, namely, that an initial condition where all of the energy is in one pendulum does not correspond to a normal mode of the system.

Equation (27) makes it possible to determine the size of these ripples, or to determine the twist rate ρ which will make them smaller than a specified value. If the medium is excited as shown in Fig. 10(a), the worst axial ratio (e_b/e_a) will be

$$v \cong 2\left|\frac{C_b}{C_a}\right| = 2\left(\frac{\sqrt{\Delta^2 + \rho^2} - \Delta}{\rho}\right) \tag{29}$$

for $|C_b/C_a|^2 \ll 1$. Solving for ρ we obtain

$$\rho \cong v\Delta \tag{30}$$

Now $\Delta = \pi/\lambda_\Delta$ where λ_Δ is the birefringent wavelength equal to $2\pi/(\beta_a - \beta_b)$. Therefore

$$\rho = \frac{\pi v}{\lambda_\Delta} \tag{31}$$

For a complete power transfer

$$\rho z = \frac{\pi}{2}$$

Finally,

$$z = \frac{\lambda_\Delta}{2v} \tag{32}$$

This gives the length of the coupler in birefringent wavelengths which

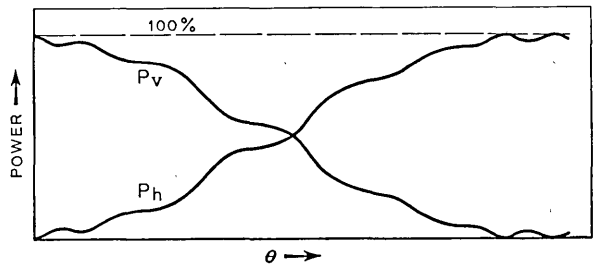


Fig. 11 — Power transfer in a rapid twist.

will insure that the axial ratio does not exceed v. (A birefringent wavelength is comparable to a beat wavelength λ_b defined earlier, and is the distance for which a 360° difference in phase shift exists for the two principal polarizations.)

Thus, to take a specific example, suppose we wish to employ a constant pitch twist which will completely transfer the power from vertical to horizontal polarization, and we wish to keep the power in the unwanted output polarization (vertical) more than 20 db down from the wanted polarization. Then $v = .1$, and $z = 5\lambda_\Delta$.

Since a birefringent wavelength is equal to several wavelengths in the medium, it is clear that the bandwidth of such a coupler has been purchased at the expense of considerable length. As a matter of fact we may note that a conventional coupler with matched phase constants must be made one half a beat wavelength long to provide complete power transfer. Thus, if the twist had the same beat wavelength as the conventional coupler, it would be ten times as long in order to keep the crosstalk more than 20 db down over a very broad band.

Even though the field will in general be elliptically polarized, there will be points along the coupler where the two normal modes are in phase and the polarization will be linear. Thus, by making the coupler the right length, the power transfer can be made correct even though $|C_b/C_a|$, and hence v, is quite appreciable. This makes it possible to make the coupler shorter and still couple power as desired; but the price paid is increased frequency dependence. In other words, by making the coupler shorter for the same total twist angle, the ripples of Fig. 11 will become more pronounced; but by choosing exactly the right length relative to the ripple position, all of the power can still be transferred to the desired polarization. It can be shown that the shortest twist section which can be made in this way is $\sqrt{3}$ times as long as a non-twisted birefringent medium having the same birefringence $(\beta_a - \beta_b)$. It will be

only slightly less frequency sensitive than the non-twisted medium. However, by making it longer the bandwidth for the twisted medium increases greatly, while the bandwidth of the non-twisted medium will decrease if its length is increased in multiples of λ_Δ .

VARIATION OF TWIST RATE TO OBTAIN MODE-MATCHING

We have seen that a more detailed analysis of the twist prototype coupler shows that the true normal modes do not exactly correspond to the local normal modes for finite rates of twist. This means that if the coupler is excited at the ends by fields corresponding to the local normal modes, both normal modes will be excited and small ripples in the power transfer characteristic will result.

One method which naturally suggests itself for avoiding the modal mismatch at the ends is to make the twist rate zero at the ends and to let it increase to a maximum at the center. Thus, it appears from (28) that if $\rho = 0$ at the ends, the normal mode will be linearly polarized at those points. Actually, however, this is not a complete cure, since it can be shown that the variation of the twist rate will also perturb the normal modes so as to make them elliptically polarized to a small degree at the ends. Nevertheless, it appears that the ellipticity can indeed be reduced by such a technique.

The effect of varying the twist rate may be determined as follows. Let us identify the local normal modes on the twist as Mode I and Mode II according to whether the polarization is principally along the A or B axis respectively. These modes are elliptically polarized and the ellipticity depends upon the rate of twist ρ as defined by (27). Now if ρ varies we can break the twist up into differential elements, each of which has a constant ρ in a manner analogous to the technique used for Fig. 10. In this way we can show that there is effectively a coupling produced between modes I and II and that the coupling coefficients are

$$K_{I\,II} = j\frac{\left(\frac{d\alpha}{dz}\right)}{1+\alpha^2} = K_{II\,I} \tag{33}$$

where α is the magnitude of the ratio given in (27)

$$\alpha = \frac{\Delta - \sqrt{\Delta^2 + \rho^2}}{\rho^2} \tag{34}$$

Now if we cause ρ to vary in such a way as to keep K (where $K = |K_{I\,II}|$) a constant independent of z, we can repeat the analytical

process of (22) to (28) and determine a new pair of normal modes which are composites of the now coupled modes I and II. For this to be valid, however, it is necessary that the quantity

$$\frac{\beta_I - \beta_{II}}{2} = D = \sqrt{\Delta^2 + \rho^2} \tag{35}$$

remain constant. Actually, it does not because ρ is a variable. But if ρ remains much smaller than Δ the assumption that D is constant may be valid. In this case the compositions of the new normal modes will be given by,

$$\frac{e_{II}}{e_I} = \frac{\Delta - \sqrt{\Delta^2 + K^2}}{K} \tag{36}$$

for the almost I mode, and

$$\frac{e_I}{e_{II}} = \frac{\Delta - \sqrt{\Delta^2 + K^2}}{K} \tag{37}$$

for the almost II mode.

In the meanwhile, the requirement that K be a constant permits us to solve (33), and if we let $\rho = 0$ when $z = 0$, we obtain

$$\alpha = \tan Kz. \tag{38}$$

Substituting for α from (34), and solving for ρ, we obtain

$$\rho = \frac{d\theta}{dz} = \Delta \tan 2Kz \tag{39}$$

The solution of this is

$$\theta = \frac{\Delta}{2K} \ln \frac{1}{\cos 2Kz} \tag{40}$$

Let us now take a special case to see what these results mean. We will assume a complete transfer coupler where

$$\begin{aligned} \theta &= \frac{\pi}{2} \\ \Delta &= 0.2\beta_0 \\ K &= 0.001\beta_0 \end{aligned} \tag{41}$$

Since the twist rate will be made symmetrical about the center point, one half the coupler will have a $\theta = \pi/4$. Substituting this in (40) we obtain

$$z = 10\,\lambda_0 \tag{42}$$

where λ_0 is the average wavelength in the medium corresponding to β_0. Thus, the whole coupler will be twenty wavelengths long. From (39) we can determine that

$$\rho_{max} = 0.025\,\beta_0$$

and this means that the phase constant difference between modes I and II which we assumed was constant actually varies by less than one per cent. At the ends of the coupler, the fact that $\rho = 0$ means that modes I and II are linearly polarized along the axes of birefringence. However, the normal modes require some of both I and II, and from (36) we find that for the almost I mode

$$\left|\frac{e_{II}}{e_I}\right| = 0.0025 \tag{43}$$

This ratio is a measure of the excitation of the undesired II mode which will be produced if the input polarization is the local normal mode polarized along the A axis. Since a similar mismatch occurs at the opposite end and also in the middle where the rate of twist begins to decrease, the total crosstalk between the output polarizations may be four times the above figure. Thus, at the output end

$$\left|\frac{e_a}{e_b}\right|_{max} = 0.01 \tag{44}$$

and the worst crosstalk into the undesired polarization is thus 40 db down from the desired output.

Let us compare this with a uniform twist which produces the same crosstalk. From (32) we obtain,

$$z = 250\,\lambda_0$$

Therefore, the uniform twist would have to be 12.5 times as long as the non-uniform twist to do the same job. While this analysis for the variable twist is only approximate, it indicates a marked superiority for the variable twist coupler.

This conclusion is also borne out by a study of the possibility of producing wave coupling by a series of step twists in a birefringent medium, where each step is one half a birefringent wavelength in length. This analysis is given in the appendix. It is shown there that if there are a large enough number of steps so that the total angle between steps is small, then the angles between steps starting at one end and passing through to the other end of the coupler should have a binomial distribution. This will result in the slowest departure from the desired power

division as frequency changes. If we regard such a long multi-step twist as an approximation to a smooth twist of the same length, we again conclude that the smooth twist should start off and finish with a zero twist rate, and should have the maximum twist rate at the center.

CONCLUSION

It has been shown that Cook's scheme for producing broadband directional couplers by variation of the phase constants, may be generalized by simultaneously varying the coupling coefficient. For the simplest case, the difference between the phase constants should vary cosinusoidally and the coupling coefficient should vary sinusoidally with distance along the coupler. Such a programming of the coupling parameters corresponds directly to a twisted birefringent medium (such as a metal waveguide having a flattened cross section) where the rate of twist is constant. Since this medium is easy to analyze and to visualize physically, it has been used as a prototype for the design of a number of different types of coupler, all of which are based on the same principle. This principle, which in retrospect sounds rather obvious, is simply that in order to avoid interference effects between two modes of propagation in a multimode waveguide system we should avoid exciting more than one of the normal modes. Also by gently warping the waveguide structure, it is possible to warp the field configuration of the desired normal mode so that it will produce the required power division at the terminals of the system without appreciably scattering power into other unwanted modes.

By avoiding wave interference, such couplers should in principle be independent of frequency. However, the requirement that warping be smooth and gradual also dictates that these couplers must be many wavelengths long. It may turn out that they will be most useful in the millimeter wavelength range where such electrical lengths are physically short. Expressions have been given for the uniform twist which allow one to compute how long the coupler must be in order to meet specified requirements on power division at the terminals. These may of course be applied equally well to any of the other types of coupler by making use of the equivalence equations (17), (18), and (19).

The importance of varying coupling coefficient as well as the phase constant difference for a pair of coupled waves is particularly evident at the ends of a complete power transfer coupler. If we let $k = 0$ in (13) and (14), we see that the local normal modes correspond to all of the power in one or the other of the two waveguide terminals. If k is not zero, it is

necessary that the phase constant difference be infinite to achieve the same objective, and this is not possible practically.

Finally, it has been noted that a constant rate of twist for the twist prototype coupler is not necessarily the best. In fact a coupler may be made shorter for specified performance, or will give better performance for a specified length if the twist rate is maximum in the middle and approaches zero at the ends.

Appendix

Multi-Step Mode Transformation

We will show the way in which power can be transferred from one to another of two coupled modes by means of a series of step-mode transformations. The analysis will be made for the special case of a birefringent medium and the results can then be translated for any other type of medium.

Specifically, we assume that the modes to be coupled are vertically and horixontally polarized dominant waves in a round waveguide. Initial excitation e_0 is vertically polarized. This is launched into a first Δ 180° section* set at an angle $\theta_1/2$ from the vertical. As a result, the output of this section is a rotated polarization at angle θ_1 from the vertical. This is introduced into a second Δ 180° section set at an angle $\theta_1 + \theta_2/2$ from the vertical. The output of this section is then at angle $\theta_1 + \theta_2$, and so forth. The polarizations between sections will then appear as

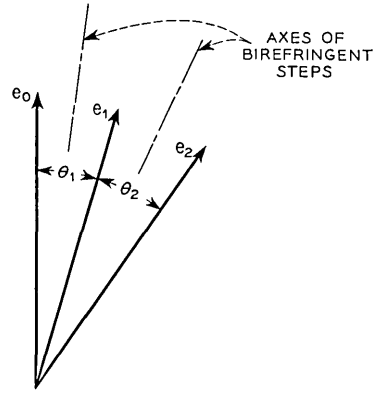


Fig. 12 — Step-twist birefringent medium.

* A section of linearly birefringent medium for which the difference in phase shift for the two principal polarization is 180°.

shown in Fig. 12 by e_1, e_2, $\cdots$, and the dashed lines represent the orientations of the principal axes of the Δ 180° sections. If the Δ 180° sections work perfectly, the output power has pure linear polarization at some arbitrary angle $\theta_1 + \theta_2 + \cdots + \theta_n$ which gives the desired final distribution of power between vertical and horizontal polarizations.

Now if the frequency varies from its design value causing a departure of the differential phase shift from Δ 180° by a small angle δ, it can be shown that a portion of the input polarization e_0 will not be rotated to the angle θ_1. Thus, at the output of the first section e_1 will be split into two components

$$e_1 = e_0 \cos \delta \, \underline{|\theta_1} + je_0 \sin \delta \, \underline{|0}$$

where $\underline{|\theta}$ means, "polarized at angle θ". This splitting recurs at each successive section. At the end of the second section, we will have

$$e_2 = e_0 \cos^2 \delta \, \underline{|\theta_1 + \theta_2} + je_0 \sin \delta \cos \delta \, \underline{|\theta_1} + je_0 \sin \delta \cos \delta \, \underline{|2\theta_1 + \theta_2} - e_0 \sin^2 \delta \, \underline{|0}$$

etc.

If now we take the total field polarized at right angles to the desired output polarization, we will have:

1 Section:

$$je_0 \sin \delta \sin \theta_1$$

2 Section:

$$- e_0 \sin^2 \delta \sin (\theta_1 + \theta_2)$$
$$+ je_0 \sin \delta \cos \delta (\sin \theta_2 - \sin \theta_1)$$

3 Section:

$$- je_0 \sin^3 \delta \sin (\theta_1 + \theta_2 + \theta_3)$$
$$- \sin^2 \delta \cos \delta \begin{bmatrix} \sin (-\theta_1 - \theta_2) \\ \sin (-\theta_1 + \theta_3) \\ \sin (\theta_2 + \theta_3) \end{bmatrix}$$
$$+ j \sin \delta \cos^2 \delta \begin{bmatrix} \sin (\theta_1) \\ \sin (-\theta_2) \\ \sin (\theta_3) \end{bmatrix}$$

4 Section:

$$+ e_0 \sin^4 \delta \sin (\theta_1 + \theta_2 + \theta_3 + \theta_4)$$

$$-je_0 \sin^3 \delta \cos \delta \begin{cases} \sin (-\theta_1 - \theta_2 - \theta_3) \\ \sin (-\theta_1 - \theta_2 + \theta_4) \\ \sin (-\theta_1 + \theta_3 + \theta_4) \\ \sin (\theta_2 + \theta_3 + \theta_4) \end{cases}$$

$$-e_0 \sin^2 \delta \cos^2 \delta \begin{cases} \sin (\theta_1 + \theta_2) \\ \sin (\theta_1 - \theta_3) \\ \sin (\theta_1 + \theta_4) \\ \sin (-\theta_2 - \theta_3) \\ \sin (-\theta_2 + \theta_4) \\ \sin (\theta_3 + \theta_4) \end{cases}$$

$$je_0 \sin \delta \cos^3 \delta \begin{cases} \sin (-\theta_1) \\ \sin (\theta_2) \\ \sin (-\theta_3) \\ \sin (\theta_4) \end{cases}$$

5 Section:

$$je_0 \sin^5 \delta \sin (\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5)$$

$$e_0 \sin^4 \delta \cos \delta \begin{cases} \sin (-\theta_1 - \theta_2 - \theta_3 - \theta_4) \\ \sin (-\theta_1 - \theta_2 - \theta_3 + \theta_5) \\ \sin (-\theta_1 - \theta_2 + \theta_4 + \theta_5) \\ \sin (-\theta_1 + \theta_3 + \theta_4 + \theta_5) \\ \sin (\theta_2 + \theta_3 + \theta_4 + \theta_5) \end{cases}$$

$$-je_0 \sin^3 \delta \cos^2 \delta \begin{cases} \sin (\theta_1 + \theta_2 + \theta_3) \\ \sin (\theta_1 + \theta_2 - \theta_4) \\ \sin (\theta_1 + \theta_2 + \theta_5) \\ \sin (\theta_1 - \theta_3 - \theta_4) \\ \sin (\theta_1 - \theta_3 + \theta_5) \\ \sin (\theta_1 + \theta_4 + \theta_5) \\ \sin (-\theta_2 - \theta_3 - \theta_4) \\ \sin (-\theta_2 - \theta_3 + \theta_5) \\ \sin (-\theta_2 + \theta_4 + \theta_5) \\ \sin (+\theta_3 + \theta_4 + \theta_5) \end{cases}$$

$$-e_0 \sin^2 \delta \cos^3 \delta \begin{cases} \sin(-\theta_1 - \theta_2) \\ \sin(-\theta_1 + \theta_3) \\ \sin(-\theta_1 - \theta_4) \\ \sin(-\theta_1 + \theta_5) \\ \sin(+\theta_2 + \theta_3) \\ \sin(+\theta_2 - \theta_4) \\ \sin(+\theta_2 + \theta_5) \\ \sin(-\theta_3 - \theta_4) \\ \sin(-\theta_3 + \theta_5) \\ \sin(+\theta_4 + \theta_5) \end{cases}$$

$$+je_0 \sin \delta \cos^4 \delta \begin{cases} \sin(\theta_1) \\ \sin(-\theta_2) \\ \sin(\theta_3) \\ \sin(-\theta_4) \\ \sin(\theta_5) \end{cases}$$

Now upon examining these voltage components representing power in the undesired polarization, we find that certain of the bracketed sums will vanish merely by making the structure symmetrical. Thus, in the 5 section case, if $\theta_1 = \theta_5$, $\theta_2 = \theta_4$, the $\sin^4 \cos$ and $\sin^2\cos^3$ brackets vanish. By setting each of the remaining brackets equal to zero, we obtain several simultaneous equations which allow us to solve for the remaining θ ratios. In general, this may be an algebraically complicated procedure. However, the reader can verify by inspection that if all of the angles are assumed very small so that

$$\sin \theta \rightarrow \theta ,$$

a binomial distribution will cause all remaining brackets to vanish. The only unwanted voltage term to remain will be

$$e_0 \sin^n \delta \sin(\theta_1 + \theta_2 + \cdots + \theta_n).$$

Since the sum of the θ's is simply the total twist angle which we assume is held constant, then clearly the voltage departure from linear output polarization as a function of δ (or frequency) varies as $\sin^n \delta$ where n is the number of sections used. This demonstrates the bandwidth advantage to be gained from using a large number of steps.

We may conclude that if a total twist angle is achieved by a very large number of very small step twists between Δ 180° sections, this may be considered an approximation of a long continuous twist. Since the optimum design for the step twists calls for a binomial distribution, it is evident that the continuous twist should also approximate this pro-

portioning by having a very small $d\theta/dz$ at the ends, and a maximum $d\theta/dz$ in the center. This conclusion bears out what was said earlier about starting the continuous twist with $d\theta/dz = 0$.

As a special case where four sections are used to obtain a total twist angle of 90°, an exact solution gives

$$\theta_1 = 10.2°$$
$$\theta_2 = 34.8°$$
$$\theta_3 = 34.8°$$
$$\theta_4 = 10.2°$$

whereas the binomial solution would have been

$$\theta_1 = 11.25°$$
$$\theta_2 = 33.75°$$
$$\theta_3 = 33.75°$$
$$\theta_4 = 11.25°$$

If a step twist design of a certain number of steps is chosen as a model, the design for some other type of coupler may be determined by using equations 17, 18, and 19 to solve for the phase constants and coupling coefficient in each step. The β_a and β_b are the phase constants for the parallel and perpendicular polarizations in the Δ 180° steps. The values of θ for the several steps will be $\theta_1/2$, $\theta_1 + \theta_2/2$, $\theta_1+\theta_2+\theta_3/2$, etc., since these are the angles at which the Δ 180° plates were oriented. Each section of the coupler will of course be of the same length as the equivalent Δ 180 section in the twist coupler.

Analysis of the Single Tapered Mode Coupler

By W. H. LOUISELL

(Manuscript received February 17, 1955)

Broadband directional couplers in which the phase constants and coupling coefficients vary with distance along two coupled transmission lines as suggested in the two preceding papers are analyzed. A criterion is given for the allowed variations in the line parameters for a given bandwidth. Parameters which describe the performance of such couplers are given. It is found that such couplers give much greater bandwidth than conventional couplers of the distributed coupling type but they must be longer than conventional couplers.

I. INTRODUCTION

Conventional directional couplers[1] of the distributed coupling type may be thought of as two coupled uniform ideal lossless transmission lines. Such a system can in general support two forward normal modes of propagation. In order to effect power transfer between two lines of such a system, both forward normal modes must be present. Since the two normal modes in general travel with a different phase velocity, they can interfere with one another to set up a standing or "beat" wave pattern. Power transfer is thus effected by the interference of these two normal modes and for this reason such couplers are called mode interference directional couplers. If for example a 100 per cent coupler is desired, that is, complete power transfer from one line to the other, the coupler is ended after half a "beat" wave length. Similarly, if a three db coupler is desired, that is, a 50–50 power division between the two lines, the coupler is ended after one-quarter of a beat wave length. Since such couplers depend on interference of normal modes to effect power transfer they are innately narrow band.

J. S. Cook[2] and A. G. Fox[3] in the two preceding articles have proposed a new principle for broad-banding directional couplers in which the phase constant difference between the two lines and the coupling coef-

ficient vary with distance along the lines. Strictly speaking, the concept of normal modes is not applicable to non-uniform coupled lines, but if the phase constant difference and the coupling coefficient vary slowly enough with distance, it is useful to introduce the idea of "quasi-normal modes." A quasi-normal mode is a field configuration at a given point in the lines when the line parameters vary slowly compared with the local beat wave length. It differs only slightly from the field configuration of a normal mode that would have existed at the same point if the line parameters were constant at their local values. Since the line parameters vary, the quasi-normal modes are actually tapered, so that the same quasi-normal mode may correspond to all the power in line 1 at the beginning of the coupler and practically all the power in line 2 at the end. The quasi-normal modes in non-uniform lines are coupled, but the coupling is small if the variation of line parameters is gradual. Hence directional couplers may be made in which essentially all the power remains in one of the quasi-normal modes while passing from one line to the other. Such couplers will be called single tapered mode couplers.

A study will be made of the restrictions that must be placed on the phase constant difference and the coupling coefficient variation in order that power transfer may be effected between two lines in a single tapered mode coupler. It is found that the phase constant difference and the coupling coefficient must vary slowly compared with the local beat wave length. In general, the coupler must be several beat wave lengths long to effect complete power transfer, in contrast to the mode interference coupler which must be only a half beat wave length long. Also it is found that the greater the length of the single tapered mode coupler, the greater the bandwidth.

Several classes of couplers with different assumed variations of phase constants and coupling coefficient will be studied. The bandwidths of a few couplers having the same length, maximum phase constant difference, and maximum coupling coefficient will be compared.

II. NORMAL MODES OF UNIFORM COUPLED TRANSMISSION LINES

If two ideal uniform lossless transmission lines are coupled together, it is to be expected that the system can support four normal modes — two in the forward and two in the backward direction. In the present work, the two backward normal modes will be disregarded.* We proceed to find the two forward normal modes for such a system. S. E. Miller[1] has

* The labor of including the backward modes in our calculation is not expected to be compatible with the additional obtainable physical insight.

shown that the wave amplitudes for the two coupled lines may be written in the form

$$\begin{aligned} \frac{dE_1}{dz} &= -j(\beta_1 + c)E_1 + jcE_2 \\ \frac{dE_2}{dz} &= jcE_1 - j(\beta_2 + c)E_2 \end{aligned} \tag{1}$$

in which

$E_{1,2}(z)$ = wave amplitudes in lines 1 and 2, respectively

$\beta_{1,2}$ = uncoupled phase constants of lines 1 and 2, respectively

c = mutual and self-coupling coefficient between the lines

Thus the backward waves are disregarded. Further, it is assumed that the lines are lossless, and that the mutual and self-coupling coefficients are identical. Energy conservation requires that c be real. Also it is assumed that the characteristic impedances are normalized so that the power in either line is equal to the square of the wave amplitude.

Although it is easy to solve (1) directly, we shall begin by making a transformation which will be useful in what follows. Taking out a common phase factor and introducing the normal coordinates $w_1(z)$ and $w_2(z)$, we let

$$\begin{aligned} E_1(z) &= e^{-j(\beta+c)z}[\cos \tfrac{1}{2}\,\xi\, w_1(z) - \sin \tfrac{1}{2}\,\xi\, w_2(z)] \\ E_2(z) &= e^{-j(\beta+c)z}[\sin \tfrac{1}{2}\,\xi\, w_1(z) + \cos \tfrac{1}{2}\,\xi\, w_2(z)] \end{aligned} \tag{2}$$

where we define*

$$\begin{aligned} \beta &= \tfrac{1}{2}(\beta_1 + \beta_2) \\ \varphi &= \tfrac{1}{2}(\beta_2 - \beta_1) \\ \Gamma &= \sqrt{\varphi^2 + c^2} \end{aligned} \tag{3}$$

$$\begin{aligned} \cos\frac{\xi}{2} &= \sqrt{\frac{\Gamma + \varphi}{2\Gamma}} \\ \sin\frac{\xi}{2} &= \sqrt{\frac{\Gamma - \varphi}{2\Gamma}} \end{aligned} \tag{4}$$

$$\cot \xi = \frac{\varphi}{c}$$

* For convenience to the reader, it may be noted that ξ in the present paper is equal to 2ϑ in the preceding paper.

Substituting (2) in (1), we find that the normal coordinates satisfy the uncoupled equations:

$$\begin{aligned} \frac{dw_1}{dz} - j\Gamma w_1 &= 0 \\ \frac{dw_2}{dz} + j\Gamma w_2 &= 0 \end{aligned} \tag{5}$$

The normal mode solutions may be written down immediately as

$$\begin{aligned} w_1(z) &= w_1(0)e^{j\Gamma z} \\ w_2(z) &= w_2(0)e^{-j\Gamma z} \end{aligned} \tag{6}$$

where $w_1(0)$ and $w_2(0)$ are arbitrary constants. $w_1(z)$ is called the fast normal mode and $w_2(z)$ is called the slow normal mode. The voltages in the two lines are given by substituting equations (6) into equations (2). $|w_1(0)|^2$ represents the amount of power excited in the fast normal mode and $|w_2(0)|^2$ represents the amount of power excited in the slow normal mode. The voltages are normalized so that $|w_1(0)|^2 + |w_2(0)|^2 = 1$. The voltage amplitudes in the two lines for the *fast normal mode* are

$$\begin{aligned} E_1'(z) &= \cos \tfrac{1}{2}\,\xi\, e^{-j(\beta+c-\Gamma)z} \\ E_2'(z) &= \sin \tfrac{1}{2}\,\xi\, e^{-j(\beta+c-\Gamma)z} \end{aligned} \tag{7}$$

while the voltage amplitudes for the *slow normal mode* are

$$\begin{aligned} E_1''(z) &= -\sin \tfrac{1}{2}\,\xi\, e^{-j(\beta+c+\Gamma)z} \\ E_2''(z) &= \cos \tfrac{1}{2}\,\xi\, e^{-j(\beta+c+\Gamma)z} \end{aligned} \tag{8}$$

(The fast normal mode has the same phase in each line and is called the in-phase normal mode, while the slow mode is called the out-of-phase normal mode.*)

III. MODE INTERFERENCE DIRECTIONAL COUPLERS

The two coupled uniform transmission lines† treated above can be used as a directional coupler.[1] From equations (2) and (6) it is seen that the power in line 1 is given by

$$\begin{aligned} P_1(z) &= |E_1(z)|^2 \\ &= |\, w_1(0)\,|^2 \cos^2 \tfrac{1}{2}\xi + |\, w_2(0)\,|^2 \sin^2 \tfrac{1}{2}\xi \\ &\qquad - \sin \xi Re(w_1(0)w_2^*(0)e^{2j\Gamma z}) \end{aligned} \tag{9}$$

* There may be cases in which the slow mode is the in-phase mode.

† The present work is assumed to be applicable to coupled wave guides, coupled helices, etc.

and the power in line 2 is given by

$$P_2(z) = |E_2(z)|^2$$
$$= |w_1(0)|^2 \sin^2 \tfrac{1}{2}\xi + |w_2(0)|^2 \cos^2 \tfrac{1}{2}\xi + \sin \xi Re(w_1(0)w_2^*(0)e^{2j\Gamma z}) \quad (10)$$

where Γ is defined in (3).

From (9) and (10) it is seen that if only one of the normal modes is present at $z = 0$, then *there is no power transfer between the two lines.* In order for the two coupled lines to act as a directional coupler, both normal modes must be excited at $z = 0$. Power transfer between the lines is effected by interference of the two normal modes. Since directional couplers[1] using distributed coupling utilize the interference of two normal modes, they will be called *mode interference directional couplers.* Both normal modes must be excited to effect power transfer between the two elements of the coupler.

The beat wave length of the coupler is defined as the minimum distance between two points along the lines at which the power in a given line has its maximum value. For example, if $\beta_1 = \beta_2$, $w_1(0) = -w_2(0) = 1/\sqrt{2}$, the beat wave length is given by $\lambda_{b_0} = \pi/c$ in the example treated.*

The transfer loss τ may be defined as the ratio of the power in line 2 at $z = \ell$, where ℓ is the length of the coupler, to the amount of power excited in line 1 at $z = 0$, assuming no excitation of line 2 at $z = 0$. Thus,

$$\tau = \frac{P_2(\ell)}{P_1(0)} = \text{Transfer Loss} \quad (11)$$

If, for example, $P_1(0) = 1$ (all the power initially in line 1) and $\tau = 1$ then all the power is transferred from line 1 to line 2. This will be called a 100 per cent or zero-db coupler. If $P_1(0) = 1$ and $\tau = \frac{1}{2}$, half the power is transferred from line 1 to line 2. This will be called a 50 per cent, or 3 db coupler. τ can thus have any value from 0 to 1 and serves as a parameter to describe "conventional" mode interference couplers.

If $\beta_1 = \beta_2$, $P_1(0) = 1$, $w_1(0) = 1/\sqrt{2} = -w_2(0)$, then from (9) and (10), it is seen that

$$P_1(z) = \cos^2 cz$$

* Reasons could be given for calling $\lambda_{b_0} = 2\pi/c$, but in this paper one half of this value has been chosen as representing the beat wave length in order to conform with the preceding papers.

and

$$P_2(z) = \sin^2 cz$$

so that $\lambda_{b_0} = \pi/c$. The transfer loss is given by

$$\tau = \sin^2 \frac{\pi \ell}{\lambda_{b_0}}$$

To get a 3 db coupler in this case, the coupler is made ¼ of a beat wave length long while a 100 per cent coupler is made ½ of a beat wave length long.

In general, when $\beta_1 \neq \beta_2$ if both modes are equally excited, so that $w_1(0) = -w_2(0) = 1/\sqrt{2}$, then from (9) and (10), it is seen that the power in the two lines becomes

$$2P_1(z) = 1 + \frac{1}{\sqrt{\left(\frac{\varphi}{2c}\right)^2 + 1}} \cos 2\Gamma z$$

and (12)

$$2P_2(z) = 1 - \frac{1}{\sqrt{\left(\frac{\varphi}{2c}\right)^2 + 1}} \cos 2\Gamma z$$

Miller[1] has plotted the transfer loss for several values of $\varphi/2c$. However, it can be seen from (12) that if $(\varphi/2c) \gg 1$, there is practically no power transfer, while if $(\varphi/2c) \ll 1$, there is practically complete periodic power transfer between the lines.

Since "conventional" couplers depend on the interference of two normal modes, it is to be expected that they will be frequency sensitive. This can be seen from the fact that in general, the beat wave length will depend on frequency, and since for a desired transfer loss the coupler must be made some definite fraction of a beat wave length long, such couplers are innately frequency-sensitive. By adding more coupling elements and by means of an ingenious variation of the strength of the coupling, Mumford[1] and Miller[1] have shown that the bandwidth may be increased, although there is a fundamental limit to the bandwidth obtained by such schemes. The proposals of Cook[2] and Fox[3] will now be shown to yield couplers which are at least an order of magnitude more broadband than mode interference couplers.

IV. QUASI-NORMAL MODES IN TAPERED COUPLED LINES

The problem we should now like to consider is this: Can a coupler in which, as before, power is injected into one transmission line only, but in which only one of the quasi-normal modes is excited throughout, and in which the propagation "constants" of the two lines β_1 and β_2, as well as the coupling coefficient c, vary with distance along the lines? It will be shown in the following that the answer to this question is in the affirmative, approximately, provided that the variations of the line parameters are sufficiently gradual.

In order to see what restrictions must be placed on the variations of β_1, β_2 and c, consider the following. For symmetry, assume the variation of β_1 and β_2 with z can be expressed by

$$\begin{aligned} \beta_1 &= \beta - \varphi(z) \\ \beta_2 &= \beta + \varphi(z) \\ c &= c(z) \end{aligned} \tag{13}$$

where β = constant and $\varphi(0) \geqq 0$. The equations for the wave amplitudes in the two lines are given by (1) with β_1, β_2 and c given by (13). In analogy with the transformation used to reduce (1) to normal form for the uniform coupled lines, we shall now introduce *local normal coordinates* $w_1(z)$ and $w_2(z)$. The local normal coordinates are related to $E_1(z)$ and $E_2(z)$ by

$$\begin{aligned} E_1(z) &= \exp\left(-j\left[\beta z + \int_0^z c(\zeta)\, d\zeta\right]\right)\{\cos \tfrac{1}{2}\xi(z)w_1(z) \\ &\qquad - \sin \tfrac{1}{2}\xi(z)w_2(z)\} \\ E_2(z) &= \exp\left(-j\left[\beta z + \int_0^z c(\zeta)\, d\zeta\right]\right)\{\sin \tfrac{1}{2}\xi(z)w_1(z) \\ &\qquad + \cos \tfrac{1}{2}\xi(z)w_2(z)\} \end{aligned} \tag{14}$$

where all symbols are defined by (3) and (4) but where it is to be understood that Γ, φ, c and ξ are now functions of z. Then, substituting (14) into equations (1) where c, β_1 and β_2 have the form of (13), we find after simple manipulation that w_1 and w_2 must satisfy

$$\begin{aligned} \frac{dw_1}{dz} - j\Gamma(z)w_1 &= \frac{1}{2}\frac{d\xi}{dz}w_2 \\ \frac{dw_2}{dz} + j\Gamma(z)w_2 &= -\frac{1}{2}\frac{d\xi}{dz}w_1 \end{aligned} \tag{15}$$

Equations (15) are coupled in general through the terms proportional to $d\xi/dz$. They reduce to the uncoupled (5) when $d\xi/dz = 0$, i.e., when ξ is constant, or equivalently when $\cot \xi = \varphi/c$ is constant. Such will be the case with uniform lines in which $\varphi(z)$ is everywhere proportional to $c(z)$. However, the condition $\varphi/c =$ constant leads only to mode interference couplers which are of no interest in the present discussion.

It is clear, then, that there will be some coupling between the quasi-normal modes in a tapered mode coupler. Such coupling between quasi-normal modes will be called "hypercoupling" to distinguish it from ordinary electromagnetic coupling between two transmission lines (as represented by the parameter c). A "hypercoupling coefficient" $\eta(z)$ may be defined by

$$\eta(z) = \frac{1}{2\Gamma(z)} \frac{d\xi}{dz} \equiv \frac{1}{2} \frac{d\xi}{d\rho} \tag{16}$$

which gives a measure of the strength of the coupling between the quasi-normal modes.

Now if $\varphi(z)$ and $c(z)$ vary slowly compared to the local beat wavelength $\lambda_b(z)$, where

$$\lambda_b(z) = \frac{\pi}{\Gamma(z)}$$

then $\eta(z) \ll 1$ for all z, and the quasi-normal modes have very little hypercoupling. We can then write down approximate solutions of (15) which proceed essentially in powers of the hypercoupling coefficient. Thus the *In-Phase Quasi-Normal Mode* is given approximately by

$$\begin{aligned} w_1(z) \cong e^{j\rho(z)} \Bigg(& w_1(0) + \frac{1}{2} w_2(0) \int_0^z \frac{d\xi}{dz'} e^{-2j\rho(z')}\, dz' \\ & - \frac{1}{4} w_1(0) \int_0^z \frac{d\xi}{dz'} e^{-2j\rho(z')} \int_0^{z'} \frac{d\xi}{dz''} e^{2j\rho(z'')}\, dz''\, dz' \Bigg) \end{aligned} \tag{17}$$

and the *Out-of-Phase Quasi-Normal Mode* by

$$\begin{aligned} w_2(z) \cong e^{-j\rho(z)} \Bigg(& w_2(0) - \frac{1}{2} w_1(0) \int_0^z \frac{d\xi}{dz'} e^{2j\rho(z')}\, dz' \\ & - \frac{1}{4} w_2(0) \int_0^z \frac{d\xi}{dz'} e^{2j\rho(z')} \int_0^{z'} \frac{d\xi}{dz''} e^{-2j\rho(z'')}\, dz''\, dz' \Bigg) \end{aligned} \tag{18}$$

where

$$\rho(z) = \int_0^z \Gamma(\zeta)\, d\zeta \tag{19}$$

If $d\xi/dz = 0$, it is seen that these become the ordinary normal modes of (6).

V. SINGLE TAPERED MODE DIRECTIONAL COUPLERS

The power in the two tapered lines [by (14)] is

$$\begin{aligned} P_1(z) &= |E_1(z)|^2 = \cos^2\frac{\xi(z)}{2}|w_1(z)|^2 + \sin^2\frac{\xi(z)}{2}|w_2(z)|^2 \\ &\qquad - \sin\xi(z)Re(w_1(z)w_2^*(z)) \\ P_2(z) &= |E_2(z)|^2 = \sin^2\frac{\xi(z)}{2}|w_1(z)|^2 + \cos^2\frac{\xi(z)}{2}w_2(z)|^2 \\ &\qquad + \sin\xi(z)Re(w_1(z)w_2^*(z)) \end{aligned} \tag{20}$$

If only one quasi-normal mode is excited by putting power in only one line provided $c(z)$ and $\varphi(z)$ are chosen so that $\eta(z) \ll 1$ for all z, $\xi(0) = 0$, and $w_1(0) = 1$, $w_2(0) = 0$, then it is seen that the power in the two lines is approximate [by (17)–(20)]

$$\begin{aligned} P_1(z) &\cong \cos^2\tfrac{1}{2}\xi(z)\{1+\nu(z)\} + \sin^2\tfrac{1}{2}\xi(z)\mu(z) + \sin\xi(z)\delta(z) \\ P_2(z) &\cong \sin^2\tfrac{1}{2}\xi(z)\{1+\nu(z)\} + \cos^2\tfrac{1}{2}\xi(z)\mu(z) - \sin\xi(z)\delta(z) \end{aligned} \tag{21}$$

where

$$\begin{aligned} \mu(z) &= \frac{1}{4}\left|\int_0^z \frac{d\xi}{dz'}e^{2j\rho(z')}dz'\right|^2 \\ \nu(z) &= -\tfrac{1}{2}Re\left(\int_0^z \frac{d\xi}{dz'}e^{-2j\rho(z')}\int_0^{z'}\frac{d\xi}{dz''}e^{2j\rho(z'')}dz''dz'\right) \\ \delta(z) &= \tfrac{1}{2}Re\left(e^{2j\rho(z)}\int_0^z\frac{d\xi}{dz'}e^{-2j\rho(z')}dz'\right) \end{aligned} \tag{22}$$

Since power must be conserved, we must have $\mu(z) + \nu(z) = 0$. This requirement may easily be verified in the specific examples treated. If, furthermore, $\varphi(z)$ and $c(z)$ are chosen so that $\xi(\ell) = \pi$ (a 100 per cent coupler), it is seen from (21) that power can be transferred almost completely from line 1 in the in-phase quasi-normal mode to line 2 in the same quasi-normal mode by exciting power in one line only. In fact, by (17)–(22) we have for a 100 per cent coupler

$$\begin{aligned} P_1(\ell) &= |w_2(\ell)|^2 = \mu(\ell) \\ P_2(\ell) &= |w_1(\ell)|^2 = 1 + \nu(\ell) \end{aligned} \tag{23}$$

where μ and ν are given in equation (22). Thus, $\mu(\ell)$ gives a measure of the error involved in making a complete power transfer coupler of length ℓ in which only the w_1-mode is excited if $\varphi(z)$ and $c(z)$ are selected so that: (1) $\eta(z) \ll 1$, all z, and (2) $\xi(0) = 0$ and $\xi(\ell) = \pi$. Since $\mu(\ell) = |w_2(\ell)|^2$, $\mu(\ell)$ also gives a measure of the power in the initially non-excited mode that is present at the end of the coupler. It is therefore appropriate to call it the "mode crosstalk."

For other power divisions, although $\mu(\ell)$ (with appropriate boundary condition on $\xi(\ell)$) gives the power in the non-excited mode present at the end of the coupler, it does not necessarily give the error in making the desired power of division. For example, for a 3 db coupler ($\xi(\ell) = \pi/2$), we find

$$\begin{aligned} P_1(\ell) &\cong \tfrac{1}{2}(1 + \nu(\ell)) + \tfrac{1}{2}\mu(\ell) + \frac{1}{\sqrt{2}}\,\delta(\ell) \\ P_2(\ell) &\cong \tfrac{1}{2}(1 + \nu(\ell)) + \tfrac{1}{2}\mu(\ell) - \frac{1}{\sqrt{2}}\,\delta(\ell) \end{aligned} \tag{24}$$

Since $\mu(\ell) + \nu(\ell) = 0$ (energy conservation, $\delta(\ell)$ would be the more appropriate parameter for describing the performance of a 3 db coupler, although $\mu(\ell)$ is the mode crosstalk. $\delta(\ell)$ might be called the "interference error power" in this case. Although no actual examples will be worked out for 3 db couplers, by comparing $\mu(\ell)$ and $\delta(\ell)$ in equations (22), it is seen that the extension from the 100 per cent coupler cases considered is very easy.

Another case of interest might be that of sampling only a very small amount of power. In this case $\xi(\ell)$ is very small and

$$\begin{aligned} P_1(\ell) &\cong 1 + \nu(\ell) + \xi(\ell)\delta(\ell) \\ P_2(\ell) &\cong \mu(\ell) - \xi(\ell)\delta(\ell) \end{aligned} \tag{25}$$

In this case, $\mu(\ell)$, $\delta(\ell)$, and $\nu(\ell)$ are all needed to describe the coupler performance.

VI. FREQUENCY SENSTIVITITY OF SINGLE TAPERED MODE COUPLERS

The "mode crosstalk" is a parameter which measures the "goodness" or quality of a 100 per cent coupler. Since, in general φ and c depend on frequency, $\mu(\ell)$ will also depend on frequency. Thus if a 100 per cent coupler is required with a "mode cross-talk" less than or equal to ε (say $\varepsilon \leqq 0.01$), the frequency range over which $\mu(\ell,\omega) \leqq \varepsilon$ determines the bandwidth. Since, in general, the frequency dependence of φ and c is

not precisely known, we can also obtain a semi-quantitative estimate of the performance if we keep ω fixed and vary ℓ.

A 100 per cent coupler has strictly zero "mode crosstalk" only if it is infinitely long, except at isolated frequencies. Of course such a coupler would be flat over an infinite bandwidth. As a general principle, we infer that the longer the coupler the greater the bandwidth. This is obviously not the case for mode interference couplers where any addition to the optimum length causes increasingly serious deterioration of performance. To give a plausibility argument, the "mode crosstalk" may be integrated by parts to yield

$$\mu(\ell) \cong \frac{1}{4}\left|\frac{1}{2j}\left(\left(\frac{d\xi}{d\rho}\right)_{\rho(\ell)} - \left(\frac{d\xi}{d\rho}\right)_{\rho(0)}\right) - \left(\frac{1}{2j}\right)^2\left(\left(\frac{d^2\xi}{d\rho^2}\right)_{\rho(\ell)} - \left(\frac{d^2\xi}{d\rho^2}\right)_{\rho(0)}\right) + \cdots\right|^2 \tag{26}$$

If the series in (26) converges, then $\mu(\ell) \equiv 0$ if all derivatives, $d^n\xi/d\rho^n$, vanish at both ends of the coupler. Since $\mu(\ell)$ is identically zero, such a coupler would be flat for infinite bandwidth. Thus, to increase the bandwidth we make ξ vary slowly at the ends. However, the coupler must be made longer, in general, so that the weak hypercoupling approximation is not violated.

VII. COMPARISON OF BANDWIDTHS FOR SEVERAL CLASSES OF TAPERED MODE COUPLERS

In general, a compromise must be made between bandwidth and length of tapered mode couplers. Several classes of tapered mode couplers will be considered to illustrate this. Perhaps other variations of $\varphi(z)$ and $c(z)$ will eventually prove better than those considered here but until they are discovered we must be content with what we have. The ones illustrated here are chosen primarily for mathematical simplicity, but it will be shown that they should be useful for bandwidths of the order of 3:1, although physical length of no more than about three minimum local beat wave lengths are required.

Two classes of couplers will be considered which will be called, respectively, uniform single tapered mode couplers and constant local beat wave length couplers.

Class 1. Uniform Single Tapered Mode Couplers

This class is characterized by a constant taper (constant hypercoupling coefficient); i.e.,

$$\eta(z) = \frac{1}{2}\frac{d\xi}{d\rho} = p \tag{27}$$

where p is a constant. For couplers to satisfy (27), it can be shown that $\varphi(z)$ and $c(z)$ must be related by

$$\frac{\varphi(z)}{c(z)} = \frac{1 - 2\sigma(z)}{2\sqrt{\sigma(z) - \sigma^2(z)}} \tag{28}$$

where

$$\sigma(z) = \frac{\int_0^z c(\zeta)\, d\zeta}{\int_0^\ell c(\zeta)\, d\zeta} \tag{29}$$

and

$$p = -\frac{1}{\int_0^\ell c(\zeta)\, d\zeta} \tag{30}$$

In order that the weak hypercoupling requirement is satisfied, $p \ll 1$. For a 100 per cent coupler $\xi(\ell) = \pi$, p must satisfy (30).

TABLE I — MODE CROSSTALK FOR TWO SPECIAL CASES OF UNIFORM TAPERED MODE COUPLERS

	$\varphi(z)$	$c(z)$	$\sigma(z)$	$\xi(z)$	$\frac{1}{2}\left\lvert\frac{d\xi}{d\rho}\right\rvert$	Crosstalk	Range of Validity*
a)	$\frac{c(1-2\sigma)}{2\sqrt{\sigma-\sigma^2}}$	$c = \frac{\pi}{\lambda_{b_0}}$	$\frac{z}{\ell}$		$+\frac{\lambda_{b_0}}{\ell\pi}$	$\frac{\sin^2 \pi\left(\frac{\pi\ell}{2\lambda_{b_0}}\right)}{4\left(\frac{\pi\ell}{2\lambda_{b_0}}\right)^2}$	$\frac{\ell}{\lambda_{b_0}} \geqq 1$
b)	$\frac{\pi}{\lambda_{b_0}}\cos\xi(z)$	$\frac{\pi}{\lambda_{b_0}}\sin\xi(z)$	$\sin^2\frac{\pi z}{2\ell}$	$\frac{\pi z}{\ell}$	$\frac{\lambda_{b_0}}{2\ell}$	$\frac{\sin^2\left(\frac{\pi\ell}{\lambda_{b_0}}\right)}{4\left(\frac{\ell}{\lambda_{b_0}}\right)^2}$	$\frac{\ell}{\lambda_{b_0}} \geqq \frac{\pi}{2}$

* The criterion for establishing the range of validity is somewhat arbitrary, but it is taken the same for all cases considered.

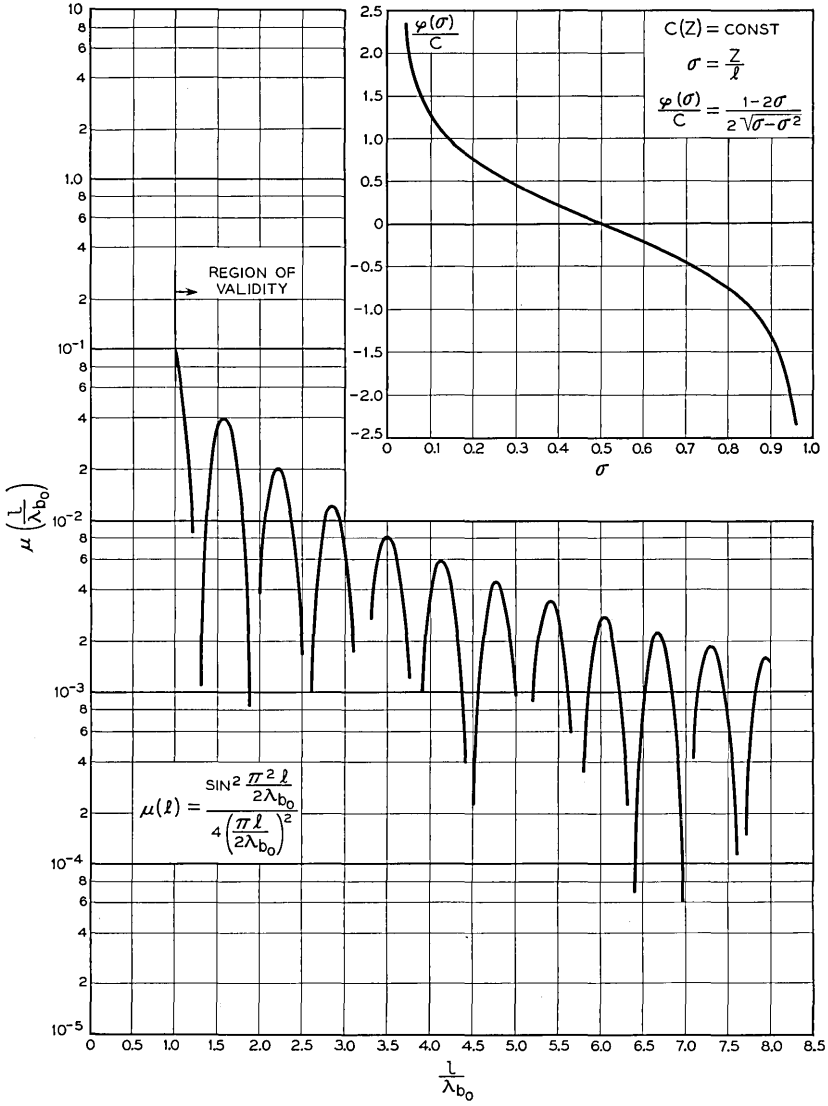


Fig. 1 — Mode crosstalk, $\mu(\ell/\lambda_{b_0})$, for a uniform tapered mode coupler as a function of coupler wavelength in minimum local beat wavelength units (ℓ/λ_{b_0}). Insert — Phase "constant" variation $\varphi(z/\ell)$ as a function of distance along coupler (z/ℓ) in units of coupler length for constant coupling coefficient $c(z)$.

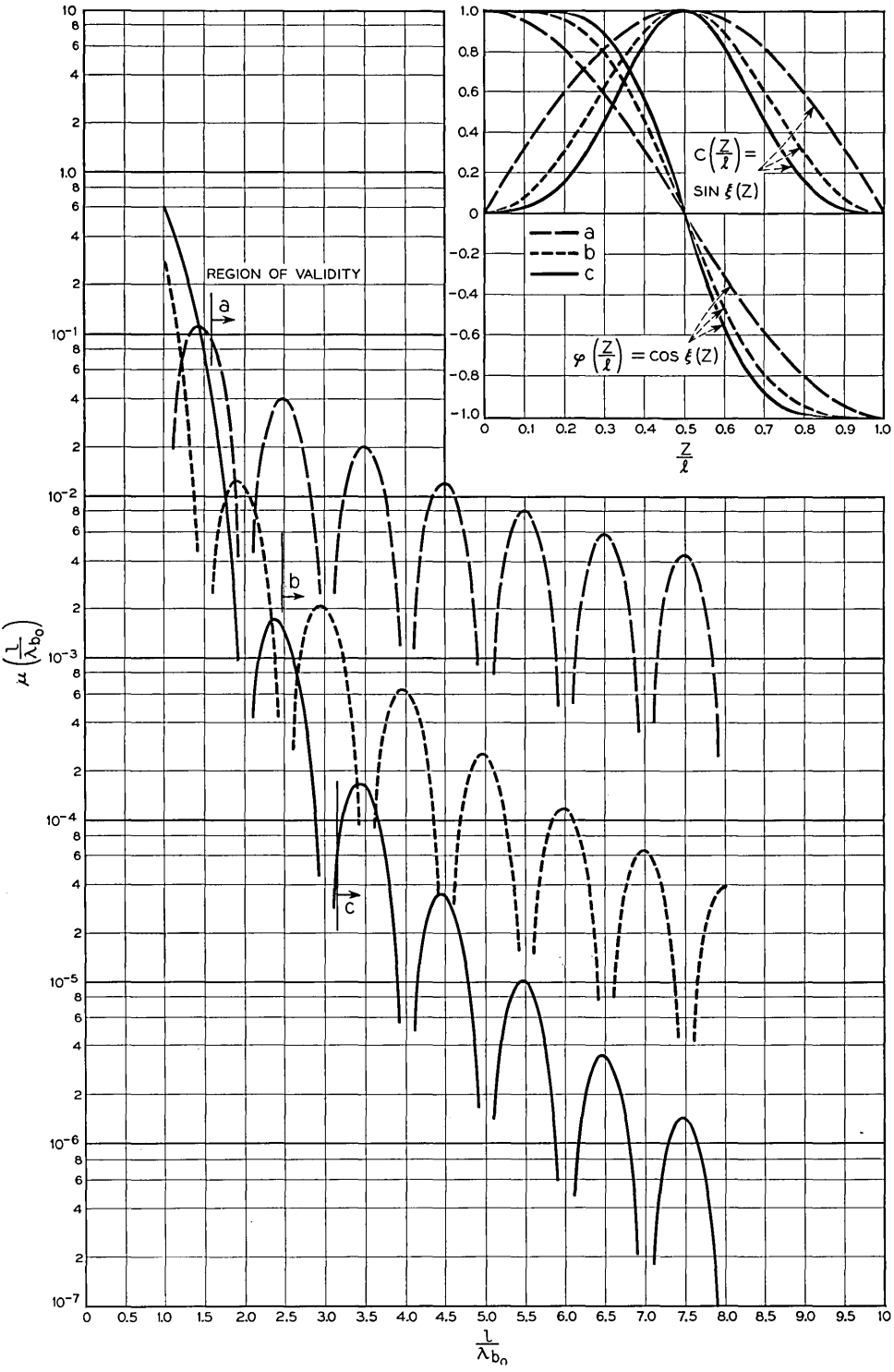

REGION OF VALIDITY
a
b
c
$\mu\left(\frac{l}{\lambda_{b_0}}\right)$
$\frac{l}{\lambda_{b_0}}$
$C\left(\frac{Z}{l}\right) =$ SIN $\xi(Z)$
$\varphi\left(\frac{Z}{l}\right) =$ COS $\xi(Z)$
$\frac{Z}{l}$

Fig. 2 — Mode crosstalk, $\mu(\ell/\lambda_{b_0})$ for three constant local beat wavelength couplers as a function of the coupler length in minimum local beat wavelength units (ℓ/λ_{b_0}). Cot $\xi(z)$ = ratio of phase "constant" variation to coupling coefficient variation. Insert — Phase "constant" variations $\varphi(z/\ell)$ and coupling coefficient variations $c(z/\ell)$ as functions of distance along the couplers in units of coupler length.

$$\text{Curve (a): } \xi(z) = \frac{\pi z}{\ell}$$

$$\text{Curve (b): } \xi(z) = \pi \sin^2 \frac{\pi z}{2\ell}$$

$$\text{Curve (c): } \xi(z) = \pi \left[\frac{z}{\ell} - \frac{1}{2\pi} \sin \frac{2\pi z}{\ell} \right]$$

The crosstalk for these three cases is given in Table 2.

$$P_1(\ell) = \frac{1}{1 + y^2} \sin^2 \frac{\pi}{2} \sqrt{1 + y^2}$$

$$P_2(\ell) = \cos^2 \frac{\pi}{2} \sqrt{1 + y^2} + \frac{y^2}{1 + y^2} \sin^2 \frac{\pi}{2} \sqrt{1 + y^2}$$

where $y = \ell/\lambda_{b_0}$.

The mode crosstalk for this class of couplers is seen to be

$$\mu(\ell) = \frac{\sin^2\left[\pi \int_0^\ell c(\zeta)\, d\zeta\right]}{\left[\int_0^\ell c(\zeta)\, d\zeta\right]^2} \tag{31}$$

Two special cases are considered:

$$\text{a)} \qquad c(z) = c \qquad \text{(Treated by Cook)} \tag{32}$$

$$\text{b)} \qquad \varphi(z) = \frac{\pi}{\lambda_{b_0}} \cos \xi(z) \qquad \text{(Treated by Fox)}$$

$$c(z) = \frac{\pi}{\lambda_{b_0}} \sin \xi(z) \tag{33}$$

$$\xi(z) = \frac{\pi z}{\ell}$$

The results for these two cases are summarized in Table I. The variations

TABLE II — MODE CROSSTALK FOR THREE SPECIAL CASES OF CONSTANT LOCAL BEAT WAVELENGTH COUPLERS

$\xi(z)$	Mode Crosstalk	Range of Validity*
a) $\frac{\pi z}{l}$	$\frac{\sin^2\left(\frac{\pi \ell}{\lambda_{b_0}}\right)}{4\left(\frac{\ell}{\lambda_{b_0}}\right)^2}$	$\frac{\ell}{\lambda_{b_0}} \geqq \frac{\pi}{2}$
b) $\pi \sin^2 \frac{\pi z}{2\ell}$	$\pi^2 \frac{\cos^2\left(\frac{\pi \ell}{\lambda_{b_0}}\right)}{64\left(\frac{\ell}{\lambda_{b_0}}\right)^4}$	$\frac{\ell}{\lambda_{b_0}} \geqq \frac{\pi^2}{4}$
c) $\pi\left[\frac{z}{\ell} - \frac{1}{2\pi} \sin^2 \frac{\pi z}{2\ell}\right]$	$\frac{\sin^2\left(\frac{\pi \ell}{\lambda_{b_0}}\right)}{4\left(\frac{\ell}{\lambda_{b_0}}\right)^6}$	$\frac{\ell}{\lambda_{b_0}} \geqq \pi$

* The criterion for establishing the range of validity is somewhat arbitrary, but it is taken the same for all cases considered.

of $\varphi(z)$, $c(z)$ and the mode crosstalk are plotted in Figs. 1 and 2 for these cases.*

Class 2. Constant Local Beat Wavelength Couplers

This class is characterized by

$$\Gamma(z) = \sqrt{\varphi^2(z) + c^2(z)} = \pi/\lambda_{b_0} = \text{constant}. \tag{34}$$

For couplers to have constant local beat wavelengths, the phase constant difference and coupling coefficient must satisfy the following:

$$\begin{aligned} \varphi(z) &= \frac{\pi}{\lambda_{b_0}} \cos \xi(z) \\ c(z) &= \frac{\pi}{\lambda_{b_0}} \sin \xi(z) \\ \Gamma(z) &= \frac{\pi}{\lambda_{b_0}} \equiv \Gamma. \end{aligned} \tag{35}$$

The mode crosstalk is seen to be [by (22)]

$$\mu(\ell) = \frac{1}{4} \left| \int_0^{\ell} \frac{d\xi}{dz'} e^{2j\Gamma z'} dz' \right| \tag{36}$$

Several examples of this class will be considered†

$$\text{a)} \qquad \xi(z) = \frac{\pi z}{\ell} \tag{37}$$

It is seen that this is also a member of Class 1 so it will not be discussed further.

$$\text{b)} \qquad \xi(z) = \pi \sin^2 \frac{\pi z}{2\ell} \tag{38}$$

$$\text{c)} \qquad \xi(z) = \pi \left[\frac{z}{\ell} - \frac{1}{2\pi} \sin 2 \frac{\pi z}{\ell} \right] \tag{39}$$

The mode crosstalk and region of validity for these cases are given in Table II and the variations $\varphi(z)$, $c(z)$ and the crosstalk are plotted in Fig. 2.

As can be seen from Table II, for Case a, $d\xi/d\rho \neq 0$ at either end of the coupler, while for Case b, $d\xi/d\rho = 0$ at both ends of the coupler and

* It may be noted in case (b) in which $\Gamma = \pi/\lambda_{b_0}$ and $d\xi/dz = \pi/\ell$ that (15) may be solved exactly. There are pure normal modes which are elliptically polarized in the normal coordinates $w_1(z)$ and $w_2(z)$. If power is injected in line 1 only, it is found that the power in the two lines at $z = \ell$ is

† It may easily be verified directly in these cases that $\mu(\ell) + \nu(\ell) = 0$ in agreement with energy conservation. See remarks following (22).

for Case c, $d\xi/d\rho$ and $d^2\xi/d\rho^2$ are zero at both ends of the coupler. Comparison of the mode crosstalk for these three cases illustrates the general remarks made concerning (26), viz., as higher order derivatives of the taper vanish at the ends of the coupler, the mode crosstalk becomes less. Again by comparing the range of validity in the three cases of Table II, it is seen the coupler length must be increased to satisfy the weak hypercoupling requirement, so that the process cannot be carried infinitely far in this direction. Presumably at very short wavelengths where physical size is not as important, this principle can be used most advantageously.

ACKNOWLEDGMENTS

The author is indebted to J. R. Pierce, L. R. Walker, R. Kompfner, J. S. Cook, and S. P. Morgan, for very helpful discussions concerning this problem, and to Mrs. C. A. Lambert for the computations of Figs. 1 and 2.

BIBLIOGRAPHY

1. W. W. Mumford, Proc. I.R.E., **35**, p. 160, Feb., 1947; S. E. Miller and W. W. Mumford, Proc. I.R.E., **40**, p. 1071, Sept., 1952; H. J. Riblet, Proc. I.R.E., **35**, p. 1307, Nov., 1947; and S. E. Miller, B.S.T.J., **33**, p. 661, May, 1954, and references contained therein.
2. J. S. Cook, page 807 of this issue.
3. A. G. Fox, page 823 of this issue.

Bell System Technical Papers Not Published in This Journal

ANDERSON, O. L.,[1] and BOMMEL, H. E.[1]

Ultrasonic Absorption in Fused Silica at Low Temperatures and High Frequencies, J. Am. Ceramic Soc., **38,** pp. 125–131, April, 1955.

BOMMEL, H. E., see Anderson, O. L.

BUEHLER, E., See Fuller, C. S.

CALLAWAY, JOSEPH[1]

Orthogonalized Plane Wave Method, Phys. Rev., **97,** pp. 933–936, Feb. 15, 1955.

COOPER, H. G.,[1] KOEHLER, J. S.,[5] and MARX, J. W.[6]

Irradiation Effects in Cu, Ag, and Au Near 10°K, Phys. Rev., **97,** pp. 599–607, Feb. 1, 1955.

CUTLER, C. C.,[1] and HINES, M. E.[1]

Thermal Velocity Effects in Electron Guns, Proc. I.R.E., **43,** pp. 307–315, March, 1955.

CUTLER, C. C.,[1] and SALOOM, J. A.[1]

Pin-Hole Camera Investigation of Electron Beams, Proc. I.R.E., **43,** pp. 299–306, March, 1955.

DAVEY, J. R.,[1] and HANLEY, F. H.,[2] and PURVIS, M. R.[1]

A New Telegraph Serviceboard Using Electronic Circuits, A.I.E.E. Commun. and Electronics, **17,** pp. 30–37, March, 1955.

DITZENBERGER, J. A., see Fuller, C. S.

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company
[5] University of Illinois, Urbana, Illinois
[6] Phillips Petroleum Company, Bartlesville, Oklahoma

EDER, M. MISS,[1] WARNER, R.,[1] and KEENE, F.[1]

Statistically Designed Experiment of the Factorial Type Applied to Point-Contact Transistors, Proc. I.R.E. National Symposium on Quality Control and Reliability in Electronics, Nov. 12–13, 1954.

FULLER, C. S.,[1] DITZENBERGER, J. A.,[1] HANNAY, N. B.,[1] and BUEHLER, E.[1]

Resistivity Changes in Silicon Single Crystals Induced by Heat Treatment, Letter to the Editor, Acta Metallurgica, **3,** pp. 97–99, Jan., 1955.

GAMBRILL, R. D.[3]

Controlling Extrusion of Foam Plastic on Wire, Electronics, **28:** pp. 144–145, April, 1955.

GELLER, S.,[1] MATTHIAS, B. T.,[1] and GOLDSTEIN, R.[1]

Some New Intermetallic Compounds with the "β-Wolfram" Structure, J. Am. Chem. Soc., **77,** pp. 1502–1504, Mar. 20, 1955.

GELLER, S.[1]

Note on the Structure of Dimethylamine-Boron Trifluoride, Letter to the Editor, Acta Crystallographica, **8,** p. 120, Feb., 1955.

GELLER, S.,[1] and WOLONTIS, V. M.[1]

The Crystal Structure of CO_2Si, Acta Crystallographica, **8,** p. 83, Feb., 1955.

GILMAN, G. W., see Kelly, Mervin J.

GOLDSTEIN, R., see Geller, S.

HALSEY, R. J., see Kelly, Mervin J.

HANLEY, F. H., see Davey, J. R.

HANNAY, N. B., see Fuller, C. S.

HEARN, A. H.[1]

Creosote Retention as Determined by Toluene Extraction of Treated Wood, Annual Proc. of A.W.P.A., Feb., 1955.

HEIDENREICH, R. D.[1]

Transition Structure in Lead-Silver Alloys and a Dislocation Mechanism, Acta Metallurgica, **3,** pp. 79–86, Jan., 1955.

[1] Bell Telephone Laboratories, Inc.
[3] Western Electric Company, Inc., Baltimore, Md.

HINES, M. E., see Cutler, C. C.

HOHN, FRANZ[1]

Some Mathematical Aspects of Switching, Am. Math. Monthly, **62,** pp. 75–90, Feb., 1955.

INSKIP, L. S.,[1] and WATSON, H. N.[9]

Grounding of Portable Electric Equipment, Elec. Engg., **74,** pp. 286–291, April, 1955.

JAYCOX, E. K.[1]

Spectrochemical Procedure of General Applicability, Anal. Chem., **27,** pp. 347–350, Mar., 1955.

KELLY, MERVIN J.,[1] RADLEY, SIR GORDON,[4] GILMAN, G. W.,[1] and HALSEY, R. J.[4]

A Transatlantic Telephone Cable, A.I.E.E. Commun. and Electronics, **17,** pp. 124–136, March, 1955.

KEENE, F., see Eder, M. Miss.

KOEHLER, J. S., see COOPER, H. G.

KOHN, W., see Luttinger, J. M.

KOHN, W.[8] and LUTTINGER, J. M.[7]

Hyperfine Splitting of Donor States in Silicon, Phys. Rev., **97,** pp. 883–888, Feb. 15, 1955.

LOUISELL, W. H.,[1] and PIERCE, J. R.[1]

Power Flow in Electron Beam Devices, Proc. I.R.E., **43,** pp. 425–427, April, 1955.

LUTTINGER, J. M., see Kohn, W.

LUTTINGER, J. M.,[7] and KOHN, W.[8]

Motion of Electrons and Holes in Perturbed Periodic Fields, Phys. Rev., **97,** pp. 869–883, Feb. 15, 1955.

[1] Bell Telephone Laboratories, Inc.
[4] Post Office of the United Kingdon, London, England
[7] University of Michigan, Ann Arbor, Michigan
[8] Carnegie Institute of Technology, Pittsburgh, Pa.
[9] General Electric Co., Bridgeport, Conn.

MARX, J. W., see Cooper, H. G.

MATTHIAS, B. T., see Geller, S.

MCCARTHY, JOHN A.[1]

Search For Double Beta Decay in Ca, Phys. Rev., **97,** pp. 1234–1236, Mar. 1, 1955.

MCSKIMIN, H. J.[1]

Transducer Design for Ultrasonic Delay Lines, J. Acous. Soc., **27,** pp. 302–309, Mar., 1955.

MENDEL, J. T.[1]

Magnetic Focusing of Electron Beams, Proc. I.R.E., **43,** pp. 327–331, March, 1955.

MERTZ, PIERRE[1]

Transmission Line Characteristic and Effects on Pulse Transmission, Symposium on Information Networks, pp. 85–114, Jan., 1955.

PIERCE, J. R.[1]

Propagation in Linear Arrays of Parallel Wires, I.R.E. Trans. on Electron Devices, PGED-2, pp. 13–24, Jan., 1955.

PIERCE, J. R.[1]

Orbital Radio Relays, Jet Propulsion, **25,** pp. 76–78, Feb., 1955.

PIERCE, J. R., see Louisell, W. H.

PURVIS, M. R., see Davey, J. R.

RADLEY, SIR GORDAN, see Kelly, Mervin J.

RAISBECK, G.[1]

Order of Magnitude of the Fourier Coefficients in Functions Having Isolated Singularities, Am. Math. Monthly, **62,** pp. 149–154, Mar., 1955.

READ, W. T., JR.[1]

Scattering of Electrons by Charged Dislocations in Semiconductors, Phil. Mag., **373,** pp. 111–131, Feb., 1955.

[1] Bell Telephone Laboratories, Inc.

ROWEN, JOHN, H.[1]

Ferromagnetism at Microwave Frequencies and Its Applications, Rad. Electronics Engg., **24,** pp. 26–28 and 40–41, April, 1955.

SALOOM, J. A., see Cutler, C. C.

SCALES, Miss E. M.,[1] and CHAPANIS, A.[1]

Effect on Performance of Tilting the Toll-Operator's Keyset, J. Appl. Psychology, **38,** pp. 452–456, Dec., 1954. (listed in March issue as published by the Journal of Applied Physics.)

SOUTHWORTH, G. C.[1]

The Challenge, I.R.E. Trans. on Microwave Theory and Techniques, PGMTT-3, No. 1, pp. 2–3, Jan., 1955.

TIEN, P. K.,[1] WALKER, L. R.,[1] and WOLONTIS, V. M.[1]

A Large Signal Theory of Traveling-Wave Amplifiers, Proc. I.R.E., **43,** pp. 260–277, March, 1955.

TALPEY, T. E.[1]

The Nature of the Uncorrelated Component of Induced Grid Noise, Proc. I.R.E., **43,** pp. 449–454, April, 1955.

WALKER, L. R., see Tien, P. K.

WARNER, R., see Eder, M. Miss.

WATSON, H. N., see Inskip, L. S.

WOLONTIS, V. M., see Geller, S.

WOLONTIS, V. M., see Tien, P. K.

VOGEL, F. L., JR.[1]

Dislocations in Polygonized Germanium, Letter to the Editor, Acta Metallurgica, **3,** pp. 95, Jan., 1955.

[1] Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BRATTAIN, W. H., and GARRETT, C. G. B.

Surface Properties of Semiconductors, Monograph 2412.

CUTLER, C. C., and HINES, M. E.

Thermal Velocity Effects in Electron Guns, Monograph 2379.

CUTLER, C. C., and SALOOM, J. A.

Pin-Hole Camera Investigation of Electron Beams, Monograph 2409.

DECOSTE, J. B., and WALLDER, V. T.

Weathering of Polyvinyl Chloride, Monograph 2376.

FINE, M. E.

Measuring the Elastic Moduli and Internal Friction of Solids, Monograph 2357.

GARRETT, C. G. B., see Brattain, W. H.

GELLER, S.

Crystal Structures of Rh_2Ge, Rh_5Ge_3 and RhGe, Monograph 2408.

GOSS, A. J., see Tannenbaum, M.

GREEN, E. I.

Creative Thinking in Scientific Work, Monograph 2257.

HINES, M. E., see Cutler, C. C.

MAITA, J. P., see Morin, F. J.

---

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

MASON, W. P.

Magnetostriction and Anisotropic Energies in Crystals, Monograph 2316.

MASON, W. P., and WICK, R. F.

Ferroelectrics and the Dielectric Amplifier, Monograph 2326.

MCKAY, K. G.

Avalanche Breakdown in Silicon, Monograph 2292.

MCMILLAN, BROCKWAY

Absolutely Monotone Functions, Monograph 2333.

MENDEL, J. T.

Magnetic Focusing of Electron Beams, Monograph 2413.

MORIN, F. J., and MAITA, J. P.

Electrical Properties of Silicon Containing Arsenic and Boron, Monograph 2342.

MURPHY, E. J.

Surface Migration of Water Molecules in Ice, Monograph 2288.

OLIVER, B. M.

Directional Electromagnetic Couplers, Monograph 2329.

OLMSTEAD, P. S.

Distribution of Sample Arrangements for Runs Up and Down, Monograph 2289.

PFANN, W. G., see Tanenbaum, M.

PFANN, W. G., and VAN ROOSBROECK, W.

Radioactive and Photoelectric p-n Junction Power Sources, Monograph 2325.

SALOOM, J. A., see Cutler, C. C.

SULLIVAN, J. W.

Wide-Band Voltage-Tunable Oscillator, Monograph 2327.

TANENBAUM, M., GOSS, A. J., and PFANN, W. G.

Purification of Antimony and Tin by Zone Refining, Monograph 2414.

TIEN, P. K., WALKER, L. R., and WOLONTIS, V. M.

Large Signal Theory of Traveling-Wave Amplifiers, Monograph 2382.

THOMAS, D. E.

VHF Point-Contact Transistor Stability Problems Monograph 2328.

VAN ROOSRROECK, W., see Pfann, W. G.

WALKER, L. R., see Tien, P. K.

WALLDER, V. T., see DeCoste, J. B.

WICK, R. F., see Mason, W. P.

WOLONTIS, V. M., see Tien, P. K.

Contributors to this Issue

ALBERT M. CLOGSTON, B.S., 1938; Ph.D., 1941, Massachusetts Institute of Technology. He joined Bell Telephone Laboratories in 1946 and has worked on various electron tube devices, a new type of magnetron, storage tubes and other information handling devices. A member of the Research in Physical Sciences Department at Murray Hill for the last two years, Dr. Clogston is concerned with the physics of solids. Fellow of American Physical Society, member of the Institute of Radio Engineers and Sigma Xi.

J. S. COOK, B.E.E., M.S., Ohio State University, 1952. He joined Bell Telephone laboratories the same year and has been engaged principally in research on the traveling wave tube. He is now engaged in electronics research at Murray Hill as a member of the Research in High Frequency and Electronics Department. A member of the Institute of Radio Engineers, he belongs to the Professional Group on Electron Devices.

STEPHEN DOBA, JR., College of the City of New York and Cooper Union. He joined Bell Telephone Laboratories in 1926 and was concerned at first with voice operated devices in Circuit Research. In 1938 he joined the Transmission Development Department and was engaged in work on television transmission systems. From 1940 to 1945 he aided in the development of airborne radar equipment. More recently he was named Transmission Systems Development Engineer in charge of groups engaged in development of television systems, submarine cable systems, and Bell System field test equipment. Mr. Doba holds some 30 patents. He is a Senior Member of the Institute of Radio Engineers.

J. JAMES EBERS, B.S., Antioch College, 1946; M.S., Ph.D., Ohio State University, 1947, 1950. Bell Telephone Laboratories, 1951–. Prior to joining the Laboratories, Dr. Ebers served as an Instructor in Electrical Engineering at Ohio State University from 1947 to 1950, and as Assistant Professor from 1950 to 1951. He worked as a Research Foundation Assistant and Associate at Ohio State University from 1946 to 1951. His early work at the Laboratories was concerned with the development of

transistors for switching applications, and for the past 2½ years he has been concerned with the development of the alloyed junction transistor. Member of the American Physical Society, Eta Kappa Nu, and Sigma Xi. Member of the I.R.E. and past chairman of its task force to standardize methods of tests for transistors in switching applications.

A. GARDNER FOX, S.B., Massachusetts Institute of Technology, 1934; S.M., Massachusetts Institute of Technology, 1935. Bell Telephone Laboratories, 1936–. Since 1944, Mr. Fox has been concerned with the design of radio-frequency amplifiers and with research on millimeter waves at the Holmdel Radio Laboratory. From 1942–1944 he designed radio transmission filters and antennas for a fire-control radar system. Prior to this he did research on waveguides, development work on radar, and the design of mobile and airborne radio transmitters. Senior member of the I.R.E., and past chairman of its Committee on Antennas and Waveguides.

A. ROBERT KOLDING, B.S.E.E., Brooklyn Polytechnic Institute, 1940: Bell Telephone Laboratories, 1930–. In his early years as a member of the Transmission Development Department, he was concerned with the development of television terminal equipment for long-distance coaxial cable transmission. During World War II, he was engaged in the development of radar bombsights for the Armed Forces. Following the war he worked on television terminals of the L-3 coaxial cable system, and is presently concerned with the development of local wire television systems. Member of the Institute of Radio Engineers and Eta Kappa Nu.

GRACE L. LAKIN, B.Sc., New Jersey College for Women, 1942; M.A., Columbia University, 1949; Bell Telephone Laboratories, 1942–. From 1942 to 1948, Miss Lakin worked principally on the development of adjustable A and B equalizers for the L-1 coaxial system. She was later concerned with the design and development of filters and equalizers for the telephotograph level compensator, and with regulators and equalizers for a 12-channel military communication system. For the past three years, Miss Lakin has been engaged in work on the equalization of the A2A video transmission system.

WILLIAM H. LOUISELL, University of Florida; Kenyon College; B.S. in Physics, University of Michigan, 1948; M.S. in Physics, University of Michigan, 1949; Ph.D. in Physics, University of Michigan, 1953;

Engineering Research Institute of University of Michigan, 1948–1953; Bell Telephone Laboratories, 1953 — . He has been concerned with the study of a new microwave coupling principle and the application of ferrites to traveling wave tubes. Dr. Louisell is a member of the American Physical Society, Phi Eta Sigma, Phi Kappa Phi, Sigma Xi and Gamma Alpha.

C. A. LOVELL, B.A., Mississippi College, 1922; M.A., 1928, Ph.D., 1932, University of Pennsylvania; Bell Telephone Laboratories, 1929–. Prior to joining the technical staff of the Laboratories, he taught Mathematics at Mississippi College and Drexel Institute in addition to doing graduate study. His early work at the Laboratories concerned loudspeaker design in the Acoustics Research Department. During a short absence from this department he worked on television terminal equipment for the New York-Philadelphia coaxial cable. He returned to his former department to take charge of the design of mechanical means for making experimental studies of telephone traffic. He is presently Director of Switching Systems Development III. Early in World War II he directed the development of the first electrical gun director and carried out similar work throughout the War. Fellow of the Acoustical Society of America; Member, American Institute of Physics and Franklin Institute of Philadelphia. Awarded Medal for Merit by President Truman in 1946, and Howard N. Potts Medal, Franklin Institute, 1948.

J. H. MCGUIGAN, S.B. and S.M. Massachusetts Institute of Technology, 1941. Bell Telephone Laboratories, 1941–. His early work concerned radar development, and later, development of television transmission facilities. In 1946 he joined the Switching Research Department, where he engaged in the development of signaling systems and magnetic drum recording. He is presently concerned with the development of magnetic core switching circuits. Associate Member of the I.R.E., and member of Sigma Xi, Eta Kappa Nu and Tau Beta Pi.

S. L. MILLER, B.S., Webb Institute of Naval Architecture, 1944; M.A. in Physics, Columbia University, 1949; Ph.D. in Physics, Columbia University, 1952; taught at City College of New York, 1948–1950; Bell Telephone Laboratories, 1952 — . Since he has been at the Laboratories Dr. Miller has been engaged in exploratory development work on transistors. He is a member of the American Physical Society and Sigma Xi.

O. J. MURPHY, B. S. in Electrical Engineering, University of Texas, 1927; Columbia University, 1928–31. Bell Telephone Laboratories, 1927–. Mr. Murphy's early Laboratories projects included studies of voice-operated switching devices, effects of transmission delay on two-way telephone conversation, and voice-frequency signaling systems. During World War II he was concerned with the design and development of the M-9 electrical gun director and related projects. After the war he resumed his research work on signaling systems and more recently has concentrated on the design of magnetic drum digital data storage apparatus and circuits. He is a member of the A.I.E.E. and a senior member of the I.R.E.

PHILIP W. ROUNDS, A.B., Harvard University, 1929; Bell Telephone Laboratories, 1929 — . Prior to World War II, Mr. Rounds was concerned with the development of transmission networks for toll telephone, telephoto and program systems. During World War II, he developed computing networks for anti-aircraft gun directors and bombsights. During this time he also developed transmission networks for sonar systems. Since the war period, he has been working with the development of transmission networks for television systems. He is a member of the American Institute of Electrical Engineers.