The Bell System Technical Journal

Vol. XXVIII *April, 1949* *No. 2*

A Carrier System for 8000-Cycle Program Transmission

By R. A. LECONTE, D. B. PENICK, C. W. SCHRAMM, A. J. WIER

With the rapid expansion of broad-band carrier telephone systems throughout the country, the use of these facilities for program transmission has become desirable. This paper describes a carrier program system capable of transmitting a band up to about 8000 cycles wide.

Introduction

FROM the beginning of radio the Bell System has supplied the broadcasting industry the needed interconnecting links between broadcasting stations, studios, and other program originating points. For many years these facilities have been provided at audio frequency over loaded cable pairs,[6] or over open-wire lines.[8] Because present growth of message facilities over main traffic routes is predominantly in broad-band carrier telephone circuits, it has become desirable to adapt these new carrier facilities for the transmission of high-quality program material.

The carrier program system to be described operates in conjunction with message circuits and can be used to provide a band width of either 5000 or 8000 cycles. It can be applied to type K multipair cable,[9] type L coaxial cable,[16] and type J open-wire carrier systems.[11] Use of the 8000-cycle band of course requires more complete equalization than the 5000-cycle band, and requires the frequency space normally occupied by three message channels. It is expected that the 5000-cycle band can be accommodated by displacing two message channels. The carrier program system was developed by 1942 but, owing to the war, its first commercial application was not made until early in 1946 on the transcontinental type K route west of Omaha. It is now in use in all sections of the country, particularly the west and south, on type L as well as type K systems and has been successfully tested on type J. In general, a band width of 5000 cycles is used in these applications.

Objectives

Existing audio-frequency program circuits may be as long as 7000 miles, may have 100 or more dropping or bridging points, any one of which may occasionally transmit to all of the others, and may be arranged for automatic reversal of the direction of transmission by means of a control signal.

In order to coordinate with these existing circuits and studio loops, a carrier program system must be capable of duplicating this flexibility while maintaining the desired standards of quality of transmission.

In setting an objective for the standards of transmission quality of this new system the trend towards wider band widths has been recognized. Most of the major networks now use a 100 to 5000-cycle band width. A large part of the present audio-frequency cable facilities, however, can be arranged to transmit a band from 50 to 8000 cycles. It was decided to match this grade of transmission in the design of the new carrier system. For the cases where still higher quality is desired, a 15-kilocycle carrier program system has been developed and is now available.

Design Features

The 12-channel bank of message circuits forms the basic building block of the broad-band carrier telephone systems. In the channel bank, each of the 12 voice-frequency channels modulates one of 12 carriers spaced 4 kilocycles apart from 64 to 108 kilocycles. The lower sideband resulting from each modulation is selected by a band filter and combined with the other 11 lower sidebands to give a channel group occupying the frequency space from 60 kilocycles to 108 kilocycles. This channel group is then further modulated as a unit to its appropriate place on a broad-band spectrum for transmission over the line.

In order to arrange a channel bank for program transmission, message channels, 6, 7, and 8 are disabled, clearing a space from 76 kilocycles to 88 kilocycles in the group-frequency spectrum. In a program terminal separate from the channel bank, an audio frequency program modulates an 88-kilocycle carrier derived from the message channel carrier supply. Its lower sideband is selected by a band filter and, combined with the lower sidebands of message channels 1 to 5 and 9 to 12, gives a group-frequency spectrum shown diagrammatically in Fig. 1. This figure also shows the same spectrum after it has been modulated with a 120-kilocycle group carrier for transmission over a type K line. Other line-frequency spectra are similarly produced in type J and type L group modulators.

The reversing and control signal in an audio-frequency program circuit is a d-c. signal superimposed on the program pair. It may be applied at the studio which originates the program, and conditions all of the amplifiers along the line to transmit away from the originating studio. As long as the signal is applied, the direction of transmission is locked so that no other control station can inadvertently break the network. When the transmission from this studio ends, the signal is removed, and the next originating point applies it. This effects such reversals as are required for transmission and again locks all amplifiers. By this means it is possible to use

a single pair of wires and one set of amplifiers for transmission in either direction as required. The carrier system over which the carrier program channel is transmitted is constantly in operation in both directions simultaneously, and therefore requires no reversal. The program terminal itself, however, must be switched between transmitting and receiving lines if equipment is not to be needlessly duplicated for transmitting and receiving. In any case, a control signal must be carried through the carrier circuit and delivered to connecting audio-frequency circuits at the receiving end as a d-c. signal. This is accomplished by means of a 78-kilocycle control signal (42 kilocycles at K line frequencies) which is transmitted along

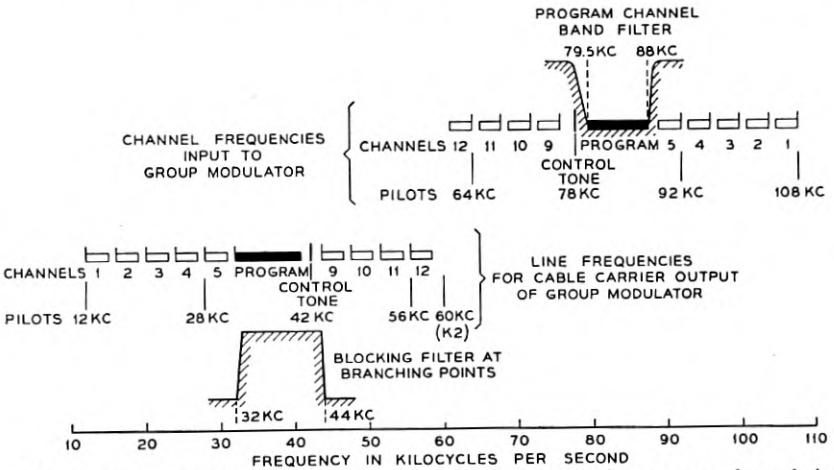


Fig. 1—Frequency allocation for one program channel and nine message channels in cable carrier systems.

with the program channel outside of its frequency band. This signal is generated in the transmitting program terminal whenever the d-c. signal is impressed from the transmitting audio-frequency circuit. At the receiving program terminal, the tone is converted into a d-c. signal which is impressed on the receiving voice-frequency facility. When there is no transmitted d-c. signal, there is no high-frequency signal and no received d-c. signal. Each program terminal, then, is ready either to receive d-c. from the voice circuit and send out 78 kilocycles to the carrier circuit or to receive 78 kilocycles from the carrier circuit and send out d-c. to the voice circuit. The program transmission path is maintained in the last established direction, regardless of the presence or absence of control, until a reversing signal is received.

The arrangement of the circuit elements in a carrier program terminal is shown in the block schematic of Fig. 2. The transmission circuit wiring is

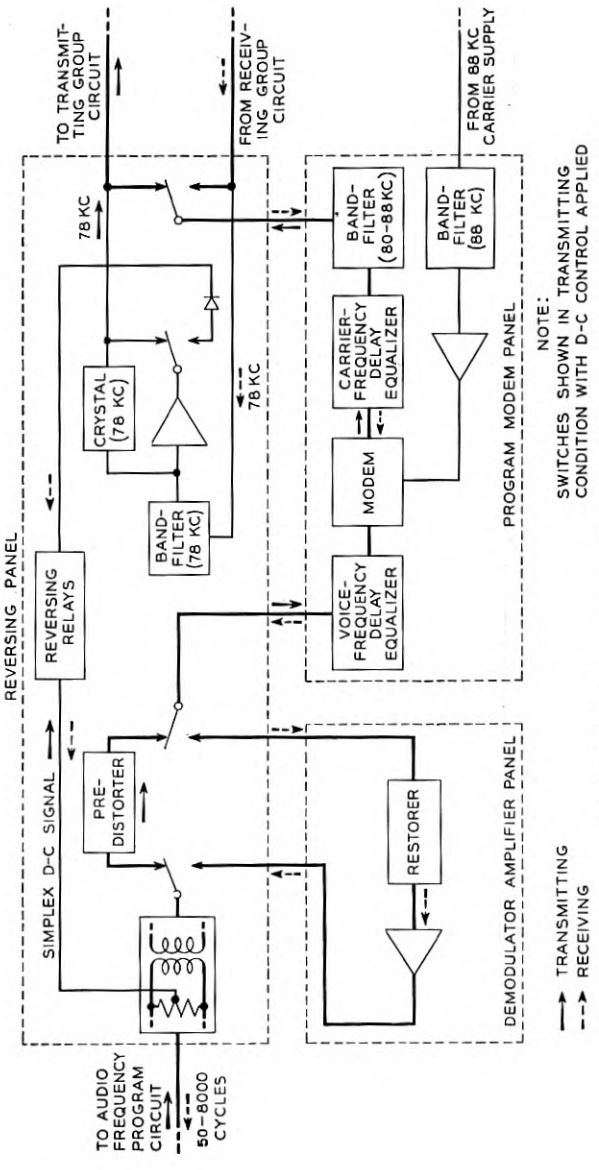


Fig. 2—Block schematic of carrier program terminal.

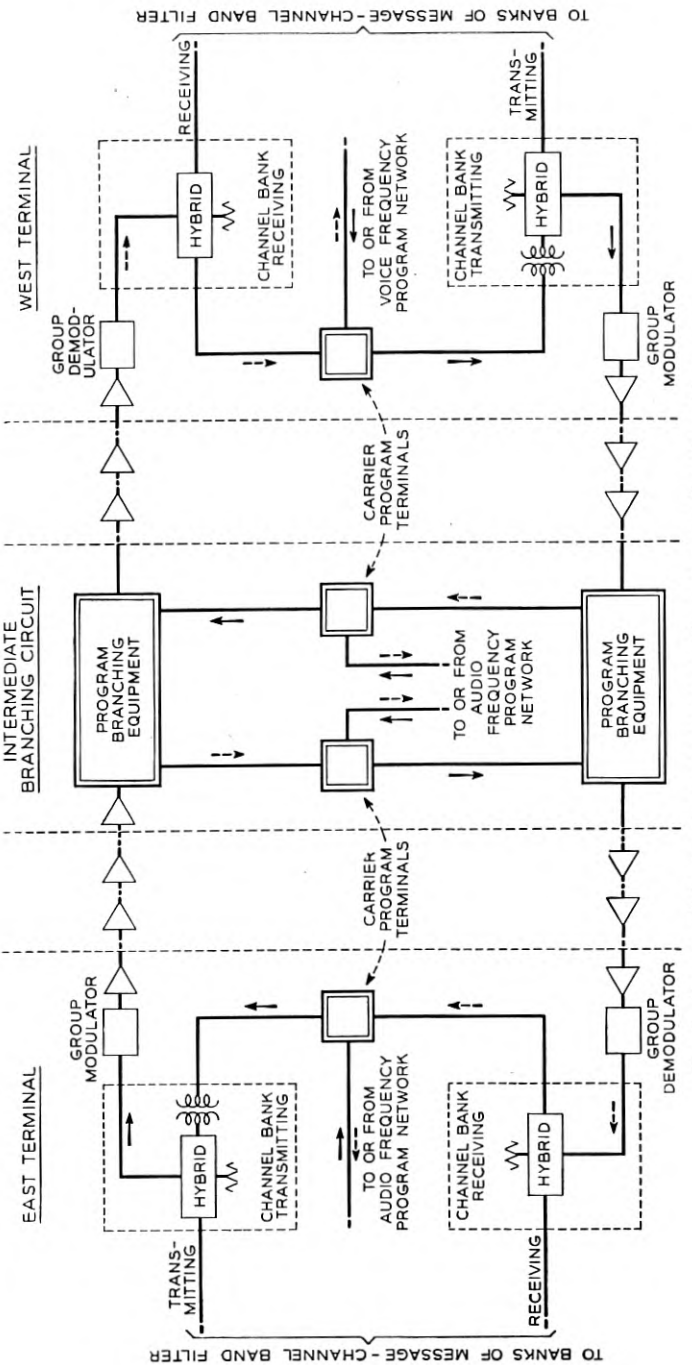


Fig. 3—Program link of cable carrier system with one intermediate branching point.

shown in heavy lines. The reversing and control circuits, indicated in light lines, are permanently connected to the external audio-frequency circuit and to the transmitting and receiving carrier line circuits regardless of the condition of the switching relays. Figure 3 shows a carrier program system including two terminals and a branching point as it is connected to a type K system. The program equipment is identified by double-line blocks. The carrier program terminals are connected into the networks in the same way as the audio-frequency facilities, through equalizers, amplifiers, bridges, and reversing circuits. Connected as one leg of a reversible bridge, a carrier program circuit may feed or be fed by any of the other legs, which may include cable, open-wire, studio loop, or other carrier circuits.

Terminal Circuit

As Fig. 2 indicates, a carrier program terminal consists of three elements: a modulator-demodulator or modem, a demodulator amplifier, and a reversing and control circuit. The heart of the terminal is the modem, which translates the program material from its original audio band to its desired position in the carrier-frequency spectrum or vice versa. It consists essentially of the non-linear varistor to which the carrier and program material are applied, and the band filter which selects the desired sideband from the modulation products. The varistor is connected in the double-balanced bridge arrangement in which the signal, carrier, and sideband circuits are each balanced against the other two. It is composed of copper-oxide elements and, in order to meet the conflicting requirements for high carrier-to-signal ratio and low transmitted carrier leak, a high degree of balance between the varistor bridge arms must be maintained. This is accomplished by building up each bridge arm of 16 copper-oxide elements connected in series-parallel. This modulator as compared to one using single-element bridge arms, has the same impedance, 12 decibels better carrier balance, 12 decibels greater carrier power capacity, and with the higher carrier power, 12 decibels lower non-linear distortion products. An amplifier provides the required power and a narrow-band filter gives additional suppression to carrier frequencies of other channels which are fed from the same carrier supply.

The band filter, which represents a major development in itself and is described in another paper,[17] introduces a considerable amount of delay distortion. This is corrected by delay equalizers incorporated in the modem circuit as shown in Fig. 2. Most of the delay correction is done in the audio-frequency branch of the circuit by a 31-section network which also includes equalization for the small residual attenuation distortion of the filter in its pass band. At the lower end of the audio-frequency band, however, attainment of the required phase characteristic with audio-frequency elements is

more difficult. Consequently, the delay correction for that portion of the band below 1000 cycles is actually done at sideband frequency, using quartz crystal elements. The design of these delay equalizers is described in another paper.[18] Transmission through the resulting modem unit is essentially constant in both attenuation and delay over the usable frequency range.

The demodulator amplifier is a conventional two-stage resistance-coupled amplifier. It is stabilized by 25 decibels of feedback to a nominal gain of 38 decibels, variable over a 12-decibel range by a potentiometer in the feedback circuit. The transmission characteristic is flat within 0.3 decibel over the 35 to 15,000-cycle frequency range. The output impedance is stabilized by the use of an output bridge for obtaining the feedback voltage. This amplifier feeds a -10 vu point in the circuit and can deliver up to $+18$ decibels above one milliwatt of output. Noise is kept to a minimum by operating the input stage vacuum tube at reduced voltages, mounting it and the magnetically shielded input transformer on a vibration-reducing suspension, and providing heavy filtering for the A and B battery circuits.

The limiting source of noise in any communication system is usually the transmission medium. In the carrier program system, the transmission medium is a carrier system which introduces noise energy equally distributed over the program band. The program energy being transmitted, however, is heavily concentrated at the lower frequencies. In order to increase the signal-to-noise ratio without an increase in total transmitted power, a predistorting network is introduced ahead of the modem, which attenuates the lower frequencies relative to the higher. The total discrimination is about 18 decibels, distributed symmetrically on a logarithmic frequency scale above and below 1500 cycles. A restoring network having an inverse characteristic is inserted in the receiving program path to return the program energy distribution to normal. The noise improvement thus obtained is about 7 decibels.

The reversing circuit consists of a set of five relays and a 78-kilocycle amplifier-oscillator. Two of these relays, as shown in Fig. 4, set up the transmission circuits for transmitting or receiving. The transmitting relay connects the predistorter in the audio-frequency circuit and connects the modem output to the transmitting high-frequency line. The receiving relay connects the modem to the receiving high-frequency line and inserts the restorer and demodulator amplifier in place of the predistorter. These relays are interlocked so that only one at a time can be operated. Their operation is supervised by two other relays, one transmitting and one receiving, which respond to the transmitting and receiving control signals respectively. The supervisory relays are similarly interlocked so that the control signal from only one direction at a time can be effective. They are

so connected to the transmission relays that, when no control signal is applied, both supervisory relays are released and the transmission relays maintain the circuit condition established at the last reversal.

A two-stage, tuned, feedback-stabilized amplifier is used to raise the level of the 78-kilocycle receiving control signal selected from the receiving high-frequency line by a narrow-band crystal filter. A copper-oxide rectifier converts the amplified signal to d-c. to operate a sensitive relay connected

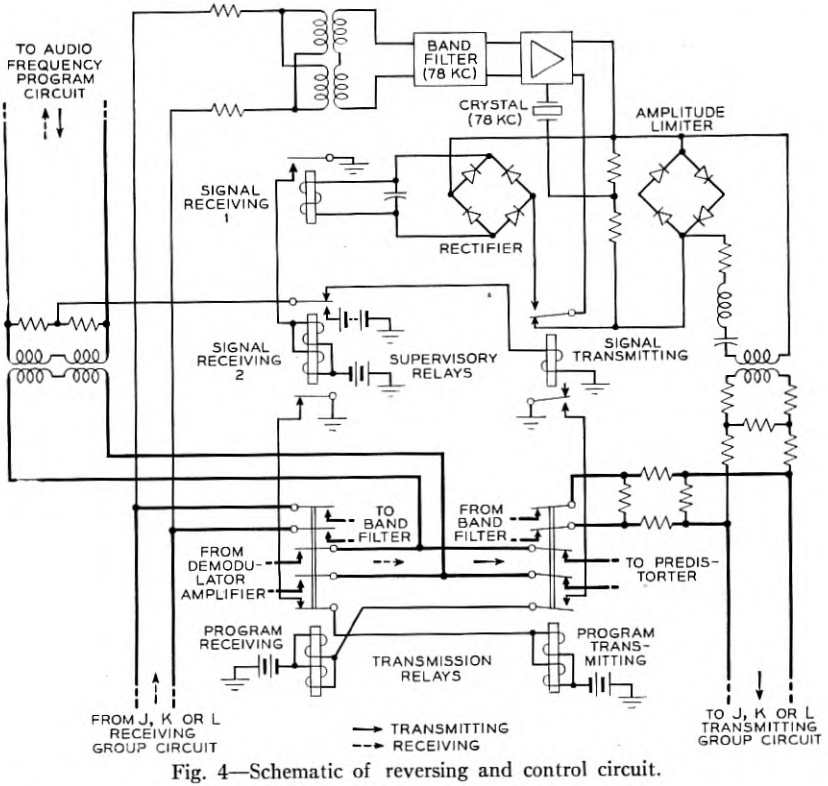


Fig. 4—Schematic of reversing and control circuit.

to the receiving supervisory relay. The supervisory relay, besides controlling the transmission circuits, also sends on a control signal as a d-c. simplex on the audio-frequency pair leaving the program terminal.

The same 78-kilocycle amplifier used for receiving the control signal is also used as an oscillator to generate the high-frequency control signal in the transmitting direction. The transmitting supervisory relay, under the control of a d-c. control signal coming in on the audio-frequency pair, disconnects the receiving control signal rectifier and connects instead a varis-

tor limiter across the output and a 78-kilocycle crystal from the output to the input, phased for positive feedback.

Type K Branching Circuit

Because of the operating requirements of a radio broadcasting network there is need for complete switching flexibility. On a national network there may be scores of intermediate points, where the program must be tapped for local broadcasting and may originate in the case of special events. If, to obtain this flexibility, the network were made up of short carrier links, bringing the program down to audio frequencies at the end of each link, there would be in some cases, between the originating studio and the most distant broadcasting station, 50 to 100 or more links in tandem, involving double that number of band filters. Terminal phase and attenuation distortion, however, are proportional to the number of links. By means of advanced filter and equalizer designs, the present system has been made suitable for about 10 to 13 links in tandem. Additional arrangements therefore are needed at intermediate points to serve local broadcasters, without breaking in on the through program transmission. The branching circuit serves this need. End branching circuits which split off a program circuit from a carrier message route are needed for some network branches and are less elaborate than the through branching circuits which provide full switching service at intermediate points on a main trunk route.

The flexibility of the through branching circuit is illustrated in the following functions which it performs under remote control:

1. Provides a receiving leg on a reversible through program circuit.
2. Splits the network to provide independently reversible links in each direction with the same or different program material on each link.

These functions are performed with negligible reaction on the associated through-message circuits. A block schematic for one direction of transmission on a type K system is shown in Fig. 5.

For splitting the network a band elimination filter[17] blocks frequencies in the program assignment (32–44 kilocycles) while passing the remaining message frequencies. As network rearrangements are made during the program switching interval transmission may be rerouted through the phase simulating network[17] which is substituted for the BEF when the program is to go through instead of being blocked. The simulation of phase, which must be close in order to avoid disturbance of voice-frequency telegraph superimposed on any of the message channels, extends over all but the two channels adjacent to the program. The transfer from one transmission path to the other is accomplished by a chain of make-before-break relays in such a manner that transmission on the message channels is virtually unaffected.

Junctions of transmission circuits are made with resistance hybrid con-

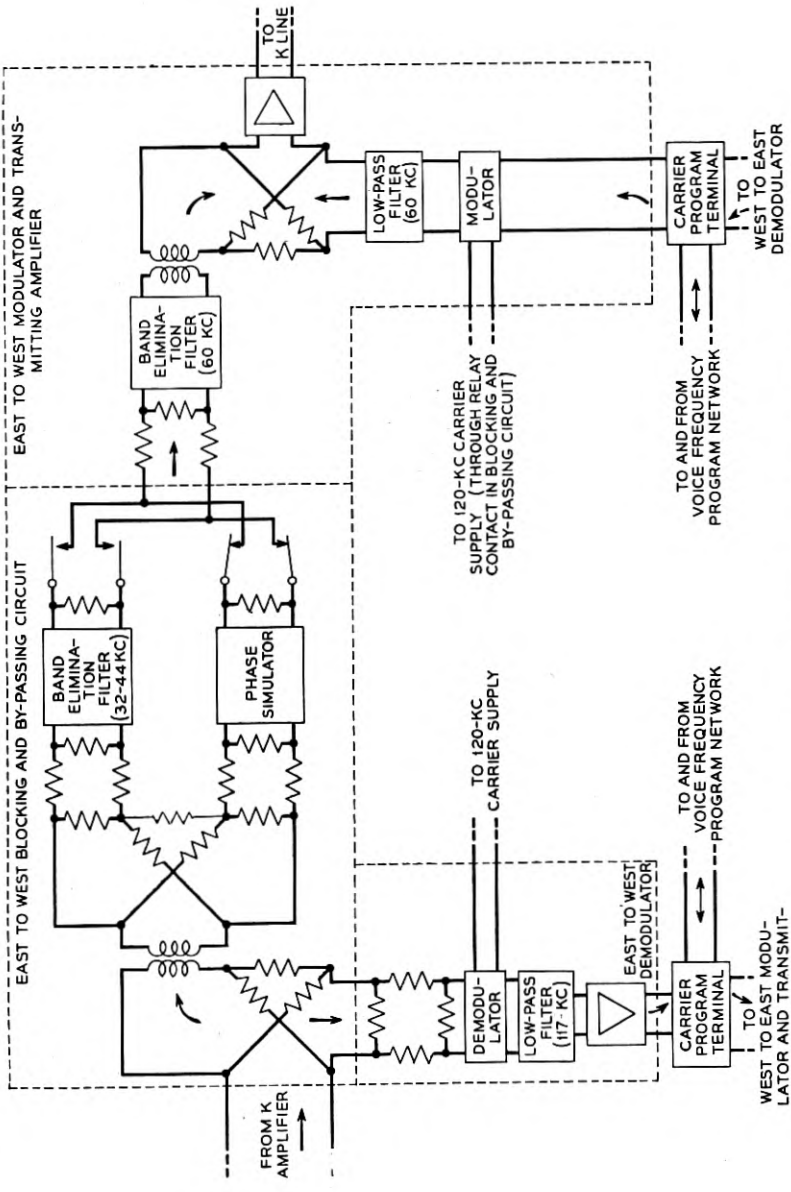


Fig. 5—Block schematic of type K branching circuit (one direction of transmission represented).

figurations to minimize transmission distortion and to give some directional discrimination.

Frequency translation from the 32–40 kilocycle range of the program on the line to the 80–88-kilocycle range of the program terminal is provided by modulating and demodulating circuits having 120-kilocycle carrier. It is of interest to note that the transmitting 120-kilocycle carrier is supplied through relay contacts which are normally open so that spurious noise and transmission will not interfere with the through program. The relay contacts are closed when the blocking filter is in circuit, thus permitting a local program origination only when the through circuit is cut off.

Relay control circuits have been arranged to coordinate with existing control practices and circuits so that reversibility may be under studio control and network splitting under local control.

Gain to offset circuit losses is supplied at the output by a transmitting amplifier so that the over-all loss of the through circuit is zero. Patching to spare circuits is thus facilitated.

In a K2 carrier system[15] the transmitting amplifier has unique properties in that it is self-oscillating at 60 kilocycles at an amplitude which complements the signal amplitude to produce a constant total output power. This feature is used as a carrier system line regulation control, and when a new program originates at the branching point it is necessary to generate another 60-kilocycle signal to complement the new total signal output, and to effectively block all 60-kilocycle received from the previous line section. A 60-kilocycle BEF is provided for that purpose.

These branching arrangements, developed for type K systems, have also been adapted for use with type L groups. Blocking and bridging functions are provided as they are for type K and in addition to the complete branching circuits, include simplified arrangements which make use of otherwise idle groups for carrier program circuits without message channels.

Branching Circuit Performance

The performance characteristics of the blocking and by-passing circuits are shown in Fig. 6, which represents transmission vs. frequency over the type K range of line frequencies. The solid line gives the normal transmission characteristic for through transmission of the program and the nine message channels. The dotted line is the program blocking characteristic indicating 80-decibel minimum suppression over most of the 32–44-kilocycle frequency range. The dashed line is the characteristic effective during the brief interval in the switching process when both branches of the circuit are connected. Its similarity to the other two characteristics is a measure of the effectiveness of the phase simulation over most of the message channel spectrum. Its departure from the other characteristics in the channel 5

and channel 9 allocations marks the end of the region in which the phase of the blocking filter can be successfully simulated. Outside of this region, in parts of channels 5 and 9, the switching operation shifts both phase and amplitude of the transmission and precludes the use of these channels for voice-frequency telegraph or telephoto services.

Equipment

As previously stated a carrier program terminal consists of a modem unit, a demodulator-amplifier panel, and a reversing panel. As shown in

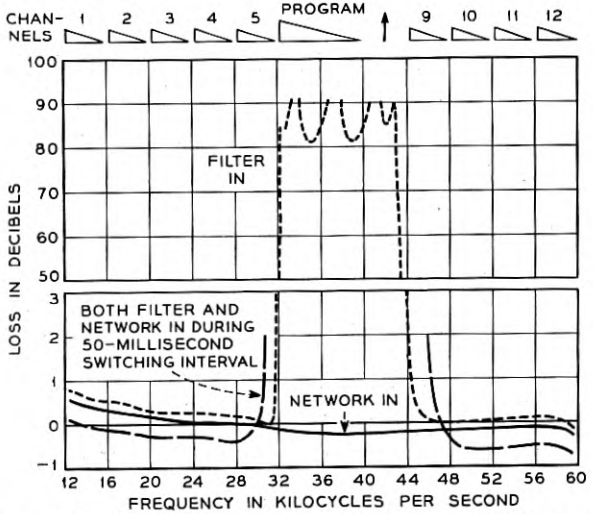


Fig. 6—Transmission characteristics of blocking and by-passing circuit under normal conditions and during switching.

Fig. 7 two such terminals, together with fuse panels for 24 and 130-volt battery supply for several bays, are mounted in one standard cable duct type bay 19″ wide and 11′6″ high. The equipment is mounted on the bay in a group, from top down, in the order mentioned. The front or wiring side of the equipment is provided with three separate covers which furnish the necessary electrical shielding as well as physical protection. Connections to carrier systems are made through carrier program high-frequency patching jacks on a 4-wire basis and to the program circuits through audio frequency testing jacks on a 2-wire basis from which point the carrier program channel is lined up for program service at the proper transmission levels for both directions of transmission.

A through branching circuit consists of two sets of line bridging equipment and two carrier program terminals mentioned above. As shown in

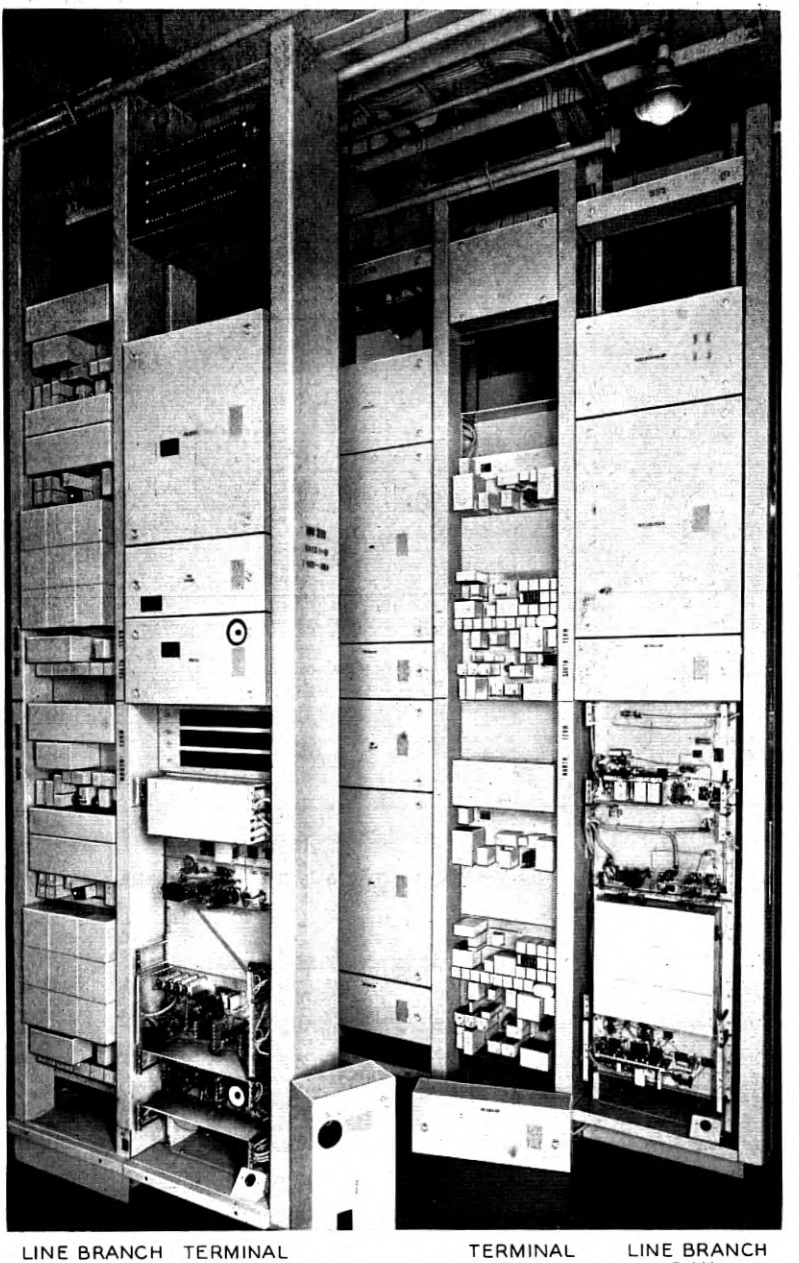


Fig. 7—Photograph of an early installation of carrier program terminals and associated type K branching circuits.

Fig. 7 the line bridging equipment is mounted in one standard cable duct type bay 19″ wide and 11′ 6″ high. The equipment for each direction of transmission is mounted on the bay in a group consisting of a modulator and transmitting amplifier, blocking filter and by-passing network with switching relays, and demodulator and demodulator-amplifier. The rear or wiring side of this equipment is also provided with three separate covers. The apparatus or front side of two of the panels, modulator and demodulator-amplifiers, is equipped with parallel vacuum tube sockets so that the vacuum tubes can be removed from the circuit for testing purposes without

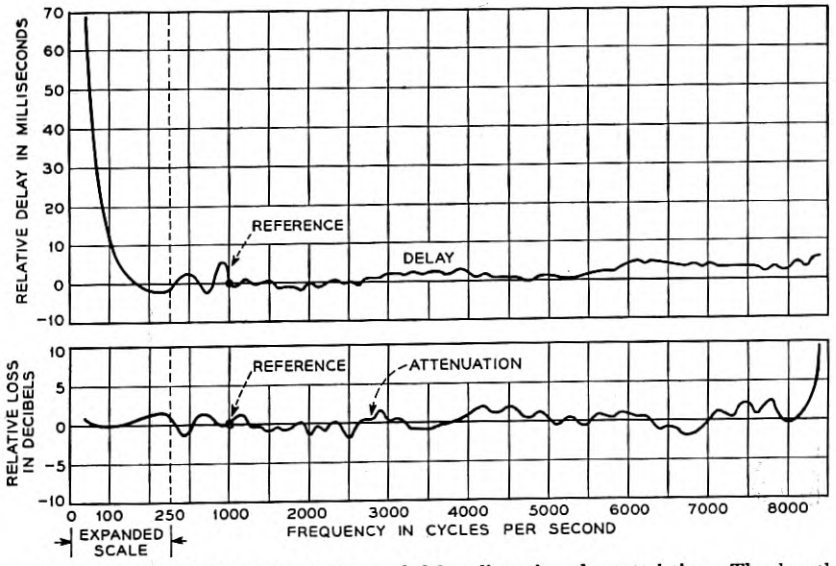


Fig. 8—Typical 10-link attenuation and delay distortion characteristics. The length of this circuit was 7300 miles.

interfering with service in the same manner as is done on K2 carrier telephone equipment.[15]

The necessary carrier supply for this equipment is obtained from regular standard carrier supply bays used for other broadband carrier equipment. The presence of undesired residual harmonics of 4-kilocycles from this carrier supply necessitates modification of the carrier circuits at several points to provide additional suppression to undesired components which would otherwise appear as 4 or 8-kilocycle tones in the program channel.

System Performance

The longest commercial network circuits now in operation are in the order of 7000 miles long, including the transcontinental backbone route

and feeder circuits along the Atlantic and Pacific coasts. On the assumption that these routes may some day be largely in carrier, exhaustive tests were made in 1947 of carrier program transmission applied to type K systems between Omaha and Los Angeles, looping back and forth as required to build up long circuits. Live program material transmitted around a 7300-mile loop consisting of 10 carrier links in tandem was judged to be of excellent quality by juries composed of experienced and critical observers. Attenuation and delay characteristics of this circuit relative to the 1000-cycle point are shown in Fig. 8 and indicate that design objectives are met with enough margin to justify practical operation over about 13 links in tandem. Background noise was about 53 decibels below the peak signal. Frequency shift due to differences in carrier frequency at the 20 terminals was less than 2 cycles. The time required for a complete reversal counting from the initial control signal release was 3 seconds. Shorter lengths will, of course, have even better performance. These characteristics, while they do not represent perfection in transmission quality, strike a balance between the various engineering limitations, which makes this system compare favorably with the best facilities previously available.

Conclusion

At the end of 1948, three years after the first installation, there were approximately 75,000 miles of carrier program circuits in service, about 70 per cent of them established full time. This is a substantial proportion of the total mileage of all grades of program service, which is in the order of 175,000 miles. The portions of the main transcontinental routes formerly carried by open-wire lines are now in carrier cables.

References

1 "Use of Public Address System with Telephone Lines," W. H. Martin and A. B. Clark, *A. I. E. E. Journal*, April 1923, pp. 359–366.
2. "High-Quality Transmission and Reproduction of Speech and Music," W. H. Martin and H. Fletcher. *A. I. E. E. Journal*, March 1924, pp. 230–238.
3. "Telephone Circuits used as an adjunct to Radio Broadcasing," H. S. Foland and A. F. Rose. *Electrical Communication*, January 1925, pp. 194–202.
4. "Telephone Circuits for Program Transmission," F. A. Cowan. *A. I. E. E. Transactions*, 1929, pp. 1045–1049.
5. "Wire Line Systems for National Broadcasting," A. B. Clark. *Bell Sys. Tech. Jour.*, January 1930, pp. 141–149.
6. "Long Distance Cable Circuit for Program Transmission," A. B. Clark and C. W. Green. *Bell Sys. Tech. Jour.*, July 1930, pp. 567–594.
7. "Auditory Perspective" (A symposium), "Transmission Lines," H. A. Affel, R. W. Chesnut and R. H. Mills, *Electrical Engineering*, January 1934, pp. 9–32, 216–218.
8. "Wide Band Open-Wire Program System," H. S. Hamilton, *Electrical Engineering*, April 1934, pp. 550–562.
9. "A Carrier Telephone System for Toll Cables," C.W. Green and E. I. Green. *Bell Sys. Tech. Jour.*, January 1938, pp. 80–105.
10. "Cable Carrier Telephone Terminals," R. W. Chestnut, L. M. Ilgenfritz and A. Kenner. *Bell Sys. Tech. Jour.*, January 1938, pp. 106–124.
11. "A Twelve-Channel Carrier Telephone System for Open-Wire Lines," B. W. Kendall and H. A. Affel. *Bell Sys. Tech. Jour.*, January 1939, pp. 119–142.

12. "Engineering Requirements for Program Transmission Circuits," F. A. Cowan, R. G. McCurdy and I. E. Lattimer, *Electrical Engineering*, April 1941, pp. 142–147.
13. "Wide-Band Program Transmission Circuits," E. W. Baker. *Electrical Engineering*, March 1945, pp. 99–103.
14. "Transmission Networks for Frequency Modulation and Television," H. S. Osborne. *Electrical Engineering*, November 1945, pp. 392–397.
15. "An Improved Cable Carrier System," H. S. Black, F. A. Brooks, A. J. Wier and I. G. Wilson, *Electrical Engineering*, *A. I. E. E. Transactions*, 1947, Vol. 66, pp. 741–746.
16. "Frequency Division Techniques for a Coaxial Cable Network," R. E. Crane, J. T. Dixon and G. H. Huber, *A. I. E. E. Transactions*, 1947, Vol. 66, pp. 1451–1459,
17. "Band-pass Filter, Band Elimination Filter, and Phase Simulator Network for Carrier Program Systems." A companion paper by F. S. Farkas, F. J. Hallenbeck and F. E. Stehlik. This issue of *BSTJ*.
18. "Delay Equalization of 8-Kc Carrier Program Circuits," A companion paper by C. H. Dagnall and P. W. Rounds. This issue of *BSTJ*.

Delay Equalization of Eight-Kilocycle Carrier Program Circuits

By C. H. DAGNALL and P. W. ROUNDS

This paper describes the equalization of delay in 8-kc program systems transmitted over broad-band carrier telephone facilities. Use is made of a condenser-plate potential analog which provides a ready method for blocking out the basic design and arriving at the final equalizer constants. Most of the equalization is accomplished at audio frequencies, and the remainder at carrier frequencies with quartz-crystal equalizers.

IN TRANSMITTING programs for radio broadcasting over the United States, an extensive network of wire circuits has been established by the Bell System. Most of the additions to this network since the war have employed a single-sideband carrier system[5] applicable to broad-band carrier facilities. The selection of a single sideband requires sharp frequency discrimination, and when this discrimination is achieved with minimum-phase structures, it is of necessity accompanied by delay distortion.[3]

In one or two carrier links, each including a transmitting and receiving terminal, the delay distortion is sufficiently small so that no deterioration in the program is noticeable. However, flexibility of maintenance and operation of an extensive program network requires that the network be built up of a large number of links in tandem. When this is done, the effects of delay distortion become quite conspicuous and equalization of the delay is necessary. Furthermore, if the equalization is to be satisfactory between any two points in the network, each link must be independently equalized.

Most of the delay distortion arises in the carrier-frequency band-pass filter which selects the lower sideband, the small remaining portion being contributed by the amplifiers and repeating coils. Figure 1 illustrates the unequalized delay in one terminal. Equalizers have been added to each terminal to make the phase characteristic approach linearity and so permit at least ten links to be operated in tandem without excessive distortion.

Theory of Design

The equalization of delay distortion is accomplished through the use of all-pass networks which, in their most general form,[2] may be constructed as a tandem set of lattice sections of the type shown in Fig. 2. An electrostatic analogy, developed during the late thirties,[4] has been found to be of great assistance in visualizing the performance of these networks and in indicating a rational method of design.

Considering the single section shown in Fig. 2 as the basic building block, the loss and phase may be expressed in the form

$$e^{A+jB} = \frac{(p + k_n - j\omega_n)(p + k_n + j\omega_n)}{(p - k_n - j\omega_n)(p - k_n + j\omega_n)} \tag{1}$$

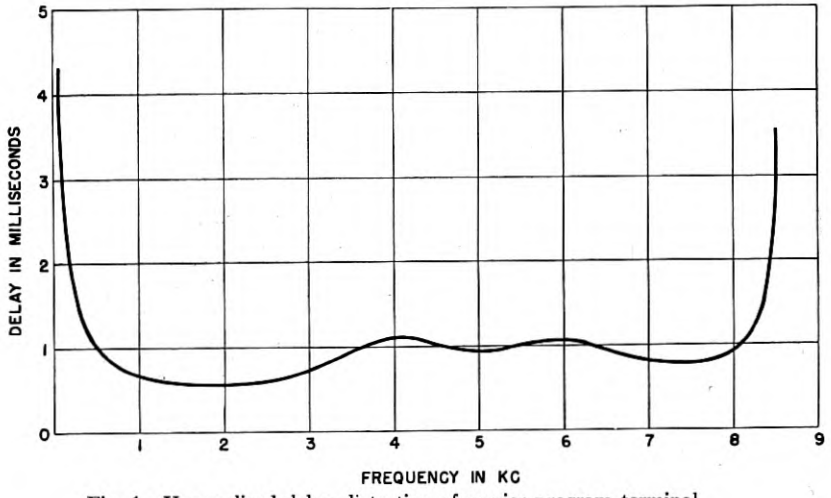


Fig. 1—Unequalized delay distortion of carrier program terminal.

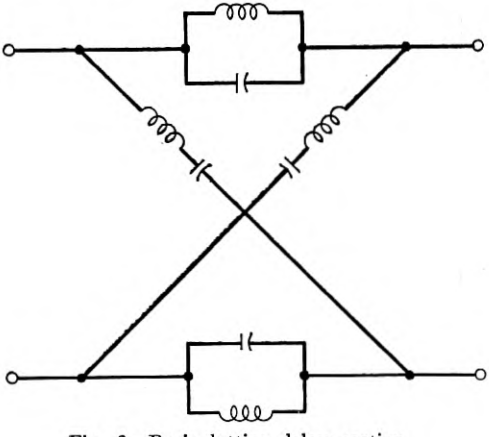


Fig. 2—Basic lattice delay section.

where

A = insertion loss in nepers
B = insertion phase in radians
$p = j\omega$
ω = frequency in radians per second
k_n, ω_n = real positive constants

An examination of equation (1) indicates that there are zeros at the two points:

$$p = -k_n + j\omega_n \text{, and } p = -k_n - j\omega_n$$

and poles at the two points

$$p = +k_n + j\omega_n \text{, and } p = +k_n - j\omega_n$$

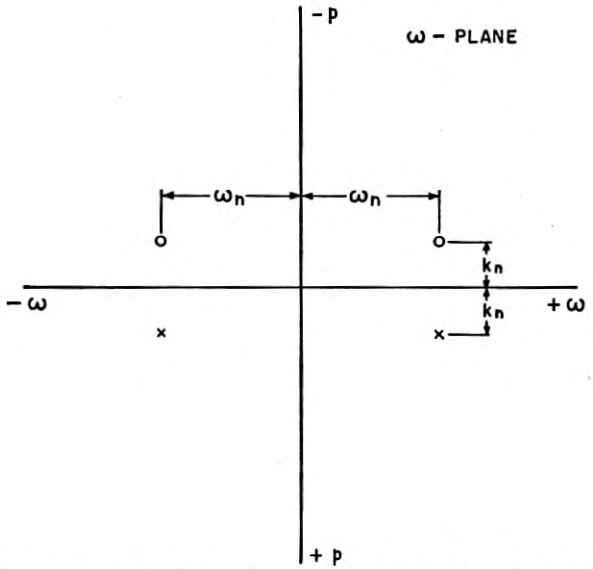


Fig. 3—Plot of zeros and poles of the network of Figure 2 on the complex-frequency plane.

The first zero in equation (1) contributes a delay (defined as the derivative of the phase with respect to frequency) of the form

$$T_1 = \frac{dB_1}{d\omega} = \frac{1}{k_n\left[1 + \frac{(\omega - \omega_n)^2}{(k_n)^2}\right]} \tag{2}$$

Similar expressions may be obtained for the other zeros and poles, the total delay of the section then being equal to the sum of the delays contributed by each zero and pole.

The four zeros and poles of equation (1) may be plotted on the complex-frequency plane as shown in Fig. 3, where the circles indicate zeros and the crosses poles. The four points are seen to be symmetrically disposed with respect to the origin. With reference to this figure, it will be noted that, since $\omega = -jp$, positive real values of p correspond to negative imaginary

values of ω and negative real values of p correspond to positive imaginary values of ω. The axes of Fig. 3 have been labelled accordingly.

It is at this point that the electrostatic analogy begins to come into play. Assume that an infinite wire filament, positively charged throughout its length, is run through the zero $p = -k_n + j\omega_n$ perpendicular to the plane of the paper and that a unit positive charge is placed at an arbitrary point, ω, along the real frequency axis. The component of the force normal to the ω axis exerted on the unit charge may be written in the form

$$F = \frac{1}{k_n \left[1 + \frac{(\omega - \omega_n)^2}{(k_n)^2} \right]} \tag{3}$$

When distances in equation (3) are identified with frequencies in equation (2), the two expressions are identical. A similar argument applies to the other zero, and also to the two poles provided that the filaments passing through the poles have charges of the opposite polarity. Thus we may say that the network of Fig. 2 will have a delay proportional to that component of the electric field intensity which is normal to the ω axis, when a positive filament passes through each zero and a negative filament through each pole. Fig. 4 indicates the character of the delay as a function of frequency. Parenthetically we may note that the component of the field intensity parallel to the ω axis is proportional to the derivative of the loss. Since this component is zero, the loss will be constant at all frequencies. In the case of the reactance networks with which we are dealing here, the loss is zero.

Although the usefulness of the electrostatic analogy lies principally in its application to more complex networks, several conclusions may be drawn from Fig. 4. The right-hand zero and pole, because of their symmetrical spacing and opposite charges, make equal contributions to the total delay. The same statement holds true for the left-hand zero and pole combination. As the zeros and poles approach the real-frequency axis, the delay peaks become sharper and higher because of the increased local field intensity. The figure also shows that the slope of the delay curve is zero at zero frequency and that, unless ω_n is large compared to k_n, the delay at zero frequency is of appreciable magnitude. These isolated facts will be exploited later in considering more complex networks.

Assume, now, a tandem series of sections of the type shown in Fig. 2, in which the zeros and poles are so selected that they are evenly spaced at intervals, a, along straight lines parallel to the real-frequency axis as shown in Fig. 5. It was pointed out by H. W. Bode[1] that the resulting field intensity may be approximated by distributing the total of the discrete charges on the plates of an equivalent condenser passing through the zeros and poles and extending a distance of $a/2$ beyond the extreme zeros and

poles. This approximation ignores the ripples caused by the granularity of the filament spacings, but it does permit the average delay to be determined in a particularly simple manner. The field intensity at any point, ω,

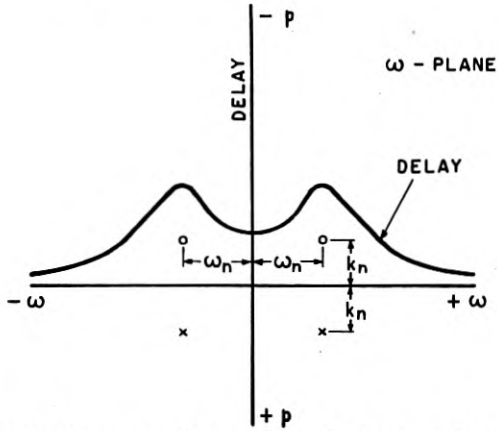


Fig. 4—Delay-frequency characteristic of the network of Fig. 2.

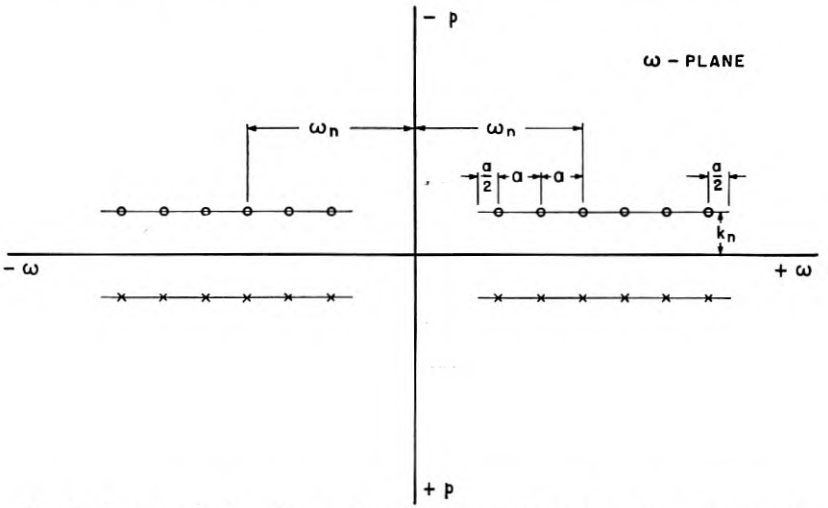


Fig. 5—Zeros and poles of a complex delay network based on the condenser-plate design.

resulting from the condenser charge is proportional to the angle subtended by the plates at that frequency. It is also proportional to the charge per unit length of plate or, in other words, to the density, $1/a$, of the filament spacings. The geometry is illustrated in Fig. 6, where $2(c + d)$ is the

angle subtended by the plates at the frequency ω. From this figure it may be seen that the field intensity in the region between the plates will have a fairly uniform value which falls off sharply as the edges are reached and becomes vanishingly small at frequencies remote from the plates.

Along with this simple determination of the average delay characteristic, D. F. Tuttle in an unpublished memorandum has derived expressions for the magnitude of the delay ripple. As shown in the appendix, the field

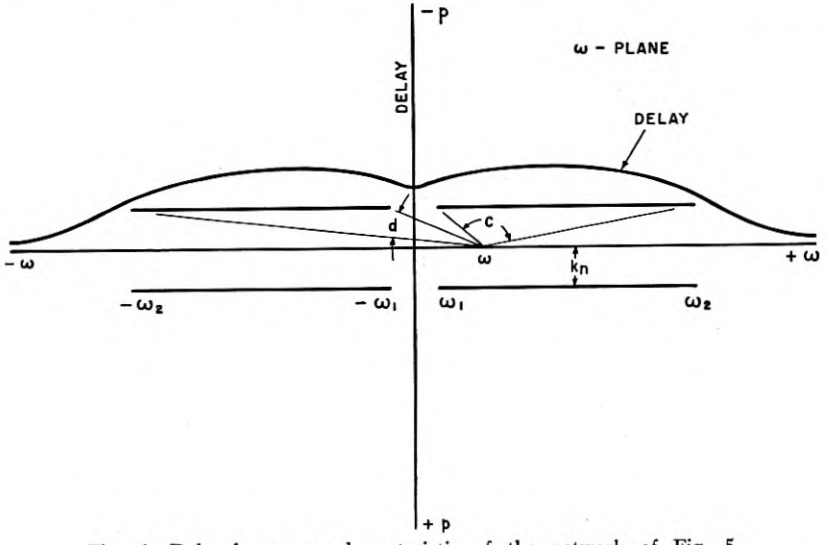


Fig. 6—Delay-frequency characteristic of the network of Fig. 5.

intensity or delay for an infinitely long set of charged filaments may be expressed in the form

$$T = \frac{2\pi}{a} \tanh \frac{2\pi k_n}{a} \left[1 - \left(\frac{\cos \frac{2\pi\omega}{a}}{\cosh \frac{2\pi k_n}{a}} \right) + \left(\frac{\cos \frac{2\pi\omega}{a}}{\cosh \frac{2\pi k_n}{a}} \right)^2 - \cdots \right] \tag{4}$$

For reasonably large values of $2\pi k_n/a$, this relation may be replaced by the approximate expression

$$T \approx T_0[1 - \delta \cos T_0\omega] \tag{5}$$

where $T_0 = 2\pi/a$ is the average delay and $\delta = 2e^{-T_0 k_n}$ is the percentage ripple about the average value. The ratio k_n/a may thus be determined from the percentage delay ripple in accordance with the formula

$$\frac{k_n}{a} = \frac{1}{2\pi} \log_e \frac{2}{\delta} \tag{6}$$

To equalize the low-frequency and high-frequency filter delay shown in Fig. 1, a condenser plate of the form shown in Fig. 6 might be visualized. Although the high-frequency delay approximates that desired, the low-frequency delay shows insufficient shaping to be complementary to the filter characteristic because of the contribution of the negative-frequency plates. By bringing the plates closer to the frequency axis, that is by decreasing the ratio k_n/ω_1, a sharper-breaking low-frequency characteristic could be obtained. However, to achieve a sufficiently small delay ripple, the spacing, a, as determined from equation (6) would then have to be decreased with the result that the number of sections would be correspondingly increased.

In attempting to reduce the total number of sections required, it was observed that a carrier-frequency delay equalizer would not be subject to

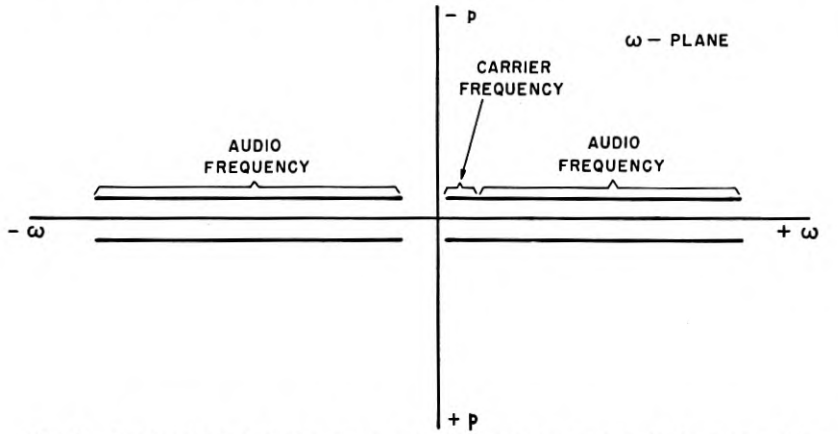


Fig. 7—Condenser-plate design for a combined carrier and audio-frequency delay equalizer.

the same low-frequency limitation, since the negative-frequency plates would be removed from the single-sideband signal by approximately twice the carrier frequency of 88 kc. However, since high-frequency delay sections are more expensive to construct than those operating at audio frequencies, a compromise is made in which the first few sections are built to operate at carrier frequencies and the remaining sections at audio frequencies. The equivalent condenser plates, referred to the audio-frequency signal, are shown in Fig. 7.

A condenser-plate design has thus been achieved which allows the low-frequency and high-frequency delay to be equalized at least approximately. Further modifications must be made in the design, particularly in the middle of the band, to shape the characteristic so that a more accurate complement of the filter delay may be obtained. The delay in a condenser-

plate design is directly proportional to the charge density along the plate. Up to now, this density has been assumed to be uniform. When the charge is located on discrete filaments, the restriction of uniform density is no longer necessary and it is possible to modify the delay characteristic as desired by changing the spacing of the filaments in inverse proportion to the desired change in delay.

The assumption of a flat plate is also useful in simplifying the analysis; in the actual design the equivalent plate is bowed out over the major portion of the frequency range to reduce the delay ripple. The final zeros and poles obtained are shown in Fig. 8, in which the carrier-frequency zeros and poles are plotted on an equivalent audio-frequency basis. A total of 29

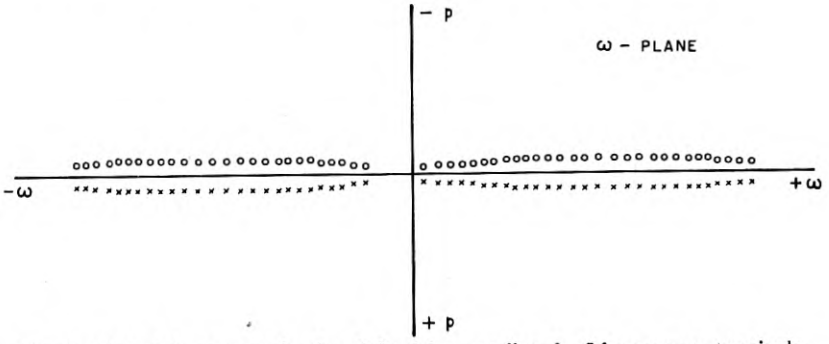


Fig. 8—Plot of the zeros and poles of the delay equalizer for 8-kc program terminals.

delay sections are required, of which three are assigned to the carrier-frequency equalizer and 26 to the audio-frequency equalizer.

Audio-Frequency Equalizer

To complete the design of the audio-frequency equalizer some means must be found for absorbing the effects of dissipation in the coils and condensers so that the final dissipative network will exhibit the theoretical non-dissipative performance plus a loss which is constant with frequency. It can be shown that a non-dissipative all-pass section plus a flat-loss pad can be replaced with a dissipative all-pass section (of modified constants) in tandem with a minimum-phase loss equalizer as in Fig. 9. It would be uneconomical to associate a loss equalizer with every phase section; and it is in fact unnecessary, since any minimum-phase device accomplishing the same result will exhibit the same performance.[3] The problem is then reduced to equalizing the loss of the network composed of dissipative delay sections.

The dissipative loss of these sections may be determined from the approximate relation

$$\text{Dissipative Loss in nepers} \approx \frac{1}{2}\left(\frac{R}{L} + \frac{G}{C}\right) T \tag{7}$$

where

$\frac{R}{L}$ = resistance-inductance ratio of coils in ohms per henry

$\frac{G}{C}$ = conductance-capacitance ratio of condensers in micromhos per microfarad

T = delay of network in seconds

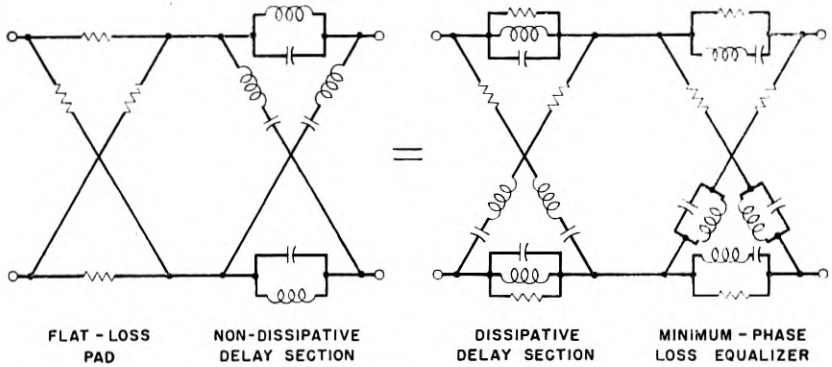


Fig. 9—Four-terminal equivalence showing the method of absorbing the effects of dissipation in the audio-frequency equalizer sections.

This expression indicates that, when the quantity $(R/L + G/C)$ is nearly constant with frequency, the shape of the loss characteristic will be generally similar to that of the delay characteristic. The ripples in the delay characteristic have been made sufficiently small so that the corresponding loss ripples may be ignored and only the general trend considered. A schematic of the resulting equalizer is shown in Fig. 10. The attenuation equalizer sections, in tandem with the delay sections, produce a loss characteristic complementary to that of the band filter over the 8000-cycle program range. Resistors have been added to the crossarms of each lattice delay section to allow the dissipative losses to be adjusted to the nominal values assumed in the design. For manufacturing convenience, the sections are assembled in seven separate containers which are mounted on an $8\frac{3}{4}$ inch by 19 inch relay-rack panel as illustrated in Fig. 11.

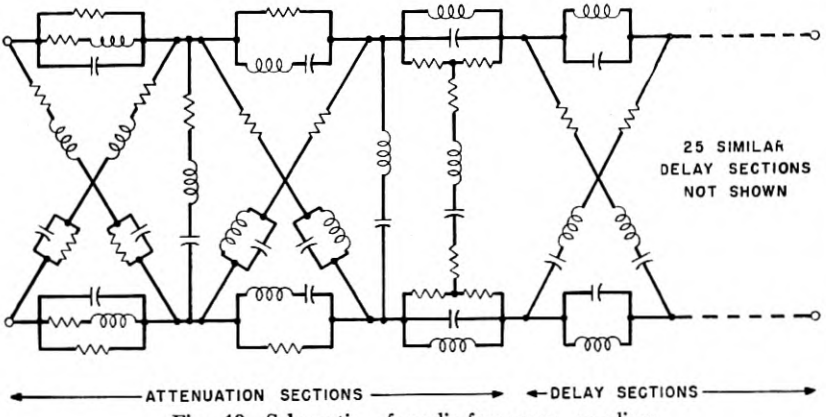


Fig. 10—Schematic of audio-frequency equalizer.

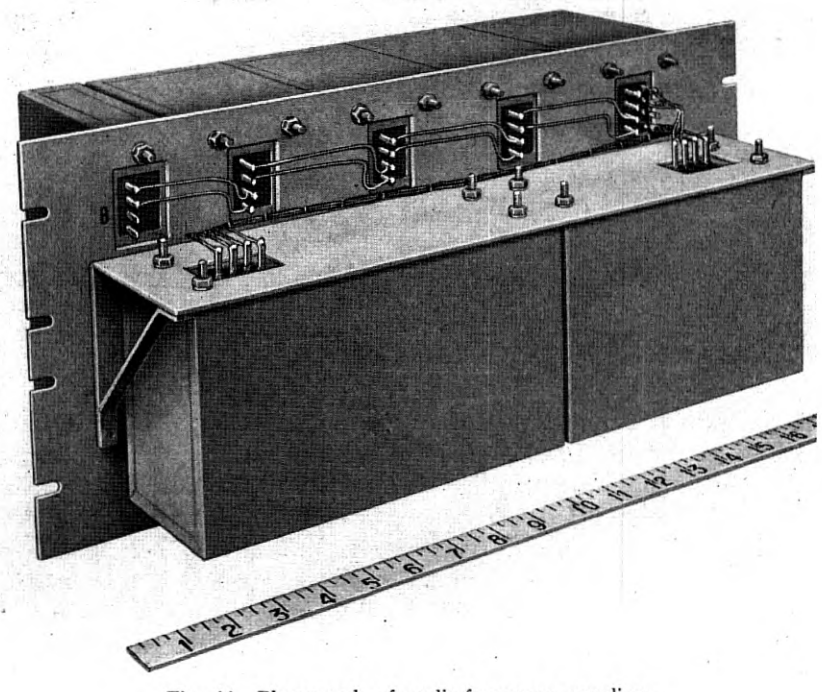


Fig. 11—Photograph of audio-frequency equalizer.

Carrier-Frequency Equalizer

The critical frequencies of the carrier-frequency equalizer are located at 318, 610 and 890 cycles on an audio basis. Since the carrier is at 88 kc

and the lower sideband is transmitted, the corresponding carrier frequencies are 87682, 87390 and 87110 cycles, respectively.

The required change of phase per cycle is the same as at audio frequencies, but the percentage rate of change is eleven times that of the audio-frequency sections operating at 8000 cycles. This requires that the arms of the sections have proportionately stiffer reactances, higher Q's, and greater tem-

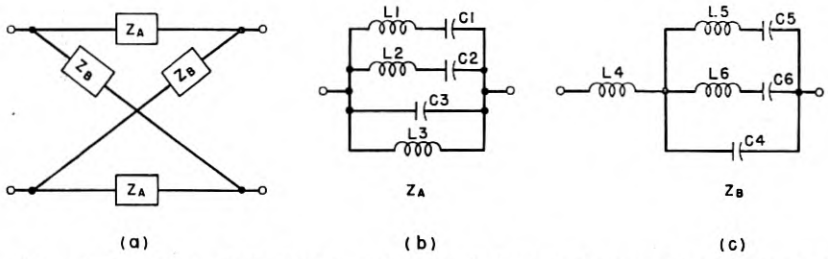


Fig. 12—Schematic of the lattice equivalent of three tandem sections of the type shown in Fig. 2.

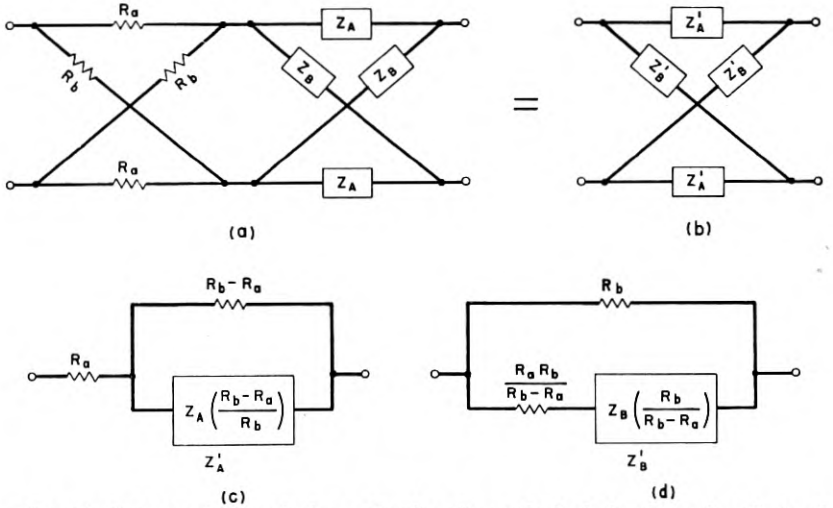


Fig. 13—Four-terminal equivalence showing the method of absorbing the effects of dissipation in the carrier-frequency equalizer.

perature stability. The only available elements meeting such requirements are piezo-electric crystals.

The approximate equivalent electrical circuit of a crystal is a capacitance in parallel with a series combination of an inductance and capacitance, and is not adaptable to the section of Fig. 2. However, when three such sections in tandem are combined into a single lattice, the configuration of

Fig. 12 is obtained. The stiffness of the reactances of arms Z_A and Z_B depends principally on the branches numbered 1, 2, 5 and 6. Each of these branches may be combined with a portion of C_3 or C_4 and replaced with a crystal.

One other restriction must be overcome before crystals can be used. Inductances L_1 and L_2 are in the order of 0.7 henry while L_5 and L_6 are in the order of 5000 henries, both inductance values being impractical for crystals. Two three-winding repeating coils are used to transform these inductances to values that may be provided by crystals. The two balanced windings of one repeating coil replace arms Z_B of Fig. 12(a), the third winding being connected to a reactance arm of the form of Fig. 12(c).

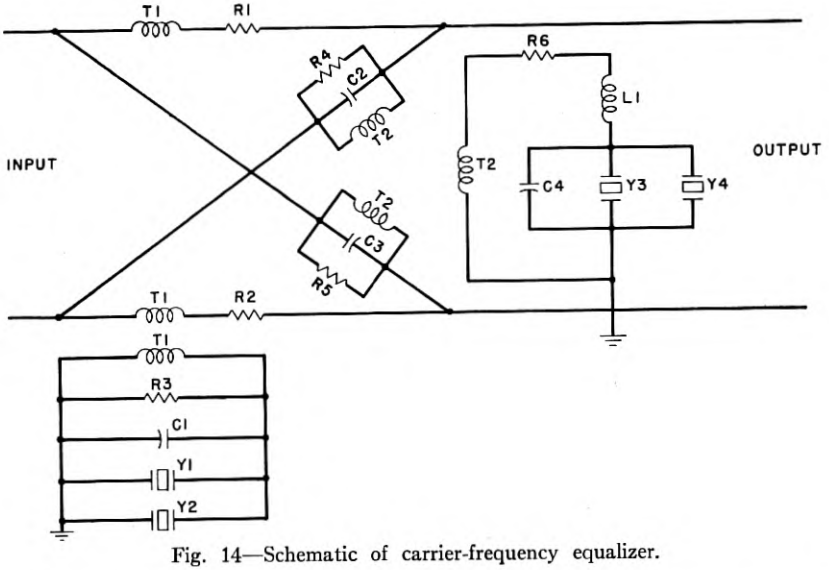


Fig. 14—Schematic of carrier-frequency equalizer.

The other repeating coil is similarly used for arms Z_A, the inductance L_3 being provided by the repeating coil. The repeating coils unavoidably introduce parasitic inductances, a small one in series with Z_A and a larger one in parallel with Z_B, the effects of which are made negligible by the addition of a capacitance in parallel with Z_B.

Dissipation in the elements of the carrier-frequency equalizer is taken into account by making use of the equivalence of Fig. 13, in which (a) represents the non-dissipative equalizer in tandem with a pad and (b) a structure similar to the non-dissipative structure but with resistances added to its arms, as shown by (c) and (d). The dissipation in each coil, condenser or crystal can be associated to a close degree of approximation with

one of these resistances. Physical resistances are added to compensate for deficiencies in dissipation. The loss of the pad is made large enough to allow for manufacturing deviations.

The complete schematic is shown in Fig. 14, and the equalizer with the shield removed in Fig. 15. Below the panel in Fig. 15, from left to right are arranged the retardation coil L1, adjusting condensers C1 and C4, the

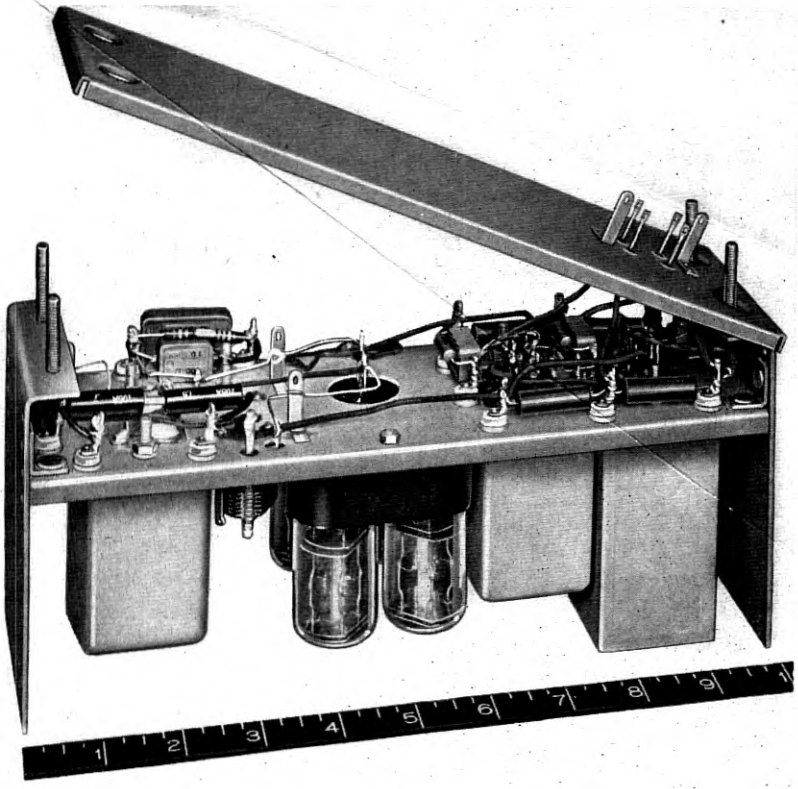


Fig. 15—Photograph of carrier-frequency equalizer.

crystals Y1 to Y4, inclusive, and the repeating coils T1 and T2. The fixed condensers and resistances are mounted above the panel.

Results

Curves *A* and *B* in Fig. 16 show the delay-frequency characteristics of the audio-frequency equalizer and the carrier-frequency equalizer, respectively. Curve *C* shows the equalized delay of one terminal, which is the sum of the delays of the equalizers added to the unequalized delay of Fig. 1.

Listening tests over ten carrier links in tandem indicate that the design objectives are sound and that a satisfactory reduction in delay distortion has been achieved.

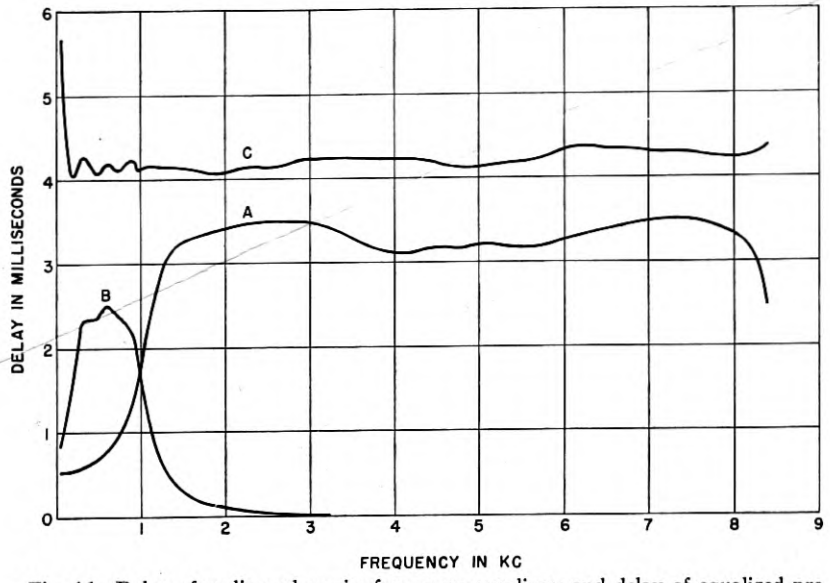


Fig. 16—Delay of audio and carrier-frequency equalizers and delay of equalized program terminal.

APPENDIX

For an infinitely long set of charged filaments of the type shown in Fig. 3 and located at $\omega = a/2,\ 3a/2,\ 5a/2$, etc., the insertion loss and phase may be expressed by the infinite-product expansion of equation (1),

$$e^{A+jB} = \prod_{n=1}^{\infty} \frac{[p + k_n - j(n - \tfrac{1}{2})a][p + k_n + j(n - \tfrac{1}{2})a]}{[p - k_n - j(n - \tfrac{1}{2})a][p - k_n + j(n - \tfrac{1}{2})a]}$$

$$= \prod_{n=1}^{\infty} \frac{1 - \left[\dfrac{i2(p + k_n)\pi/a}{(2n - 1)\pi}\right]^2}{1 - \left[\dfrac{i2(p - k_n)\pi/a}{(2n - 1)\pi}\right]^2} \tag{8}$$

Expression (8) is a standard form of product expansion and may be written

$$e^{A+jB} = \frac{\cos j(p + k_n)\pi/a}{\cos j(p - k_n)\pi/a} \tag{9}$$

or

$$A + jB = \log \cos j(p + k_n)\pi/a - \log \cos j(p - k_n)\pi/a \tag{10}$$

Substituting $j\omega$ for p and differentiating with respect to ω, we obtain

$$\frac{dA}{d\omega} + j\frac{dB}{d\omega} = j\frac{2\pi}{a}\frac{\sinh 2\pi k_n/a}{\cosh 2\pi k_n/a + \cos 2\pi\omega/a} \tag{11}$$

from which $dA/d\omega$ is zero. Equation (11) may be written

$$\frac{dB}{d\omega} = \frac{2\pi}{a}(\tanh 2\pi k_n/a)\left(\frac{1}{1 + \dfrac{\cos 2\pi\omega/a}{\cosh 2\pi k_n/a}}\right) \tag{12}$$

which, when expanded, gives equation (4).

References

1. "Wave Transmission Network," H. W. Bode, *United States patent* 2,342, 638.
2. "Network Analysis and Feedback Amplifier Design" (book) Chapter 11, H. W. Bode, D. Van Nostrand Co., Inc., New York, N. Y., 1945.
3. Reference 2, Chapter 14.
4. "Network Theory Comes of Age," R. L. Dietzold, *Electrical Engineering*, Volume 67, Number 9, September 1948, page 898.
5. "A Carrier System for 8000-cycle Program Transmission," R. A. Leconte, D. B. Penick, C. W. Schramm, A. J. Wier. A companion paper. This issue of *BSTJ*.
6. "Band Pass Filter, Band Elimination Filter and Phase Simulating Network for Carrier Program Systems," F. S. Farkas, F. J. Hallenbeck, F. E. Stehlik. A companion paper. This issue of *BSTJ*.

Band Pass Filter, Band Elimination Filter and Phase Simulating Network for Carrier Program Systems

By F. S. FARKAS, F. J. HALLENBECK, F. E. STEHLIK*

A paper by Leconte, Penick, Schramm and Wier[1] discusses the system aspects of 8-kc program circuits over carrier facilities and outlines the functions of several filters and networks. This paper describes in detail two of the filters and one network. These are:

1. The channel selecting crystal band pass filter used at program terminals of all broad-band carrier systems,
2. The band elimination filter which blocks the program at branching points on type K carrier systems,
3. The network used at branching points on type K carrier systems to simulate the phase shift of the band elimination filter.

Channel Selecting Crystal Band Pass Filter

AN IMPORTANT component of the modulator-demodulator circuit at the carrier program terminal is the band pass filter which selects the lower side band resulting from modulation of the audio frequency program material with the 88-kc carrier. This step of modulation locates the program frequencies in their allotted position in the carrier frequency spectrum of the standard broad-band terminal.

System flexibility requires that long program circuits be established by tandem connections of carrier links. A link consists of a transmitting and a receiving carrier program terminal connected by the appropriate transmission medium. The original objectives were based on a ten-link carrier circuit. This means that each terminal must introduce no more than five per cent of the total allowable system distortion. Assuming the band filter introduces the major part of the terminal distortion it is seen that the requirements placed on each band filter are extremely severe.

One of the transmission objectives of the system is to transmit audio frequencies as low as 50 cps. Hence the band filter must transmit the wanted carrier frequency sideband to within 50 cps of the carrier and must suppress the unwanted sideband beginning at 50 cps above the carrier. This sharp cut-off and the need for low distortion in the pass band requires the use of filter elements with so little dissipation that the only possibility of realizing the desired performance is by the use of quartz crystal elements.

In addition to suppressing the unwanted sideband above the carrier the filter must also provide sufficient discrimination above and below the pass

* Phase simulating network by F. S. Farkas.
Band elimination filter by F. J. Hallenbeck.
Band pass filter by F. E. Stehlik.

band to prevent crosstalk of the adjacent message channels into the program channel.

The necessity of using quartz crystal elements limits the maximum band width of filter which can be realized. This limitation is the result of the comparatively poor electromechanical coupling of quartz.[2] The resulting filter band width is 8.5 kc with the upper cut-off located near the 88-kc carrier. This is slightly greater than the 8-kc nominal band width of the system.

The crystal band pass filter designed for the single sideband program channel weighs approximately 30 lbs. and occupies 7 inches of mounting space on a standard 19 inch relay rack. A total of 44 filter components are required for its construction, half of which are balanced quartz crystal plates. The remaining components consist of eight adjustable air capacitors, three fixed mica capacitors, seven balanced retardation coils, three of which are adjustable, and four resistors. A schematic which shows the relative placement of these parts in the filter is given in Fig. 1.

The measured insertion loss characteristic of the filter between resistive terminations is shown in Fig. 2. The pass band and the vicinity of the upper cut-off are given in greater detail in the enlarged characteristics of Figs. 3 and 4. The extreme sharpness of the upper cut-off is evident in the latter figure. At 40 cps above the 88-kc carrier the discrimination has reached 20 db while the slope of the insertion loss versus frequency curve through this point is about 1 db per cps. Since at least two filters are connected in tandem in any program circuit a minimum of 40 db discrimination is provided to all frequencies in the unwanted sideband. The loss realized at frequencies outside the band also is shown in Fig. 2.

The delay distortion in the pass band of the filter, computed from the slope of its measured insertion phase characteristic, is given in Fig. 5. For short program systems, where no more than six filters are used in tandem, the delay distortion would not exceed the limits set for a high quality system. For longer systems it is necessary to equalize this delay distortion. The design and performance of the delay equalizers for this purpose are given in a separate paper.[3] These equalizers also include some attenuation equalization to correct for the systematic distortion of the filter.

Figure 6 shows an exterior view of the filter. On both sides of the mounting panel are metal containers which are provided with covers that can be soldered on to make a hermetic-sealed enclosure. In a corner of one can is a terminal box which contains the input and output terminals. These terminals are of the metal glass seal type which are vacuum tight. Mounted on brackets in each of the containers is a brass panel supporting the filter elements. One side of one of these panels is visible in Fig. 6, the other side is shown in Fig. 7.

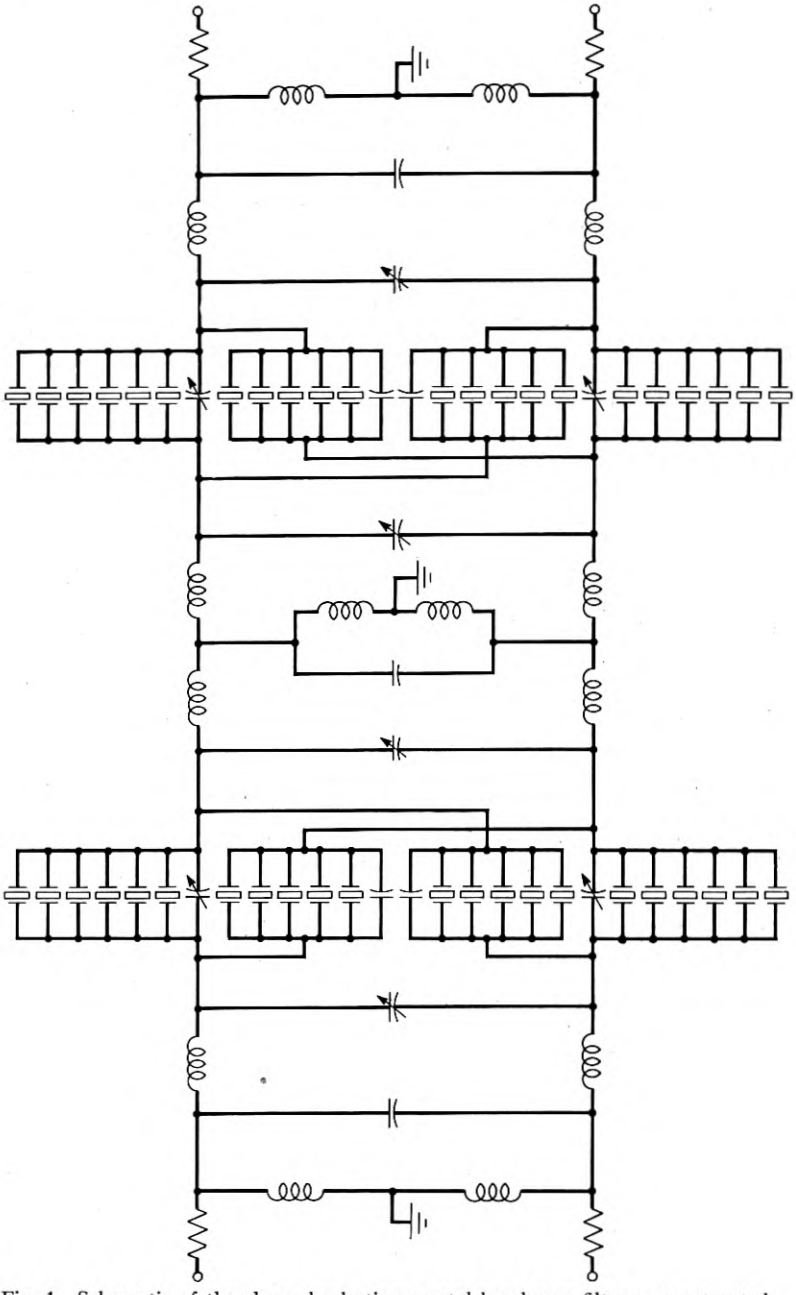


Fig. 1—Schematic of the channel selecting crystal band pass filter as constructed.

In Fig. 7 the two large cylindrical containers parallel to the panel house eleven of the balanced quartz crystal elements. The smaller cylindrical cans contain adjustable retardation coils while the rectangular cans house fixed coils. Adjustable air capacitors can be seen mounted on the hard rubber plate between the two crystal units.

The adjustment side of the brass panel is exposed in Fig. 6. Screwdriver adjustment of the retardation coils is possible through the circular holes at

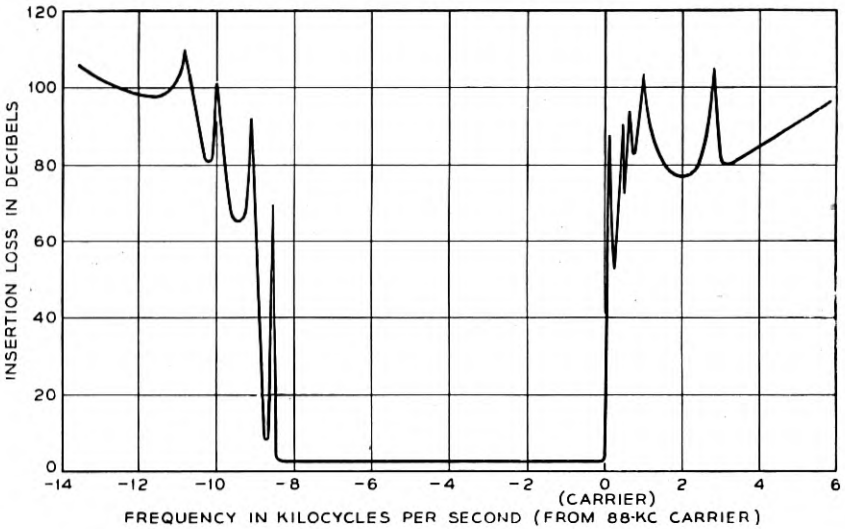


Fig. 2—The insertion loss-frequency characteristic of the filter.

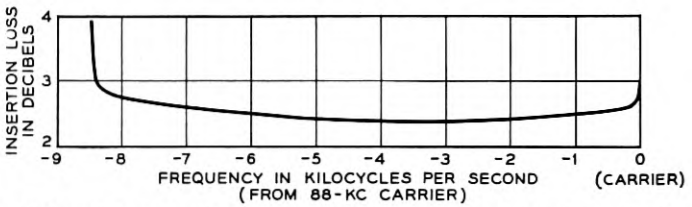


Fig. 3—Enlarged insertion loss-frequency characteristic of the filter pass band.

the top left and right of the panel. The rotors of three of the four air capacitors are visible inside the square cut-out in the panel. The panel in the lower half of the filter contains the remaining elements mounted and wired in a similar manner.

The schematic which was found to be most useful during the design of the filter is shown in Fig. 8. Thus the electrical circuit consists essentially of two complex lattice sections separated by one constant-k ladder section

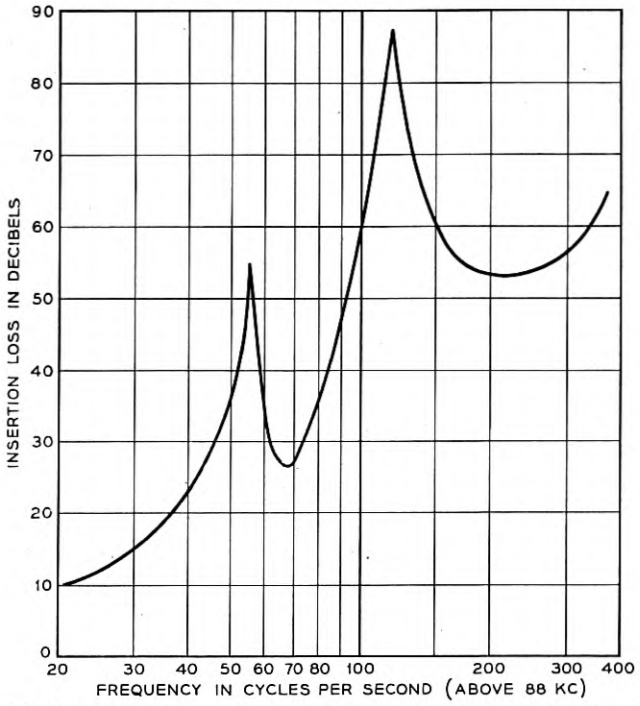


Fig. 4—The sharpness of the upper cut-off of the filter is shown in this enlarged loss characteristic.

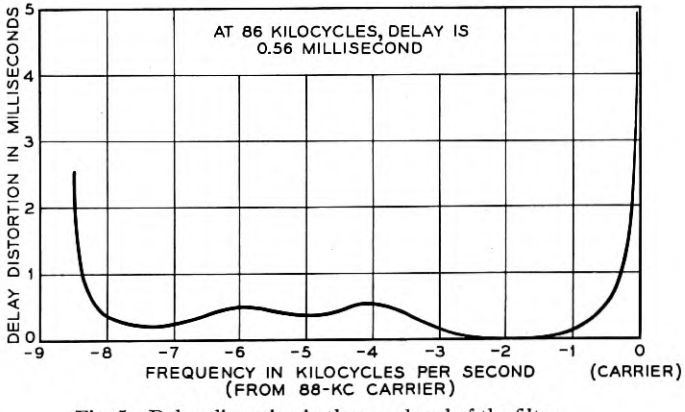


Fig. 5—Delay distortion in the pass band of the filter.

and terminated at each end by half-sections of the constant-k ladder type. The performance of the filter results almost entirely from the lattice sections since they control the flatness of the pass band, the sharpness of the cut-off

and give practically all the discrimination required. It will be noted that the filter uses the equivalent of 130 electrical elements consisting of 63 inductors, 63 capacitors and 4 resistors.

The use of complex filter sections permits the realization of filter characteristics which have low distortion in the pass band and high discrimination outside the pass band with a more efficient utilization of elements than is possible with a larger number of simpler sections. Improved mathematical methods of network analysis developed in the past several years

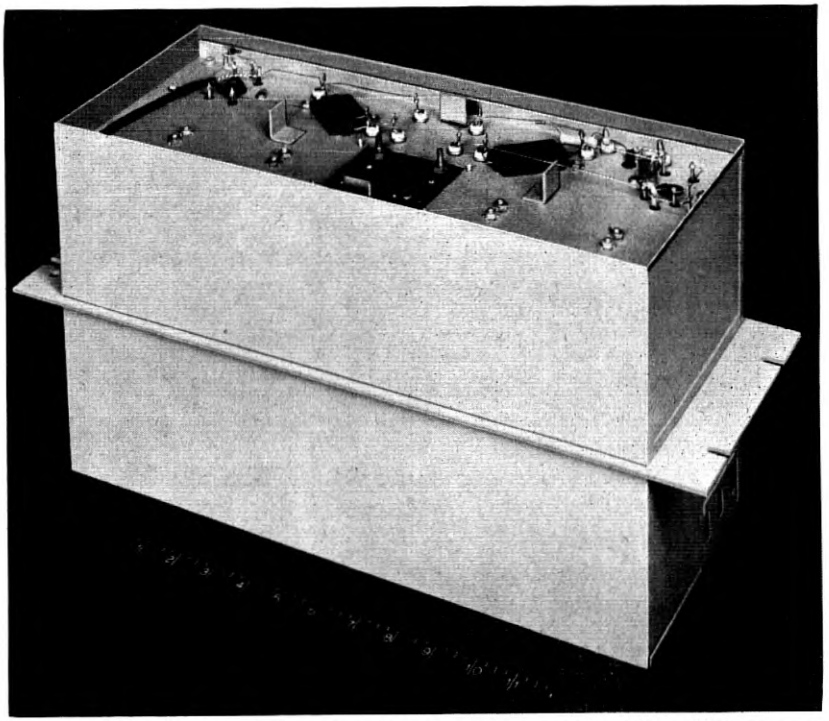


Fig. 6—Exterior view of the filter with one cover removed. After adjustments are completed the cover is soldered on to seal the assembly.

have made the design of such complex filter sections possible while recent developments of precise and stable filter elements and improved measuring circuits have made it possible to manufacture such filters.

It has been mentioned before that the use of quartz crystal elements restricts the filter band width which can be realized. In the frequency location selected for this filter (lower sideband of an 88-kc carrier frequency) a complex lattice section of the type shown in Fig. 8, when used alone, will permit the use of physical crystal elements for bands not over 7300 cps

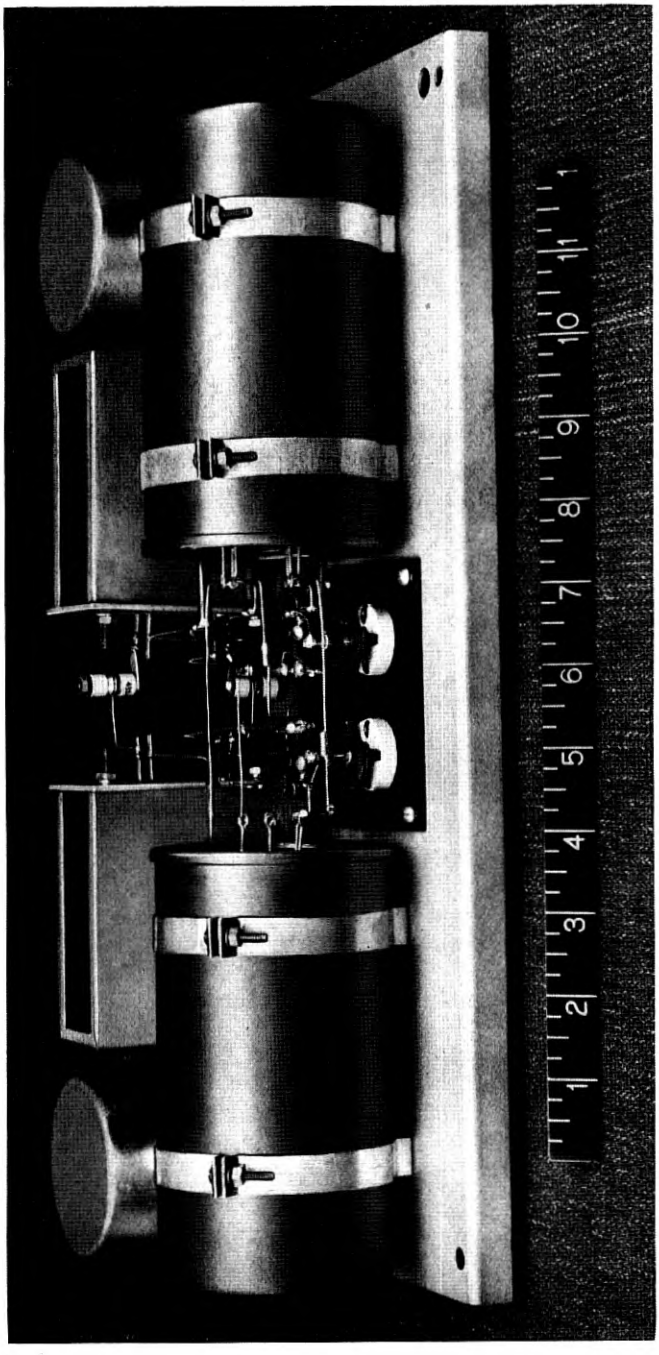


Fig. 7—Another view of one of the two panels which support the filter elements.

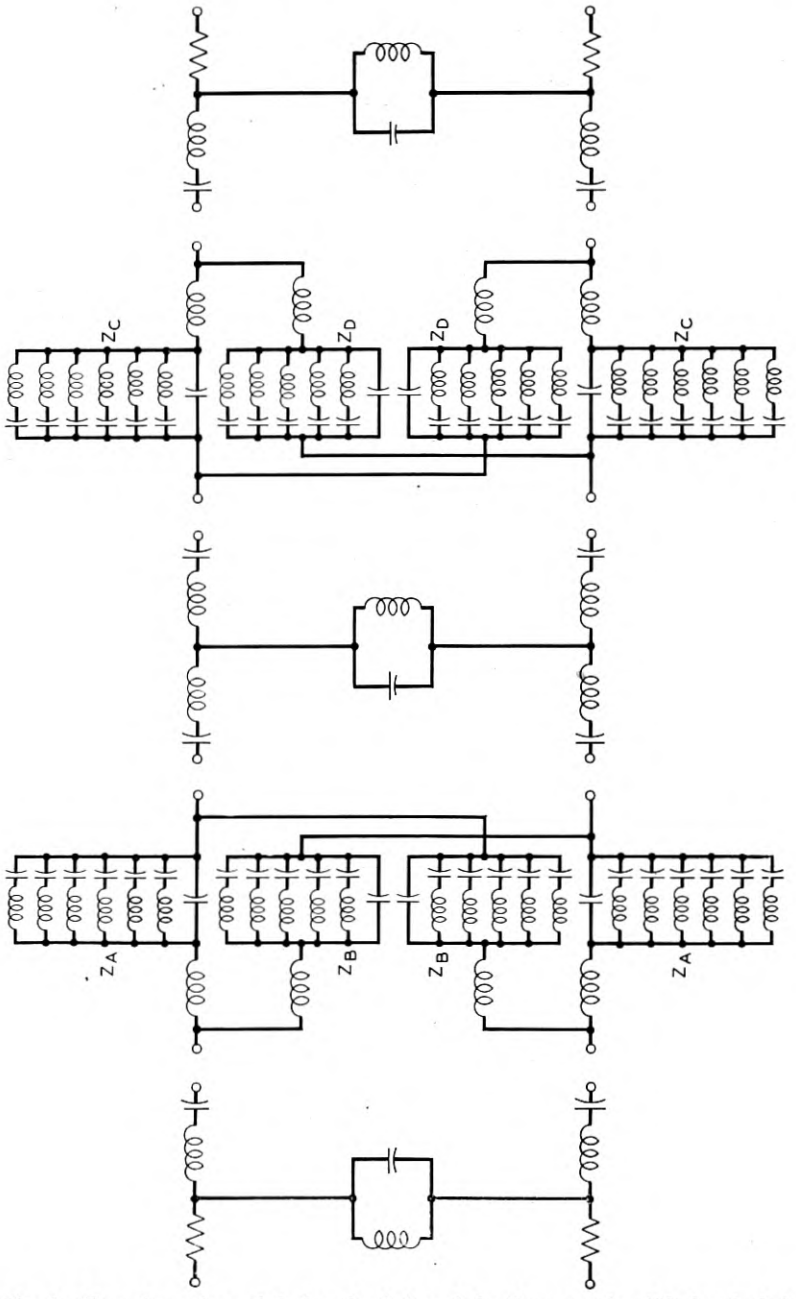


Fig. 8—The schematic used during the design of the filter contains 130 electrical elements.

wide. A wider band was obtained in this case by combining the complex lattice sections with ladder sections as shown in the figure. For this filter a combination of sections was designed which gave physically realizable crystal elements for a band width of 8.5 kc. This was the maximum band width possible without increasing the distortion in the band.

Summarizing, the filter design process consists of:

1. Design of wide band lattice sections which have quartz elements which cannot be realized in practice.
2. Design of electrical ladder sections of still wider band which introduce little distortion at the pass band frequencies of the lattice section. At this point the schematic is as shown in Fig. 8.
3. Combination of like elements, electrical transformations, and replacement of groups of elements consisting of an inductor and capacitor in series shunted by a second capacitor by their equivalent crystal elements. This gives the final schematic shown in Fig. 1, in which the crystal elements are physically realizable.

The general steps in the design of lattice filters[2,4] are as follows:

1. Choice of filter cut-offs.
2. Determination of number and location of impedance controlling frequencies to give a good match of image impedance to the termination.
3. Location of peaks of infinite attenuation to give the necessary transfer loss at frequencies removed from the pass band.
4. Determination of impedance level which gives the most reasonable element values.

Theoretically a filter could be designed which contains only one lattice section. The decision to split the filter into two sections was based on a desire to simplify the design to ease the manufacturing problems. The attenuation burdens of each section were reduced sufficiently to allow wider tolerances to be placed on the filter components. The last design steps are to determine the schematic of each section and to compute the theoretical element values in accordance with previously described methods.[2,4]

Although the filter elements computed were physically realizable they represented such extreme values as to introduce difficult problems. This was true especially of the crystal elements where the equivalent inductances of the eleven crystal elements in one section varied from 16 to 465 henries, a range of 1:29. A similar situation existed in the other section.

Crystal elements of the +5 degree X-cut type vibrating in their fundamental longitudinal mode are used in this filter. The equivalent inductance of such crystal elements varies directly with the thickness and inversely with the width of the plate. Therefore the high inductance plates are thick and narrow and the low inductance plates are thin and wide. In one section of the filter the dimensions of the plates required varied in width from

0.67 to 0.17 inch, in thickness from 0.119 to 0.012 inch and in length from 1.40 to 1.23 inches. The small variation in length is due to the fact that the length is determined primarily by the frequency of resonance of the plate and this change is small across the filter band. The temperature coefficient of the +5 degree X-cut quartz crystal element used in this filter is superior to the −18 degree X-cut longitudinal type which has been used in many other crystal filters but otherwise they are similar in use and in manufacture.

The filter attenuation distortion in the vicinity of the cut-offs is dependent on the dissipation in the elements which resonate there. In order to minimize this distortion, it has been found necessary to impose minimum Q requirements of 80,000 on the high-impedance crystal elements which resonate near the cut-offs. This high Q is realized by suspending the quartz crystal plates from fine wires[5] and operating them inside of evacuated containers. The low-impedance crystal elements which resonate at frequencies removed from the cut-offs require a minimum Q of 15,000. This comparatively low Q is realized by quartz crystal elements vibrating in air at atmospheric pressure.

In the equivalent electrical circuit of a quartz crystal element the large ratio of the shunt capacitance to the internal capacitance is a measure of the poor electromechanical coupling of quartz. For the +5 degree X-cut quartz crystal element this ratio of capacitances is about 140 for a plated blank before fabrication. It is obvious that fabrication, wiring and parasitic capacitances which may be in parallel with the quartz plate will make this ratio still higher and thus will reduce further the filter band width obtainable. For this reason it is important to keep to a minimum any capacitances which appear across any arms of the crystal lattices. One method used to minimize these capacitances was to assemble the eleven crystal elements required for each section in two containers instead of eleven separate ones. The five high-impedance elements requiring minimum Q's of 80,000 are assembled in one evacuated metal container and the six low-inductance elements having the lower Q's are assembled in another hermetic sealed container filled with dry air. A photograph showing the method of assembly used in given in Fig. 9.

A method was found to reduce the ratio of capacitances of the crystal elements. This method consists of dividing the plating on the surface of the quartz so that the driving voltage is removed from the end portions of the quartz plates. This plating division increases the equivalent inductance of the quartz plate but also decreases the direct capacitance between the plated surfaces. It has been found that the decrease in shunt capacitance with removal of plating is more rapid than the increase in equivalent inductance up to a certain point. If the plating is removed up to this optimum point it has been found possible to reduce the shunt capacitance about

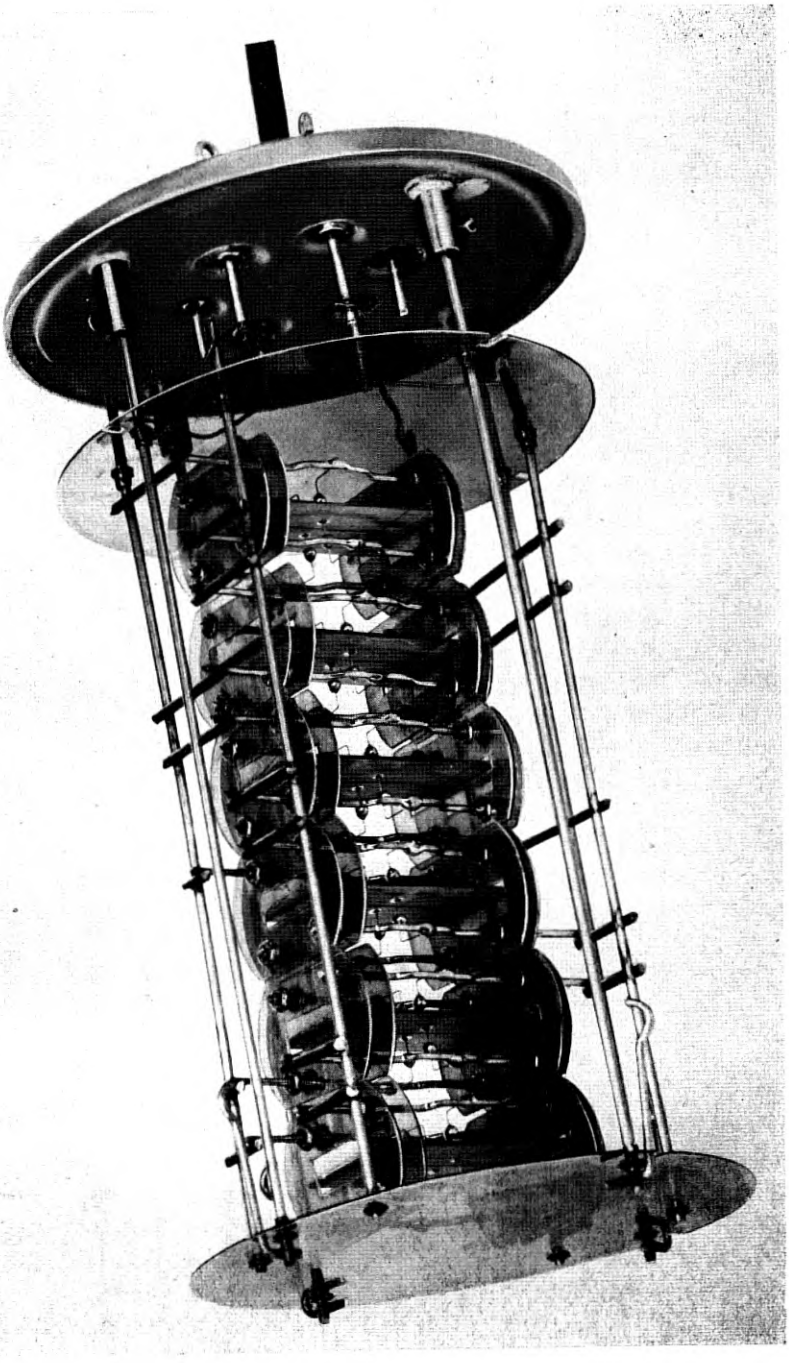


Fig. 9—Method of assembly of the quartz crystal elements.

17% below what it would be with a fully plated crystal element having the same inductance. This method of capacitance reduction was used on the six low-inductance crystal elements in each section. Another step in minimizing the unwanted capacitances was to design the retardation coils which connect to the terminal ends of each lattice to have as little capacitance as possible. Finally precautions were taken to keep the wiring capacitances to a minimum and the air condensers used inside the lattice for adjustment purposes are of special design having a minimum capacitance of only 0.5 mmf.

The resonant frequencies of each of the twenty-two crystal elements must be adjusted to the desired nominal frequencies within very close tolerances. On the ten high-impedance crystal elements the tolerance is ± 2 cps while on the 12 low-impedance crystal elements the tolerance is ± 5 cps. This precise frequency adjustment is accomplished by careful grinding of the length of the quartz plate.

The equivalent inductance of each of the 22 quartz crystal elements is required to be within two per cent of its nominal value. This specification is met primarily by close dimensional tolerances in the manufacture of the quartz plate. Any small adjustments which are necessary to meet this requirement are done by the aforementioned method of isolation of a small amount of plating from near the end of the quartz plate.

The four fixed retardation coils are adjusted to be within two per cent of their nominal inductance values. The variations from nominal are partially absorbed in the filter adjustment procedure where the coils are tuned with their associated variable capacitors to give the desired resonance frequency. The fixed mica capacitors are manufactured to be within 0.5 per cent of the desired nominal value. The three adjustable retardation coils are constructed to permit an inductance variation of five per cent on either side of their nominal values. This is done by moving a permalloy core in the field of the coil. Adjustment of these coils in the filter is accomplished by tuning them with their associated precision capacitor to give the desired resonance frequency within ± 25 cps. This type of adjustment procedure gives the correct LC product. The correct L/C quotient is obtained also since C is accurate to ± 0.5 per cent. The two resistors at each end of the filter compensate for the dissipation in the end retardation coils and thus restore the terminating impedance to the value required for optimum filter performance.

Each lattice of crystal elements and capacitors is a four-terminal bridge which is adjusted for maximum bridge balance at a particular frequency by means of the variable air capacitors in two of the arms. The precision of inductance adjustment of the crystal elements insures that the other peaks of attenuation will be sufficiently close to their nominal locations.

To obtain maximum loss at the filter peaks it is necessary to secure a conductance balance in each lattice section as well as a susceptance balance. This can be done if care is exercised in the choice of materials used in fabricating the crystal elements and capacitors which appear inside the lattice. In this case the crystal element insulators and dielectrics consist of glass, mica, quartz and clean dry air or vacuum while the air capacitors use glass, ceramic and air for their insulators and dielectrics. If these materials are clean and dry they have very low conductance and do not influence the bridge balance. A complete discussion of the effects of impedance unbalances on crystal lattice performance has been given in a recent paper by E. S. Willis.[6]

To further insure that dirt and moisture will not influence its performance the filter is adjusted, tested and hermetically sealed in an air conditioned room where the relative humidity does not exceed 40 per cent. Since manufacture started about the beginning of 1946 several hundred of these filters have been made and are functioning satisfactorily in the telephone plant.

Band Elimination Filter at Branching Points

When broad-band carrier systems are equipped for the transmission of a carrier program channel, it is frequently necessary to provide between carrier terminals intermediate or branching points at which the program may also be received. If only receiving facilities are involved, rather simple bridging arrangements can be provided. However, program network needs often require a more flexible arrangement at the branching point so that a line may be cleared of the program originating at one terminal and a new program introduced for transmittal toward the next terminal.

To do this without affecting the message channels also being transmitted on the line, a filter has been developed to block the program channel already on the line while freely transmitting the message channels. With this filter in the circuit the high-frequency line between the branch point and the following terminal is free of program frequencies and the program originating at the branch point may be sent toward that terminal.

Since the program channel occupies frequency space near the center of the 12-channel message group, the remaining message channels appear above and below the program frequencies. Therefore the blocking filter at the branching points must be of the band elimination type. The circuit employing this filter may be designed to block either at line frequencies or at basic group frequencies. The latter method, of course, requires that a demodulation process be provided to translate line frequencies to basic group frequencies before the blocking filter is inserted in the circuit.

The band elimination filter described herein was developed for the type *K* carrier system (Carrier-on-Cable) for which the first option mentioned above was chosen. This filter operating at the line frequencies of the type *K* system is required to transmit frequencies from 12 to 31.6 kc and 44.2 to 60 kc while blocking those from 32 to 43.2 kc. Actually the filter will transmit frequencies below 12 kc and above 60 kc but these do not appear on the type *K* line and therefore there are no requirements in these ranges.

The filter which performs these functions is shown schematically in Fig. 10. Several factors made its design difficult. A high level of discrimination of the order of 75 db is required over a wide frequency range of about 12 kc. Also the allowable waste interval between wanted and unwanted frequencies is very small. The filter must transmit with a maximum distortion of 0.2 db to within 97.5% of the first unwanted frequencies at which a discrimination level of 75 db is required.

Because of the severe requirements the familiar image parameter design method was not employed. In this, as is well known, the composite filter first presented by Zobel[7] is made up of sections with matched image impedances but different transfer constants depending upon the attenuation requirements. Instead, it was felt that a design method proposed by Darlington[8] offered a better possibility of meeting the requirements with a reasonably sized filter. This procedure known as the *insertion loss* method is based upon the determination of a four-terminal transducer of reactances which, when inserted between definite resistance terminations, will produce a specified loss characteristic. A filter so designed has an advantage over image parameter filters in that the attenuation obtainable is greater for the same effective cut-off and an equal number of elements. *Effective cut-off* as used here means the last frequency of interest in the transmitted band. It is possible, therefore, with an *insertion loss* filter to use fewer elements for a given attenuation, or to obtain a wider transmission band with the same number of elements.

The advantage inherent in the newer design method is not derived from a difference in structure. In configuration there is no way to distinguish such a filter from one of conventional image design. The difference lies solely in the element values. A simple way to visualize how the *insertion loss* design varies from image design is to consider that the newer method removes an arbitrary restriction placed upon the image theory to simplify the mechanics of design. The restriction is that the nondissipative image attenuation must be identically zero over continuous frequency ranges including the transmitted bands and other than zero everywhere else. This leads to the familiar ladder image filter composed of matched sections, or the lattice filter with coincident critical frequencies.

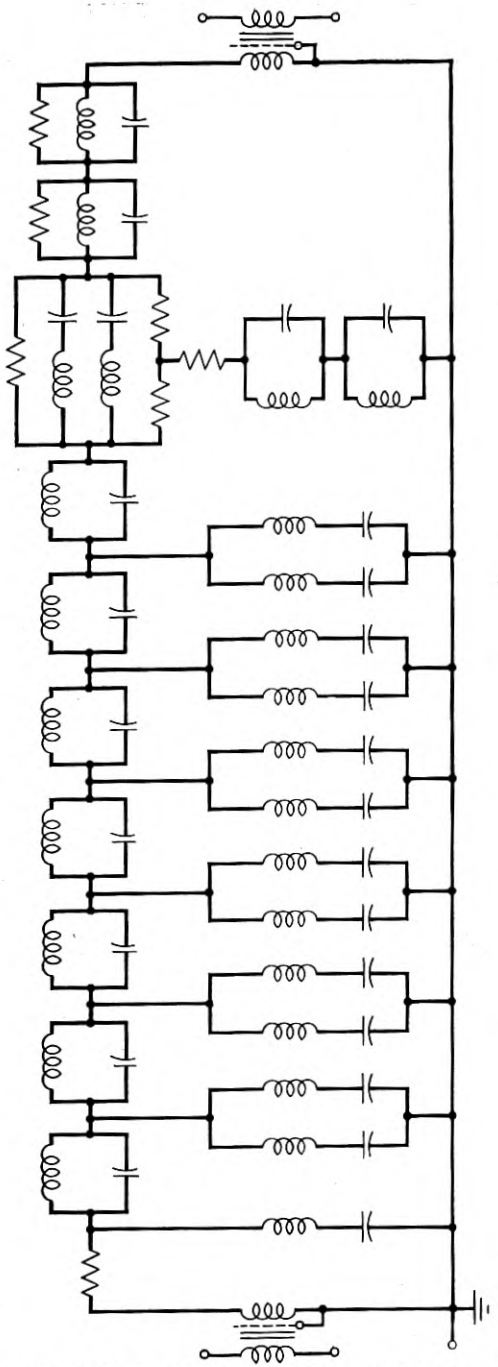


Fig. 10—Schematic of the band elimination filter used at branching points of the type K system.

Analysis of an *insertion loss* ladder filter shows that it may be considered a composite of image sections which are not matched in image impedance. As a composite filter this means that the effective pass band has been split into a number of pass bands each separated by a small attenuation region. Darlington has formulated the process by which these bands can be so arranged that advantage can be taken of the fact that the image attenuations in these bands for small mismatch are comparable to the terminal effects and that reflection gains up to 6 db are possible in the same regions. The combination of these effects, which can be controlled up to and including the cut-off, gives the *insertion loss* filter its improved performance since the *effective* and *theoretical* cut-offs can be made identical, with no

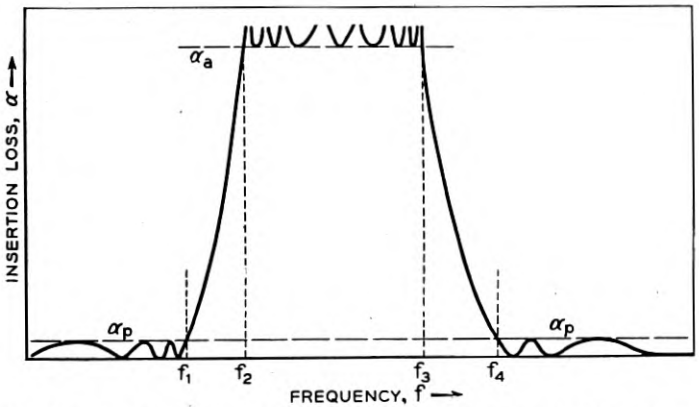


Fig. 11—Non-dissipative filter characteristic obtained by use of Tchebycheff parameters in pass bands and attenuation band.

frequency space needed for the rounding due to the terminal effects in image filters.

In general the mathematical steps required to design a filter by this method are as follows: An insertion loss frequency function is chosen which will satisfy the filter requirements and will lead to a structure economical of elements. From this are found the open and short circuit impedances of the proposed network which is normally of the standard lattice or ladder forms. Finally from these expressions the element values are determined.

The particular form of *insertion loss* design employed for the filter described here is a special case of the general theory. The filter requirements lent themselves to the use of Tchebycheff parameters simultaneously in the pass bands and attenuation band. The application of these parameters was first described by Cauer.[9] The typical non-dissipative characteristic resulting from their use is shown on Fig. 11. It is seen that the

pass band characteristic is of the *ripple* type with equal maxima and equal minima. In the attenuation region the valleys of loss are of equal value.

The general form for the insertion power ratio to obtain the desired characteristic is:

$$e^{2\alpha} = \frac{4R_1R_2}{(R_1+R_2)^2}[1 + (e^{2\alpha_p} - 1)\cosh^2\theta_I]$$

In this equation R_1 and R_2 are the resistive terminations and α_p is the maximum ripple in the pass band as shown in Fig. 11.

θ_I represents a function of frequency so chosen that $\cosh\theta_I$ is an odd or even rational function of frequency. Also θ_I must be a pure imaginary throughout the passing band and must be of the form $(\alpha_I + n\Pi i)$ in the attenuation region. The term α_I is real at all attenuation frequencies becoming infinite at those required by the specification of minimum α_a in Fig. 11.

Darlington further showed that θ_I closely conforms with the image transfer constant of an image parameter filter if the effective pass band of the insertion loss filter coincides with the theoretical pass band of the image filter. Based on this conclusion a design method was formulated which permits a reference filter derived from image parameters to be used as the basis of the *insertion loss* filter. There is, of course, no correspondence between the elements of the reference image filter and the insertion filter. This reference filter is not a requisite to the development of the insertion theory but it does offer a convenient and well known transfer constant which is the right functional form for use in the insertion power ratio stated above.

Referring again to Fig. 11, the approximate minimum loss, α_a, determines the number of peak sections required in the reference filter from the relationship:

$$\alpha_a = 20\log(e^{2\alpha_p} - 1) - 10(2m+1)\log q - 18$$

where "m" is the number of peaks required and α_p is the band ripple function as before. The new term introduced here is "q" which is directly tied up with the selectivity demanded of the filter, i.e., the amount of frequency space avilable between the last useful frequency or *effective* cut-off and the first frequency at which attenuation equal to α_a is needed. The relationships are as follows:

$$q = \frac{1}{2}\left[\frac{1-\sqrt{K'}}{1+\sqrt{K'}}\right] + \frac{1}{16}\left[\frac{1-\sqrt{K'}}{1+\sqrt{K'}}\right]^5$$

where $K' = \sqrt{1-K^2}$

and $K = \dfrac{f_3 - f_2}{f_4 - f_1}$

The filter described here actually consists of two filters connected in tandem, each derived from a different power ratio. This step was taken because of the relatively low dissipation factor realizable with coils of reasonable size. By dividing the total attenuation between two power ratios, lower overall distortion due to dissipation was achieved. The distortion represented by the non-dissipative ripple "α_p" was minimized by so assigning the frequencies of infinite attenuation to the two functions that phasing in of the ripples was avoided as far as possible.

The two power ratios selected are:

$$e^{2\alpha_1} = 1 + (e^{2\alpha_p} - 1)\cosh^2 \theta_{I_1}$$

$$e^{2\alpha_2} = \frac{4R_1R_2}{(R_1 + R_2)^2}\,[1 + (e^{2\alpha_p} - 1)\cosh^2 \theta_{I_2}]$$

For these the peak frequencies were assigned on an alternate basis as follows:

$$\text{To } \theta_{I_1} : m_1, m_3, m_5 \text{ and } m_7$$

$$\text{To } \theta_{I_2} : m_1, m_2, m_4 \text{ and } m_6$$

with the value of "m" decreasing from m_1 to m_7 and $m_1 = 1$. The parameter "m" has the same meaning as in image filter theory.

The next step in the process is the finding of the roots of the two power ratios. These may be obtained from the following expansions:

For $e^{2\alpha_1}$ representing a reference filter of $3\frac{1}{2}$ sections:

$$(m_3 + x)^2(m_5 + x)^2(m_7 + x)^2(1 + x) + \left(\frac{e^{\alpha_p} - 1}{e^{\alpha_p} + 1}\right)(m_3 - x)^2(m_5 - x)^2(m_7 - x)^2(1 - x) = 0$$

which is expressed in the form

$$K_1[x^7 + a_1x^6 + a_2x^5 + a_3x^4 + a_4x^3 + a_5x^2 + a_6x + a_7 = 0]$$

For $e^{2\alpha_2}$ representing a reference filter of 4 sections:

$$(m_2 + x)(m_4 + x)(m_6 + x)(1 + x) + i\sqrt{\frac{e^{\alpha_p} - 1}{e^{\alpha_p} + 1}}\,(m_2 - x)(m_4 - x)(m_6 - x)(1 - x) = 0$$

which is expressed by

$$K_2[x^4 + a_8x^3 + a_9x^2 + a_{10}x + a_{11} = 0]$$

In the above expressions $x = \sqrt{1 + \frac{1}{p^2}}$ where $p = i\omega$ and α_p for the filter

discussed here is 0.1 db. From the roots obtained from the above equations of 7th degree and 4th degree complexity, the open and short-circuit impedances are determined which in turn lead to the element values. The complete development of the process resulted in the filter portion of the network shown on Fig. 10.

The remainder of the schematic shows the equalizer which corrects the rounding of the filter characterisic near the cut-offs due to dissipation. The equalizer is of conventional bridged "*T*" design with constant "*R*" impedance in tandem with a simple series section.

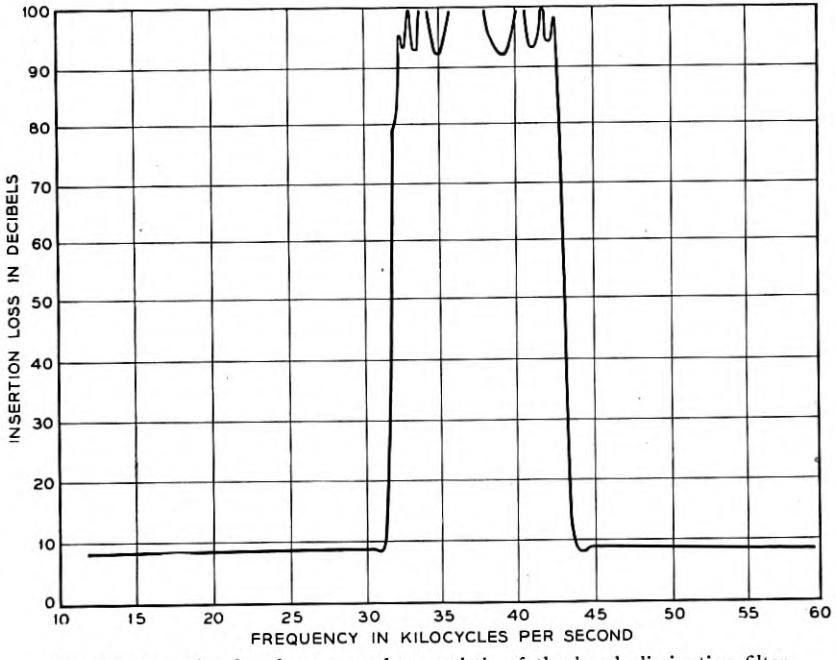


Fig. 12—Insertion loss-frequency characteristic of the band elimination filter.

Repeating coils are required as shown because the filter was designed at a 600-ohm level to give commercial elements whereas it is required to operate between 135 ohm resistances. In the schematic a resistance will be noted in series with one termination. This is needed because the "insertion" design with inverse impedance terminations as shown here requires unequal terminations to produce the specified loss characteristic. Usually this would be taken care of by proper design of the repeating coil but, in this case, economic reasons dictated the use of the same repeating coil at both ends of the structure. The termination was therefore built out with a

physical resistance. This of course introduces a flat loss but in this case enough gain was available in the circuit to permit it.

On Fig. 12 is shown a typical transmission characteristic when the filter is operating between 135-ohm resistances. A variety of component parts are required to give this performance. The filter portion employs mica condensers throughout and a mixture of molybdenum permalloy and air-core retard coils. As many permalloy coils are used as possible in order

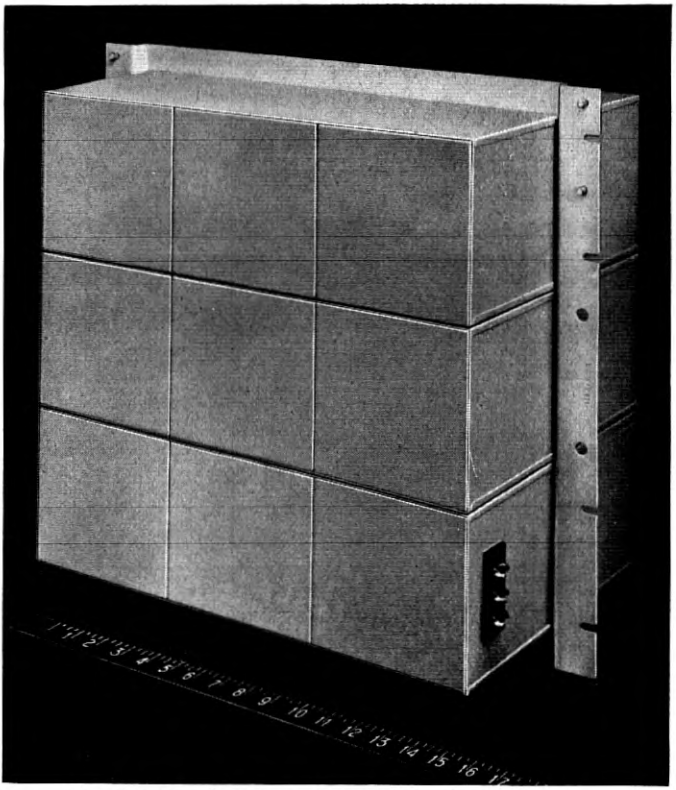


Fig. 13—Exterior view of the band elimination filter.

to obtain high "*Q*". The air-core coils, of an adjustable type, are used in those arms which control the peak frequencies near the pass band. These arms must be adjusted very accurately for resonance in order to maintain the steep slope of loss in the cut-off region. The equalizer sections employ duolaterally wound air core coils also adjustable in order to set the pass band losses accurately. Mica and paper condensers are used in the equalizer, the latter being used where capacity values make mica condensers extremely expensive.

A completed filter is shown in the photograph of Fig. 13, while the internal arrangements of one portion of the assembly are shown on Fig. 14.

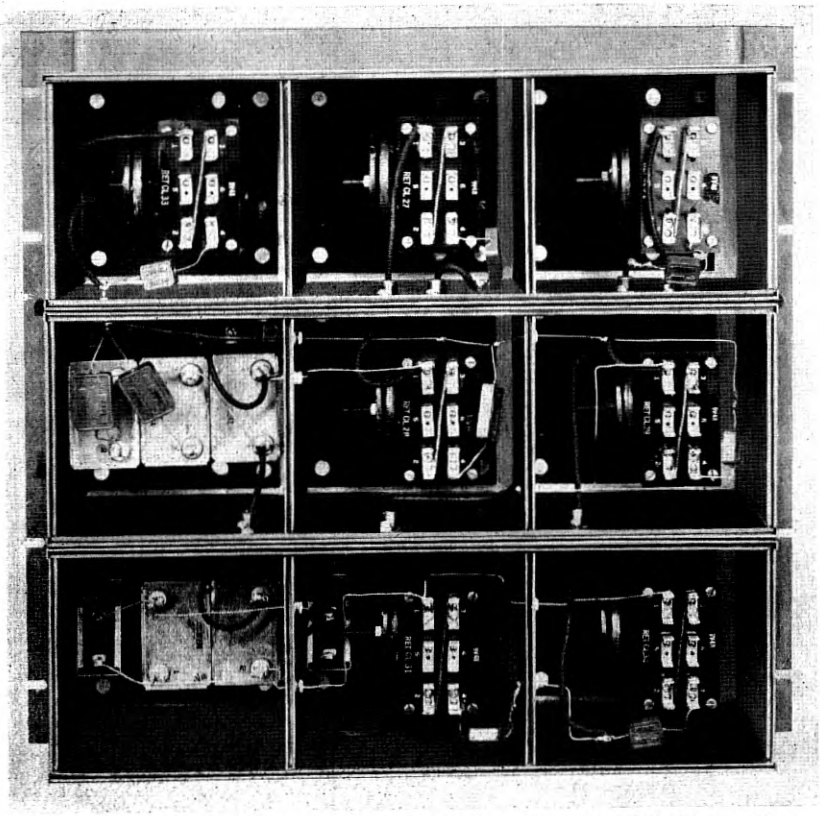


Fig. 14—Interior view of one portion of the assembly of the band elimination filter.

Phase Simulating Network

When program rearrangements at a branching point are required, the band elimination filter must be switched into or out of the through transmission path. This transfer is accomplished without opening the through path. Thus, for a brief time during the switching interval, message channels are transmitted simultaneously through the filter and the non-blocked circuit. A large phase difference between the two parallel paths is introduced by the filter which, in the absence of phase correction in the through circuit, could cause errors in the transmission of voice frequency telegraph signals. Therefore a network having phase shift similar to that of the filter over most of the message range is provided in the through circuit.

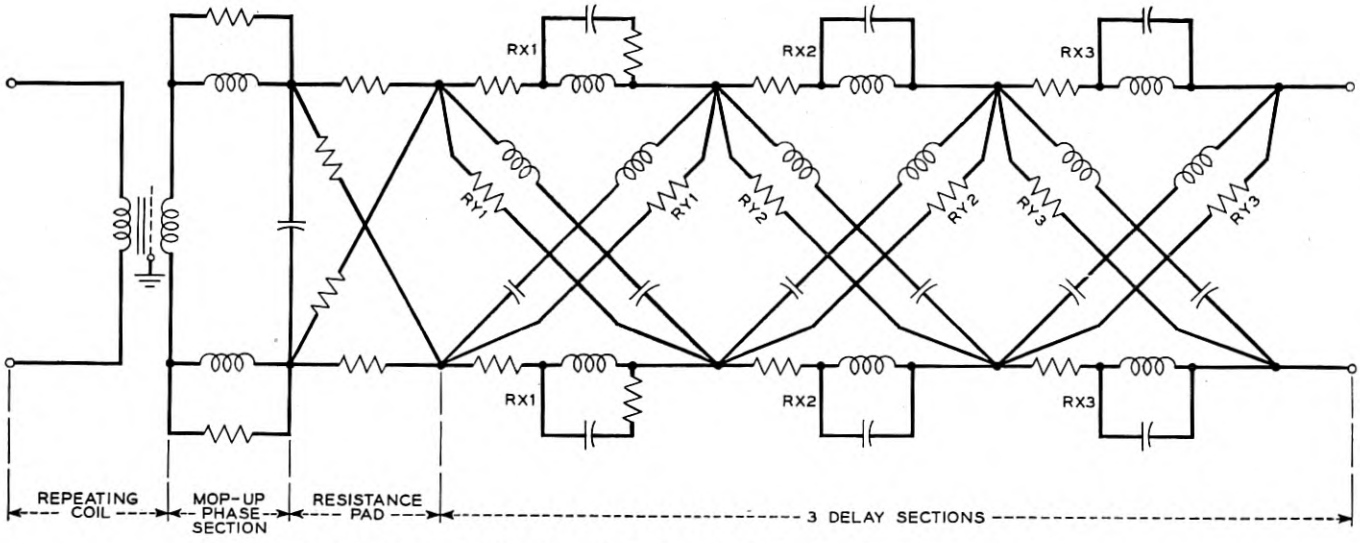


Fig. 15—Schematic of the phase simulating network showing the component sections.

The phase simulating network is shown in schematic form in Fig. 15.

The network is a balanced structure and consists of the following pieces of apparatus connected in tandem:

1. An input repeating coil to improve the longitudinal balance at the sending end,
2. A half-section high-frequency cut-off low-pass filter to mop up the phase shift introduced by two repeating coils of the band elimination filter,

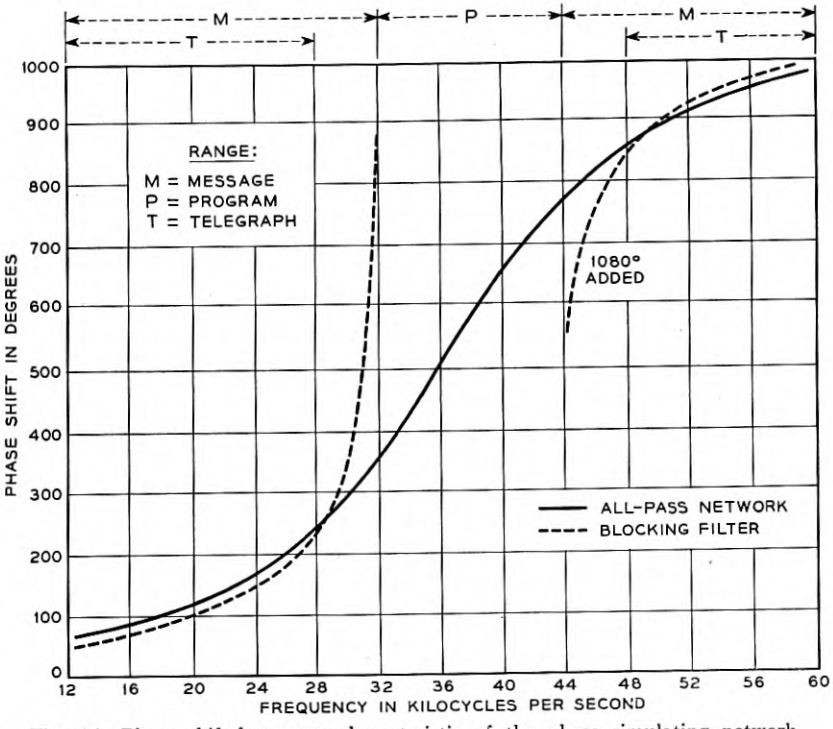


Fig. 16—Phase shift-frequency characteristic of the phase simulating network.

3. A resistance pad to equalize the over-all loss level of the all pass network to within ±0.1 db of the pass band loss of the band elimination filter, and
4. Three delay sections, self equalized for loss,[10] for simulating the phase shift of the band elimination filter.

The network simulates the phase shift of the band elimination filter over the frequency ranges covered by message channels 1 to 4 and 10 to 12 to within 20 electrical degrees as shown in Fig. 16. As phase simulation is

incomplete in the frequency ranges occupied by channels 5 and 9 due to the steep phase shift slope of the band elimination filter near its cut-off points, no telegraph channel are assigned to these channels of type "*K*" carrier circuits equipped with branching points.

The phase shift of the band elimination filter is discontinuous between its cut-off frequencies and has a positive slope with frequency in its pass bands. As the phase shift of a delay section increases continuously with frequency, it is impossible to provide the exact counterpart of the filter in a delay network. However, the addition of any multiple of 2Π radians does not change the transmission characteristic. Hence 6Π radians (3

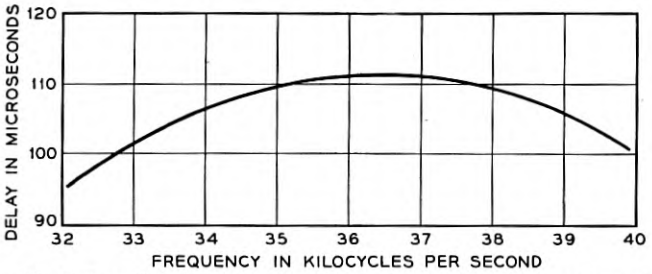


Fig. 17—Delay of the phase simulating network at program frequencies.

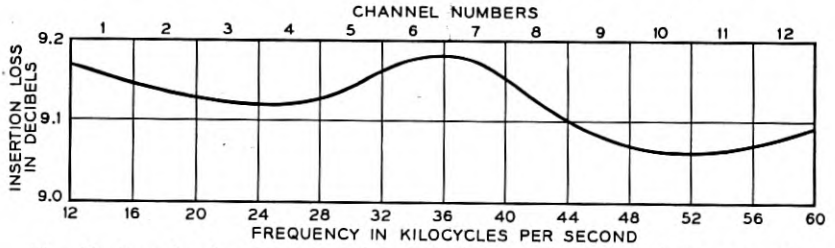


Fig. 18—Insertion loss-frequency characteristic of the phase simulating network.

revolutions) are added to the phase shift of the elimination filter above the upper cut-off to simulate its phase characteristic in the 10 to 12 message channel range, as well as to provide an almost linear phase slope in the 32 to 40 kc program channel range resulting in minimum delay distortion.

The delay distortion of the network over the program channel is approximately 16 microseconds as shown in Fig. 17. The loss distortion over the program channel is approximately 0.05 db and over any one message channel it is less than 0.05 db as shown in Fig. 18.

The self loss equalizing feature of a delay section is evaluated at zero frequency in the form of a resistance pad by making the pad loss approximate the insertion loss at the critical frequency of the delay section. The

resistance R_x located in the series branch may be evaluated from the expression

$$R_x = R\frac{\epsilon^\theta - 1}{\epsilon^\theta + 1} - R_{DC}$$

in which ϵ^θ is the transfer loss in nepers at the critical frequency, R is the 135 ohm resistance termination and R_{DC} is the DC resistance of the inductance coil. The resistance Ry located in parallel with the series resonant branch may be evaluated from the expression $Ry = \frac{R^2}{R_x}$.

By changing the loss, ϵ^θ, of the derived resistance pad at zero frequency slightly from the measured loss at the critical frequency of the delay section, a suitable loss compensation may be realized to produce an optimum loss equalization over the message and program channel ranges. It is satisfactory to follow this technique when the condenser Q factor is much greater than the coil Q factor. When this condition exists, the insertion loss about the critical frequency becomes geometrically dissymmetrical, that is, the loss falls off more rapidly for frequencies above the critical frequency because of the controlling condenser Q factor.

References

1. "A Carrier System for 8000-Cycle Program Transmission," R. A. Leconte, D. B. Penick, C. W. Schramm, A. J. Wier., a companion paper. This issue of *BSTJ*.
2. "Electromechanical Transducers and Wave Filters" (book), W. P. Mason, D. Van Nostrand Co., New York, N. Y., 1942.
3. "Delay Equalization of 8-kc Carrier Program Circuits," C. H. Dagnall and P. W. Rounds, a companion paper. This issue of *BSTJ*.
4. "Communication Networks," Vol. II (book), E. A. Guillemin, John Wiley & Sons, New York, N. Y., 1935.
5. "The Mounting and Fabrication of Plated Quartz Crystal Units," R. M. C. Greenidge, *Bell Sys. Tech. Jour.*, Vol. 23, July 1944, page 234.
6. "A New Crystal Channel Filter for Broad Band Carrier Systems," E. S. Willis, *Elec. Eng.*, Vol. 65, March 1946, Page 134.
7. "Theory and Design of Uniform and Composite Electric Wave Filters," O. J. Zobel, *Bell Sys. Tech. Jour.*, Vol. 2, 1923, Pages 1–46.
8. "Synthesis of Reactance 4-Poles which Produce Prescribed Insertion Loss Characteristics," S. Darlington, *Journal of Mathematics and Physics*, Vol. 18, No. 4, Sept. 1939.
9. "Ein Interpolationsproblem mit Funktionen mit Positiven Realteil," W. Cauer, *Mathematische Zeitschrift*, 38, 1–44, 1933.
10. "Distortion Correction in Electrical Circuits with Constant Resistance Recurrent Networks," O. J. Zobel, *Bell Sys. Tech. Jour.*, Vol. 7, July 1928, Pages 438–534.

A Precise Direct Reading Phase and Transmission Measuring System for Video Frequencies

By D. A. ALSBERG and D. LEED

THE evolution of transmission networks for communications systems progresses through three fairly well-defined phases—design, synthesis and final adjustment. The design phase ordinarily involves no problem of measurement. In the synthesis stage, during which the physical model is constructed from the paper design, precise equipment is often needed for measuring the magnitude of the various components comprising the network. The adjustment stage, in which the network is actually tested as an element in a transmission circuit, generally requires the most complex instrumentation. In the latter category we may include insertion loss, gain, and phase measurement systems.

Television and broad-band carrier facilities, such as the New York-Midwest video cable link, employ vast numbers of transmission networks. These include, for example, filters, equalizers, and repeaters. The final adjustment of these networks requires a large number of precise insertion phase and transmission measurements during both development and manufacturing stages. Consequently, the measurement equipment must combine laboratory accuracy with speed of measurement suitable for use in production testing.

The quantities measured are defined in Fig. 1. Conforming with current usage, the term *Transmission* is used herein to designate insertion loss and gain.

The performance of the system with respect to frequency range, measurement range and accuracy is as follows:

Frequency Range: 50–3600 kilocycles

Generator and Network Termination Impedance: 75 Ω

Transmission Range: +40 db to −40 db; Accuracy ±0.05 db
−40 db to −60 db; reduced accuracy.

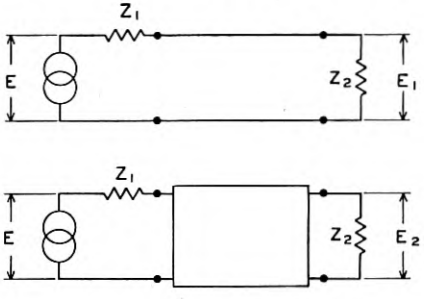
Insertion Phase Shift Range: 0–360°; Accuracy ±0.25 degree (+40 db to −40 db)

The measuring circuit is based on the heterodyne principle whereby the phase and transmission of the *unknown* are translated from the variable frequency to a constant intermediate frequency at which the phase and transmission standards operate. Accurate phase-shifters and variable attenuators with negligible phase shifts are constructed readily for fixed

frequency operation. This advantage more than offsets resulting problems of modulator design and automatic frequency control.

Conforming with the definitions of insertion phase and transmission, the measurement system compares, with respect to phase and amplitude, the outputs of two transmission channels energized from the measurement frequency source, one of which serves as a standard or reference channel, while the other contains the apparatus under test. This is illustrated by the block drawing in Fig. 2.

For loss measurements the range attenuator I (Fig. 2) is set at 0 db. Measurement frequency F from the master oscillator is applied to both standard "S" and unknown "X" channels through splitting pad I. The voltages at "S" and "X" modulator inputs, points A and B respectively



$$\left\{\begin{matrix}\text{INSERTION LOSS}\\ \text{OR}\\ \text{INSERTION GAIN}\end{matrix}\right\} = 20 \log \frac{|E_1|}{|E_2|}$$

$$\text{INSERTION PHASESHIFT} = \angle E_2 - \angle E_1$$

Fig. 1—Definition of quantities measured.

in Fig. 2, differ with respect to phase and amplitude because of the transmission differential introduced between the two channels by the apparatus under test. By frequency conversion in the "S" and "X" modulators these amplitude and phase differences at frequency F are translated at points C and D to a constant intermediate frequency, 31 kc. The second input to the "S" and "X" modulators, of frequency $F + 31$ kc, is supplied by the slave oscillator which automatically tracks at constant 31 kc difference with respect to the master oscillator. By selective filtering, only the difference frequency appears at the modulator outputs C and D. 31 kc has been chosen as the intermediate frequency, primarily on the basis of filtering requirements in the modulators. The detector (Fig. 2) compares the voltages of the "X" and "S" channels at K and L as to magnitude

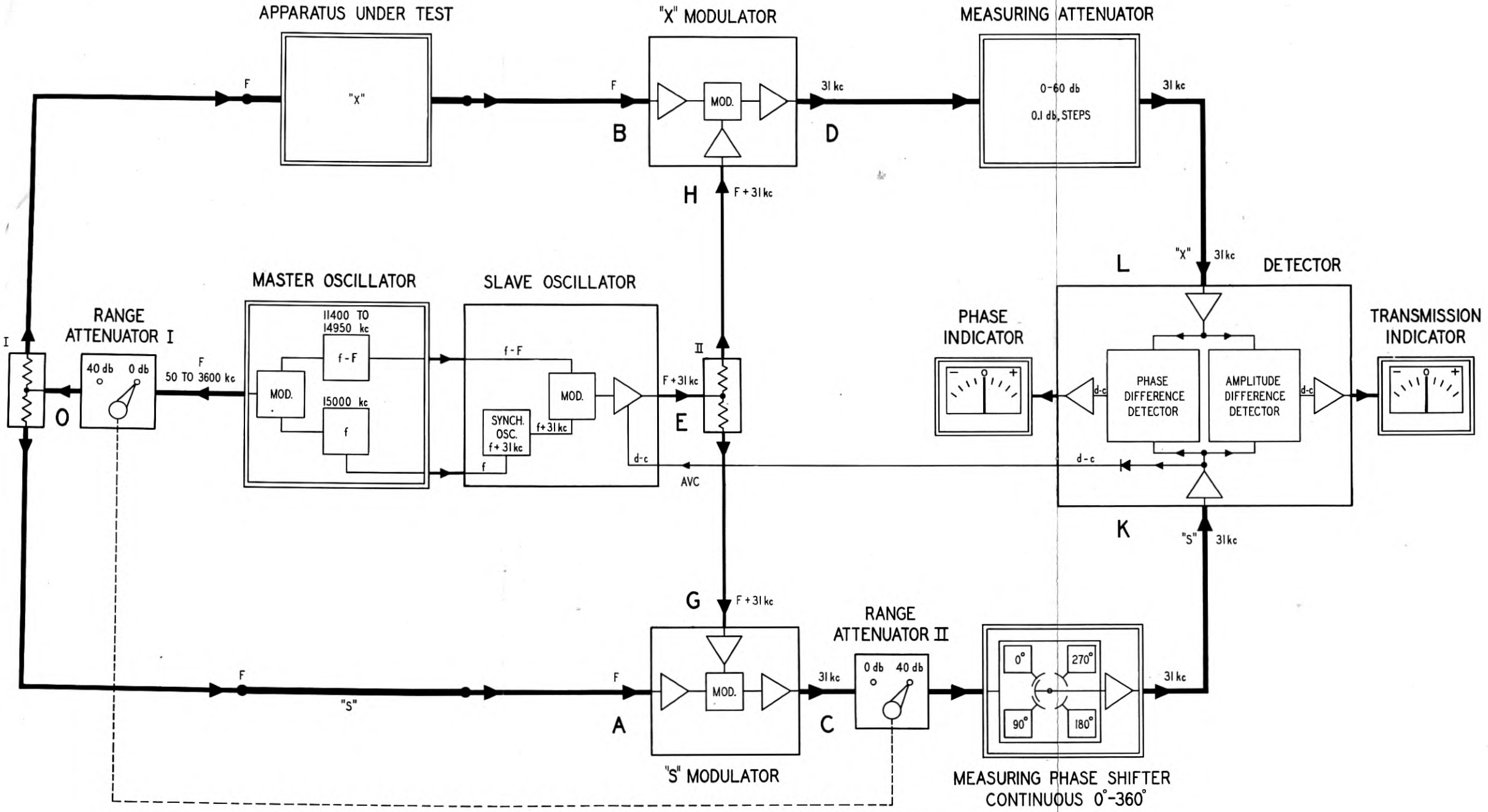


Fig. 2—Block schematic of the phase and transmission measurement system.

and phase, and indicates their difference on the direct reading scales of the indicator meters.

If the measuring attenuator is set at 60 db loss, and the range attenuator *II* at 40 db loss, "*S*" and "*X*" channels are in balance when the *apparatus under test* is replaced by a *zero loss* strap. The phase-shifter has, by design, 20 db loss; so that under these conditions "*S*" and "*X*" channels are nominally in balance, except for small residual phase and transmission differentials which may be *zeroed-out* by initial adjustment of the phase-shifter and of the relative gain between "*S*" and "*X*" channel amplifiers within the detector. Null readings on the phase and transmission difference indicating meters tell when exact phase and transmission balance between the two channels has been established. The phase-shifter and attenuator dials are arranged to read zero after this initial balance has been made. To measure apparatus transmission and phase, the strap is replaced by the *apparatus under test* and the balance restored by adjustment of the phase-shifter and the measuring attenuator. The insertion phase and transmission of the apparatus under test are then read directly from the calibrated dials of the phase shifter and attenuator.

When measuring loss, attenuation in the measuring attenuator is reduced by the amount of attenuation introduced in the high-frequency portion of "*X*" channel by the "apparatus under test." In measuring gain, the attenuation through the measuring attenuator must be increased by the amount of apparatus gain. To insure that "*S*" and "*X*" channel modulators are not overloaded by excessive input, range attenuator I is set to 40 db loss during gain measurements. This attenuator is common to both channels and therefore introduces no phase differential. Simultaneously and automatically, the range attenuator *II* immediately following the "*S*" modulator, is operated, removing 40 db loss from the 31 kc standard channel.

The measuring attenuator is self-computing and indicates directly in illuminated figures the gain or loss of the apparatus under test. A simple switching arrangement automatically controls the dial-lighting circuit of the measuring attenuator. When measuring gain the dial indications increase in one direction, and when measuring loss the indications increase in the opposite direction (Fig. 3).

In addition to the null-balance method, a deflection method of measurement using direct reading scales of the phase and transmission difference indicating meters is also possible. An automatic volume control circuit assures invariance of the indicator scale factors with either the modulator frequency-transmission characteristic, or input voltage variation at the "*S*" modulator caused by reflections from apparatus under test. The automatic volume control circuit regulates the output voltage of the slave oscillator to maintain the amplitude of the "*S*" channel input to the dif-

Fig. 3—Measuring attenuator with computing dials.

ferential detector constant. As the control action simultaneously affects both "*S*" and "*X*" modulators uniformly, the system zero is undisturbed.

Careful attention has been given to the problem of obtaining an electrical match between "*S*" and "*X*" modulators and coaxial cable lengths in the high-frequency channels. (RG 6/*U* cable contributes a phase shift of 0.2°/inch at 3600 kc.) Consequently, with the apparatus under test replaced by a *coaxial strap*, a balance indication on the phase and transmission difference indicators may be obtained which shifts less than 0.1 degree in phase and 0.02 db in transmission when the master oscillator frequency is varied over its entire band.

Because of the frequency independence of the system zero and the automatic frequency control of the slave oscillator, the master oscillator may be swept through the entire frequency band for rapid appraisal of the network performance by observation of the phase and transmission difference indicators.

The component chassis of the set are mounted in a specially designed console, shown in Fig. 4, which places all controls within easy reach of the operator. This console houses as much apparatus as three 6-foot relay racks within a floor space equal to that occupied by a 5-foot laboratory bench. Though not visible, a full bay of apparatus is mounted behind the central meter panel. Easily movable partitions and covers permit accessibility to all units, thus expediting maintenance.

Some of the significant design considerations are discussed separately under the following headings:

(1) Master Oscillator
(2) Slave Oscillator
(3) Modulators
(4) Phase and Transmission Detector
(5) Phase-shifter

Master Oscillator

As indicated in Fig. 2 and Fig. 5C, the master oscillator is of the heterodyne type. It employs 15,000 and 11,400–14,950 kc local oscillators. A high degree of frequency stability has been achieved through special oscillator circuit design. A motion picture film type scale, 300 inches in length, calibrated every 10 kc, and further subdivided every 2 kc, covers the entire frequency range 50–3600 kc without band-switching. A 0–10 kc interpolation dial with 100-cycle divisions, which operates on the fixed local oscillator frequency, is used to interpolate between adjacent 2 kc graduations on the main film scale. By oscilloscopic comparison with a 10 kc standard of frequency, the oscillator can be set within 50 cycles of any desired fre-

quency in its band. An A.V.C. circuit maintains the output power at six db above one milliwatt.

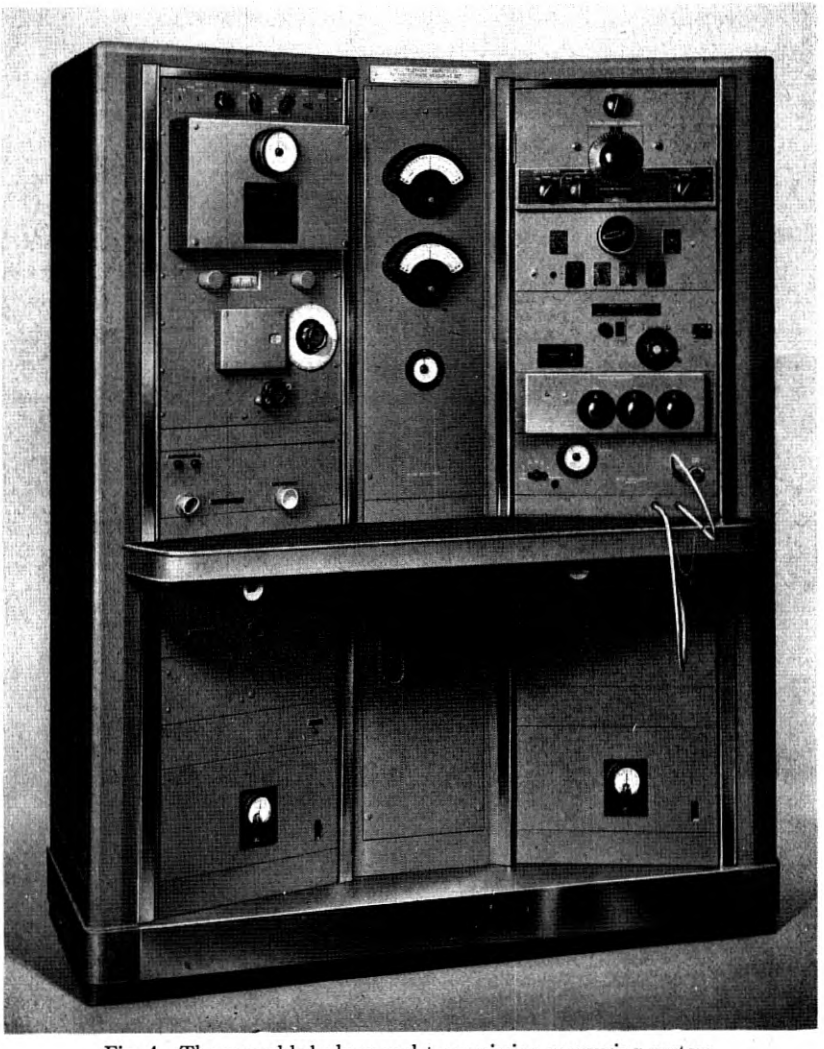


Fig. 4—The assembled phase and transmission measuring system.

Slave Oscillator

To make possible the operation of the measuring attenuator, phase-shifter, and phase and transmission difference detectors at constant frequency, the inputs to the "*S*" and "*X*" channel modulators from the master and slave

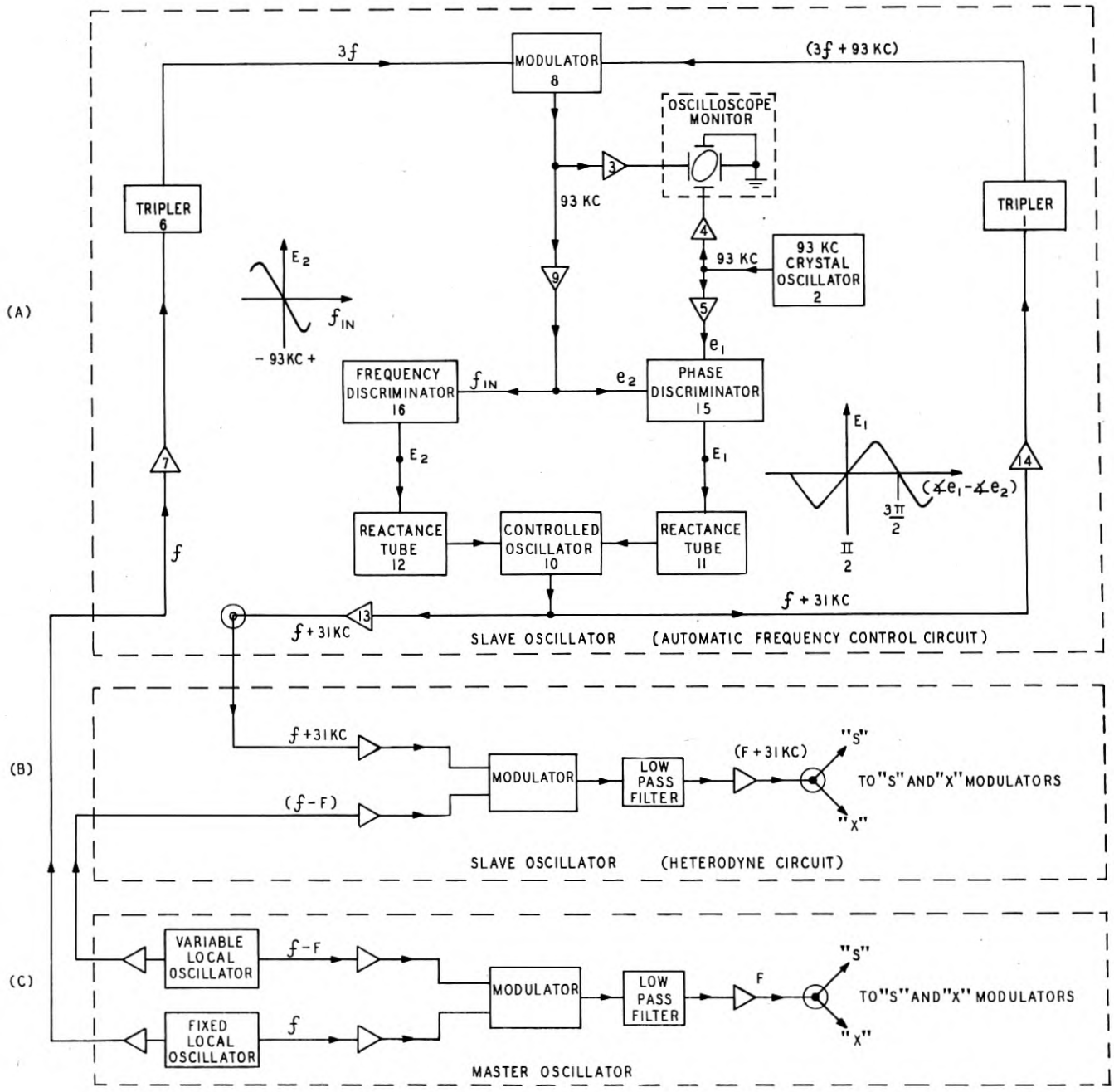


Fig. 5—Master and slave oscillators.

oscillators must always differ in frequency by a constant amount. This difference is maintained at 31 kc by the control of the master oscillator over the slave oscillator frequency.

Very briefly, the scheme consists in applying the fixed local oscillator frequency, f, of the master oscillator, to an automatic frequency control circuit which produces an output frequency $f + 31$ kc. $f + 31$ kc is then modulated with variable local oscillator frequency, $f - F$, of the master oscillator, resulting in an output of frequency $F + 31$ kc. Frequency F, formed by modulation of f and $f - F$, is the master oscillator frequency.

In the automatic control circuit, frequency f is compared with that of a controlled oscillator, by detecting their difference in a modulator. The nature of the control is such, that any deviation of this difference from 31 kc causes the frequency of the controlled oscillator to change in the direction which eliminates the deviation. While it is simpler to compare f and the controlled oscillator frequency directly, in the slave oscillator the comparison is made between the outputs of tripler circuits energized from the latter frequencies. In this way more complete isolation is realized between f and the controlled frequency than would be afforded with only buffer amplifiers. Because of the tripling, it follows that the oscillator must be controlled according to the departure of the difference between the tripler circuit frequencies from 93 kc. This, however, has the advantage of avoiding the generation of 31 kc anywhere in the automatic frequency control circuit, which could, by spurious modulation, cause the $f + 31$ kc output to be contaminated with small traces of frequency f. The necessity for exceptional purity of $f + 31$ kc output arises in the measurement of high losses where minute amounts of F at the $F + 31$ kc input to "S" and "X" modulators may produce appreciable error.

Owing to phase tracking requirements between "S" and "X" intermediate frequency channels, and to the frequency dependence of the phase-shifter calibration, it is necessary to maintain the intermediate frequency as closely as possible to the precise value, 31,000 cycles. The permissible deviation from the correct value has been limited to ± 1 cycle. This precise control is maintained in the presence of 10 kc changes in f, which may occur when the setting of the 0–10 kc interpolation dial of the master oscillator is varied in the course of measurement.

Figures 5A and 5B illustrate the automatic frequency control and heterodyne circuits of the slave oscillator. The frequency of oscillator 10 in Fig. 5A is controlled by the reactance tubes 11 and 12. Reactance tube 12 is actuated by direct voltage from frequency discriminator 16, so that it controls oscillator 10 according to frequency error. Frequency error is the difference between the input frequency to discriminator 16 from amplifier 9, and 93 kc, the frequency of zero voltage output from the dis-

criminator. The voltage from phase discriminator 15 controls oscillator 10 according to the difference of phase between the input from stage 9, and an input of reference phase from amplifier 5. This difference of phase is proportional to the time integral of the frequency error. The gross effect, therefore, is to control the oscillator 10 according to the controller law, *proportional to frequency error + time integral of frequency error*, or, in the terminology of feedback regulators, *proportional + integral control*[1]. When in equilibrium, the system operates with a static phase difference between the phase discriminator inputs, a condition which can exist only when these inputs are of equal frequency. The system is thus endowed with the property *zero frequency error*, and the frequency at the output of modulator 8 is maintained in exact equality with crystal oscillator 2 frequency. Consequently the intermediate frequency difference between input, f, and controlled oscillator 10 is held precisely at the value 31,000 cycles.

Automatic frequency control circuits of the phase sensitive type have been previously described[2,3,4].

The system of combined phase and frequency sensitive control in the slave oscillator is superior to those which use only phase or frequency sensitive control. In a control circuit which uses only a phase discriminator and associated reactance tube, the controlled oscillator may lock-in at either of two sideband frequencies. These are $f + 31$ kc, and $f - 31$ kc. Operation is at upper sideband when control stabilizes on the positive slope of the phase discriminator output voltage curve in Fig. 5A, and at lower sideband if control is along the negative slope. Thus an ambiguity of sideband exists, though the attribute of *zero frequency error* is retained. When only a frequency discriminator and reactance tube are used, lock-in is possible at only one of the two sideband frequencies, determined by the poling of the frequency discriminator output voltage. A frequency error, however, is present.

The combination in Fig. 5 of the two systems operating jointly utilizes the phase sensitive discriminator to insure close control of oscillator frequency, and the polarizing property of the coarser frequency discriminator to eliminate the possibility of synchronization at the undesired sideband.

The joint system of phase and frequency sensitive automatic control has the further virtue of possessing a far greater degree of stability than is obtainable with the phase discriminator loop acting alone.

In the heterodyne circuit of Fig. 5B, $f + 31$ kc from the automatic frequency control circuit is modulated with $f - F$, the variable local oscillator frequency of the master oscillator. The frequency at the output of the heterodyne circuit is $F + 31$ kc, and this is modulated with frequency F in the "S" and "X" modulators to produce the constant intermediate frequency, 31 kc, in the measurement portion of the set.

The Modulators

The difficulties of precise measurement over a wide frequency band essentially are concentrated in the modulator. With the precision to which measurement must be made, effects ordinarily of small concern assume importance. The following discussion is valid for any modulator, though the specific example of the vacuum tube is used.

It is the function of the modulator to convert linearly changes in amplitude and phase from the input frequency F to the output frequency 31 kc. The linear range of conversion is limited by overload at the high-level limit and by noise at the low-level limit.

Let the input x to a modulator consist of two frequencies F_1 and F_2. In the ideal square law modulator[5] perfect linearity results between changes in the input signal F_1 and the output signal $F_2 - F_1$. The output filter rejects all frequencies but $F_2 - F_1$.

In actual tubes the plate current is

$$I_p = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \cdots. \tag{1}$$

The effect of the term a_4x^4 and higher even-order terms is to contribute output currents of frequency $F_2 - F_1$ which do not vary linearly with the input.[5] In addition to this the effect of remodulation in plate, screen and suppressor circuits is that the coefficients a_2, a_4 etc. are not independent of the input x and so contribute to the distortion. Further, in presence of modulation of higher than second order, the d-c. term in even-order modulation will cause distortion if cathode bias is used. Removal of d-c. degeneration using fixed bias eliminates this effect.

The high-level limit may be defined as the signal value for which the total error due to overload equals the desired limits of modulator performance.

The lowest input level into the modulator which may be tolerated, and hence the lower limit of loss which can be measured, is determined by the effective signal-to-noise ratio at the modulator output. If no amplification exists preceding the modulator the input grid noise is usually limiting. The signal-to-noise ratio of the signal F_1 and a noise band centered on F_1 is unaffected by the modulation process as only the modulated portion of the noise band passes through the output filter. Yet for a noise band centered on the intermediate frequency $F_2 - F_1$ for which the output filter is transparent the modulator acts as a straight amplifier; hence the effective signal-to-noise ratio is degraded approximately by the ratio of amplifier gain to conversion gain of the modulator.

The low-level limit may be defined as the signal value for which the error due to noise equals the desired modulator performance limit. For example for a noise error of 0.01 db, a signal-to-noise ratio of 1000 to 1 or 60 db is required.

To obtain maximum signal-to-noise ratio, a tube must be chosen to have the lowest product of *noise multiplied by the ratio of amplifier to conversion gain*, the latter requirement being in conflict to overload requirements. When inputs below the low-level limit are to be utilized a preamplifier ahead of the modulator tube is required. This amplifier also contains a noiseband centered on $F_2 - F_1$. If the amplifier is selective and rejects this noise band or if an $F_2 - F_1$ rejection filter is inserted ahead of the modulator tube the resultant new low-level limit is determined by the signal-to-noise ratio of the preamplifier at the signal frequency F_1 only.

Dynamic range is defined as the useful range of a modulator limited by the high-level limit on one end and by the low-level limit on the other. The dynamic range of a number of pentodes was determined. It was found experimentally that differences in dynamic range between pentodes of different power ratings, such as 6AK5, 6AC7, 6AG7, 6L6, 829B, are small. A dynamic range of 30–36 db can be realized with a 6AK5 for a .01 db linearity requirement. The 6AK5 was the most suitable tube of those investigated considering all other requirements of the circuit such as band width, available signal levels, etc.

Buffer amplifiers are required ahead of the modulator tube to prevent crosstalk between measuring and reference modulator through common paths. These buffer amplifiers are of conventional video amplifier design, with phase and gain characteristics closely controlled to the order of 0.01 db and 0.1 degree.

The Phase and Transmission Detector

In the null type of phase measurement an initial *circuit zero* is made. When the circuit is rebalanced with the *apparatus under test* inserted, the phase detector must be able to verify that the same phase relationship has been reestablished as existed when the initial circuit zero balance was made. Bridge circuits yield high sensitivity and a high degree of independence of input voltage amplitudes. In Fig. 6 a four-arm resistance phase bridge is shown, which has two inputs E_1 and E_2, and two outputs E_S and E_D corresponding to the vectorial sum and difference of the input voltages E_1 and E_2.

As derived in the appendix, for the equal arm bridge, the amplitudes of the voltages E_S and E_D are equal for phase angles of $\varphi = \pi/2 + n\pi$, where n is any integer, regardless of the amplitudes of E_1 and E_2. Thus equality of $|E_S|$ and $|E_D|$ is convenient to define the circuit phase zero. Equality of $|E_S|$ and $|E_D|$ by itself does not distinguish between 90° and 270° phase shifts. This ambiguity can be resolved with a detection circuit which responds to both the amount and the sign of the difference $|E_S| - |E_D|$ and by making provision for the introduction of a small increase $\Delta\varphi$

in phase angle φ of a known direction. From equation (11) in the appendix for equal amplitudes E_1 and E_2

$$(2) \qquad |E_S| - |E_D| = |E_1|[\cos(\varphi/2) - \sin(\varphi/2)]$$

Substituting (a) $\varphi = \pi/2 + \Delta\varphi$ and (b) $\varphi = 3\pi/2 + \Delta\varphi$ into equation (2) it is evident that the sign of $|E_S| - |E_D|$ in substitution (b) is the reverse of substitution (a).

$|E_S|$ and $|E_D|$ are equal every 180° only if all arms of the phase bridge are exactly equal. If the arms are unequal, balance exists for all angles $\varphi = \pi/2 + 2n\pi + \Delta\theta_1$ and $\varphi = \pi/2 + (2n - 1)\pi + \Delta\theta_2$ where $\Delta\theta$ may be called departure angle.

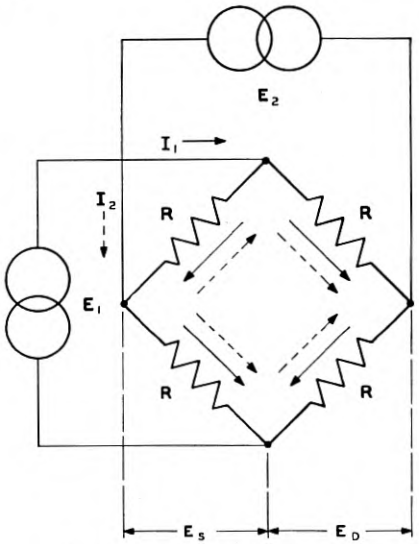


Fig. 6—The phase bridge.

The phase detector can also be used as a deflection bridge. If the phase indicating meter is calibrated according to equation (2), phase angle departures from $\pi/2 + n\pi$ may be read directly on the indicator when $|E_1| = |E_2|$ and the scale factor is adjusted for the amplitude of E_2 which is maintained constant by the overall automatic volume control circuit. Equation (2) is almost a linear function and is plotted in Fig. 7.

In using the deflection on the indicator to measure phase shift, an error $\Delta\psi$ is incurred if $|E_1| \neq |E_2|$. The maximum permissible ratio of $|E_2|/|E_1|$ for a given error $\Delta\psi$ is given by

$$(3) \qquad |E_2|/|E_1| = \cos\varphi/\cos(\varphi + \Delta\psi) + \sqrt{[\cos\varphi/\cos(\varphi + \Delta\psi)]^2 - 1}$$

Equation (3) is derived in the appendix and plotted for several values $\Delta\psi$ in Fig. 8.

The phase bridge essentially converts the measurement of phase into the measurement of voltage difference. Vacuum tube diodes are used as dif-

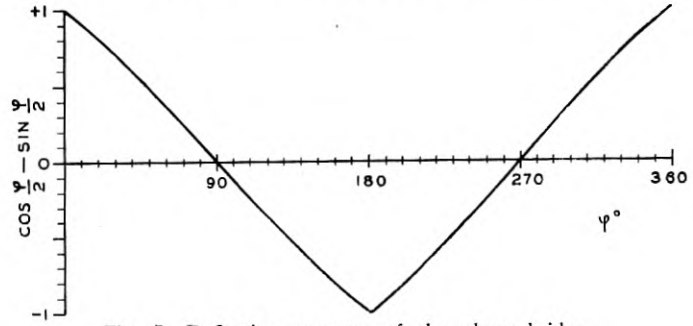


Fig. 7—Deflection response of the phase bridge.

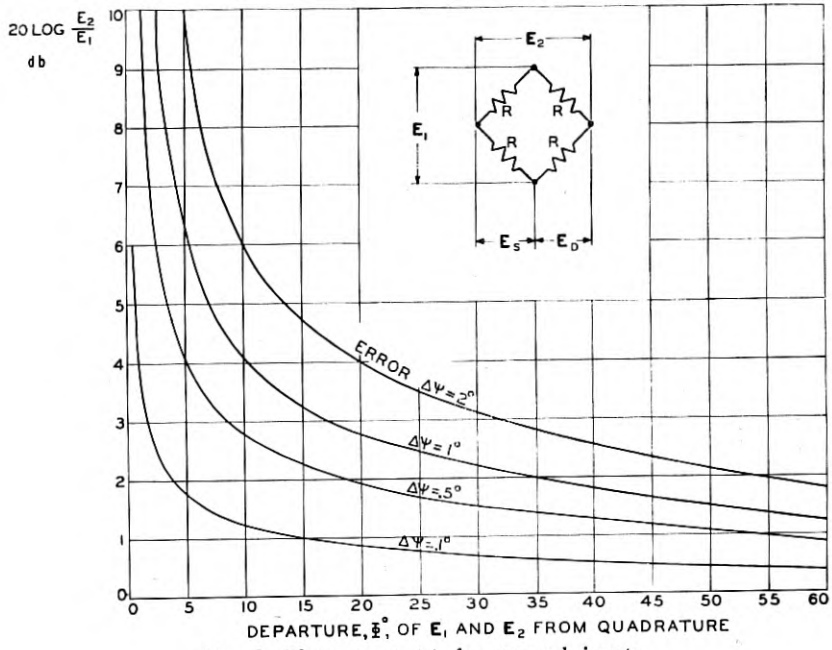


Fig. 8—Phase error $\Delta\psi$ for unequal inputs.

ferential rectifiers with a high resistance load consisting of hermetically sealed carbon deposited resistors closely matched for value and temperature coefficient and specially mounted to minimize temperature differentials. The differential output of the rectifiers is amplified in a feedback stabilized

d-c. amplifier which has adjustable gain to adjust scale factors on the indicator. The phase detection circuit is energized from the output of the phase bridge and the almost identical transmission detection circuit is energized from the inputs to the phase bridge. Thus both phase and transmission are measured simultaneously.

Each indicator (Fig. 9) has three scales, *fine* (−5° to +5°; −1 db to +1 db), *coarse* (−90° to +90°; −10 db to +3 db), and *null balance.* The *fine* and *coarse* scales are linear while the *null balance* scale has maximum

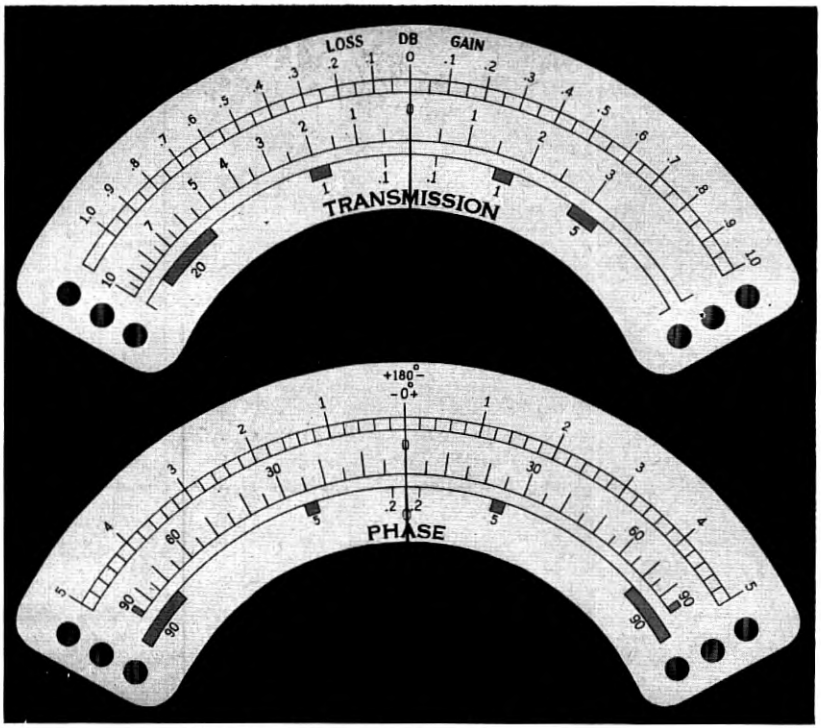


Fig. 9—Phase and transmission indicators.

sensitivity in the neighborhood of the center zero and greatly reduced sensitivity at each end. Varistor shunts across the indicators compress the null scale for large deflections. Colored pilot lights at the ends of the indicator scales, operated by the scale switch, indicate the scale in use directly.

The Phase Shifter

The phase-shifter employs a four-quadrant variable sine condenser. It has two linearly subdivided scales—*coarse* 0-360° on a cylinder and *fine*

0–10° on a dial. The fine dial is connected through reduction gearing to the shaft of the sine condenser. The construction of a phase-shifter which has a sufficiently linear correspondence of electrical phase-shift and mechanical displacement of a shaft is not practical. Instead a movable index for the fine scale permits correction of the deviation from linearity. The index position is controlled by a corrector which is fastened to the condenser shaft. As the corrector is rigidly associated with the sine condenser position and not with the scales this permits shifting the linear scales inde-

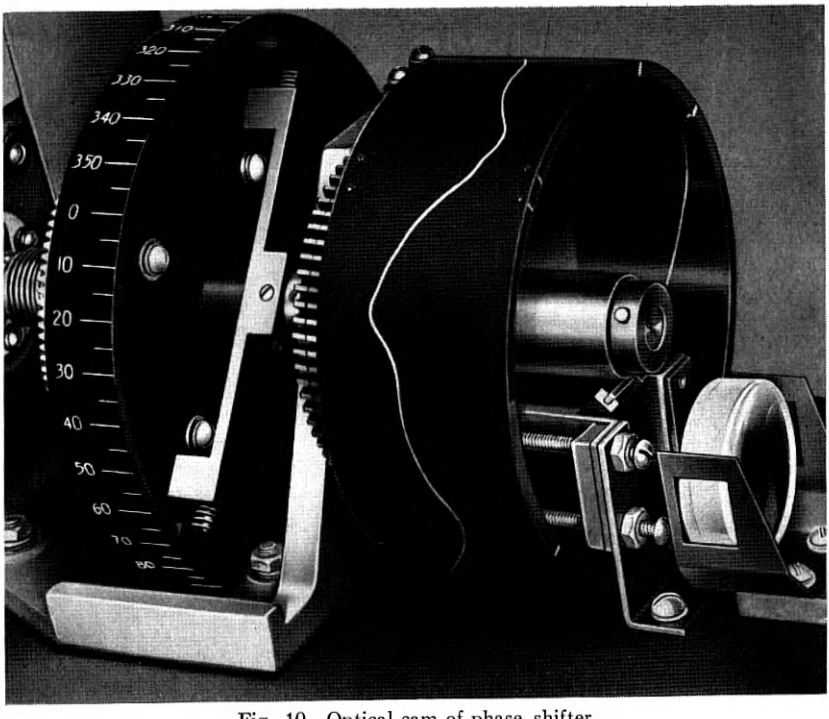


Fig. 10—Optical cam of phase shifter.

pendently without affecting the correction. The correction curve (Fig. 10) is printed on a photographic negative which is placed on a transparent lucite drum and projected optically as an index (Fig. 11) adjacent to the fine dial of the phase-shifter. The calibration curve is obtained by marking the correction at each calibrating point on a piece of cellulose acetate placed on the lucite drum. The correction point is projected on the screen adjacent to the fine scale during the calibration and problems arising from divergence or misalignment of the light beam are thus avoided. Since the index is projected upon a surface coplanar with the dial, no parallax exists.

The phase-shifter's deviation from linearity is sufficiently small that no correction is needed on the coarse scale.

Both dials can be moved with respect to their shafts by releasing friction clutches. Thus the measuring system phase zero can be established by an initial balance of the phase-shifter and restoration of the coarse and fine scales to zero. This is only possible because the scales are linear. Thus no zero readings have to be subtracted from the measurement readings and the need for a separate zero setting phase-shifter is avoided.

The phase-shifter is calibrated by a method of substitution. As discussed previously the phase-indicator indicates balance uniquely in mul-

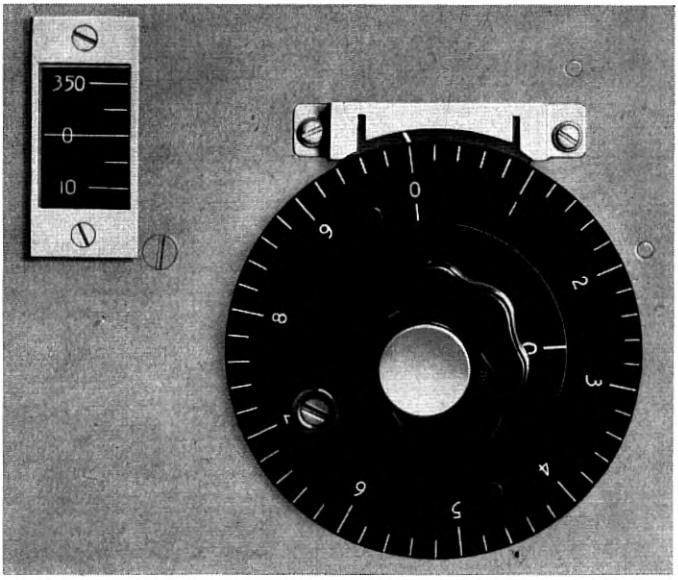


Fig. 11—Phase shifter scales and projected index.

tiples of 360° phase-shift. Exact sub-multiples of 360° can be generated and used to calibrate the phase-shifter.

For example (Fig. 12), to establish an exact 180° phase-shift the standard phase-shifter is set to an arbitrary starting point. With the switches in the position shown a null is obtained on the indicator by adjusting the auxiliary phase-shifter. The network of nominally 180° phase-shift is inserted and a null obtained on the indicator by adjusting the standard phase-shifter. Now the network is removed and the null reestablished by adjustment of the auxiliary phase-shifter. The 180° network again is inserted and a null obtained by adjustment of the standard phase-shifter, which now has been moved through twice the actual phase-shift of the

nominal 180° network. The amount the standard phase-shifter failed to return to the original starting point indicates the residual error of the 180° network. The 180° network is adjusted accordingly and the procedure repeated until an error is no longer discernible. Thus the 180° point on the phase-shifter scale can be determined. In similar fashion, by combination of the 180°, 90° and 60° networks, calibration points in multiples of 30° are obtained. The equivalent of a 10° network is obtained by use of the ±5° scale on the indicator and scale factor adjustment. Interpolation to 1° is then made using the scale divisions on the indicator. Calibration to an absolute accuracy of ±0.1° was found adequate for use in the measuring system. Much higher accuracy could be obtained if the need arose. There appears to be no inherent frequency limitation in this calibration method.

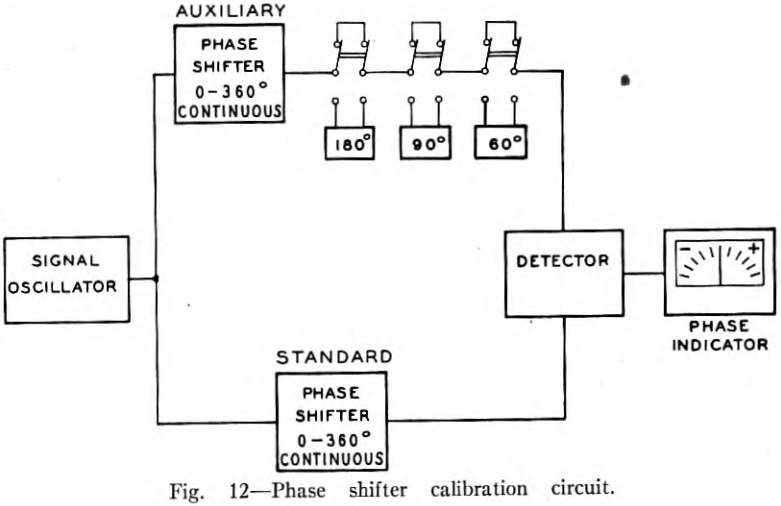


Fig. 12—Phase shifter calibration circuit.

Conclusion

The design effort has been directed toward achieving laboratory precision in measurement and at the same time maintaining the speed necessary for production testing of transmission networks.

The measurement of phase-shift is unambiguous with respect to quadrants and the measurements of insertion phase-shift and loss or gain are independent of each other. The entire frequency range is covered without band switching by use of a heterodyne signal oscillator and the system zero is independent of measurement frequency. Detector tuning is eliminated through the use of frequency conversion, employing a beating oscillator automatically controlled in frequency by the signal oscillator. Phase-shift and transmission may be read directly, without auxiliary computations,

from the dials of the phase-shifter and attenuator or from the scales of the indicators.

ACKNOWLEDGMENT

Acknowledgment is due members of the groups supervised by Mr. E. P. Felch (electrical) and Mr. W. J. Means (mechanical) for contributions to the design.

APPENDIX

THE PHASE DISCRIMINATOR BRIDGE

The general phase relationship of the discriminator is shown in Fig. 13a.

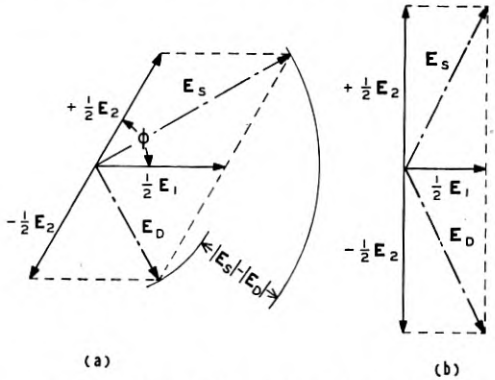


Fig. 13—Phase bridge vector relationships.

Using complex vectorial notation,

$$E_S = (1/2)(E_1 e^{j\varphi_1} + E_2 e^{j\varphi_2}) \tag{4}$$

$$E_D = (1/2)(E_1 e^{j\varphi_1} - E_2 e^{j\varphi_2}) \tag{5}$$

Hence for $\varphi = \pi/2 + n\pi$ where $\varphi = \varphi_1 - \varphi_2$ and n is an integer,

$$|E_S| = |E_D| \tag{6}$$

Stated in words: The amplitudes of E_S and E_D are equal if the relative phase angle φ is equal to 90°, 270°, 450°, etc., independently of the amplitudes of E_1 and E_2 (Fig. 13b).

Inequality of $|E_S|$ and $|E_D|$ can be utilized to measure phase departure from $\varphi = \pi/2 + n\pi$

From (4) and (5),

$$|E_S| = (1/2)\sqrt{|E_1|^2 + 2|E_1E_2|\cos\varphi + |E_2|^2} \tag{7}$$

$$|E_D| = (1/2)\sqrt{|E_1|^2 - 2|E_1E_2|\cos\varphi + |E_2|^2} \tag{8}$$

If the amplitudes $|E_1|$ and $|E_2|$ are equal, then

$$|E_S| = (|E_1|/2)\sqrt{2(1 + \cos\varphi)} = |E_1|\cos(\varphi/2) \tag{9}$$

$$|E_D| = (|E_1|/2)\sqrt{2(1 - \cos\varphi)} = |E_1|\sin(\varphi/2) \tag{10}$$

$$|E_S| - |E_D| = |E_1|[\cos(\varphi/2) - \sin(\varphi/2)] \tag{11}$$

$$|E_S| / |E_D| = \operatorname{cotan}(\varphi/2) \tag{12}$$

When $|E_1| \neq |E_2|$ determination of φ by (11) or (12) is in error by $\Delta\psi$. From (7) and (8)

$$\operatorname{cotan}\frac{\varphi + \Delta\psi}{2} = \sqrt{\frac{|E_1|^2 + 2|E_1E_2|\cos\varphi + |E_2|^2}{|E_1|^2 - 2|E_1E_2|\cos\varphi + |E_2|^2}} \tag{13}$$

Hence

$$\left|\frac{E_2}{E_1}\right|^2 + 2\,\frac{1 + \operatorname{cotan}^2[(\varphi + \Delta\psi)/2]}{1 - \operatorname{cotan}^2[(\varphi + \Delta\psi)/2]}\,\frac{|E_2|}{|E_1|}\cos\varphi + 1 = 0 \tag{14}$$

From trigonometry

$$\frac{1 + \operatorname{cotan}^2[(\varphi + \Delta\psi)/2]}{1 - \operatorname{cotan}^2[(\varphi + \Delta\psi)/2]} = -\frac{1}{\cos(\varphi + \Delta\psi)} \tag{15}$$

$$|E_2|/|E_1| = \cos\varphi/\cos(\varphi + \Delta\psi) + \sqrt{[\cos\varphi/\cos(\varphi + \Delta\psi)]^2 - 1} \tag{16}$$

References

1. "Electronic Instruments" (book), *M.I.T. Radiation Laboratory Series*, Volume 21. McGraw Hill Book Company, First Edition, Page 342.
2. "Automatic Frequency Control Systems," A. F. Pomeroy, *United States Patent* 2,288,025.
3. "The Carrier Stabilization of Frequency Modulated Transmitters," *Brown Boveri Review*, Vol. 33, August 1946, Page 193.
4. "Frequency Modulated Broadcast Transmitters for 88–108 Megacycles," Leonard Everett, *Electrical Communication*, Vol. 24, March 1947, Pages 84–86.
5. "Transconductance as a Criterion of Electron Tube Performance," T. Slonczewski: *Bell Sys. Tech. Jour.*, Vol. XVIII, April 1949, Pages 315–328.

Physical Principles Involved in Transistor Action*

By J. BARDEEN and W. H. BRATTAIN

The transitor in the form described herein consists of two-point contact electrodes, called emitter and collector, placed in close proximity on the upper face of a small block of germanium. The base electrode, the third element of the triode, is a large area low resistance contact on the lower face. Each point contact has characteristics similar to those of the high-back-voltage rectifier. When suitable d-c. bias potentials are applied, the device may be used to amplify a-c. signals. A signal introduced between the emitter and base appears in amplified form between collector and base. The emitter is biased in the positive direction, which is that of easy flow. A larger negative or reverse voltage is applied to the collector. Transistor action depends on the fact that electrons in semi-conductors can carry current in two different ways: by excess or conduction electrons and by defect "electrons" or holes. The germanium used is n-type, i.e. the carriers are conduction electrons. Current from the emitter is composed in large part of holes, i.e. of carriers of opposite sign to those normally in excess in the body of the block. The holes are attracted by the field of the collector current, so that a large part of the emitter current, introduced at low impedance, flows into the collector circuit and through a high-impedance load. There is a voltage gain and a power gain of an input signal. There may be current amplification as well.

The influence of the emitter current, I_e, on collector current, I_c, is expressed in terms of a current multiplication factor, α, which gives the rate of change of I_c with respect to I_e at constant collector voltage. Values of α in typical units range from about 1 to 3. It is shown in a general way how α depends on bias voltages, frequency, temperature, and electrode spacing. There is an influence of collector current on emitter current in the nature of a positive feedback which, under some operating conditions, may lead to instability.

The way the concentrations and mobilities of electrons and holes in germanium depend on impurities and on temperature is described briefly. The theory of germanium point contact rectifiers is discussed in terms of the Mott-Schottky theory. The barrier layer is such as to raise the levels of the filled band to a position close to the Fermi level at the surface, giving an inversion layer of p-type or defect conductivity. There is considerable evidence that the barrier layer is intrinsic and occurs at the free surface, independent of a metal contact. Potential probe tests on some surfaces indicate considerable surface conductivity which is attributed to the p-type layer. All surfaces tested show an excess conductivity in the vicinity of the point contact which increases with forward current and is attributed to a flow of holes into the body of the germanium, the space charge of the holes being compensated by electrons. It is shown why such a flow is to be expected for the type of barrier layer which exists in germanium, and that this flow accounts for the large currents observed in the forward direction. In the transistor, holes may flow from the emitter to the collector either in the surface layer or through the body of the germanium. Estimates are made of the field produced by the collector current, of the transit time for holes, of the space charge produced by holes flowing into the collector, and of the feedback resistance which gives the influence of collector current on emitter current. These calculations confirm the general picture given of transistor action.

I—Introduction

THE transistor, a semi-conductor triode which in its present form uses a small block of germanium as the basic element, has been described briefly

* This paper appears also in the *Physical Review*, April 15, 1949.

in the Letters to the Editor columns of the Physical Review.[1] Accompanying this letter were two further communications on related subjects.[2, 3] Since these initial publications a number of talks describing the characteristics of the device and the theory of its operation have been given by the authors and by other members of the Bell Telephone Laboratories staff.[4] Several articles have appeared in the technical literature.[5] We plan to give here an outline of the history of the development, to give some further data on the characteristics and to discuss the physical principles involved. Included is a review of the nature of electrical conduction in germanium and of the theory of the germanium point-contact rectifier.

A schematic diagram of one form of transistor is shown in Fig. 1. Two point contacts, similar to those used in point-contact rectifiers, are placed in close proximity (−.005–.025 cm) on the upper surface of a small block of germanium. One of these, biased in the forward direction, is called the emitter. The second, biased in the reverse direction, is called the collector. A large area low resistance contact on the lower surface, called the base electrode, is the third element of the triode. A physical embodiment of the device, as designed in large part by W. G. Pfann, is shown in Fig. 2. The transistor can be used for many functions now performed by vacuum tubes.

During the war, a large amount of research on the properties of germanium and silicon was carried out by a number of university, government, and industrial laboratories in connection with the development of point contact rectifiers for radar. This work is summarized in the book of Torrey and Whitmer.[6] The properties of germanium as a semi-conductor and as a rectifier have been investigated by a group working under the direction of K. Lark-Horovitz at Purdue University. Work at the Bell Telephone Laboratories[7] was initiated by R. S. Ohl before the war in connection with the development of silicon rectifiers for use as detectors at microwave frequencies. Research and development on both germanium and silicon rectifiers during and since the war has been done in large part by a group under J. H. Scaff. The background of information obtained in these various investigations has been invaluable.

The general research program leading to the transistor was initiated and directed by W. Shockley. Work on germanium and silicon was emphasized because they are simpler to understand than most other semi-conductors. One of the investigations undertaken was the study of the modulation of conductance of a thin film of semi-conductor by an electric field applied by an electrode insulated from the film.[3] If, for example, the film is made one plate of a parallel plate condenser, a charge is induced on the surface. If the individual charges which make up the induced charge are mobile, the conductance of the film will depend on the voltage applied to the condenser.

The first experiments performed to measure this effect indicated that most of the induced charge was not mobile. This result, taken along with other unexplained phenomena such as the small contact potential difference between n- and p- type silicon[8] and the independence of the rectifying properties of the point contact rectifier on the work function of the metal point, led one of the authors to an explanation in terms of surface states.[9] This work led to the concept that space charge barrier layers may be present at the free surfaces of semi-conductors such as germanium and silicon, independent of a metal contact. Two experiments immediately suggested were to measure the dependence of contact potential on impurity concentration[10] and to measure the change of contact potential on illuminating the surface with light.[11] Both of these experiments were successful and confirmed the theory. It was while studying the latter effect with a silicon surface immersed in a liquid that it was found that the density of surface charges and the field in the space charge region could be varied by applying a potential across an electrolyte in contact with the silicon surface.[12] While studying the effect of field applied by an electrolyte on the current voltage characteristic of a high-back-voltage germanium rectifier, the authors were led to the concept that a portion of the current was being carried by holes flowing near the surface. Upon replacing the electrolyte with a metal contact transistor action was discovered.

The germanium used in the transistor is an n-type or excess semi-conductor with a resistivity of the order of 10-ohm cm, and is the same as the material used in high-back-voltage germanium rectifiers.[13] All of the material we have used was prepared by J. C. Scaff and H. C. Theuerer of the metallurgical group of the Laboratories.

While different metals may be used for the contact points, most work has been done with phosphor bronze points. The spring contacts are made with wire from .002 to .005″ in diameter. The ends are cut in the form of a wedge so that the two contacts can be placed close together. The actual contact area is probably no more than about 10^{-6} cm^2.

The treatment of the germanium surface is similar to that used in making high-back-voltage rectifiers.[14] The surface is ground flat and then etched. In some cases special additional treatments such as anodizing the surface or oxidation at 500°C have been used. The oxide films formed in these processes wash off easily and contact is made to the germanium surface.

The circuit of Fig. 1 shows how the transistor may be used to amplify a small a-c. signal. The emitter is biased in the forward (positive) direction so that a small d-c. current, of the order of 1 ma, flows into the germanium block. The collector is biased in the reverse (negative) direction with a higher voltage so that a d-c. current of a few milliamperes flows out through the collector point and through the load circuit. It is found

that the current in the collector circuit is sensitive to and may be controlled by changes of current from the emitter. In fact, when the emitter current is varied by changing the emitter voltage, keeping the collector voltage constant, the change in collector current may be larger than the change in emitter current. As the emitter is biased in the direction of easy flow, a small a-c. voltage, and thus a small power input, is sufficient to vary the emitter current. The collector is biased in the direction of high resistance and may be matched to a high resistance load. The a-c. voltage and power in the load circuit are much larger than those in the input. An overall power gain of a factor of 100 (or 20 db) can be obtained in favorable cases.

Terminal characteristics of an experimental transistor[15] are illustrated in Fig. 3, which shows how the current-voltage characteristic of the collector is changed by the current flowing from the emitter. Transistor characteristics, and the way they change with separation between the points, with temperature, and with frequency, are discussed in Section II.

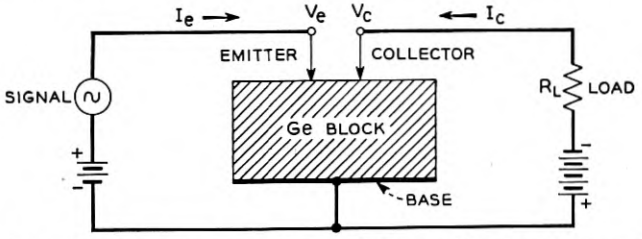


Fig. 1—Schematic of transistor showing circuit for amplification of an a-c. signal and the conventional directions for current flow. Normally I_e and V_e are positive, I_c and V_c negative.

The explanation of the action of the transistor depends on the nature of the current flowing from the emitter. It is well known that in semi-conductors there are two ways by which the electrons can carry electricity which differ in the signs of the effective mobile charges.[16] The negative carriers are excess electrons which are free to move and are denoted by the term conduction electrons or simply electrons. They have energies in the conduction band of the crystal. The positive carriers are missing or defect "electrons" and are denoted by the term "holes". They represent unoccupied energy states in the uppermost normally filled band of the crystal. The conductivity is called n- or p-type depending on whether the mobile charges normally in excess in the material under equilibrium conditions are electrons (negative carriers) or holes (positive carriers). The germanium used in the transistor is n-type with about 5×10^{14} conduction electrons per c.c.; or about one electron per 10^8 atoms. Transistor action depends on the fact that the current from the emitter is composed in

large part of *holes*; that is of carriers of opposite sign to those normally in excess in the body of the semi-conductor.

The collector is biased in the reverse, or negative direction. Current flowing in the germanium toward the collector point provides an electric

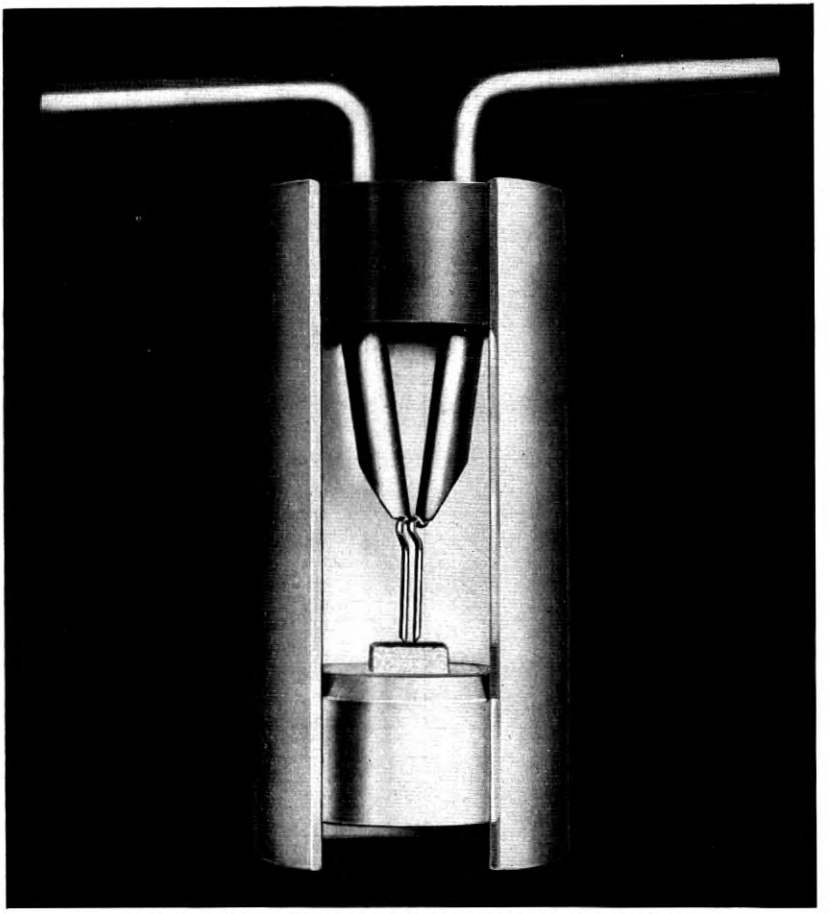


Fig. 2—Microphotograph of a cutaway model of a transistor

field which is in such a direction as to attract the holes flowing from the emitter. When the emitter and collector are placed in close proximity, a large part of the hole current from the emitter will flow to the collector and into the collector circuit. The nature of the collector contact is such as to provide a high resistance barrier to the flow of electrons from the metal to the semi-conductor, but there is little impediment to the flow of holes into

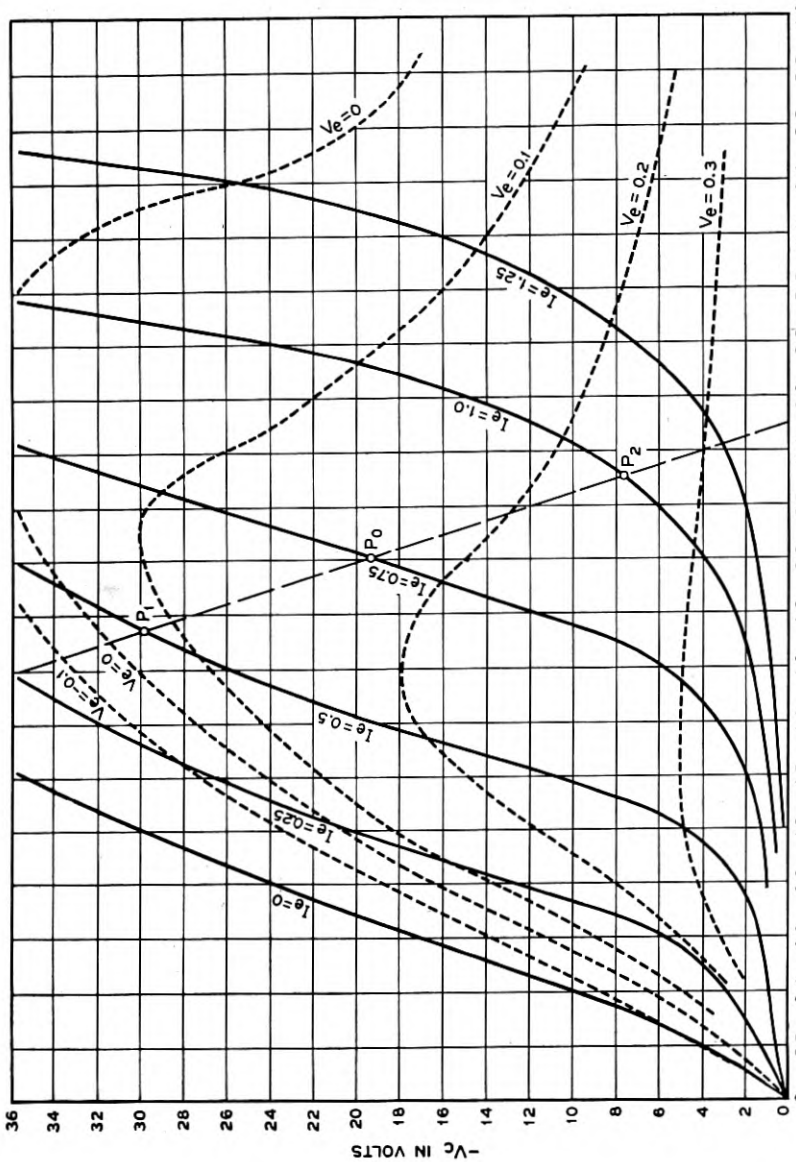


Fig. 3—Characteristics of an experimental transistor.[15] The conventional directions for current and voltage are as in Fig. 1.

the contact. This theory explains how the change in collector current might be as large as but not how it can be larger than the change in emitter current. The fact that the collector current may actually change more than the emitter current is believed to result from an alteration of the space charge in the barrier layer at the collector by the hole current flowing into the junction. The increase in density of space charge and in field strength makes it easier for electrons to flow out from the collector, so that there is an increase in electron current. It is better to think of the hole current from the emitter as modifying the current-voltage characteristic of the collector, rather than as simply adding to the current flowing to the collector.

In Section III we discuss the nature of the conductivity in germanium, and in Section IV the theory of the current-voltage characteristic of a germanium-point contact. In the latter section we attempt to show why the emitter current is composed of carriers of opposite sign to those normally in excess in the body of germanium. Section V is concerned with some aspects of the theory of transistor action. A complete quantitative theory is not yet available.

There is evidence that the rectifying barrier in germanium is internal and occurs at the free surface, independent of the metal contact.[9, 17] The barrier contains what Schottky and Spenke[18] call an inversion region; that is a change of conductivity type. The outermost part of the barrier next to the surface is p-type. The p-type region is very thin, of the order of 10^{-5} cm in thickness. An important question is whether there is a sufficient density of holes in this region to provide appreciable lateral conductivity along the surface. Some evidence bearing on this point is described below.

Transistor action was first discovered on a germanium surface which was subjected to an anodic oxidation treatment in a glycol borate solution after it had been ground and etched in the usual way for diodes. Much of the early work was done on surfaces which were oxidized by heating in air. In both cases the oxide is washed off and plays no direct role. Some of these surfaces were tested for surface conductivity by potential probe tests. Surface conductivities, on a unit area basis, of the order of .0005 to .002 mhos were found.[2] The value of .0005 represents about the lower limit of detection possible by the method used. It is inferred that the observed surface conductivity is that of the p-type layer, although there has been no direct proof of this. In later work it was found that the oxidation treatment is not essential for transistor action. Good transistors can be made with surfaces prepared in the usual way for high-back-voltage rectifiers provided that the collector point is electrically formed. Such surfaces exhibit no measurable surface conductivity.

One question that may be asked is whether the holes flow from the emitter to the collector mainly in the surface layer or whether they flow

through the body of the germanium. The early experiments suggested flow along the surface. W. Shockley proposed a modified arrangement in which in effect the emitter and collector are on opposite sides of a thin slab, so that the holes flow directly across through the semi-conductor. Independently, J. N. Shive made, by grinding and etching, a piece of germanium in the form of a thin flat wedge.[19] Point contacts were placed directly opposite each other on the two opposite faces where the thickness of the wedge was about .01 cm. A third large area contact was made to the base of the wedge. When the two points were connected as emitter and collector, and the collector was electrically formed, transistor action was obtained which was comparable to that found with the original arrangement. There is no doubt that in this case the holes are flowing directly through the n-type germanium from the emitter to the collector. With two points close together on a plane surface holes may flow either through the surface layer or through the body of the semi-conductor.

Still later, at the suggestion of W. Shockley, J. R. Haynes[20] further established that holes flow into the body of the germanium. A block of germanium was made in the form of a thin slab and large area electrodes were placed at the two ends. Emitter and collector electrodes were placed at variable separations on one face of the slab. The field acting between these electrodes could be varied by passing currents along the length of the slab. The collector was biased in the reverse direction so that a small d-c. current was drawn into the collector. A signal introduced at the emitter in the form of a pulse was detected at a slightly later time in the collector circuit. From the way the time interval, of the order of a few microseconds, depends on the field, the mobility and sign of the carriers were determined. It was found that the carriers are positively charged, and that the mobility is the same as that of holes in bulk germanium (1700 cm^2/volt sec).

These experiments clarify the nature of the excess conductivity observed in the forward direction in high-back-voltage germanium rectifiers which has been investigated by R. Bray, K. Lark-Horovitz, and R. N. Smith[21] and by Bray.[22] These authors attributed the excess conductivity to the strong electric field which exists in the vicinity of the point contact. Bray has made direct experimental tests to observe the relation between conductivity and field strength. We believe that the excess conductivity arises from holes injected into the germanium at the contact. Holes are introduced because of the nature of the barrier layer rather than as a direct result of the electric field. This has been demonstrated by an experiment of E. J. Ryder and W. Shockley.[23] A thin slab of germanium was cut in the form of a pie-shaped wedge and electrodes placed at the narrow and wide boundaries of the wedge. When a current is passed between the electrodes,

the field strength is large at the narrow end of the wedge and small near the opposite electrode. An excess conductivity was observed when the narrow end was made positive; none when the wide end was positive. The magnitude of the current flow was the same in both cases. Holes injected at the narrow end lower the resistivity in the region which contributes most to the over-all resistance. When the current is in the opposite direction, any holes injected enter in a region of low field and do not have sufficient life-time to be drawn down to the narrow end and so do not alter the resistance very much. With some surface treatments, the excess conductivity resulting from hole injection may be enhanced by a surface conductivity as discussed above.

The experimental procedure used during the present investigation is of interest. Current voltage characteristics of a given point contact were displayed on a d-c. oscilloscope.[24] The change or modulation of this characteristic produced by a signal impressed on a neighboring electrode or point contact could be easily observed. Since the input impedance of the scope was 10 megohms and the gain of the amplifiers such that the lower limit of sensitivity was of the order of a millivolt, the oscilloscope was also used as a very high impedance voltmeter for probe measurements. Means were included for matching the potential to be measured with an adjustable d-c. potential the value of which could be read on a meter. A micromanipulator designed by W. L. Bond was used to adjust the positions of the contact points.

II—Some Transistor Characteristics

The static characteristics of the transistor are completely specified by four variables which may be taken as the emitter and collector currents, I_e and I_c, and the corresponding voltages, V_e and V_c. As shown in the schematic diagram of Fig. 1, the conventional directions for current flow are taken as positive into the germanium and the terminal voltages are relative to the base electrode. Thus I_e and V_e are normally positive, I_c and V_c negative.

There is a functional relation between the four variables such that if two are specified the other two are determined. Any pair may be taken as the independent variables. As the transistor is essentially a current operated device, it is more in accord with the physics involved to choose the currents rather than the voltages. All fields in the semi-conductor outside of the space charge regions immediately surrounding the point contacts are determined by the currents, and it is the current flowing from the emitter which controls the current voltage characteristic of the collector. The voltages are single-valued functions of the currents but, because of inherent feedback, the currents may be double-valued functions of the voltages.

In reference 1, the characteristics of an experimental transistor were shown by giving the constant voltage contours on a plot in which the independent variables I_e and I_c are plotted along the coordinate axes.

In the following we give further characteristics, and show in a general way how they depend on the spacing between the points, on the temperature, and on the frequency. The data were taken mainly on experimental setups on a laboratory bench, and are not to be taken as necessarily typical of the characteristics of finished units. They do indicate in a general way the type of results which can be obtained. Characteristics of units made in pilot production have been given elsewhere.[5]

The data plotted in reference 1 were taken on a transitor made with phosphor bronze points on a surface which was oxidized and on which potential probe tests gave evidence for considerable surface conductivity. The collector resistance is small in units prepared in this way. In Fig. 3 are shown the characteristics of a unit[15] in which the surface was prepared in a different manner. The surface was ground and etched in the usual way[14], but was not subjected to the oxidation treatment. Phosphor bronze contact points made from 5 mil wire were used. The collector was electrically formed by passing large currents in the reverse direction. This reduced the resistance of the collector in the reverse direction, improving the transistor action. However, it remained considerably higher than that of the collector on the oxidized surface.

While there are many ways of plotting the data, we have chosen to give the collector voltage, V_c, as a function of the collector current, I_c, with the emitter current, I_e, taken as a parameter. This plot shows in a direct manner the influence of the emitter current on the current-voltage characteristic of the collector. The curve corresponding to $I_e = 0$ is just the normal reverse characteristic of the collector as a rectifier. The other curves show how the characteristic shifts to the right, corresponding to larger collector currents, with increase in emitter current. It may be noted that the change in collector current for fixed collector voltage is larger than the change in emitter current. The current amplification factor, α, defined by

$$\alpha = -(\partial I_c/\partial I_e)_{V_c = \text{const.}} \tag{2.1}$$

is between 2 and 3 throughout most of the plot.

The dotted lines on Fig. 3 correspond to constant values of the emitter voltage, V_e. By interpolating between the contours, all four variables corresponding to a given operating point may be obtained. The V_e contours reach a maximum for I_e about 0.7 ma. and have a negative slope beyond. To the left of the maximum, V_e increases with I_e as one follows along a line corresponding to V_c = const. To the right, V_e decreases as

I_e increases, corresponding to a negative input admittance. For given values of V_e and V_c, there are two possible operating points. Thus for $V_e = 0.1$ and $V_c = -20$ one may have $I_e = 0.3$ ma, $I_c = -1.1$ ma or $I_e = 1.0$, $I_c = -2.7$.

The negative resistance and instability result from the effect of the collector current on the emitter current.[1] The collector current lowers the potential of the surface in the vicinity of the emitter and increases the effective bias on the emitter by an equivalent amount. This potential drop is $R_F I_c$, where R_F is a feedback resistance which may depend on the currents flowing. The effective bias on the emitter is then $V_e - R_F I_c$, and we may write

$$I_e = f(V_e - R_F I_c), \tag{2.2}$$

where the function gives the forward characteristic of the emitter point. In some cases R_F is approximately constant over the operating range; in other cases R_F decreases with increasing I_e as the conductivity of the germanium in the vicinity of the points increases with forward current. Increase of I_e by a change of V_e increases the magnitude of I_c, which by the feedback still further increases I_e. Instability may result. Some consequences will be discussed further in connection with the a-c. characteristics.

Also shown on Fig. 3 is a load line corresponding to a battery voltage of -100 in the output circuit and a load, R_L, of 40,000 ohms, the equation of the line being

$$V_c = -100 - 40 \times 10^3 I_c. \tag{2.3}$$

The load is an approximate match to the collector resistance, as given by the slope of the solid lines. If operated between the points P_1 and P_2, the output voltage is 8.0 volts r.m.s. and the output current is 0.20 ma. The corresponding values at the input are 0.07 and 0.18, so that the overall power gain is

$$\text{Gain} \sim 8 \times 0.20/(0.07 \times 0.18) \sim 125, \tag{2.4}$$

which is about 21 db. This is the available gain for a generator with an impedance of 400 ohms, which is an approximate match for the input impedance.

We turn next to the equations for the a-c. characteristics. For small deviations from an operating point, we may write

$$\Delta V_e = R_{11}\,\Delta I_e + R_{12}\,\Delta I_c, \tag{2.5}$$
$$\Delta V_c = R_{12}\,\Delta I_e + R_{22}\,\Delta I_c, \tag{2.6}$$

in which we have taken the currents as the independent variables and the directions of currents and voltages as in Fig. 1. The differentials represent

small changes from the operating point, and may be small a-c. signals. The coefficients are defined by:

$$R_{11} = (\partial V_e/\partial I_e)_{I_c = \text{const.}}, \tag{2.7}$$
$$R_{12} = (\partial V_e/\partial I_c)_{I_e = \text{const.}}, \tag{2.8}$$
$$R_{21} = (\partial V_c/\partial I_e)_{I_c = \text{const.}}, \tag{2.9}$$
$$R_{22} = (\partial V_c/\partial I_c)_{I_e = \text{const.}} \tag{2.10}$$

These coefficients are all positive and have the dimensions of resistances. They are functions of the d-c. bias currents, I_e and I_c, which define the operating point. For $I_e = 0.75$ ma and $I_c = -2$ ma the coefficients of the unit of Fig. 3 have the following approximate values:

$$\begin{aligned} R_{11} &= 800 \text{ ohms}, \\ R_{12} &= 300, \\ R_{21} &= 100{,}000, \\ R_{22} &= 40{,}000. \end{aligned} \tag{2.11}$$

Equation (2.5) gives the emitter characteristic. The coefficient R_{11} is the input resistance for a fixed collector current (open circuit for a-c.). To a close approximation, R_{11} is independent of I_c, and is just the forward resistance of the emitter point when a current I_e is flowing. The coefficient R_{12} is the feedback or base resistance, and is equal to R_F as defined by Eq. (2.2) in case R_F is a constant. Both R_{11} and R_{12} are of the order of a few hundred ohms, R_{12} usually being smaller than R_{11}.

Equation (2.6) depends mainly on the collector and on the flow of holes from the emitter to the collector. The ratio R_{21}/R_{22} is just the current amplification factor α as defined by Eq. (2.1). Thus we may write:

$$\Delta V_c = R_{22}\,(\alpha \Delta I_e + \Delta I_c). \tag{2.12}$$

The coefficient R_{22} is the collector resistance for fixed emitter current (open circuit for a-c.), and is the order of 10,000–50,000 ohms. Except in the range of large I_e and small I_c, the value of R_{22} is relatively independent of I_c. The factor α generally is small when I_c is small compared with I_e, and increases with I_c, approaching a constant value the order of 1 to 4 when I_c is several times I_e.

The a-c. power gain with the circuit of Fig. 1 depends on the operating point (the d-c. bias currents) and on the load impedance. The positive feedback represented by R_{12} increases the available gain, and it is possible to get very large power gains by operating near a point of instability. In giving the gain under such conditions, the impedance of the input generator should be specified. Alternatively, one can give the gain which would exist with no feedback. The maximum available gain neglecting feedback, obtained when the load R_L is equal to the collector resistance R_{22},

and the impedance of the generator is equal to the emitter resistance, R_{11}, is:

$$\text{Gain} = \alpha^2 R_{22}/4R_{11}, \tag{2.13}$$

which is the ratio of the collector to the emitter resistance multiplied by 1/4 the square of the current amplification factor. This gives the a-c. power delivered to the load divided by the a-c. power fed into the transistor. Substituting the values listed above (Eqs. (2.11)) for the unit whose characteristics are shown in Fig. 3 gives a gain of about 80 times (or 19 db) for the operating point P_0. This is to be compared with the gain of 21 db estimated above for operation between P_1 and P_2. The difference of 2 db represents the increase in gain by feedback, which was omitted in Eq. (2.13).

Equations (2.5) and (2.6) may be solved to express the currents as functions of the voltages, giving

$$\Delta I_e = Y_{11}\,\Delta V_e + Y_{12}\,\Delta V_c \tag{2.14}$$

$$\Delta I_c = Y_{21}\,\Delta V_e + Y_{22}\,\Delta V_c \tag{2.15}$$

where

$$\begin{aligned} Y_{11} &= R_{22}/D,\ Y_{12} = -R_{12}/D \\ Y_{12} &= -R_{21}/D,\ Y_{22} = R_{11}/D \end{aligned} \tag{2.16}$$

and D is the determinant of the coefficients

$$D = R_{11}\,R_{22} - R_{12}\,R_{21}. \tag{2.17}$$

The admittances, Y_{11} and Y_{22}, are negative if D is negative, and the transistor is then unstable if the terminals are short-circuited for a-c. currents. Stability can be attained if there is sufficient impedance in the input and output circuits exterior to the transistor. Feedback and instability are increased by adding resistance in series with the base electrode. Further discussion of this subject would carry us too far into circuit theory and applications. From the standpoint of transistor design, it is desirable to keep the feedback resistance, R_{12}, as small as possible.

Variation with Spacing

One of the important parameters affecting the operation of the transistor is the spacing between the point electrodes. Measurements to investigate this effect have been made on a number of germanium surfaces. Tests were made with use of a micro-manipulator to adjust the positions of the points. The germanium was generally in the form of a slab from .05 to 0.20 cm thick, the lower surface of which was rhodium plated to form a low resistant contact, and the upper plane surface ground and etched, or other-

wise treated to give a surface suitable for transistor action. The collector point was usually kept fixed, since it is more critical, and the emitter point moved. Measurements were made with formed collector points. Most of the data have been obtained on surfaces oxidized as described below.

As expected, the emitter current has less and less influence on the collector as the separation[26], s, is increased. This is shown by a decrease in R_{21}, or α, with s. The effect of the collector current on the emitter, represented by the feedback resistance R_{12}, also decreases with increase in s. The other coefficients, R_{11} and R_{22}, are but little influenced by spacing. Figures

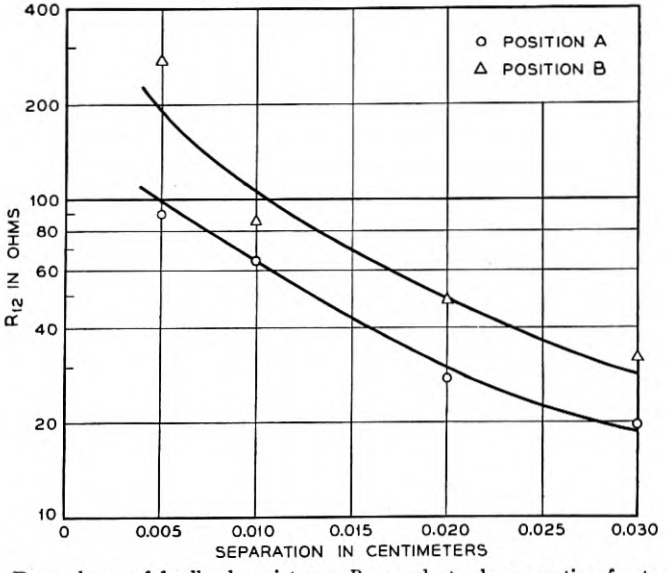


Fig. 4—Dependence of feedback resistance R_{12} on electrode separation for two different parts A and B, of the same germanium surface. The surface had been oxidized by heating in air.

4, 5 and 6 illustrate the variation of R_{12} and α with the separation. Shown are results for two different collector points A and B on different parts of the same germanium surface[25]. In making the measurements, the bias currents were kept fixed as the spacing was varied. For collector A, $I_e =$ 1.0 ma and $I_c = 3.8$ ma; for collector B, $I_e = 1.0$ ma and $I_c = 4.0$ ma. The values of R_{11} and R_{22} were about 300 and 10,000, respectively, in both cases.

Figure 5 shows that α decreases approximately exponentially with s for separations from .005 cm to .030 cm, the rate of decrease being about the same in all cases. Extrapolating down to $s = 0$ indicates that a further

increase of only about 25 per cent in α could be obtained by decreasing the spacing below .005 cm.

Figure 6 shows that the decrease of α with distance is dependent on the germanium sample used. Curve 1 is similar to the results in Fig. 5. Curve

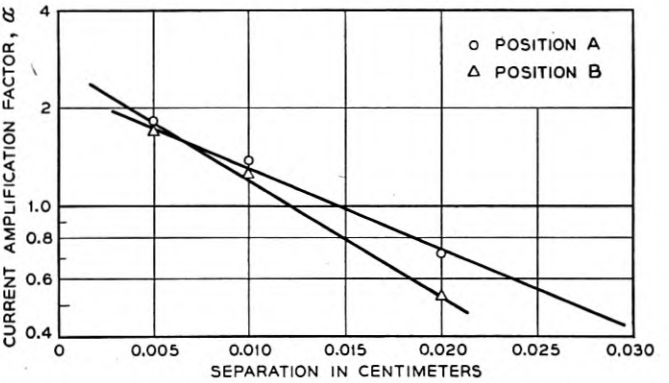


Fig. 5—Dependence of current amplification factor α on electrode separation. Positions A and B as in Fig. 4.

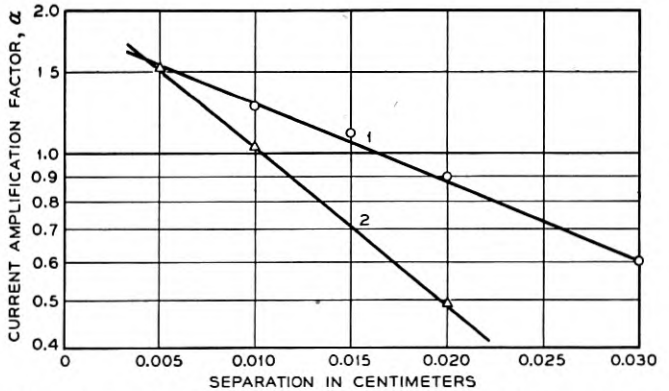


Fig. 6—Dependence of current amplification factor α on separation for germanium surfaces from two different melts, 1 and 2.

2 is for a germanium slice with the same surface treatment but from a different melt.

Figure 4 shows the corresponding results for R_{12}. There is an approximate inverse relationship between R_{12} and s.

Another way to illustrate the decreased influence of the emitter on the collector with increase in spacing is to plot the collector characteristic for fixed emitter current at different spacings. Figure 7 is such a plot for a

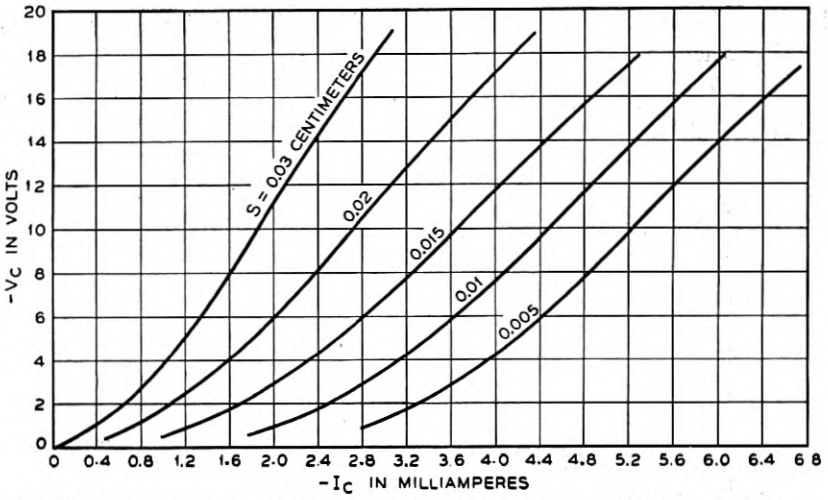


Fig. 7—Collector characteristic V_c vs I_c for fixed I_e but variable distance of separation.

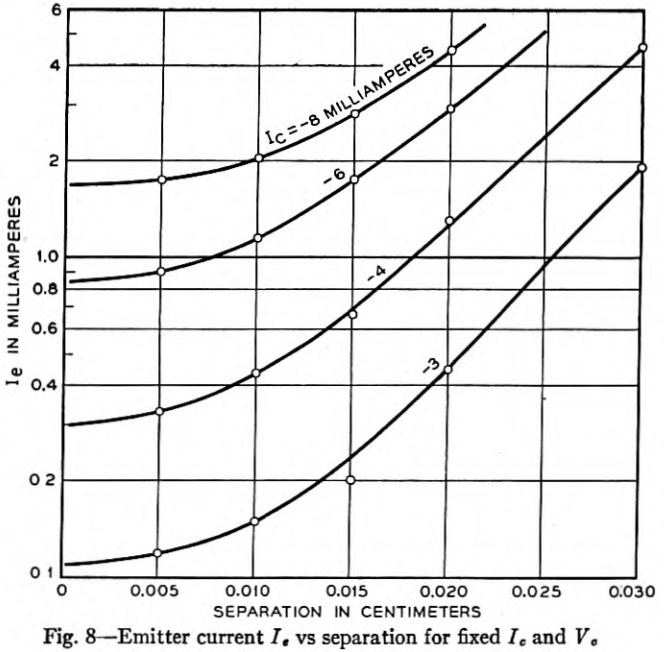


Fig. 8—Emitter current I_e vs separation for fixed I_c and V_c

different surface which was ground flat, etched, and then oxidized at 500°C in moist air for one hour. The resultant oxide film was washed off.[27] The emitter current I_e was kept constant at 1.0 ma.

Data taken on the same surface have been plotted in other ways. As the spacing increases, more emitter current is required to produce the same change in collector current. The fraction of the emitter current which is

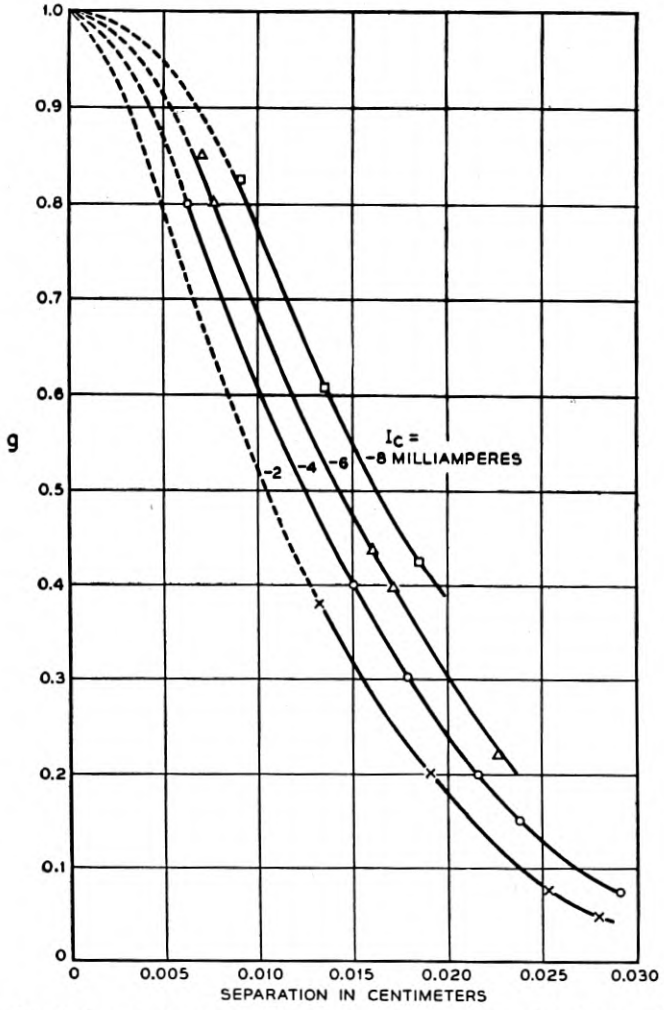


Fig. 9—The factor g is the ratio of the emitter current extrapolated to $s = 0$ to that at electrode separation s required to give the same collector current, I_c and voltage, V_c. Plot shows variation of g with s for different I_c. The factor is independent of V_c over the range plotted.

effective at the collector decreases with spacing. It is of interest to keep V_c and I_c fixed by varying I_e as s is changed and to plot the values of I_e so obtained as a function of s. Such a plot is shown in Fig. 9. The collec-

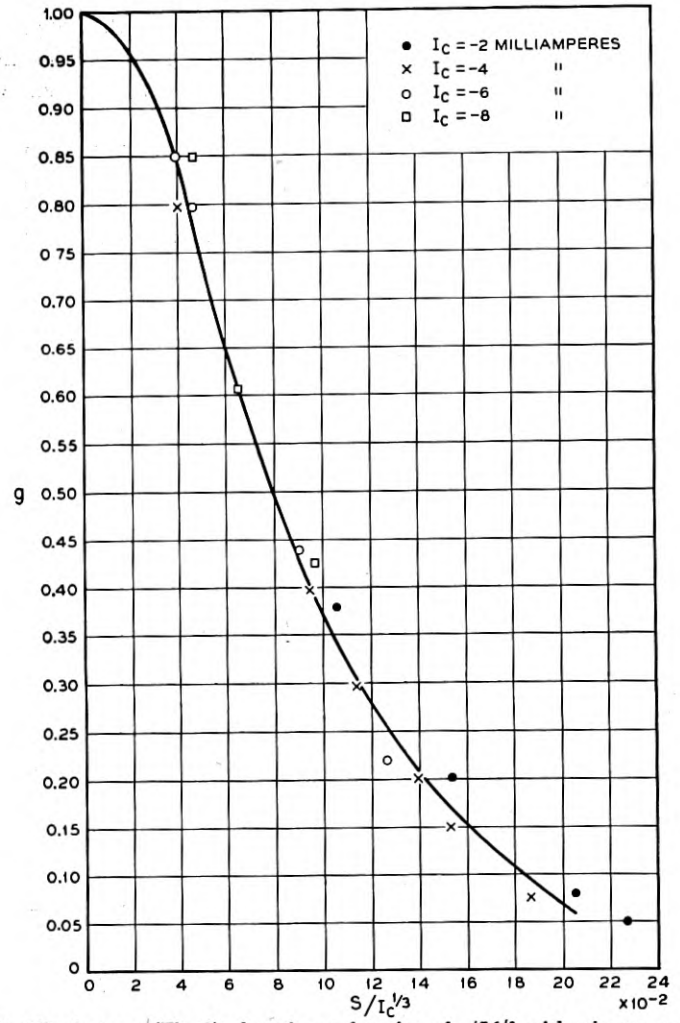


Fig. 10—The factor g (Fig. 9) plotted as a function of $s/I_c^{1/3}$, with s in cm. and I_c in amps.

tor voltage, V_c, is fixed at -15 volts. Curves are shown for $I_c = -3$, -4, -6, and -8 ma. We may define a geometrical factor, g, as the ratio of I_e extrapolated to zero spacing to the value of I_e at the separation s:

$$g(s) = (I_e(0)/I_e(s))_{V_c, I_c = \text{const.}} \tag{2.18}$$

It is to be expected that $g(s)$ will depend on I_c, as it is the collector current which provides the field which draws the holes into the collector. For the

same reason, it is expected that $g(s)$ will be relatively independent of V_c. This was indeed found to be true in this particular case and the values $V_c = -5$, -10, and -15 were used in Figure 9 which gives a plot of g versus s for several values of I_c. The dotted lines give the extrapolation to $s = 0$. As expected, g increases with I_c for a fixed s. The different curves can be brought into approximate agreement by taking $s/I_c^{1/3}$ as the independent variable, and this is done in Fig. 10. As will be discussed in Section V, such a relation is to be expected if g depends on the transit time for the holes.

Variation with Temperature

Only a limited amount of data has been obtained on the variation of transistor characteristics with temperatures. It is known that the reverse

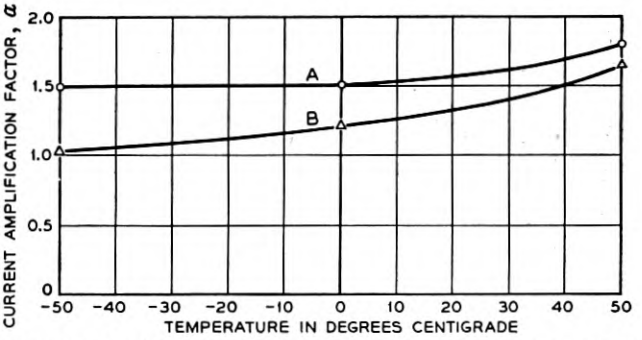


Fig. 11—Current amplification factor α vs. temperature for two experimental units A and B.

characteristic of the germanium diode varies rapidly with temperature, particularly in the case of units with high reverse resistance. In the transistor the collector is electrically formed in such a way as to have relatively low reverse resistance, and its characteristic is much less dependent on temperature. Both R_{22} and R_{11} decrease with increase in T, R_{22} usually decreasing more rapidly than R_{11}. The feedback resistance, R_{12}, is relatively independent of temperature. The current multiplication factor, α, increases with temperature, but the change is not extremely rapid. Figure 11 gives a plot of α versus T for two experimental units. The d-c. bias currents are kept fixed as the temperature is varied. The over-all change in α from -50°C to $+50$°C is only about 50 per cent. The increase in α with T results in an increase in power gain with temperature. This may be nullified by a decrease in the ratio R_{22}/R_{11}, so that the over-all gain at fixed bias current may have a negative temperature coefficient.

Variation with Frequency

Equations (2.5) and (2.6) may be used to describe the a-c. characteristics at high frequencies if the coefficients are replaced by general impedances. Thus if we use the small letters i_e, v_e, i_c, v_c, to denote the amplitude and phase of small a-c. signals about a given operating point, we may write

$$v_e = Z_{11}\, i_e + Z_{12} i_c, \tag{2.19}$$

$$v_c = Z_{21}\, i_e + Z_{22} i_c. \tag{2.20}$$

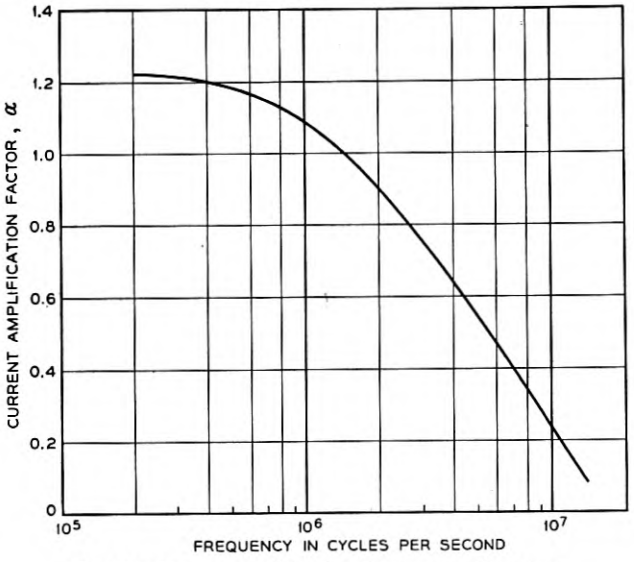


Fig. 12—Current amplification factor α vs. frequency

Measurements of A. J. Rack and others,[28] have shown that the over-all power gain drops off between 1 and 10 mc/sec and few units have positive gain above 10 mc/sec. The measurements showed further that the frequency variation is confined almost entirely to Z_{21} or α. The other coefficients, Z_{11}, Z_{12} and Z_{22}, are real and independent of frequency, at least up to 10 mc/sec. Figure 12 gives a plot of α versus frequency for an experimental unit. Associated with the drop in amplitude is a phase shift which varies approximately linearly with the frequency. A phase shift in Z_{21} of 90° occurs at a frequency of about 4 mc/sec, corresponding to a delay of about 5×10^{-8} seconds. Estimates of transit time for the holes to flow from the emitter to the collector, to be made in Section V, are of the same order. These results suggest that the frequency limitation is associated with transit time rather than electrode capacities. Because of the difference

in transit times for holes following different paths there is a drop in amplitude rather than simply a phase shift.

III—Electrical Conductivity of Germanium

Germanium, like carbon and silicon, is an element of the fourth group of the periodic table, with the same crystal structure as diamond. Each germanium atom has four near neighbors in a tetrahedral configuration with which it forms covalent bonds. The specific gravity is about 5.35 and the melting point 958°C.

The conductivity at room temperature may be either n or p type, depending on the nature and concentration of impurities. Scaff, Theuerer, and Schumacher[29] have shown that group III elements, with one less valence electron, give p-type conductivity; group V elements, with one more valence electron, give n-type conductivity. This applies to both germanium

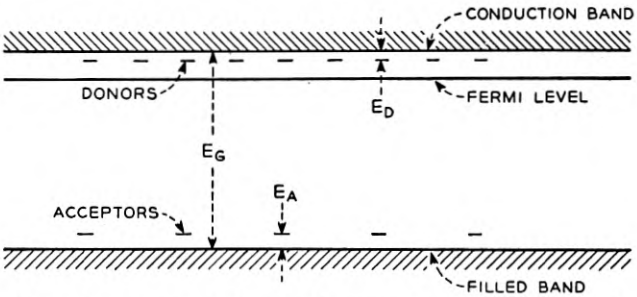


Fig. 13—Schematic energy level diagram for germanium showing filled and conduction bands and donor and acceptor levels.

and silicon. There is evidence that both acceptor (p-type) and donor (n-type) impurities are substitutional[30].

A schematic energy level diagram[31] which shows the allowed energy levels for the valence electrons in a semi-conductor like germanium is given in Fig. 13. There is a continuous band of levels, the filled band, normally occupied by the electrons in the valence bonds; an energy gap, E_G, in which there are no levels of the ideal crystal; and then another continuous band of levels, the conduction band, normally unoccupied. There are just sufficient levels in the filled band to accomodate the four valence electrons per atom. The acceptor impurity levels, which lie just above the filled band, and the donor levels, just below the conduction band, correspond to electrons localized about the impurity atoms. Donors are normally neutral, but become positively charged by excitation of an electron to the conduction band, an energy E_D being required. Acceptors, normally neutral, are negatively ionized by excitation of an electron from the filled band, an energy E_A

when evaluated for room temperature. Thus for $n_e \sim 10^{15}/cm^3$, corresponding to high-back-voltage germanium, n_h is the order of 10^{12}. The equilibrium concentration of holes is small.

Below the intrinsic temperature range, n_e is approximately constant and n_h varies as

$$n_h = (C_e C_h T^3/n_e) \exp(-E_G/kT). \tag{3.11}$$

IV—Theory of the Diode Characteristic

Characteristics of metal point-germanium contacts include high forward currents, as large as 5 to 10 ma at 1 volt, small reverse currents, correspond-

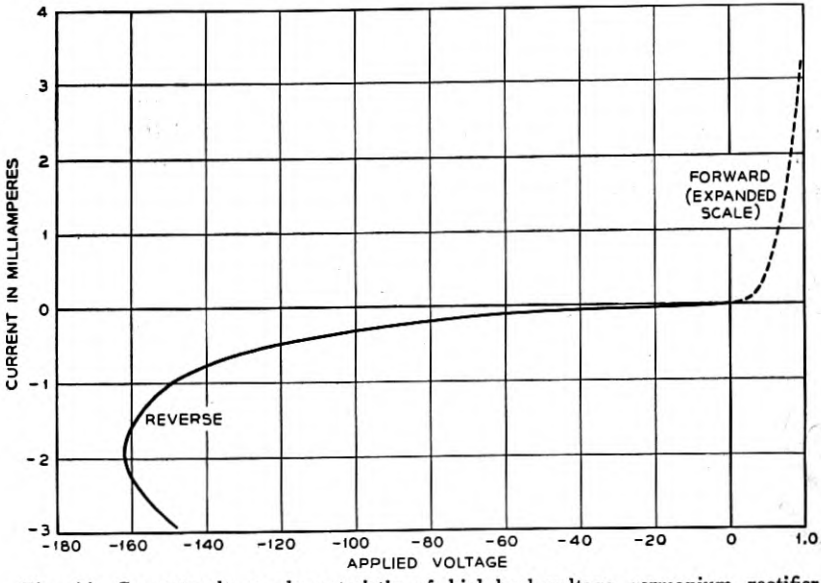


Fig. 14—Current-voltage characteristic of high-back-voltage germanium rectifier. Note that the voltage scale in the forward direction has been expanded by a factor of 20.

ing to resistances as high as one megohm or more at reverse voltages up to 30 volts, and the ability to withstand large voltages in the reverse direction without breakdown. A considerable variation of rectifier characteristics is found with changes in preparation and impurity content of the germanium, surface treatment, electrical power or forming treatment of the contacts, and other factors.

A typical d-c. characteristic of a germanium rectifier[35] is illustrated in Fig. 14. The forward voltages are indicated on an expanded scale. The forward current at one volt bias is about 3.5 ma and the differential resistance is about 200 ohms. The reverse current at 30 volts is about .02 ma

and the differential resistance about 5×10^5 ohms. The ratio of the forward to the reverse current at one volt bias is about 500. At a reverse voltage of about 160 the differential resistance drops to zero, and with further increase in current the voltage across the unit drops. The nature of this negative resistance portion of the curve is not completely understood, but it is believed to be associated with thermal effects. Successive points along the curve correspond to increasingly higher temperatures of the contact. The peak value of the reverse voltage varies among different units. Values of more than 100 volts are not difficult to obtain.

Theories of rectification as developed by Mott,[36] Schottky,[37] and others[38] have not been successful in explaining the high-back-voltage characteristic in a quantitative way. In the following we give an outline of the theory and its application to germanium. It is believed that the high forward currents can now be explained in terms of a flow of holes. The type of barrier which gives a flow of carriers of conductivity type opposite to that of the base material is discussed. It is possible that a hole current also plays an important role in the reverse direction.

The Space-Charge Layer

According to the Mott-Schottky theory, rectification results from a potential barrier at the contact which impedes the flow of electrons between the metal and the semi-conductor. A schematic energy level diagram of the barrier region, drawn roughly to scale for germanium, is given in Fig. 15. There is a rise in the electrostatic potential energy of an electron at the surface relative to the interior which results from a space charge layer in the semi-conductor next to the metal contact. The space charge arises from positively ionized donors, that is from the same impurity centers which give the conduction electrons in the body of the semi-conductor. In the interior, the space-charge of the donors is neutralized by the space charge of the conduction electrons which are present in equal numbers. Electrons are drained out of the space-charge layer near the surface, leaving the immobile donor ions.

The space charge layer may be a result of the metal-semi-conductor contact, in which case the positive charge in the layer is compensated by an induced charge of opposite sign on the metal surface. Alternatively, the charge in the layer may be compensated by a surface charge density of electrons trapped in surface states on the semi-conductor.[9] It is believed, for reasons to be discussed below, that the latter situation applies to high-back-voltage germanium, and that a space-charge layer exists at the free surface, independent of the metal contact. The height of the conduction band above the Fermi level at the surface, φ_s, is then determined by the distribution in energy of the surface states.

That the space-charge layer which gives the rectifying barrier in germanium arises from surface states, is indicated by the following:

(1) Characteristics of germanium-point contacts do not depend on the work function of the metal, as would be expected if the space-charge layer were determined by the metal contact.

(2) There is little difference in contact potential between different samples of germanium with varying impurity concentration. Benzer[39]

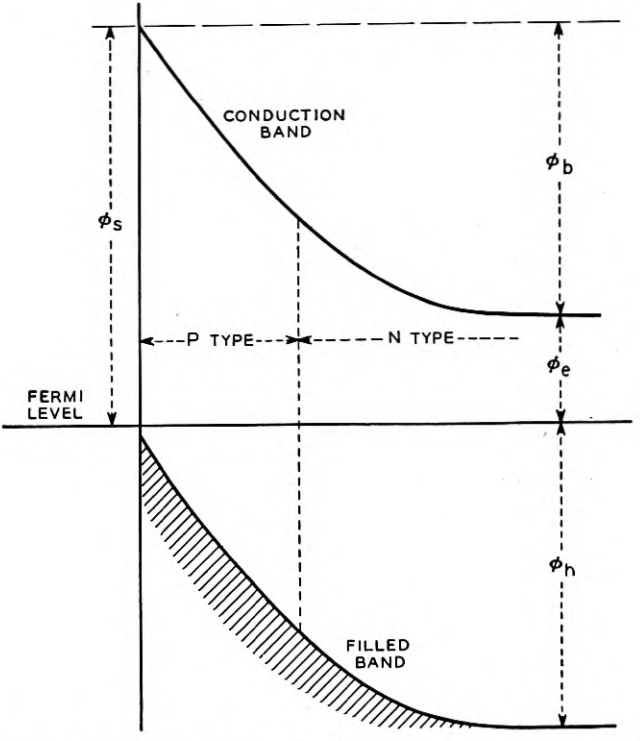


Fig. 15—Schematic energy level diagram of barrier layer at germanium surface showing inversion layer of p-type conductivity.

found less than 0.1-volt difference between samples ranging from n-type with 2.6×10^{18} carriers/cm^3 to p-type with 6.4×10^{18} carriers/cm^3. This is much less than the difference of the order of the energy gap, 0.75 volts, which would exist if there were no surface effects.

(3) Benzer[40] has observed the characteristics of contacts formed from two crystals of germanium. He finds that in both directions the characteristic is similar to the reverse characteristic of one of the crystals in contact with a highly conducting metal-like germanium crystal.

(4) One of the authors[11] has observed a change in contact potential with light similar to that expected for a barrier layer at the free surface.

Prior to Benzer's experiments, Meyerhof[8] had shown that the contact potential difference measured between different metals and silicon showed little correlation with rectification, and that the contact potential difference between n- and p-type silicon surfaces was small. There is thus evidence that the barrier layers in both germanium and silicon are internal and occur at the free surface[41].

In the development of the mathematical theory of the space-charge layer at a rectifier contact, Schottky and Spenke[18] point out the possibility of a change in conductivity type between the surface and the interior if the potential rise is sufficiently large. The conductivity is p-type if the Fermi level is closest to the filled band, n-type if it is closest to the conduction band. In the illustration (Fig. 15), the potential rise is so large that the filled band is raised up to a position close to the Fermi level at the surface. This situation is believed to apply to germanium. There is then a thin layer near the surface whose conductivity is p-type, superimposed on the n-type conductivity in the interior. Schottky and Spenke call the layer of opposite conductivity type an inversion region.

Referring to Eqs. (3.5a and 3.5b) for the concentrations, it can be seen that since C_e and C_h are of the same order of magnitude, the conductivity type depends on whether φ_e is larger or smaller than φ_h. The conductivity is n-type when

$$\varphi_e < 1/2\, E_G, \qquad \varphi_h > 1/2\, E_G, \tag{4.1}$$

and is p-type when the reverse situation applies. The maximum resistivity occurs at the position where the conductivity type changes and

$$\varphi_e \sim \varphi_h \sim 1/2\, E_G. \tag{4.2}$$

The change from n- to p-type will occur if

$$\varphi_s > 1/2\, E_G, \tag{4.3}$$

or if the overall potential rise, φ_b, is greater than

$$1/2\, E_G - \varphi_{e0}, \tag{4.4}$$

where φ_{e0} is the value of φ_e in the interior. Since for high-back-voltage germanium, $E_G \sim 0.75$ e.v. and $\varphi_{e0} \sim 0.25$ e.v., a rise of more than 0.12 e.v. is sufficient for a change of conductivity type to occur. A rise of 0.50 e.v. will bring the filled band close to the Fermi level at the surface.

Schottky[37] relates the thickness of the space charge layer with a potential rise as follows. Let ρ be the average change density, assumed constant for simplicity, in the space charge layer. In the interior ρ is compensated by

the space charge of the conduction electrons. Thus, if n_0 is the normal concentration of electrons,[42]

$$\rho = en_0 \tag{4.5}$$

Integration of the space charge equations gives a parabolic variation of potential with distance, and the potential rise, φ_b, is given in terms of the thickness of the space charge layer, ℓ, by the equation

$$\varphi_b = 2\pi e\rho\ell^2/\kappa = 2\pi e^2 n_0 \ell^2/\kappa \tag{4.6}$$

For

$$\varphi_b = \varphi_s - \varphi_{s0} \sim 0.5 \text{ e.v.} \sim 8 \times 10^{-13} \text{ ergs}$$

and

$$n_0 \sim 10^{15}/cm^3$$

the barrier thickness, ℓ, is about 10^{-4} cm. The dielectric constant, κ, is about 18 in germanium.

When a voltage, V_a, is applied to a rectifying contact, there will be a drop, V_b, across the space charge layer itself and an additional drop, IR_s, in the body of the germanium which results from the spreading resistance, R_s, so that

$$V_a = V_b + IR_s. \tag{4.7}$$

The potential energy drop, $-eV_b$, is superimposed on the drop φ_b which exists under equilibrium conditions. For this case Eq. (4.6) becomes

$$\varphi_b - eV_b = 2\pi e^2 n_0 \ell^2/\kappa \tag{4.8}$$

The potential V_b is positive in the forward direction, negative in the reverse. A reverse voltage increases the thickness of the layer, a forward voltage decreases the thickness of the layer. The barrier disappears when $eV_b = \varphi_b$, and the current is then limited entirely by the spreading resistance in the body of the semi-conductor.

The electrostatic field at the contact is

$$F = 4\pi e n_0 \ell/\kappa = (8\pi n_0(\varphi_b - eV_b)/\kappa)^{1/2} \tag{4.9}$$

For $n_0 \sim 10^{15}$, $\ell \sim 10^{-4}$ and $\kappa \sim 18$, the field F is about 30 e.s.u. or 10,000 volts/cm. The field increases the current flow in much the way the current from a thermionic emitter is enhanced by an external field.

Previous theories of rectification have been based on the flow of only one type of carrier, i.e. electrons in an n-type or holes in a p-type semi-conductor. If the barrier layer has an inversion region, it is necessary to consider the flow of both types of carriers. Some of the hitherto puzzling features

of the germanium diode characteristic can be explained by the hole current. While a complete theoretical treatment has not been carried out, we will give an outline of the factors involved and then give separate discussions for the reverse and forward directions.

The current of holes may be expected to be important if the concentration of holes at the semi-conductor boundary of the space charge layer is as large as the concentration of electrons at the metal-semi-conductor interface. In equilibrium, with no current flow, the former is just the hole concentration in the interior, n_{h0}, which is given by

$$n_{h0} = C_h T^{3/2} \exp(-\varphi_{h0}/kT), \tag{4.10}$$

where φ_{h0} is the energy difference between the Fermi level in the interior and the top of the filled band. The concentration of electrons at the interface is given by:

$$n_{em} = C_e T^{3/2} \exp(-\varphi_s/kT). \tag{4.11}$$

Since C_h and C_e are of the same order, n_{h0} will be larger than n_{em} if φ_s is larger than φ_{h0}. This latter condition is met if the hole concentration at the metal interface is larger than the electron concentration in the interior. The concentrations will, of course, be modified when a current is flowing; but the criterion just given is nevertheless a useful guide. The criterion applies to an inversion barrier layer regardless of whether it is formed by the metal contact or is of the surface states type. In the latter case, as discussed in the Introduction, a lateral flow of holes along the surface layer into the contact may contribute to the current.

Two general theories have been developed for the current in a rectifying junction which apply in different limiting cases. The diffusion theory applies if the current is limited by the resistance of the space charge layer. This will be the case if the mean free path is small compared with the thickness of the layer, or, more exactly, small compared with the distance required for the potential energy to drop kT below the value at the contact. The diode theory applies if the current is limited by the thermionic emission current over the barrier. In germanium, the mean free path (10^{-5} cm) is of the same order as the barrier thickness. Analysis shows, however, that scattering in the barrier is unimportant and that it is the diode theory which should be used.[43]

Reverse Current

Different parts of the d-c. current-voltage characteristics require separate discussion. We deal first with the reverse direction. The applied voltages are assumed large compared with kT/e (.025 volts at room temperature), but small compared with the peak reverse voltage, so that ther-

mal effects are unimportant. Electrons flow from the metal point contact to the germanium, and holes flow in the opposite direction.

Benzer[43] has made a study of the variation of the reverse characteristic with temperature. He divides the current into three components whose relative magnitudes vary among different crystals and which vary in different ways with temperature. These are:

(1) A saturation current which arises very rapidly with applied voltage, approaching a constant value at a fraction of a volt.

(2) A component which increases linearly with the voltage.

(3) A component which increases more rapidly than linearly with the voltage.

The first two increase rapidly with increasing temperature, while the third component is more or less independent of ambient temperature. It is the saturation current, and perhaps also the linear component, which are to be identified with the theoretical diode current.

The third component is the largest in units with low reverse resistance. It is probable that in these units the barrier is not uniform. The largest part of the current, composed of electrons, flows through patches in which the height of the barrier is small. The electrically formed collector in the transistor may have a barrier of this sort.

Benzer finds that the saturation current predominates in units with high reverse resistance, and that this component varies with temperature as

$$I_s = -I_0 e^{\epsilon/kT}, \tag{4.12}$$

with ϵ nearly 0.7 e.v. The negative sign indicates a reverse current. According to the diode theory,[44] one would expect it to vary as

$$I_s = -BT^2 e^{\epsilon/kT}. \tag{4.13}$$

Since ϵ is large, the observed current can be fitted just about as well with the factor T^2 as without. The value of ϵ obtained using (4.13) is about 0.6 e.v. The saturation current[43] at room temperature varies from 10^{-7} to 10^{-6} amps, which corresponds to values of B in the range of 0.01 to 0.1 amps/deg^2.

The theoretical value of B is 120 times the contact area, A_c. Taking $A_c \sim 10^{-6}$ cm^2 as a typical value for the area of a point contact gives $B \sim 10^{-4}$ amps/deg^2 which is only about 1/100 to 1/1000 of the observed. It is difficult to reconcile the magnitude of the observed current with the large temperature coefficient, and it is possible that an important part of the total flow is a current of holes into the contact. Such a current particularly is to be expected on surfaces which exhibit an appreciable surface conductivity.

Neglecting surface effects for the moment, an estimate of the saturation

hole current might be obtained as follows: The number of holes entering the space charge region per second is[45]

$$n_{hb}v_aA_c/4,$$

where n_{hb} is the hole concentration at the semi-conductor boundary of the space-charge layer and v_a is an average thermal velocity ($\sim 10^7$ cm/sec.). The hole current, I_h, is obtained by multiplying by the electronic charge, giving

$$I_h = -n_{hb}ev_aA_c/4 \tag{4.14}$$

If we set n_{hb} equal to the equilibrium value for the interior, say $10^{12}/\text{cm}^3$, we get a current $I_h \sim 4 \times 10^{-7}$ amps, which is of the observed order of magnitude of the saturation current at room temperature. With this interpretation, the temperature variation of I_s is attributed to that of n_h, which, according to Eq. (3.11), varies as exp $(-E_G/kT)$. The observed value of ϵ is indeed almost equal to the energy gap.

The difficulty with this picture is to see how n_{hb} can be as large as n_h when a current is flowing. Holes must move toward the contact area primarily by diffusion, and the hole current will be limited by a diffusion gradient. The saturation current depends on how rapidly holes are generated, and reasonable estimates based on the mean life time, $-\tau$, yield currents which are several orders of magnitude too small. A diffusion velocity, v_D, of the order

$$v_D \sim (D/\tau)^{1/2}, \tag{4.15}$$

replaces $v_a/4$ is Eq. (4.14). Setting $D \sim 25$ cm^2/sec and $\tau \sim 10^{-6}$ sec gives $v_D \sim 5 \times 10^3$, which would give a current much smaller than the observed. What is needed, then, is some other mechanism which will help maintain the equilibrium concentration near the barrier. Surface effects may be important in this regard.

Forward Current

The forward characteristic is much less dependent on such factors as surface treatment than the reverse. In the range from 0 to 0.4 volts in the forward direction, the current can be fitted quite closely by a semi-empirical expression[46] of the form:

$$I = I_0(e^{\beta V_b} - 1), \tag{4.16}$$

where V_b is the drop across the barrier resulting from the applied voltage, as defined by Eq. (4.7). Equation (4.16) is of the general form to be expected from theory, but the measured value of β is generally less than the theoretical value e/kT (40 volts^{-1} at room temperature). Observed values

of β may be as low as 10, and in other units are nearly as high as the theoretical value of 40. The factor I_0 also varies among different units and is of the order 10^{-7} to 10^{-6} amperes. While both experiment and theory indicate that the forward current at large forward voltages is largely composed of holes, the composition of the current at very small forward voltages is uncertain. Small areas of low φ_s, unimportant at large forward voltages, may give most of the current at very small voltages. Currents flowing in these areas will consist largely of electrons.

Above about 0.5 volts in the forward direction, most of the drop occurs across the spreading resistance, R_s, rather than across the barrier. The theoretical expression for R_s for a circular contact of diameter d on the surface of a block of uniform resistivity ρ is:

$$R_s = \rho/2d \tag{4.17}$$

Taking as typical values for a point contact on high-back-voltage germanium, $\rho = 10$ ohm cm. and $d = .0025$ cm, we obtain $R_s = 2000$ ohms, which is the order of ten times the observed.

As discussed in the Introduction, Bray and others[21, 22] have attempted to account for this discrepancy by assuming that the resistivity decreases with increasing field, and Bray has made tests to observe such an effect. The authors have investigated the nature of the forward current by making potential probe measurements in the vicinity of a point contact.[2] These measurements indicate that there may be two components involved in the excess conductivity. Some surfaces, prepared by oxidation at high temperatures, give evidence for excess conductivity in the vicinity of the point in the reverse as well as in the forward direction. This ohmic component has been attributed to a thin p-type layer on the surface. All surfaces investigated exhibit an excess conductivity in the forward direction which increases with increasing forward current. This second component is attributed to an increase in the concentration of carriers, holes and electrons, in the vicinity of the point with increase in forward current. Holes flow from the point into the germanium and their space charge is compensated by electrons.

The ohmic component is small, if it exists at all, on surfaces treated in the normal way for high-back-voltage rectifiers (i.e., ground and etched). The nature of the second component on such surfaces has been shown by more recent work of Shockley, Haynes[20], and Ryder[23] who have investigated the flow of holes under the influence of electric fields. These measurements prove that the forward current consists at least in large part of holes flowing into the germanium from the contact.

It is of interest to consider the way the concentrations of holes and electrons vary in the vicinity of the point. An exact calculation, including the

effect of recombination, leads to a non-linear differential equation which must be solved by numerical methods. A simple solution can be obtained, however, if it is assumed that all of the forward current consists of holes and if recombination is neglected.

The electron current then vanishes everywhere, and the electric field is such as to produce a conduction current of electrons which just cancels the current from diffusion, giving

$$n_e F = -(kT/e) \text{ grad } n_e. \tag{4.18}$$

This equation may be integrated to give the relation between the electrostatic potential, V, and n_e,

$$V = (kT/e) \log (n_e/n_{e0}). \tag{4.19}$$

The constant of integration has been chosen so that $V = 0$ when n_e is equal to the normal electron concentration n_{e0}. The equation may be solved for n_e to give:

$$n_e = n_{e0} \exp(eV/kT). \tag{4.20}$$

If trapping is neglected, electrical neutrality requires that

$$n_e = n_h + n_{e0}. \tag{4.21}$$

Using this relation, and taking n_{e0} a constant, we can express field F in terms of n_h

$$F = -(kT/e(n_h + n_{e0})) \text{ grad } n_h \tag{4.22}$$

The hole current density, i_h, is the sum of a conduction current resulting from the field F and a diffusion current:

$$i_h = n_h e\mu_h F - kT\mu_h \text{ grad } n_h \tag{4.23}$$

Using Eq. (4.22), we may write this in the form

$$i_h = -kT\mu_h((2n_h + n_{e0})/(n_h + n_{e0})) \text{ grad } n_h \tag{4.24}$$

The current density can be written

$$i_h = -\text{ grad } \psi, \tag{4.25}$$

where

$$\psi = kT\mu_h(2n_h - n_{e0} \log ((n_h + n_{e0})/n_{e0})) \tag{4.26}$$

Since i_h satisfies a conservation equation,

$$\text{div } i_h = 0, \tag{4.27}$$

ψ satisfies Laplace's equation.

If surface effects are neglected and it is assumed that holes flow radially in all directions from the point contact, ψ may be expressed simply in terms of the total hole current, I_h, flowing from the contact:

$$\psi = -I_h/2\pi r \tag{4.28}$$

Using (4.26), we may obtain the variation of n_h with r. We are interested in the limiting case in which n_h is large compared with the normal electron concentration, n_{e0}. The logarithmic term in (4.26) can then be neglected, and we have

$$n_h = I_h/4\pi r\mu_h kT. \tag{4.29}$$

For example, if $I_h = 10^{-3}$ amps, $\mu_h = 10^3$ cm^2/volt sec, and $kT/e = .025$ volts, we get, approximately,

$$n_h = 2 \times 10^{13}/r. \tag{4.30}$$

For $r \sim .0005$ cm, the approximate radius of a point contact,

$$n_h \sim 4 \times 10^{16}/\text{cm}^3, \tag{4.31}$$

which is about 40 times the normal electron concentration in high-back-voltage germanium. Thus the assumption that n_h is large compared with n_{e0} is valid, and remains valid up to a distance of the order of .005 cm, the approximate distance the points are separated in the transistor.

To the same approximation, the field is

$$F = kT/er, \tag{4.32}$$

independent of the magnitude of I_h.

The voltage drop outside of the space-charge region can be obtained by setting n_e in (4.19) equal to the value at the semi-conductor boundary of the space-charge layer. This result holds generally, and does not depend on the particular geometry we have assumed. It depends only on the assumption that the electron current i_e is everywhere zero. Actually i_h will decrease and i_e increase by recombination, and there will be an additional spreading resistance for the electron current.

If it is assumed that the concentration of holes at the metal-semi-conductor interface is independent of applied voltage and that the resistive drop in the barrier layer itself is negligible, that part of the applied voltage which appears across the barrier layer itself is:

$$V_b = (kT/e) \log (n_{hb}/n_{h0}), \tag{4.33}$$

where n_{hb} is the hole concentration at the semi-conductor boundary of the space charge layer and n_{h0} is the normal concentration. For $n_{hb} \sim 5 \times 10^{16}$ and $n_{h0} \sim 10^{12}$, V_b is about 0.35 volts.

The increased conductivity caused by hole emission accounts not only for the large forward currents, but also for the relatively small dependence of spreading resistance on contact area. At a small distance from the contact, the concentrations and voltages are independent of contact area. The voltage drop within this small distance is a small part of the total and does not vary rapidly with current.

We have assumed that the electron current, I_e, at the contact is negligible compared with the hole current, I_h. An estimate of the electron current can be obtained as follows: From the diode theory,

$$I_e = (en_{eb}v_aA_c/4) \exp(-(\varphi_b - eV_b)/kT), \tag{4.34}$$

since the electron concentration at the semi-conductor boundary of the space-charge layer is n_{eb} and the height of the barrier with the voltage applied is $\varphi_b - eV_b$. For simplicity we assume that both n_{eb} and n_{hb} are large compared with n_{e0} so that we may replace n_{eb} by n_{hb} without appreciable error. The latter can be obtained from the value of ψ at the contact:

$$\psi = I_h/4a \tag{4.35}$$

Expressing ψ in terms of n_{hb}, we find

$$n_{hb} = I_h/8kT\mu_h a \tag{4.36}$$

Using (4.33) for V_b, and (3.5b) for n_{h0} we find after some reduction,

$$I_e = I_h^2/I_{crit}, \tag{4.37}$$

where

$$I_{crit} = \frac{256\, C_h\, (kT\mu_h)^{\,2} T^{3/2}}{\pi e v_a} \exp(-\varphi_{hm}/kT) \tag{4.38}$$

The energy difference φ_{hm} is the difference between the Fermi level and the filled band at the metal-semi-conductor interface. Evaluated for germanium at room temperature, (4.38) gives

$$I_{crit} = 0.07 \exp(-\varphi_{hm}/kT) \text{ amps},$$

which is a fairly large current if φ_{hm} is not too large compared with kT. If I_h is small compared with I_{crit}, the electron current will be negligible.

V—Theoretical Considerations about Transistor Action

In this section we discuss some of the problems connected with transistor action, such as:

(1) fields produced by the collector current,
(2) transit times for the holes to flow from emitter to collector,
(3) current multiplication in collector,
(4) feedback resistance.

We do no more than estimate orders of magnitude. An exact calculation, taking into account the change of conductivity introduced by the emitter current, loss of holes by recombination, and effect of surface conductivity, is difficult and is not attempted.

To estimate the field produced by the collector, we assume that the collector current is composed mainly of conduction electrons, and that the electrons flow radially away from the collector. This assumption should be most nearly valid when the collector current is large compared with the emitter current. The field at a distance r from the collector is,

$$F = \rho I_c / 2\pi r^2 \tag{5.1}$$

For example, if, $\rho = 10$ ohm cm, $I_c = .001$ amps, and $r = .005$ cm, F is about 100 volts/cm.

The drift velocity of a hole in the field F is $u_h F$. The transit time is

$$T = \int \frac{dr}{\mu_h F} = \frac{2\pi}{\mu_h \rho I_c} \int_0^s r^2 \, dr, \tag{5.2}$$

where s is the separation between the emitter and collector. Integration gives,

$$T = \frac{2\pi s^3}{3\mu_{h0} I_c} \tag{5.3}$$

For $s = .005$ cm, $\mu_h = 1000$ cm^2/volt sec, $\rho = 10$ ohm cm, and $I_c = .001$ amps, T is about 0.25×10^{-7} sec. This is of the order of magnitude of the transit times estimated from the phase shift in α or Z_{21}.

The hole current, I_h, is attenuated by recombination in going from the emitter to the collector. If τ is the average life time of a hole, I_h will be decreased by a factor, $e^{-T/\tau}$. In Section II it was found that the geometrical factor, g, which gives the influence of separation on the interaction between emitter and collector, depends on the variable $s/I_c^{1/3}$. This suggests that the transit time is the most important factor in determining g. An estimate[47] of τ, obtained from the data of Fig. 10, is 2×10^{-7} sec.

Because of the effect of holes in increasing the conductivity of the germanium in the vicinity of the emitter and collector, it can be expected that the field, the life time, and the geometrical factor will depend on the emitter current. The effective value of ρ to be used in Eqs. (5.1) and (5.2) will decrease with increase in emitter current. This effect is apparently not serious with the surface used in obtaining the data for Figs. 8 to 10.

Next to be considered is the effect of the space charge of the holes on the barrier layer of the collector. An estimate of the hole concentration can be obtained as follows: The field in the barrier layer is of the order of 10^4

volts/cm. Multiplying by the mobility gives a drift velocity, V_d of 10^7 cm/sec, which is approximately thermal velocity.[48] The hole current is

$$I_h = n_h e V_d A_c \tag{5.4}$$

where A_c is the area of the collector contact, and n_h the concentration of holes in the barrier. Solving for the latter, we get

$$n_h = I_h / e V_d A_c \tag{5.5}$$

For $I_h = .001$ amps $V_d = 10^7$ cm/sec, and $A_c = 10^{-6}$ cm, n_h is about $.6 \times 10^{15}$, which is of the same order as the concentration of donors. Thus the hole current can be expected to alter the space charge in the barrier by a significant amount, and correspondingly alter the flow of electrons from the collector. It is believed that current multiplication (values of $\alpha > 1$) can be accounted for along these lines.

As discussed in Section II, there is an influence of collector current on emitter current of the nature of a positive feedback. The collector current lowers the potential of the surface in the vicinity of the emitter by an amount

$$V = \rho I_c / 2\pi s \tag{5.6}$$

The feedback resistance R_F as used in Eq. (2.2) is

$$R_F = \rho / 2\pi s \tag{5.7}$$

For $\rho = 10$ ohm cm and $s = .005$ cm, the value of R_F is about 300 ohms, which is of the observed order of magnitude. It may be expected that R_F will decrease as ρ decreases with increase in emitter current.

The calculations made in this section confirm the general picture which has been given of the way the transistor operates.

VI—Conclusions

Our discussion has been confined to the transistor in which two point contacts are placed in close proximity on one face of a germanium block. It is apparent that the principles can be applied to other geometrical designs and to other semi-conductors. Some preliminary work has shown that transistor action can be obtained with silicon and undoubtedly other semi-conductors can be used.

Since the initial discovery, many groups in the Bell Laboratories have contributed to the progress that has been made. This work includes investigation of the physical phenomena involved and the properties of the materials used, transistor design, and measurements of characteristics and circuit applications. A number of transistors have been made for experimental use in a pilot production. Obviously no attempt has been made

to describe all of this work, some of which has been reported on in other publications[5].

In a device as new as the transistor, various problems remain to be solved. A reduction in noise and an increase in the frequency limit are desirable While much progress has been made toward making units with reproducible characteristics, further improvement in this regard is also desirable.

It is apparent from reading this article that we have received a large amount of aid and assistance from other members of the Laboratories staff, for which we are grateful. We particularly wish to acknowledge our debt to Ralph Bown, Director of Research, who has given us a great deal of encouragement and aid from the inception of the work and to William Shockley, who has made numerous suggestions which have aided in clarifying the phenomena involved.

References

1. J. Bardeen and W. H. Brattain, *Phys. Rev.*, 74, 230 (1948).
2. W. H. Brattain and J. Bardeen, *Phys. Rev.*, 74, 231 (1948).
3. W. Shockley and G. L. Pearson, *Phys. Rev.*, 74, 232 (1948).
4. This paper was presented in part at the Chicago meeting of the American Physical Society, Nov. 26, 27, 1948. W. Shockley and the authors presented a paper on "The Electronic Theory of the Transistor" at the Berkeley meeting of the National Academy of Sciences, Nov. 15–17, 1948. A talk was given by one of the authors (W. H. B.) at the National Electronics Conference at Chicago, Nov. 4, 1948. A number of talks have been given at local meetings by J. A. Becker and other members of the Bell Laboratories Staff, as well as by the authors.
5. Properties and characteristics of the transistor are given by J. A. Becker and J. N. Shive in Elec. Eng. 68, 215 (1949). A coaxial form of transistor is described by W. E. Kock and R. L. Wallace, Jr. in Elec. Eng. 68, 222 (1949). See also "The Transistor, A Crystal Triode," D. G. F. and F. H. R., Electronics, September (1948) and a series of articles by S. Young White in Audio Eng., August through December, (1948).
6. H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*, McGraw-Hill, New York (1948).
7. J. H. Scaff and R. S. Ohl, *Bell System Tech. Jour.* 26, 1 (1947).
8. W. E. Meyerhof, *Phys. Rev.*, 71, 727 (1947).
9. J. Bardeen, *Phys. Rev.*, 71, 717 (1947).
10. W. H. Brattain and W. Shockley, *Phys. Rev.*, 72, p. 345(L) (1947).
11. W. H. Brattain, *Phys. Rev.*, 72, 345(L) (1947).
12. R. B. Gibney, formerly of Bell Telephone Laboratories, now at Los Alamos Scientific Laboratory, worked on chemical problems for the semi-conductor group, and the authors are grateful to him for a number of valuable ideas and for considerable assistance.
13. J. H. Scaff and H. C. Theuerer "Preparation of High Back Voltage Germanium Rectifiers" NDRC 14-555, Oct. 24, 1945—See reference 6, Chap. 12.
14. The surface treatment is described in reference 6, p. 369.
15. The transistor whose characteristics are given in Fig. 3 is one of an experimented pilot production which is under the general direction of J. A. Morton.
16. See, for example, A. H. Wilson *Semi-Conductors and Metals*, Cambridge University Press, London (1939) or F. Seitz, *The Modern Theory of Solids*, McGraw-Hill Book Co., Inc., New York, N.Y., (1940), Sec. 68.
17. The nature of the barrier is discussed in Sec. IV.
18. W. Schottky and E. Spenke, *Wiss. Veroff. Siemens-Werke*, 18, 225 (1939).
19. J. N. Shive, *Phys. Rev. 75*, 689 (1949).
20. J. R. Haynes, and W. Shockley, *Phys. Rev. 75*, 691 (1949).
21. R. Bray, K. Lark-Horovitz and R. N. Smith, *Phys. Rev.*, 72, 530 (1948).
22. R. Bray, *Phys. Rev.*, 74, 1218 (1948).
23. E. J. Ryder and W. Shockley, *Phys. Rev. 75*, 310 (1949).

24. This instrument was designed and built by H. R. Moore, who aided the authors a great deal in connection with instrumentation and circuit problems.
25. The surface had been oxidized, and potential probe measurements (ref. (2)) gave evidence for considerable surface conductivity.
26. Measured between centers of the contact areas.
27. Potential probe measurements on the same surface, given in reference (2), gave evidence of surface conductivity.
28. Unpublished data.
29. J. H. Scaff, H. C. Theuerer, and E. E. Schumacher, "P-type and N-type Silicon and the Formation of the Photovoltaic Barrier in Silicon" (in publication).
30. G. L. Pearson and J. Bardeen, *Phys. Rev.* March 1, 1949.
31. See, for example, reference 6, Chap. 3.
32. K. Lark-Horovitz, A. E. Middleton, E. P. Miller, and I. Walerstein, *Phys. Rev.* 69, 258 (1946).
33. Hall and resistivity data at the Bell Laboratories were obtained by G. L. Pearson on samples furnished by J. H. Scaff and H. C. Theuerer. Recent hall measurements of G. L. Pearson on single crystals of n- and p-type germanium give values of 2600 and 1700 cm^2/volt sec. for electrons and holes, respectively at room temperature. The latter value has been confirmed by J. R. Haynes by measurements of the drift velocity of holes injected into n-type germanium. These values are higher, particularly for electrons, than earlier measurements on polycrystalline samples. Use of the new values will modify some of the numerical estimates made herein, but the orders of magnitude, which are all that are significant, will not be affected. W. Ringer and H. Welker, Zeits. f. Naturforschung, **1**, 20 (1948) give a value of 2000 cm^2/volt sec. for high resistivity *n*-type germanium.
34. See R. H. Fowler, *Statistical Mechanics*, 2nd ed., Cambridge University Press, London (1936).
35. From unpublished data of K. M. Olsen.
36. N. F. Mott, *Proc. Roy. Soc.*, 171A, 27 (1939).
37. W. Schottky, *Zeits. f. Phys.*, 113, 367 (1939), *Phys. Zeits.*, 41, p. 570 (1940), *Zeits f. Phys.*, 118, p. 539 (1942). Also see reference 18.
38. See reference 6, Chap. 4.
39. S. Benzer, Progress Report, Contract No. W-36-039-SC-32020, Purdue University, Sept. 1–Nov. 30, 1946.
40. S. Benzer, *Phys. Rev.*, 71, 141 (1947).
41. Further evidence that the barrier is internal comes from some unpublished experiments of J. R. Haynes with the transistor. Using a fixed collector point, and keeping a fixed distance between emitter and collector, he varied the material used for the emitter point. He used semi-conductors as well as metals for the emitter point. While the impedance of the emitter point varied, it was found that equivalent emitter currents give changes in current at the collector of the same order for all materials used. It is believed that in all cases a large part of the forward current consists of holes.
42. The space charge of the holes in the inversion region of the barrier layer is neglected for simplicity.
43. Reference 6, Chap. 4.
44. S. Benzer "Temperature Dependence of High Voltage Germanium Rectifier D.C. Characteristics," *N.D.R.C.* 14-579, Purdue Univ., October 31, 1945. See reference 6, p. 376.
45. See, for example, E. H. Kennard, *Kinetic Theory of Gases*, McGraw-Hill, Inc., New York, N. Y. (1938) p. 63.
46. Reference 6, p. 377.
47. Obtained by plotting log g versus s^3/I_c. This plot is not a straight line, but has an upward curvature corresponding to an increase in τ with separation. The value given is a rough average, corresponding to s^3/I_c the order of 10^{-3} cm^3/amp.
48. One may expect that the mobility will depend on field strength when the drift velocity is as large as or is larger than thermal velocity. Since ours is a borderline case, the calculation using the low field mobility should be correct at least as to order of magnitude.

Lightning Current Observations in Buried Cable

By H. M. TRUEBLOOD and E. D. SUNDE

Results are given of observations of lightning currents, voltages, and charges in a buried cable over most of three lightning seasons. These are compared with theoretical expectations. Data regarding the incidence of lightning strokes to ground, as observed with automatic recording equipment, are also reported, together with comparisons with similar data published previously.

Introduction

LIGHTNING currents in buried telephone cable are of considerable importance in that they may cause excessive voltages between the cable sheath and the conductors with resultant insulation failure, and may also cause severe damage by crushing the cable or fusing holes in the sheath. The incidence of lightning strokes to buried cable, the resulting voltages, and lightning trouble expectancy, have therefore been subjects of theoretical, experimental, and field studies, which, together with operating experience, have pointed the way to improvements in the design of communication cable to minimize its liability to lightning damage, and in the application of remedial measures where excessive lightning trouble has been experienced.[1]

The territory around Atlanta, Georgia, has appeared to be particularly severe with respect to these lightning hazards, because of high earth resistivity and high thunderstorm rate. Buried cables initially installed in this territory were accordingly provided with protective measures in the form of increased core-sheath insulation and shield wires buried with the cable. In spite of these measures, however, a substantially higher rate of lightning damage than anticipated was experienced, as a result of which a new design was adopted for the transcontinental coaxial cable westward from Atlanta. In this cable, the lead sheath was insulated from an outside corrugated 10-mil copper shield by a 100-mil layer of thermoplastic insulation intended to prevent the entrance of lightning currents into the sheath and thereby to minimize voltages between the sheath and the cable conductors.

Simulative tests with surge currents, believed to have a wave shape representative of lightning stroke current, had indicated satisfactory agreement between measured and calculated voltages between sheath and cable conductors. It appeared, therefore, that the departure from predicted

[1] References are listed at end of paper.

lightning trouble expectancy in the earlier cables was due to one or more of the following conditions: a higher rate of occurrence of lightning strokes during thunderstorms, higher stroke currents than in other parts of the country, a longer duration of the lightning currents than assumed, or a higher incidence of strokes to buried cable than predicted theoretically. The observations described here, the larger part of which have extended over a period of three lightning seasons, were intended to secure information on these points. The data forming the principal subject of this paper were obtained from a section of the coaxial cable mentioned above, which for a number of reasons was particularly suitable for the purpose.

I. Theoretical Expectations

1.0 *General*

As mentioned above, the observations were made on a cable route through territory of high thunderstorm rate and high earth resistivity, both of which facilitate measurements of currents along the cable. As a result of the high thunderstorm rate, the incidence of strokes to ground is high, and because of the high earth resistivity, the number of strokes to ground near the cable which flash to latter is also high. Another result of the high earth resistivity is that the attenuation of current along the cable is relatively low, so that currents and voltages may be observed at appreciable distances along the cable from the points of the lightning strokes. The rate of attenuation is, furthermore, smaller the longer the duration of the lightning current, that is, the longer the time to half-value. Since lightning trouble experience in this territory indicated the possibility of currents of rather long durations, this was an additional factor favorable to the purposes of the tests, although, like the others, it increases the liability of cables to lightning damage.

Though the relationships of the various factors mentioned above to earth resistivity and to lightning current wave shape have been dealt with in detail in the study[1] referred to above, they are briefly reviewed here to facilitate comparisons with and discussions of the observations.

1.1 *Incidence of Strokes to Buried Cable*

The current in a lightning stroke to ground spreads in all directions from the point where it enters the earth. If a cable is in the vicinity, it will provide a low resistance path, so that much of the current will flow to the cable and in both directions along the sheath to remote points. The current in the ground between the lightning channel and the cable may give rise to a voltage drop along the surface of the earth sufficient to exceed the breakdown gradient of the soil, particularly when the earth resistivity is

high. The lightning stroke will then arc directly to the cable from the point where it enters the ground, often at the base of a tree. Furrows exceeding 100 feet in length have been found along the ground path of such an arc.

For a crest current J in the lightning stroke, the arcing distance in meters is given by[1]

$$r = k(J\rho)^{1/2} \tag{1}$$

where J is in kiloamperes, ρ is the earth resistivity in meter-ohms and k is a constant depending on the surface breakdown gradient of the soil. Low resistivity soil, up to $\rho = 100$ meter-ohms, has an average breakdown gradient of about 2500 volts/cm, and the corresponding value of k is about .08. For high resistivity soil, $\rho = 1000$ meter-ohms and up, the average breakdown gradient is about 5000 volts/cm and $k \cong .047$. Thus, for an earth resistivity of 2000 meter-ohms, and $J = 100$ ka, $r \cong 21$ meters or 70 feet.

The number of strokes arcing to a cable of length s may conveniently be expressed as

$$N = 2\bar{r}sn \tag{2}$$

where n is the number of strokes to ground per unit area, $\bar{r}$ is an equivalent arcing distance, and $2\bar{r}s$ an equivalent area near the cable within which the cable is assumed to attract all lightning strokes. In obtaining $\bar{r}$, the number of strokes arcing to the cable from various distances r as given by (1), depending on the current in the strokes, must be evaluated. This number and the equivalent arcing distance will thus depend on the crest current distribution of lightning strokes. For the distribution curve shown by Curve 1 in Fig. 1, the effective distance in meters is[1]

$$\begin{aligned} \bar{r} &\cong .36\rho^{1/2} \text{ when } \rho \leq 100 \text{ meter-ohms} \\ &\cong .22\rho^{1/2} \text{ when } \rho \geq 1000 \text{ meter-ohms} \end{aligned} \tag{3}$$

Thus, for soil having an average resistivity near the surface (to a depth of at least 10 meters) of 2000 meter-ohms, $\bar{r} \cong 10$ meters $= 33$ feet.

A cable will thus collect direct lightning strokes for an effective distance $\bar{r}$ to either side of it, and when the incidence of strokes to ground per unit area is known, the number of strokes to a cable of a given length may readily be calculated. The incidence of strokes to ground has been estimated, on the average, as about 2.4 per square mile for each 10 thunderstorm days, and the corresponding expectancy of strokes to a buried cable, per 100 miles of length, is about 2.1 for 10 thunderstorm days when the earth resistivity is 1000 meter-ohms and 3.0 when it is 2000 meter-ohms.

The distribution of the crest currents in direct strokes to a cable may be

obtained by use of Curve 2 in Fig. 1, which is a theoretical curve derived from Curve 1. Thus for a total of 2.1 for 10 thunderstorm days, the incidence of strokes exceeding 60 ka is $2.1 \cdot 0.2 = 0.42$. In Fig. 2 are shown crest current distribution curves for cable currents due to direct strokes obtained in this manner, together with distribution curves for currents due to both direct strokes and strokes to ground not arcing to the cable. The latter curves may be obtained by methods similar to those used in evaluating curves for the lightning trouble expectancy of buried cable, which are

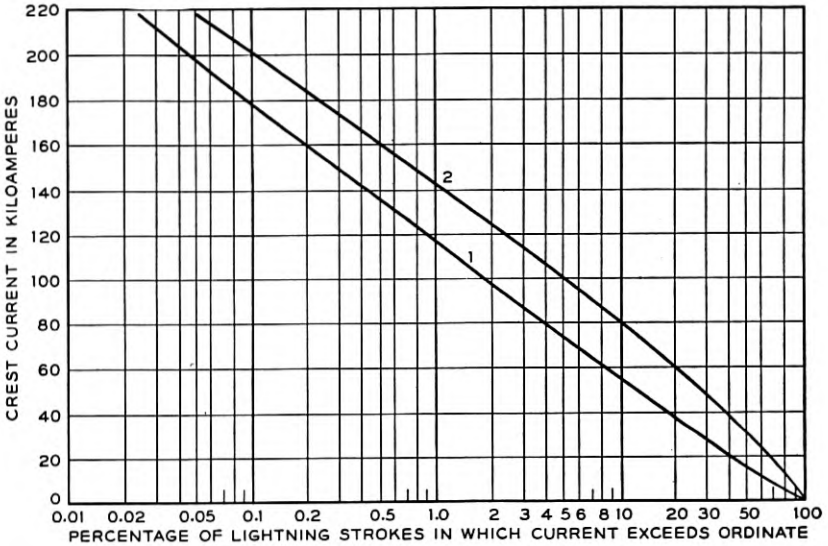


Fig. 1—Distribution of crest currents in lightning strokes.

Curve 1. Currents in strokes to transmission line ground structures, based on 4410 measurements, 2721 in U. S. and 1689 in Europe.

Curve 2. Currents in strokes to buried structures derived from Curve 1.

shown in Fig. 3 for cable having a dielectric strength of 2 kv between the sheath and the cable conductors.[1] The latter curves may, in fact, be used to find the incidence of cable currents of various crest values due to direct strokes and strokes to ground, by calculating the cable currents required to produce 2 kv between the sheath and the core corresponding to various sheath resistances shown in Fig. 3. Thus for a sheath resistance of 2 ohms per mile and an earth resistivity of 1000 meter-ohms, this current is 14.2 ka (see Section 1.3) and for a sheath resistance of 1 ohm, it is 28.4 ka, etc., as plotted in Fig. 2 for an earth resistivity of 1000 meter-ohms.

From the above it follows that a verification of the distribution curves in Fig. 2 by observations of lightning currents in buried cable would apply

equally well to the curves in Fig. 3, which have been used to evaluate the liability of such cable to lightning damage.

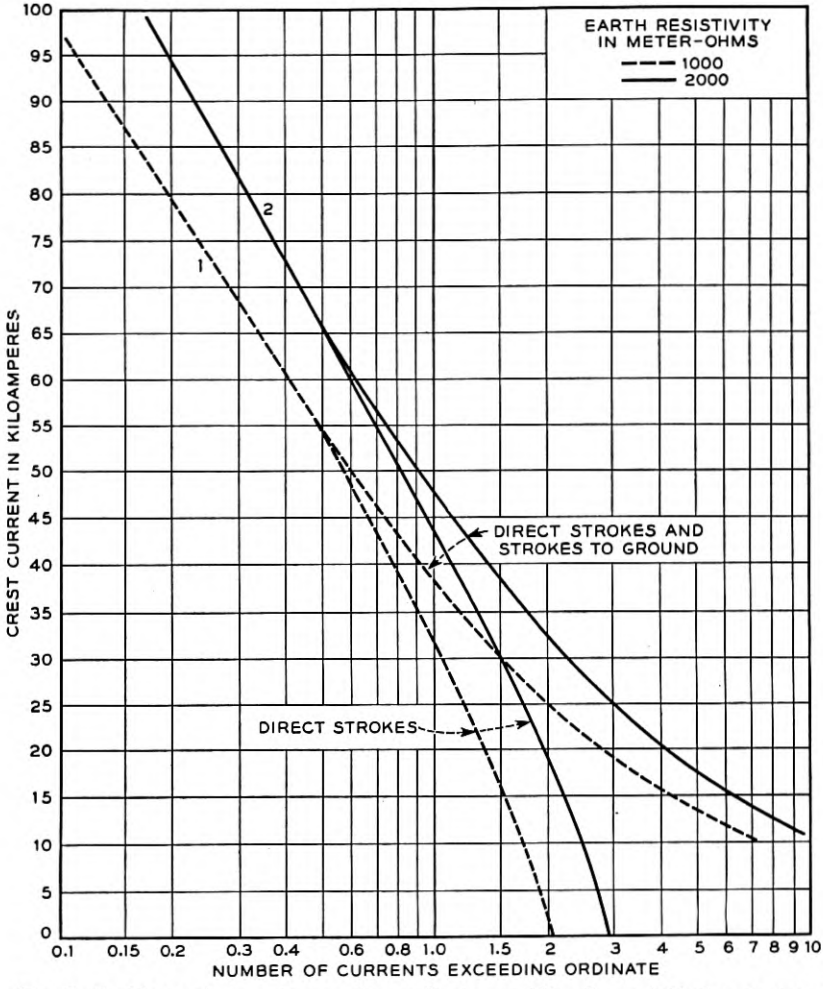


Fig. 2—Incidence of currents exceeding ordinate per 100 miles and 10 thunderstorm days.

1.2 *Propagation of Currents Along Cable*

The current entering the cable at or near the stroke point, depending on whether a direct stroke or a nearby stroke to ground is involved, is attenuated as it travels along the sheath towards remote points. The current leaving the sheath must flow through the adjacent soil, and the leak-

age current is therefore smaller the higher the soil resistivity. Thus the current will travel farther the higher the earth resistivity. It has been shown elsewhere[1, 2] that a sinusoidal current would be propagated as

$$I(x) = I(0)\, e^{-\Gamma x} \tag{4}$$

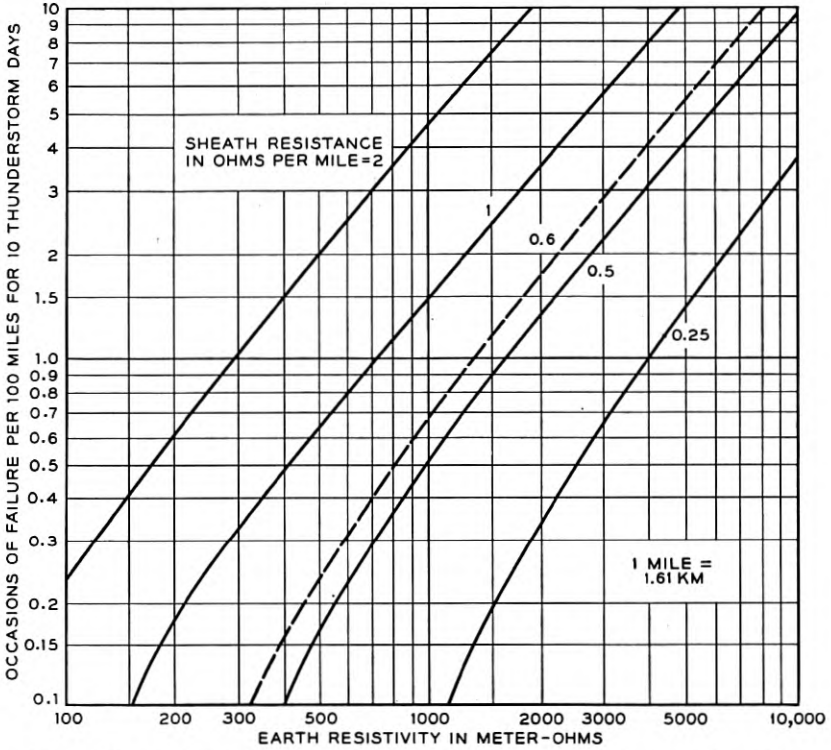


Fig. 3—Theoretical lightning trouble expectancy curves showing number of times insulation failures due to excessive voltages would be expected per 100 miles for 10 thunderstorm days, for cables having sheath resistances as indicated on curves and 2000 volts insulation between core and sheath. Dashed line represents full size cable.

where $I(0)$ is the current in the sheath in one direction from the stroke point, $I(x)$ the current at the distance x, and the propagation constant Γ per meter is given by:

$$\Gamma = \sqrt{i\omega\nu/2\rho} \tag{5}$$

where $\omega = 2\pi f$, ν = inductivity of the earth = $1.257 \cdot 10^{-6}$ henry/meter, and ρ is the earth resistivity in meter-ohms.

Let it be assumed that the current at the distance x has been evaluated for a given earth resistivity ρ and radian frequency ω. If the earth re-

sistivity is increased by a factor k, or if ω is decreased by the same factor, it is evident from (4) and (5) that the same current will be obtained at the distance $x_1 = k^{1/2}x$. Thus, if the earth resistivity is increased by $k = 4$, $x_1 = 2x$ and the current attenuation will be half as great as before. This rule applies to surge-currents of a given wave-shape as well, since they may be regarded as composed of sinusoidal components, each of which would

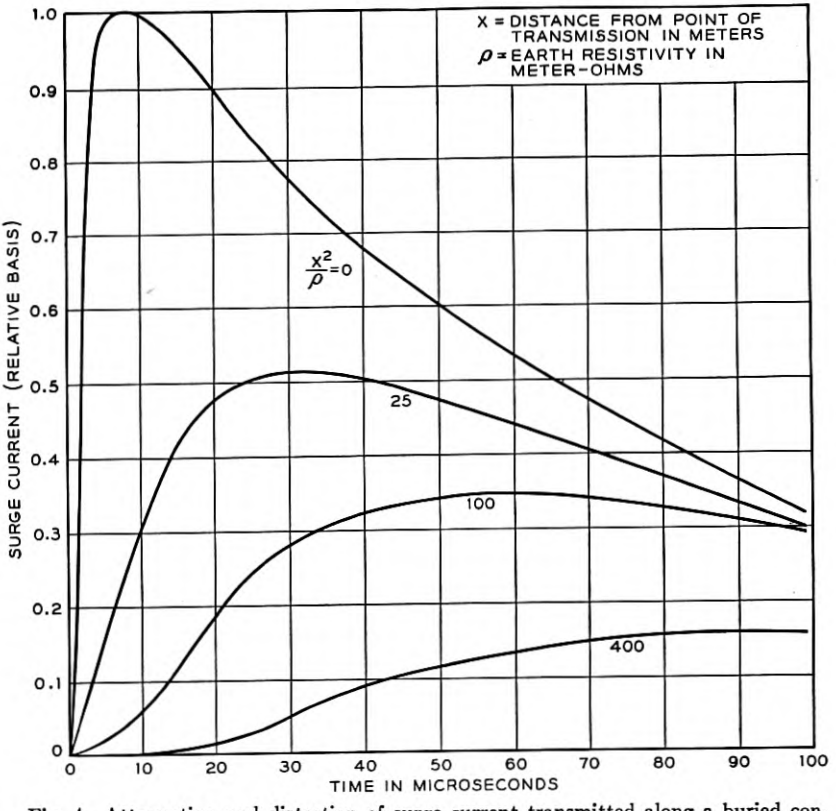


Fig. 4—Attenuation and distortion of surge current transmitted along a buried conductor.

travel farther by the factor $k^{1/2}$. Furthermore, it follows by the same reasoning that for surge-currents of congruent wave-shapes, but different durations, the rate of attenuation is inversely proportional to the square root of the duration. That is to say, let in one case the current reach its crest value at the stroke point in 5 microseconds and its half-value in 65 microseconds, and let the crest current at a given distance x have a certain value. Then, if in another case the current reaches the same stroke-point crest

value in 20 microseconds, with half-value in 260 microseconds, the same crest current as that found before at x is obtained at twice the distance x. This follows because all component frequencies of the first surge are related to corresponding components of the second surge by the same factor, viz. 4.

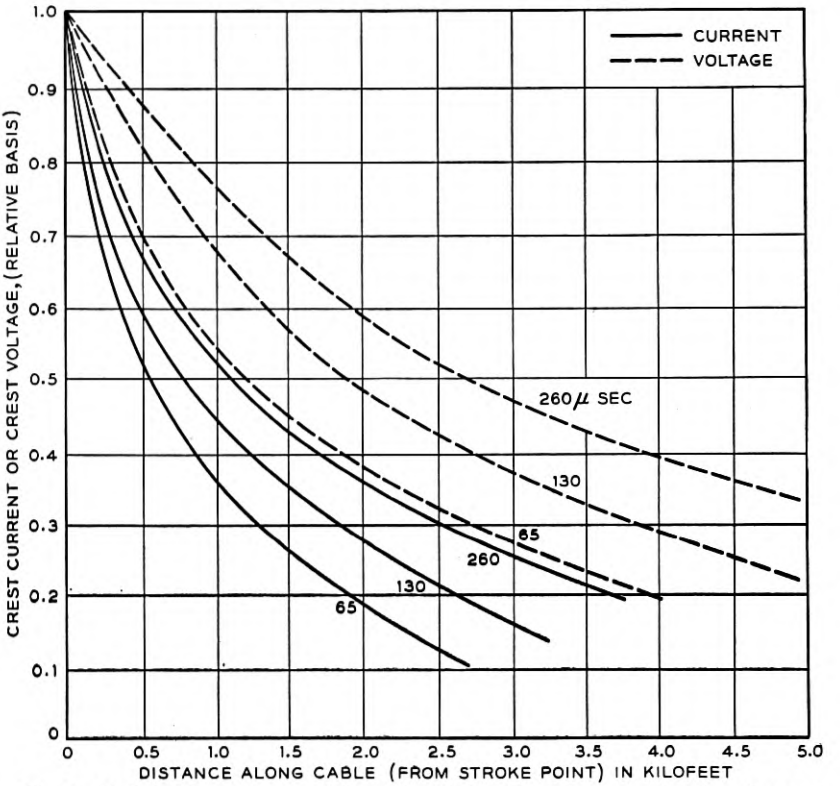


Fig. 5—Variation in crest current in cable and in voltage between sheath and copper shield for stroke currents having various times to half-value as indicated on curves, for an earth resistivity of 1000 meter-ohms.

In Fig. 4 are shown curves[2] from which the attenuation may be obtained for a surge-current reaching its crest value in 5 microseconds and its half-value in 65 microseconds, as shown by the curve for $x^2/\rho = 0$. The crest current has attenuated by 50 per cent when $x^2/\rho = 25$ or $x = 5\rho^{1/2}$. Thus, for an earth resistivity of 1000 meter-ohms, $x = 160$ meters $= 525$ feet when the time to half-value of the current at the stroke point is 65 microseconds, while $x = 1050$ feet when the current at the stroke point reaches its half-value in 260 microseconds. In Fig. 5 are shown crest current

attenuation curves obtained in this manner, together with similar curves for the voltage, between the sheath and the core conductor of an ordinary cable, or between the copper shield and the lead sheath of the cable on which these observations were made.

1.3 *Crest Values and Attenuation of Voltages*

The current along the copper shield produces a voltage between this shield and the lead sheath, due to the resistance drop along the shield from the stroke point to a point sufficiently remote for the current in the shield to have become negligible. This voltage is proportional to the unit-length resistance of the copper shield. From the considerations of the preceding section, it follows that the voltage will be proportional to the square root of the earth resistivity and, if the wave-shape remains congruent but the duration of the current is changed, that it will be proportional to the square root of the duration or of the time to half-value. These two propositions follow from the fact that the voltage is proportional to the distance traveled by the current before it is attenuated to a given value.

The crest voltage between the sheath and the cable conductors of an ordinary cable, or between the copper shield and the insulated lead sheath of a cable of the type on which these observations were made, is given by the following expression for a current reaching its half-value in 65 microseconds:

$$V = 2.25\, JR\rho^{1/2} \tag{6}$$

where V is in volts and J is the crest current in kiloamperes, R the resistance per mile of the outer envelope (in this case the copper shield), and ρ the earth resistivity in meter-ohms. This formula follows from expressions given in the paper referred to previously, which also contains curves from which the voltage attenuation along the cable shown in Fig. 5 may be obtained. For a resistance of .7 ohm/mile, which is that of the copper shield, and $\rho = 1000$ meter-ohms, the crest voltage for a current of 1 ka would thus be 50 volts; and, for a crest current of 200 ka, 10,000 volts. If the dielectric strength of the thermoplastic insulation exceeds 10 kv, the liability of such cable to lightning damage would thus be small, unless the time to half-value of the current substantially exceeds 65 microseconds.

II. Experimental Installation

2.0 *General*

From the preceding discussion it is seen that a lightning current dropping to half-value in some 50 to 75 microseconds, which is of the wave-shape ordinarily assumed, would attenuate to half its crest value in 500 to 1000 feet when the earth resistivity is from 1000 to 2000 meter-ohms. With

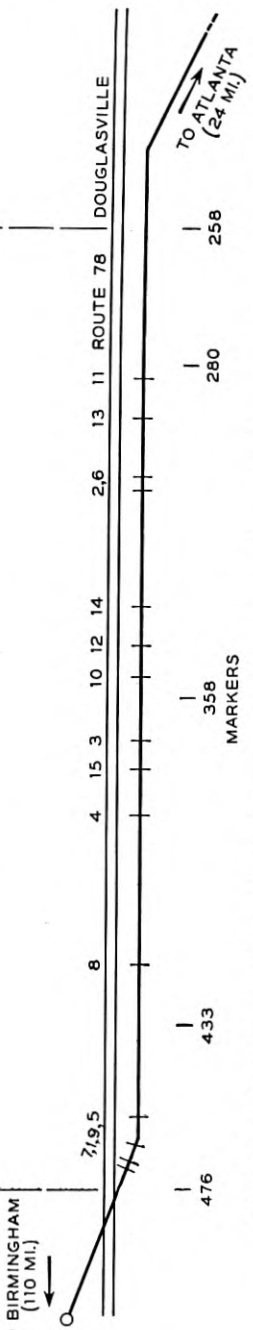


Fig. 6—Location of test section with points of 15 numbered maximum observed cable currents indicated (Table I).

test points along the cable at intervals of about 2300 feet, as employed in the observations, it should thus be possible to secure substantial indications at a number of points along the test section, although closer spacings would of course be desirable.

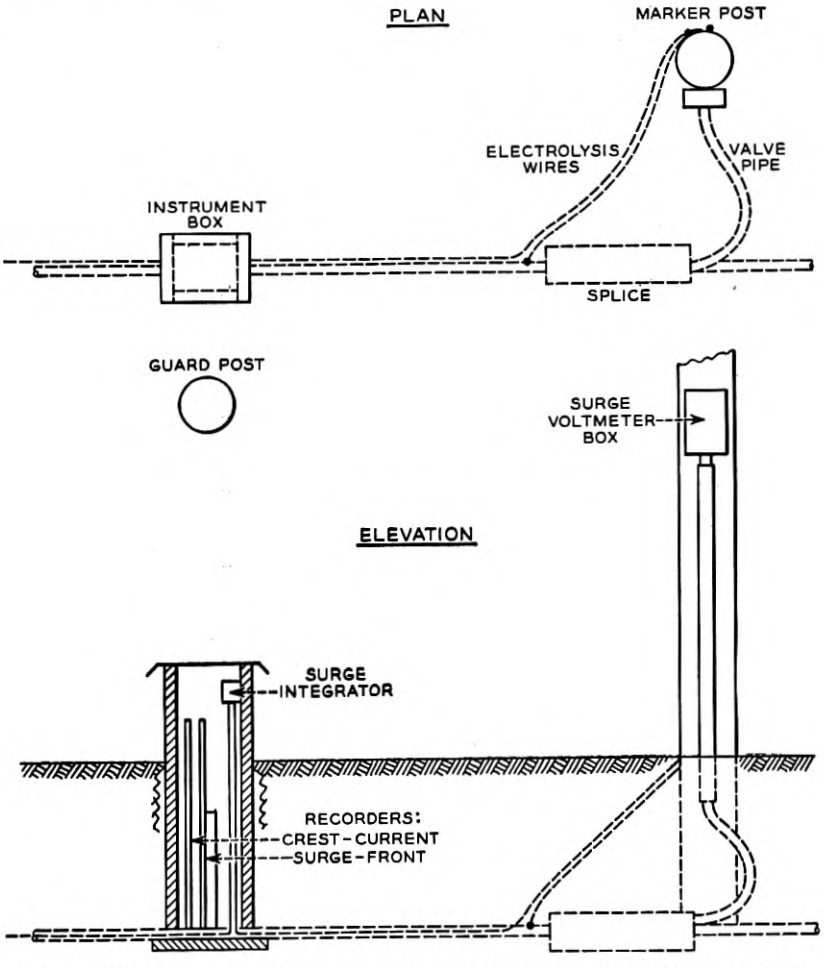


Fig. 7—Typical instrument arrangement at splice points about 2300 feet apart.

In Fig. 6 is shown the 21.5-mile test section and in Fig. 7 a typical installation at every second splice in the cable. At these points the lead sheath is accessible through a lead gas-pressure pipe extending to a marker post, an arrangement which was utilized in making measurements of voltage between the sheath and the outside copper shield. At the same points an

insulated wire is installed along the outside of the copper shield for measurement of voltage drop in the copper jacket, in connection with routine corrosion surveys. This facilitated measurements of the charge transferred along the shield by lightning currents, as described in the following. Magnetic link instruments intended to measure the steepness of the wave front were also employed, but lacked the sensitivity required for accurate measurements and are not discussed further here.

In addition to these measurements of current, charge, and voltage, involving the cable structure, observations were also made of the incidence of strokes to ground as described later in this paper.

2.1 *Crest Current Measurements*

To measure crest currents in the cable, magnetic links[3] were mounted at distances of 1.6, 5.7, 12.7, and 36.4 inches from the center of the cable, to cover a current range from 1 to 220 ka. The readings on these magnetic links indicated the total current in the cable, that is, the sum of the currents in the copper shield, the lead sheath, and inside cable conductors.

2.2 *Measurements of Charge*

These measurements were made by means of *surge integrators.*[4] In principle this instrument consists of a shunt R_0 across which is connected a coil of inductance L. When a surge current $I_0(t)$ passes through the shunt, the current $I(t)$ in the coil is given by:

$$L \frac{dI(t)}{dt} = R_0 I_0(t)$$

$$I(t) = \frac{R_0}{L} \int_0^t I_0(t)\, dt$$

$$= \frac{R_0}{L} Q_0(t)$$

where $Q_0(t)$ is the charge which has passed through the shunt up to the time t. By measuring the crest value $\hat{I}$ of the current $I(t)$, the total charge may be obtained from the relation:

$$\hat{Q}_0 = \frac{L}{R_0} \hat{I}$$

This relation is always valid if the surge current rises to a peak value and then decays uniformly. The relation should provide a good approximation to the total charge conveyed by natural lightning strokes, even if there are small oscillations.

The shunt R_0 consisted of about 26 feet of the copper shield over the cable,

which had a resistance of about 3.5 milliohms. The inductance L consisted of two coils connected in series, each containing a magnetic link. The larger of these coils had 187 turns of copper wire, the smaller 50 turns. The larger coil provided the greater sensitivity, on account of the more intense magnetization of the link. The relation between the current $\hat{I}$ in the coil and the deflection on the magnetic link meter[3] used to measure the intensity of magnetization was obtained by calibration with direct current.

The inductance of the two coils in series was about 700 microhenries and the d-c resistance about .39 ohms. The time constant of the coils L/R was thus about 1800 microseconds, which is large compared to the duration of the main surge of a lightning stroke, which may last for 100 to 500 microseconds. The instrument will thus effectively integrate the main surge, but will not record the charge which may be caused by a small tail current of much longer duration.

The measurements of charge were made mainly to determine the time to half-value of the currents. The theoretical curves in Part I and elsewhere in this paper are based on a current of the form:

$$J(t) = 1.15\hat{J}\ (e^{-at} - e^{-bt})$$

where $a = .013 \cdot 10^6$, $b = .5 \cdot 10^6$ for a current reaching its crest value $\hat{J}$ in about 5 microseconds and decaying to its half-value in 65 microseconds. If $\alpha = R/L = .00056 \cdot 10^6$ for the surge integrator, the total charge recorded for a current of the above wave shape is

$$\hat{Q} = \hat{J}1.15\left[\frac{1}{a+\alpha} - \frac{1}{b+\alpha}\right] = \hat{J} \cdot 83 \cdot 10^{-6}$$

for a current decaying to its half-value in 65 microseconds. The relationship between $\hat{Q}/\hat{J}$ and the time to half-value, $t_{1/2}$, is as follows for currents reaching their half-values in 65, 130, 260, and 520 microseconds:

$\hat{Q}/\hat{J}$:	83	160	295	540 microseconds
$t_{1/2}$:	65	130	260	520 microseconds

From a curve of $\hat{Q}/\hat{J}$ versus $t_{1/2}$, the time to half-value may be obtained from the observed ratio of charge to crest current. The values given later on, in Table I, were obtained in this manner. If a triangular wave shape had been assumed, the times to half-value would have been $\hat{Q}/\hat{J}$ and therefore somewhat longer.

2.3 *Measurements of Voltage Between Sheath and Copper Shield*

These measurements were made by means of a magnetic link voltmeter consisting of a solenoid of inductance L containing the magnetic link and a

series resistance, R. When a constant voltage E is suddenly applied the current through the coil is

$$I = \frac{E}{R}(1 - e^{-(R/L)t})$$

If the time constant L/R is small in comparison with the time to crest value of a variable voltage to be measured, there will be no material delay between the crest value of the voltage and that of the current in the coil. The applied voltages may therefore be obtained by multiplying the coil current as obtained from the magnetic link reading by the series resistance, provided the latter is much greater than the impedance of the circuit to which the instrument is connected. Since the voltmeter in the present case was designed to measure the voltage between the sheath and the copper shield, and the surge impedance of this test circuit is less than 3 ohms, comparatively low values of series resistance could be used. Three separate solenoids and series resistances were used, to provide three voltage ranges, from 0 to 1.5 kv, 0 to 4 and 0 to 9 kv.

2.4 *Magnetic Link Calibrations*

When several magnetic links, which have been exposed to the same field, are inserted in the magnetic link meter, considerable differences in the deflections may be observed due to variations in the material of the links. For this reason, all links used in this installation were placed in a magnetic field of 257 gauss and were then classified according to their response in the magnetic link meter. This field was such as to produce deflections in the most useful part of the meter range, centering around mid-scale.

By this method the magnetic links used in the installation were divided into four classes, in accordance with the ratio of the deflection observed on the magnetic link meter for the link in question to the average deflection for all links. The maximum deviation from the average in each class was about $\pm$ 3 per cent. Instruments of the same type at all installations were provided with links of the same class, to minimize inaccuracies.

2.5 *Observations of Incidence of Strokes to Ground*

To obtain data on the incidence of strokes to ground, observations were made at one location within the test section, by means of an automatically operated magnetic wire recorder arranged to record thunder picked up by a microphone. The recorder was provided with a triggering arrangement which put it in operation on the approach of a thunderstorm, by action of the voltage induced in an antenna by lightning current. The induced voltage from a given lightning stroke was also made to record itself upon the magnetic wire; and, from the delay between this indication and the

recorded thunder, the distance to the lightning stroke could be determined upon play-back of the wire record. In this manner the number of strokes to ground within areas of various radii around the observation point could be ascertained, and thus the incidence of strokes to ground. These observations were made during the 1947 and 1948 lightning seasons.

III. Results of Observations

3.0 *General*

From the preceding discussion of theoretical expectations and of the experimental arrangement, it is evident that considerable attenuation would take place between the stroke point and the nearest test points on either side, for a stroke midway between the latter. Accurate evaluation of the maximum current, voltage, and charge, and of the current wave-shape, would thus be rather difficult for strokes nearly midway between test points, since these quantities would have to be evaluated by extrapolation from the observations at the points along the cable. Such extrapolation is rendered more accurate by employing the theoretical attenuation curves given in Fig. 5. This has been done for the currents, by trial and error, until the current wave-shape derived at the stroke point approximately coincided with that assumed for the attenuation curve used in the extrapolation.

These observations involving the cable structure extended over the greater part of three lightning seasons, and included a total of 108 thunderstorm days, 35 in 1946, 38 in 1947 and 35 in 1948. The average number of thunderstorm days per year as recorded by the Atlanta Weather Bureau is 49, which compares with about 60 given on the U. S. Weather Bureau map.[5] The more significant data are tabulated in Table I.

In the following, the observations made of currents, voltages and charges along the cable are discussed for a number of the more important strokes and compared with theoretical expectations. This is followed by a discussion of the observed incidence of cable currents of substantial magnitude and of the incidence of strokes to ground observed at one location along the route and at a second point in the northern part of the country.

3.1 *Wave-Shapes and Attenuation of Currents*

In Fig. 8 is shown the distribution along the cable of the crest currents, the crest voltages, and the charges, for the most severe direct stroke measured, which had a crest value of 70 ka and a total charge of 11 coulombs. This stroke occurred to a 35-foot antenna connected to the cable and used in oscillographic observations of induced voltages due to strokes to ground, as another means of securing data on stroke currents. At this same point

TABLE I
SUMMARY OF CURRENTS EXCEEDING 10 KA

Stroke No. *	Year	Date	Extrapolated Crest Current (Kiloamperes)	Extrapolated Max. Charge (Coulombs)	Derived Time to Half-value (Microseconds)	Shown In
1	1946	April 7	30	7	430	
2	35	June 21	20	4	170	
3	Thunderstorm	June 21	15	14	950	
4	Days	Aug. 3	16	3.6	190	Fig. 11
5		Aug. 3	16	4	240	
6		Aug. 25	50	11.2	190	Fig. 9
7	1947	May —	20	12	580	
8	38	July 28	14	4	240	
9	Thunderstorm	Aug. 5	14	3.8	180	
10	Days	Aug. 18	10	6.4	620	
11	1948	April 19 to 23	20	8	370	
12	35	July 14	15	7.6	530	
13	Thunderstorm	July 28	70**	11.2	130	Fig. 8
14	Days	July 28	50	8.8	140	Fig. 10
15		Aug. 4	12	12	1000	

* For location of strokes, see Fig. 6.
** Measured at stroke point.

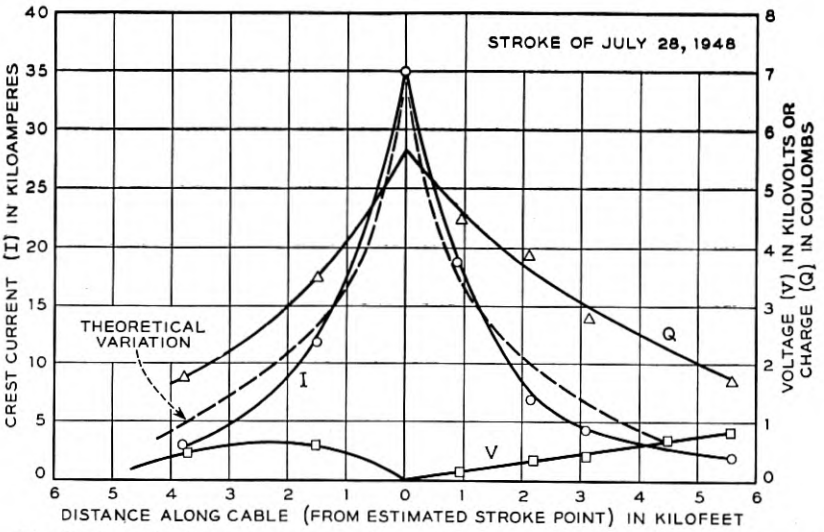


Fig. 8—Variation in crest current, voltage, and charge along cable for max. observed stroke current of 70 ka, having a time to half-value of 130 microseconds. Dashed curve shows theoretical variation of current for this time to half-value and an earth resistivity of 1200 meter-ohms. Variation in voltage between sheath and copper shield indicates breakdown between sheath and copper shield near stroke point.

simulative surge tests had been made three years before to obtain data on voltages in the cable due to surge currents, and tests had also been made of the dielectric strength of the thermoplastic insulation between the sheath and the copper shield. These latter tests disclosed low dielectric strength in the thermoplastic insulation at the location of the antenna referred to above, a defect which was repaired at the time. The voltage curve in Fig.

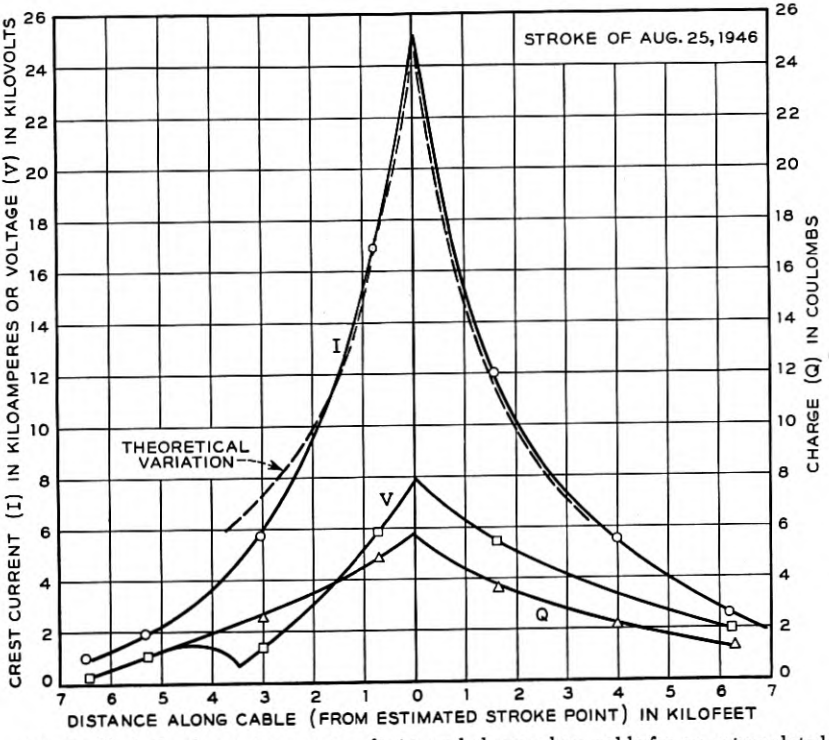


Fig. 9—Variation in crest current, voltage, and charge along cable for an extrapolated stroke current of 50 ka, having an estimated time to half-value of 190 microseconds. Dashed curve shows theoretical variation of current for this time to half-value and an earth resistivity of 1700 meter-ohms.

8 indicates that breakdown of the thermoplastic insulation occurred as a result of excessive voltage between the sheath and the copper jacket, but no other damage to the cable resulted. In Fig. 8 is also shown the theoretical variation in crest current along the cable, for a uniform earth of 1200 meter-ohms resistivity, for a stroke current reaching its half-value in 130 microseconds, as obtained from the crest current and charge at the stroke point.

In Fig. 9 are shown similar curves for an extrapolated stroke current of 50 ka and 190 microseconds to half-value, together with the theoretical

attenuation curve for the current for 1700 meter-ohms, which appears to provide a fairly satisfactory check on the extrapolation. The maximum observed voltage obtained by extrapolation is about 8 kv, which compares with 5.6 calculated as outlined in Section 1.3. The higher observed voltage may be due to a long duration tail current of small value, which may increase the voltage appreciably because of its slow rate of attenuation along the cable.

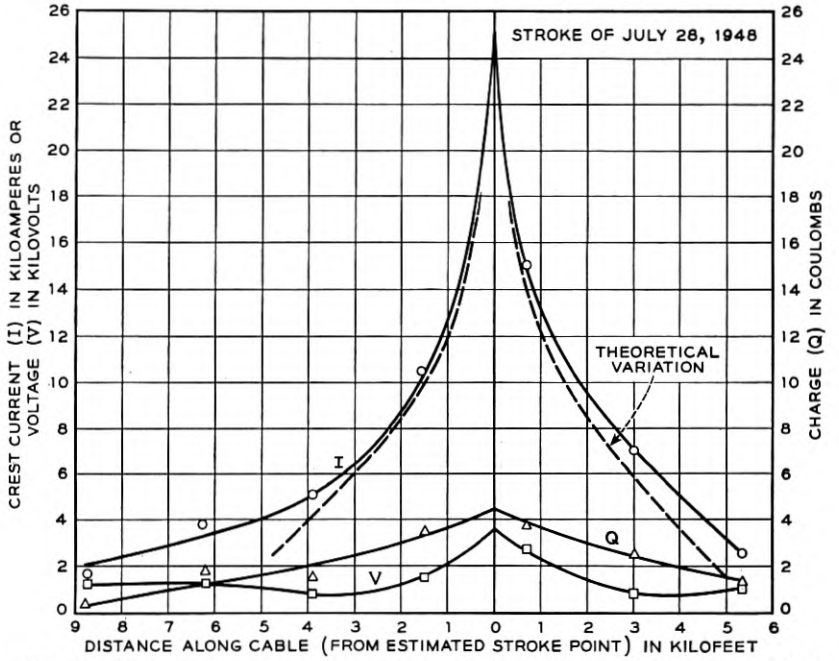


Fig. 10—Variation in crest current, voltage, and charge along cable for an extrapolated stroke current of 50 ka, having a time to half-value of 140 microseconds. Dashed curve shows theoretical variation of current for an earth resistivity of 1200 meter-ohms.

In Fig. 10 are shown similar curves for an extrapolated stroke current of 50 ka, reaching its half-value in 140 microseconds, together with theoretical attenuation curve for the current, for an earth resistivity of 1200 meter-ohms. The maximum extrapolated voltage is in this case about 3.5 kv, as compared with 4.1 calculated for 1200 meter-ohms.

The curves in Fig. 11 are for a fairly low extrapolated current, 16 ka, reaching its half-value in 190 microseconds. Again satisfactory agreement between observed and calculated attenuation is obtained. The maximum observed voltage of 1.5 kv in this case agrees with that calculated for an earth resistivity of 1200 meter-ohms.

Some of the other observations, not reproduced here, were less consistent than those shown, probably due to the combined effects of more than one stroke; but they permitted fairly satisfactory determinations of crest currents and times to half-value.

From Table I it appears that the minimum duration to half-value is about 130 microseconds, the maximum about 1000 and that the average for the three most severe strokes is about 150 microseconds. In general the duration appears to be longer the lower the crest currents, the average for currents of 20 ka and less being about 500 microseconds to half-value.

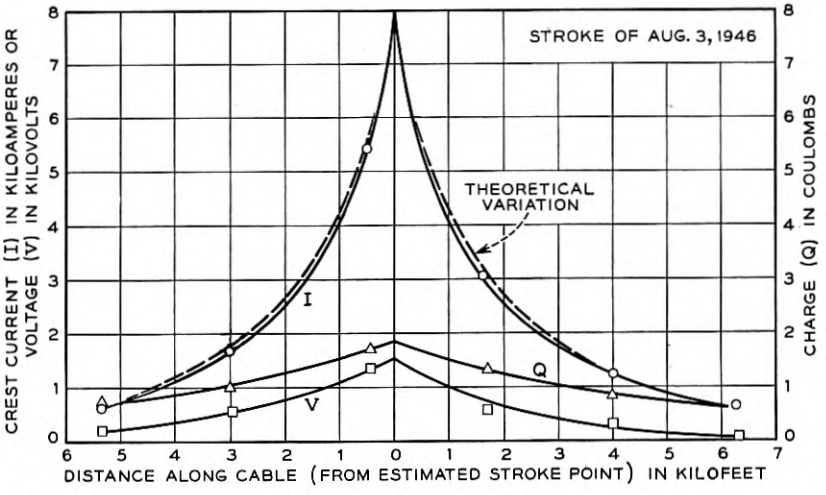


Fig. 11—Variation in crest current, voltage, and charge along cable for an extrapolated stroke current of 16 ka, having a time to half-value of 190 microseconds. Dashed curve shows theoretical variation of current for this time to half-value and an earth resistivity of 1200 meter-ohms.

In the above derivations a simple wave-shape was assumed, although actually it is likely to be rather complex with many fluctuations. The use of an equivalent simple wave-shape is, however, permissible in evaluating the liability to lightning damage, since statistical results rather than the wave-shapes of individual currents are of main importance.

Regarding the cause for the long duration of the currents, it appears to be inherent in meteorological rather than geological conditions, as for the wave-shapes of lightning currents in general. The significance of meteorological conditions is also indicated by the observations discussed in Section 3.3. In the case of strokes to the cable, the latter provides a path of very low impedance compared to that of the lightning channel, so that the wave-shape is determined by the impedance of the lightning channel and the distribution of charge along the leader and in the cloud. This is also true

for strokes to ground not arcing to the cable, at least during the time required for the tip of the channel to propagate from the earth towards the cloud, which may be of the order of 50 to 100 microseconds, depending on the height of the cloud. During this inverval ionization of the soil around the base of the channel provides a path in the earth of low impedance compared to that of the channel, as shown in the paper referred to previously. It is possible of course that, during later stages of the discharge, the resistivity of the earth to some extent may limit the current, as the impedance of the completely ionized channel will then be lower and that in the earth higher because of the lower current in the earth with resultant decrease in ionization. This, however, would tend to reduce the current and thereby decrease rather than increase the time to half-value, and at the same time it would tend to cause a long duration tail current of low magnitude.

3.2 *Incidence of Cable Currents of Various Crest Values*

In Fig. 12 is shown the number of observed currents exceeding various crest values, together with curves of the crest current distribution expected on the basis of the theoretical curves given in Fig. 2. The latter curves, together with those in Fig. 3, are based on an incidence of strokes to ground of 2.4 per square mile for 10 thunderstorm days, a value derived from the rate of strokes to transmission line ground structures, as outlined in the paper referred to previously. Although the observations appear to be in fairly satisfactory agreement with theoretical expectations, a total of 15 currents is hardly sufficient as a check of the theoretical curves, particularly since the latter presume a uniform earth structure.

The intersections of the theoretical curves (Fig. 12) with the axis of abscissae indicate that from five to seven of the currents were due to direct strokes. Actually, visual evidence of direct strokes was found in but two cases, in which the strokes occurred to and damaged test equipment. This does not preclude the possibility of additional direct strokes, as evidence thereof in the absence of cable damage may easily escape detection.

3.3 *Incidence of Strokes to Ground*

In Fig. 13 is shown the incidence of strokes to ground observed during 1947 and 1948 from one point within the test section, by the method described in Section 2.5. In the same figure are shown the results of similar observations by the same method, made at one location in New Jersey during 1948 for purposes of comparison. Published data obtained from direct visual and aural observations at one locality in Massachusetts[6] are also shown in the same figure.

As shown by the curves in Fig. 13, the observed or apparent incidence of strokes to ground diminishes as the radius of the observation area increases,

for the reason that more of the remote than of the near strokes of low intensity escape observation. To find the actual incidence, a curve of the apparent incidence versus the radius of the observation area may be extrapolated to an area of zero radius. On account of the comparatively few observations for small radii, however, such extrapolation is rather inaccu-

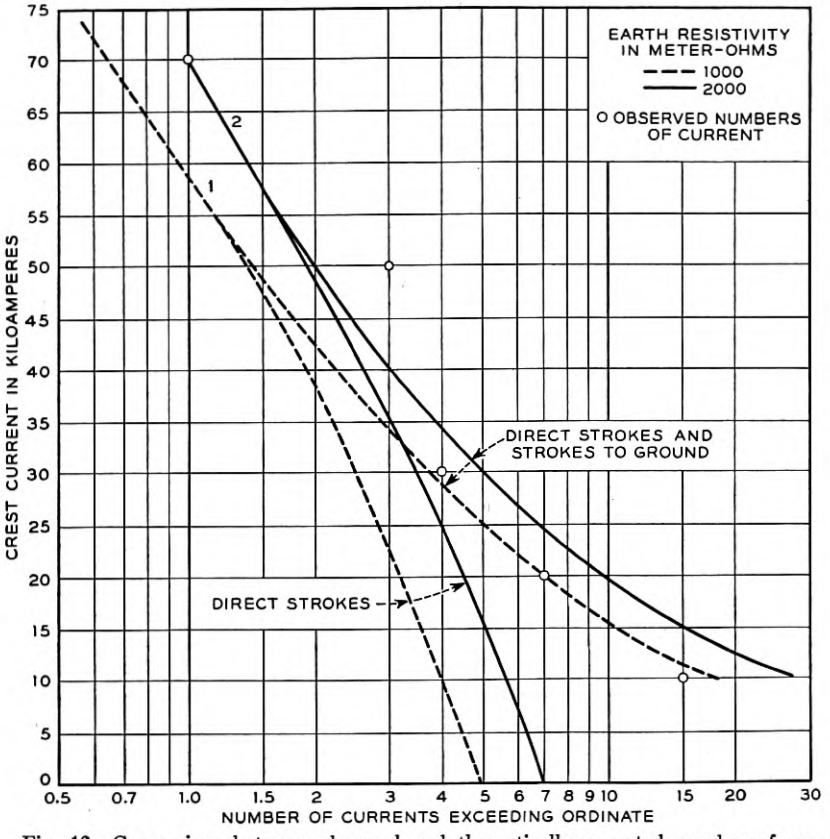


Fig. 12—Comparison between observed and theoretically expected number of currents exceeding given crest values during 108 thunderstorm days in a 21.5-mile section.

rate. Improved accuracy is obtained by using theoretical expectancy curves in the extrapolation as indicated by the curves in the figure. These curves are derived on the assumption that the proportion of currents exceeding a given crest value I is given by $P(I) = e^{-kI}$ where k is a constant —a relation in substantial agreement with observations[1]—and that the energy in the electromagnetic wave from the stroke current, as well as that in the sound wave due to the thunder, is proportional to I^2/r^2, where r

is the distance to the stroke. If the trigger arrangement in the apparatus mentioned in Section 2.5 is sufficiently sensitive to be operated by strokes so remote that the thunder cannot be distinguished above noise on the wire record, then the energy in the sound wave would be controlling, in the sense that it would determine the making of a usable record. On the other hand if the triggering should occur only for strokes of such substantial intensity

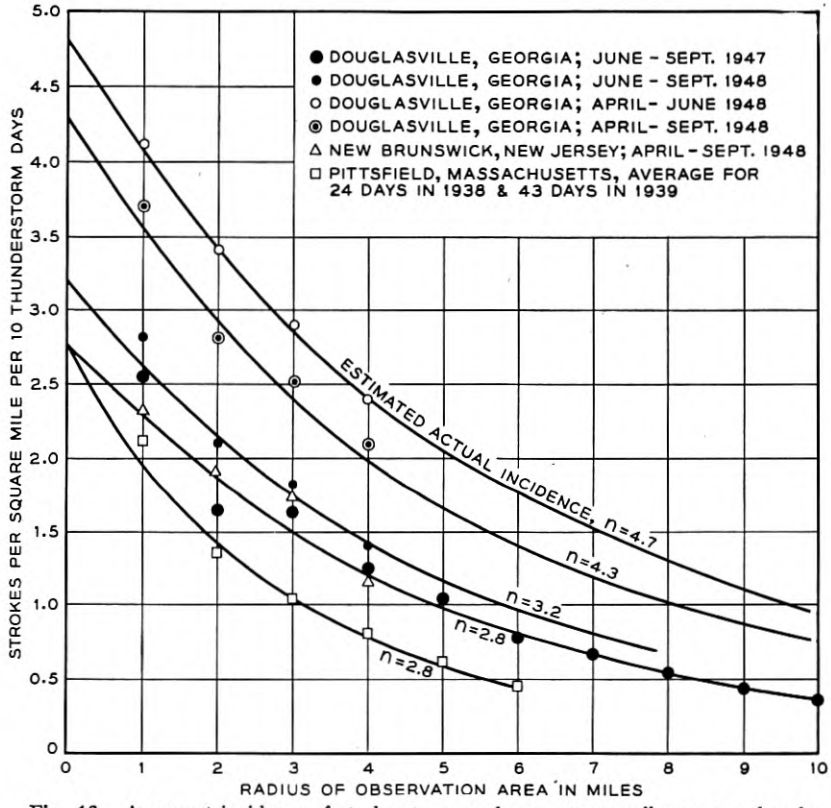


Fig. 13.—Apparent incidence of strokes to ground, per square mile per ten thunder storm days, as a function of radius of observation area.

that some of the more remote strokes of low intensity would not be recorded, then the energy in the electromagnetic wave would be controlling. Similarly, in the case of direct visual-aural observation, the light radiation from the stroke may be assumed proportional to I^2/r^2. If, for any of these methods the energy density is taken as $u = cI^2/r^2$ where c is a constant and the minimum energy required for observation is u_0, then only stroke currents in excess of $I = (u_0/c)^{1/2}r$ would be observed. The observed or

apparent number of strokes to ground within an area of radius r would then be, for an actual incidence n of strokes to ground and with $P(I) = e^{-kI}$:

$$N_a = 2\pi n \int_0^r re^{-\alpha r} dr$$

$$= \frac{2\pi n}{\alpha^2} [1 - e^{-\alpha r} (1 + \alpha r)]$$

where $\alpha = k(u_0/c)^{1/2}$.

The apparent incidence of strokes to ground $n_a = N_a/\pi r^2$ is accordingly

$$n_a = \frac{2n}{(\alpha r)^2} [1 - e^{-\alpha r} (1 + \alpha r)]$$

By choice of a proper value of α in the latter expression a theoretical curve, varying in substantially the same manner with r as a given observed curve, may be obtained. The actual incidence is next obtained by taking a value of n such that the two curves substantially coincide. This value of n also corresponds to the incidence given by the theoretical curve for $r = 0$, i.e. the value that would be expected if a sufficient number of observations were available for small values of r to permit extrapolation of the observed curves to $r = 0$. The value of n obtained in the above manner is about 2.8 for the New Jersey and Massachusetts and 1947 Georgia observations. The latter extended over the last half of the lightning season, while the 1948 Georgia observations, which indicate a higher incidence, extended over the entire season. The comparison, shown in the figure, between the observations in Georgia during the first and second halves of the 1948 season, indicates that the incidence during the first half is about 50 per cent greater than during the second half. A similar comparison of the New Jersey observations, not shown in the figure, indicates the opposite trend, i.e. a somewhat smaller incidence during the first half. The difference, however, is less marked than in the Georgia case.

There is reason to believe that this change in Georgia with the advance of the season is due to a change in the character of the lightning storms. During several years the more severe lightning damage on cable routes in this territory has occurred during early-season thunderstorms, which ordinarily are of the "frontal" type extending over fairly wide areas where hot and cold masses of air come together. These storms appear to be of greater extent, duration, and severity than the "convection" type of storm, ordinarily experienced later in the season, which occur more frequently as the result of local air convection currents but are of more limited extent and duration than storms of the frontal type.[7]

As mentioned before, the theoretical expectancy of lightning damage and of strokes to the cable discussed in this paper has been based on an incidence of 2.4 strokes per square mile for 10 thunderstorm days, a value derived from magnetic link observations of the rate of stroke occurrence to the aerial supporting structures of transmission lines, on the assumptions that they attract lightning strokes in accordance with laws established from laboratory observations on small-scale models, and that the average height of the ground wires is 70 feet above the earth or adjacent trees.[1] If this height had been taken as 60 feet instead the incidence would have been 2.8, in substantial agreement with that obtained from our observations for northern territory—in the main the territory traversed by the transmission lines from which the data were obtained.

The curves shown in Fig. 13 include substantial areas and a rather large amount of data and should, therefore, be fairly representative. Thus a radius of four miles corresponds to an observation area of 50 square miles. Within this area a total of 342 strokes was recorded during 1948 at the observation point in the test section near Atlanta. One of the storms during this period, in which the antenna was struck, passed directly over the observation point and provided a considerable amount of the data. However, even if the observations during this storm were omitted, the total for the season would have been reduced by less than 10 per cent, while the observations during July, August, and September would have been reduced about 20 per cent and would have been slightly lower than in the same 1947 period. The data thus indicate that the yearly incidence per square mile of strokes to ground is about 2.8 per 10 thunderstorm days in northern parts of the country, but may be as high as 4.3 in those southern parts where more severe types of thunderstorms occur. Considering, however, both the 1947 and 1948 observations in Georgia, it appears that an incidence of 3.7 would be a reasonable expectation for an entire season. With this incidence, rather than 2.4 as assumed in Fig. 12, Curves 1 and 2 in that figure would approximately correspond to earth resistivities of 500 and 1000 meter-ohms, respectively.

Conclusions

The observations indicate that the duration of lightning currents in the southern territory under observation is substantially longer than the average ordinarily assumed. The time to half-value of intense currents, which are of main importance as regards liability to lightning damage, is of the order of 150 microseconds, and for lower currents even larger. This, together with the higher incidence of strokes to ground and the high earth resistivity, would appear to account for the high rate of lightning damage experienced in this territory in cables of earlier design than the copper-

jacketed cable upon which measurements were made. The incidence of cable currents of various intensities, their rate of attenuation, and the resultant voltages appear to be in satisfactory agreement with theoretical expectations.

Acknowledgements

The field observations were made possible by the cooperation of the Long Lines Department of the A. T. and T. Company, and the Southern Bell Telephone and Telegraph Company, both in the installation and the operation of the equipment. The observations were conducted by our associate Mr. D. W. Bodle, who was also responsible for the design of the automatic recording equipment used to measure the rate of strokes to ground, and who suggested, from some of the observations discussed here, the greater intensity of early-season storms.

References

1. E. D. Sunde: "Lightning Protection of Buried Toll Cable," *B. S. T. J.*, Vol. 24, April 1945.
2. E. D. Sunde: "Earth Conduction Effects in Transmission Systems," D. Van Nostrand Company, Inc., New York, London, Toronto, 1949.
3. C. M. Foust and H. P. Kuehni: "The Surge-crest Ammeter," *General Electric Review*, Vol. 35, December 1932.
4. C. F. Wagner and G. D. McCann: "New Instruments for Recording Lightning Currents," *Trans. A. I. E. E.*, Vol. 58, 1939.
5. W. H. Alexander: "The Distribution of Thunderstorms in the United States—1904–1933"—*Monthly Weather Review*, Vol. 63, 1935.
6. J. H. Hagenguth: "Photographic Study of Lightning," *Trans. A. I. E. E.*, Vol. 66, 1947.
7. F. A. Berry, E. Bollay and N. R. Beers: "Handbook of Meteorology," McGraw-Hill Book Company, Inc., New York and London, 1945.

The Electrostatic Field in Vacuum Tubes With Arbitrarily Spaced Elements

By W. R. BENNETT and L. C. PETERSON

VACUUM tubes with close spacing between electrodes have become of increasing importance in recent years. The higher transconductances and lower electron transit times thus obtained combine with other features to raise both the frequency and band width at which the tube may operate satisfactorily as an amplifier. Specific designs have been discussed in papers by E. D. McArthur and E. F. Peterson[1], and by Fremlin, Hall and Shatford[2]. The important contributions to structural technique made by E. V. Neher have been described in the Radiation Laboratory Series[3]. Further important advances in the art have been recently announced by J. A. Morton and R. M. Ryder of the Bell Laboratories at the recent I.R.E. Electronics Conference held at Cornell University in June, 1948. The material of the present paper represents work done by the authors over a decade ago, and naturally there has been considerable publication on related topics in the intervening years. It has been suggested by our colleagues, however, that some of the results are not available in the technical literature and are of sufficient utility to warrant a belated publication. These results have to do with the variation of the electric intensity, amplification factor, and current density which takes place along the cathode surface because of the nearby grid wires.

We shall deal mainly with the approximate solution which neglects the effect of space charge. The correction required to take account of space charge is in general relatively small as shown by both qualitative argument and experimental data in an early paper by R. W. King[15]. More recent theoretical work[19] extending into the high frequency realm has confirmed the minor nature of the modification needed. The problem is thereby reduced to one of finding solutions of Laplace's equation which reduce to constant values on the cathode, grid, and anode surfaces. The original work on this problem was done by Maxwell[4] who calculated the electrostatic screening effect of a wire grating between conducting planes long before the vacuum tube was invented. All subsequent work has followed the methods outlined by Maxwell. In particular he suggested the replacement of the conducting planes by an infinite series of images of the grid wires and described an appropriate solution in series for the case of finite size wires. The useful approximation obtained when the diameter of the grid wires is

assumed small compared to their spacing was discussed in detail only for the case of large distances between the grating and each of the conducting planes.

Figure 1 shows the assumed geometry of the grid, anode, and cathode. End effects are neglected. The origin is taken at the center of one of the grid wires which have radius c, and the X-axis is along the grid plane. The spacing of the wires between centers is a, the distance from grid to anode is d_2, and that from grid to cathode is d_1. No restrictions are placed on the sizes of a, d_2, and d_1. Above the anode and below the cathode is shown a doubly infinite set of images which may be inserted to replace the conducting planes of the anode and cathode. By symmetry the potential from the

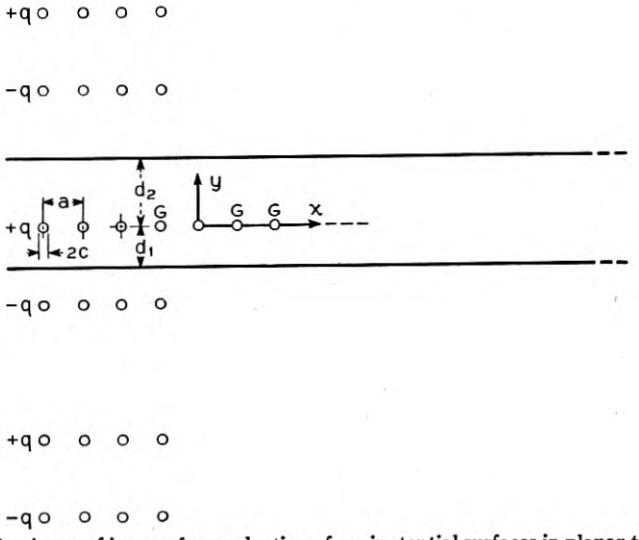


Fig. 1—Array of images for production of equipotential surfaces in planar triode.

array of charges there shown must be constant for all x when $y = d_2$ and also for all x when $y = -d_1$. The double periodicity of the array suggests immediately an application of elliptic functions. The solution of the symmetrical case was actually stated in terms of the elliptic function sn z by F. Noether[5]. The extension to the non-symmetrical case shown in Fig. 2 is fairly obvious. One of the authors worked out such a solution in terms of Jacobi's Theta functions in 1935, but abandoned any plans for publishing his analysis in view of the excellent treatment appearing shortly after that time in the Proceedings of the Royal Society by Rosenhead and Daymond[6], who applied Theta functions to both tetrodes and triodes, and both cylindrical and planar tube structures for the case of fine grid wires. Some of their formulas were later included in a book by Strutt[7]. Methods of calculating

the case of thick grid wires in terms of expansions in series of elliptic functions were discussed by Knight, Howland and McMullen[8–10]. The problem of a finite number of grid wires was treated by Barkas[11]. More recently tubes with close spacing between grid and cathode, but with anode and grid assumed far apart, have been analyzed in terms of elementary functions by

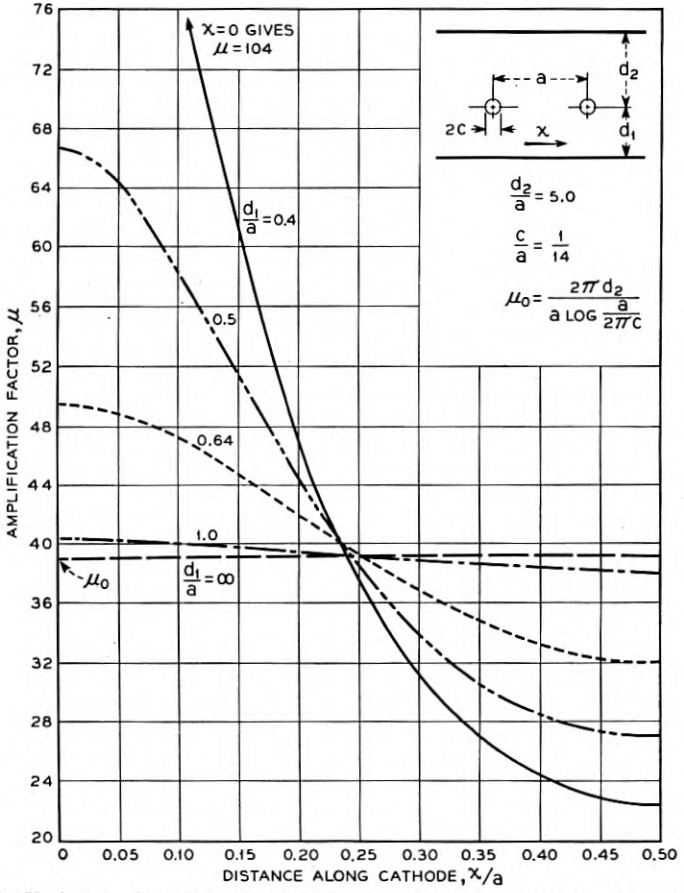


Fig. 2—Variation of amplification factor along the cathode surface of a triode.

Fremlin[12]. A solution based on the Schwartz-Christoffel transformation has been given by Herne[13] for the case of grid wires of finite size and approximately circular in shape.

Since the derivations have been adequately covered in the references cited, we merely state the final formula here and indicate how it may be verified as correct. Let $V(x, y)$ represent the potential function corresponding to

Fig. 1, the planar triode with fine grid wires. The potential of the cathode is set equal to zero. Then in the space between anode and cathode,

$$AV(x, y) = [2\pi d_2(y - d_2)/a + (d_1 + d_2)f(x, y)]V_g + [B(y + d_1) - 2\pi d_2 y/a - d_1 f(x, y)]V_p, \tag{1}$$

where

$$f(x, y) = \ell n \left| \frac{\vartheta_1[\pi(x + iy - 2i d_2)/a]}{\vartheta_1[\pi(x + iy)/a]} \right| \tag{2}$$

$$A = (d_1 + d_2)B - 2\pi d_2^2/a \tag{3}$$

$$B = \ell n \left| \frac{a\vartheta_1(2\pi i d_2/a)}{\pi c \vartheta_1'(0)} \right| \tag{4}$$

Here we have used Jacobi's notation for the ϑ_1-function, as explained by Whittaker and Watson[14], rather than the Tannery-Molk notation used by Rosenhead and Daymond. We write $\vartheta_1(\pi z)$ for their $\vartheta_1(z)$. In our notation

$$\vartheta_1(z) = 2\sum_{n=0}^{\infty} (-)^n e^{i(n+1/2)^2\tau} \sin(2n + 1)z \tag{5}$$

where the parameter τ in the above formulas is given by:

$$\tau = 2i(d_1 + d_2)/a \tag{6}$$

By $\vartheta_1'(z)$ is meant the derivative with respect to z:

$$\vartheta_1'(z) = 2\sum_{n=0}^{\infty} (-)^n(2n + 1)e^{i(n+1/2)^2\tau} \cos(2n + 1)z \tag{7}$$

Verification of the solution is straightforward. The resulting $V(x, y)$ is seen to be the real part of a function which is analytic in the complex variable $x + iy$ except for logarithmic singularities at the points where the Theta functions vanish. Hence $V(x, y)$ satisfies Laplace's equation in two dimensions in the region excluding the singular points. Since the zeros of $\vartheta_1(z)$ occur at $z = m\pi + n\pi\tau$, where m and n take on all positive and negative values as well as zero, the singular points of the solution are at

$$\left.\begin{aligned} x + iy &= ma + 2in(d_1 + d_2) - 2id_2 \\ \text{and} \qquad x + iy &= ma + 2in(d_1 + d_2) \end{aligned}\right\} \tag{8}$$

which coincide with the centers of the image circles of Fig. 1. The logarithmic singularities represent line charges with the first set arising from a ϑ_1-function in the numerator, yielding a positive charge, and the second set from the ϑ_1-function in the denominator giving a negative sign. The equipotential curves are approximately circular in the neighborhood of the

charges and hence $V(x, y)$ gives a constant potential on the surface of each grid wire if the radius of the grid wire is small compared with the spacing.

We may show by direct substitution that $V(x, y)$ becomes equal to V_p at all points of the anode and equal to zero at all points of the cathode. On the anode we have $y = d_2$ which, when substituted in the expression for $f(x, y)$, gives the logarithm of the absolute value of the ratio of conjugate complex quantities, and hence

$$f(x, d_2) = 0$$

Substituting in (1), we then readily verify that $V(x, d_2) = V_p$. On the cathode we make use of the quasi-periodicity of the ϑ_1-function, as expressed by

$$\vartheta_1(z) = -e^{i(\pi\tau+2z)}\,\vartheta_1(z + \pi\tau), \tag{9}$$

to prove

$$f(x, -d_1) = \frac{2\pi\, d_2}{a}, \tag{10}$$

from which it follows that $V(x, -d_1) = 0$. To show that all grid wires are at the same potential, we make use of the other periodicity of the ϑ_1-function,

$$\vartheta_1(z + \pi) = -\vartheta_1(z), \tag{11}$$

which shows that

$$f(x \pm ma, y) = f(x, y), m = 0, 1, 2, \cdots \tag{12}$$

It remains to prove that V actually approaches the value V_g in the neighborhood of the typical wire, which may be taken at the origin since the solution repeats periodically with the wire spacing. We let

$$x + iy = ce^{i\theta} \tag{13}$$

and assume $c/a \ll 1$. Expanding in power series in c/a, we find that the first order terms are included in:

$$f(c\cos\theta, c\sin\theta) = \ell n\left|\frac{\vartheta_1(-2\pi i\, d_2/a)}{\vartheta_1'(0)\pi ce^{i\theta}/a}\right| = B \tag{14}$$

The sign of the argument of the ϑ_1-function in the numerator is of no consequence since it does not affect the absolute value. Substituting back in (1), we then find

$$\lim_{c/a\to 0} V(c\cos\theta, c\sin\theta) = V_g \tag{15}$$

The solution is thus completely established.

The quantities in which we are specifically interested are electric field, amplification factor, and current density. The electric field is equal to the negative gradient of the potential function. The amplification factor is found by taking the ratio of partial derivatives of the electric field at the cathode with respect to grid and anode voltages. The current density may then be studied for any assumed operating values of grid and plate voltages.

To calculate the gradient we note that since $V(x, y)$ is the real part of an analytic function $W(z) = V + iU$, it follows from the Cauchy-Riemann equations,

$$W'(z) = \frac{\partial V}{\partial x} - i\frac{\partial V}{\partial y} = -E_x + iE_y \tag{16}$$

where E_x and E_y are the x- and y- components of the electric intensity. From (1),

$$\begin{aligned} AW(z) = {} & [(d_1 + d_2)\, F(z) - 2\pi d_2(iz + d_2)/a]V_g \\ & + [B(d_1 - iz) + 2\pi i\, d_2 z/a - d_1\, F(z)]V_p. \end{aligned} \tag{17}$$

where

$$F(z) = \ell n \frac{\vartheta_1[\pi(z - 2i\, d_2)/a]}{\vartheta_1(\pi z/a)} \tag{18}$$

Calculating the derivative and making use of the relation,

$$\frac{\vartheta_1'(z - \pi\tau)}{\vartheta_1(z - \pi\tau)} = \frac{\vartheta_1'(z)}{\vartheta_1(z)} + 2i, \tag{19}$$

we find at the cathode surface

$$F'(x - i\, d_1) = \frac{2\pi i}{a}[1 + C(x)] \tag{20}$$

where

$$C(x) = Im \frac{\vartheta_1'[\pi(x + i\, d_1)/a]}{\vartheta_1[\pi(x + i\, d_1)/a]} \tag{21}$$

It follows that when $y = -d_1$, we must have $E_x = 0$ and

$$\begin{aligned} aAE_y/2\pi = {} & [d_1 + (d_1 + d_2)C(x)]V_g \\ & + [d_2 - d_1 - aB/2\pi - d_1 C(x)]V_p \end{aligned} \tag{22}$$

The amplification factor is then given by

$$\mu = \frac{\partial E_y/\partial V_g}{\partial E_y/\partial V_p} = \frac{d_1 + (d_1 + d_2)C(x)}{d_2 - d_1 - aB/2\pi - d_1 C(x)} \tag{23}$$

Numerical calculation from these formulas can be made by means of (5) and (6). When d_1 and d_2 are both large compared with unity, Eq. (23)

reduces to the familiar approximate formula derived, for example, in an early paper by R. W. King[15],

$$\mu \doteq \frac{2\pi d_2}{a \ln \dfrac{a}{2\pi c}}, \tag{23a}$$

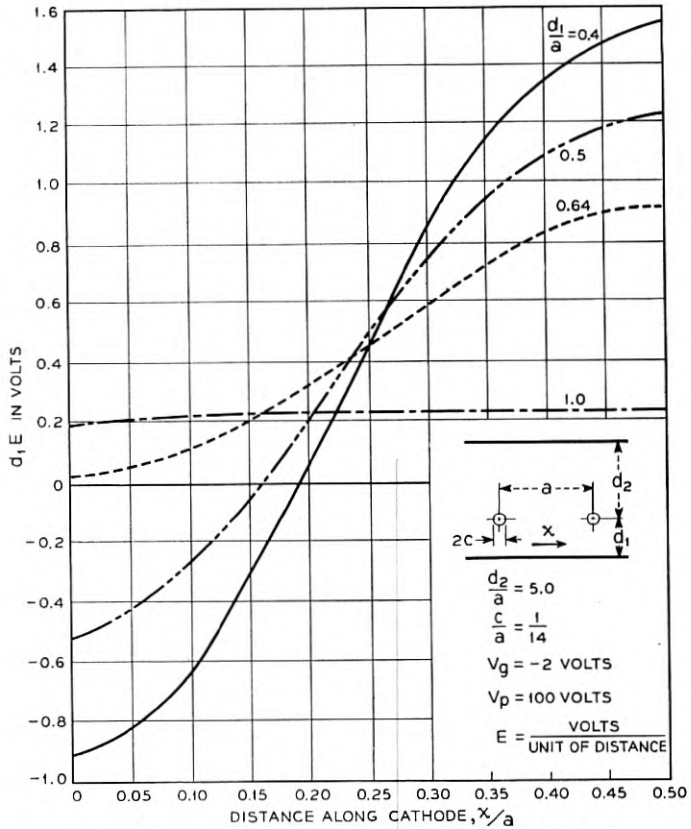


Fig. 3—Variation of cathode field strength in a triode.

Some calculated curves for μ and E_g are shown in Figs. 2 and 3. Figure 2 shows the amplification factor as a function of the distance along the cathode with the ratio of grid-cathode separation to grid wire spacing as a parameter. The ratio of grid-anode separation to grid wire spacing is held constant at five. Only half the grid spacing interval is included since the curves are symmetrical. The increase in μ-variation as the grid-cathode separation becomes small is clearly demonstrated. For negligible μ-variation we must select d_1/a of the order of 2 or greater.

Figure 3 shows the variation in field strength along the cathode for the typical operating point, $V_g = -2$ and $V_p = 100$ volts. It is to be noted that for d_1/a less than 0.6, the electric field actually changes sign as we move from a point immediately below a grid wire to the midpoint between two grid wires. In other words a part of the cathode will not emit at all in these cases while the remainder emits in a non-uniform manner. In the rather extreme case of $d_1/a = 0.4$ only about a quarter of the cathode is emitting. It is worth noting how relatively rapid the "shadow" or "island" formation increases between $d_1/a = 0.64$ and 0.5 as compared to the increase in the interval from 0.5 to 0.4.

If the equation for μ is solved for $C(x)$ and the result substituted back in the expression for E_y at the cathode we find:

$$-E_y = \frac{V_g + V_p/\mu}{d_1 + (d_1 + d_2)/\mu} \tag{24}$$

where here of course μ varies with x. This is identical with the expression derived by Benham[16] from Maxwell's approximate solution except that in the latter case μ was a constant. Our colleague, Mr. L. R. Walker, has pointed out that the equation follows directly from the assumption of small grid wires without explicit solution for the potential function. Since the charge density σ_c on the cathode is proportional to the field strength (the factor of proportionality in MKS units is the dielectric constant ϵ of vacuum or 9.854×10^{-12} farads/meter), Maxwell's capacity coefficients C_{gc} and C_{pc} may be calculated from

$$\sigma_c = \epsilon E_y = -(C_{gc}V_g + C_{pc}V_p) \tag{25}$$

The minus sign is used here because we are taking the ratio of charge to voltage at the negative plate of the condenser consisting of cathode, grid and anode surfaces. Hence

$$C_{gc} = \frac{\epsilon}{d_1 + (d_1 + d_2)/\mu} \tag{26}$$

$$C_{pc} = \frac{\epsilon/\mu}{d_1 + (d_1 + d_2)/\mu} \tag{27}$$

Since μ is variable, an integration is required to determine the total capacitance. From the periodicity of μ with grid spacing it is possible to express the result in terms of the average values of C_{gc} and C_{pc} over an interval of length a along a direction parallel to the grid plane and multiply these values by the total area of cathode surface.

Equation (24) may be interpreted in a number of different ways of which we shall mention the following two:

1. The "equivalent voltage" $V_g + V_p/\mu$ does not act at the grid but at a distance D from the cathode, where

$$D = d_1 + (d_1 + d_2)/\mu \tag{28}$$

Both the equivalent voltage and distance vary along the cathode surface.

2. The "equivalent voltage"

$$V_e = (V_g + V_p/\mu)/[1 + (1 + d_2/d_1)/\mu] \tag{29}$$

acts in the grid plane and varies with distance along the cathode surface.

As far as the cold tube is concerned the two formulas are equivalent at the cathode, but not at the grid. When the tube is heated and complete space charge is present, the two formulas also differ at the cathode. The current density in the presence of space charge is, according to (28) and Child's law:

$$I = K(V_g + V_p/\mu)^{3/2}/D^2 \tag{30}$$

while, from (29),

$$I = KV_e^{3/2}/d_1^2 \tag{31}$$

In both, $K^2 = 32\,\epsilon^2 e/81\,m$, where e/m is the ratio of electronic charge to mass. The value of current given by (31) is $[1 + (1 + d_2/d_1)/\mu]^{1/2}$ times as large as that given by (30). If $u >> 1 + d_2/d_1$ the two values are nearly the same. In tubes with close grid-to-cathode spacing the inequality may not be fulfilled. As to which viewpoint is more accurate, we note that Ferris and North in their papers [17, 18] on input loading adopted the latter, and that at high frequencies where electron transit time must be considered the second viewpoint is preferable because of the more accurate representation of effects at the grid. For a more complete discussion see Reference 19. Figure 4 shows curves of relative current density as a function of distance along the cathode as computed from Eq. (31). The transconductance for unit area of cathode surface as computed from the same equation is given by:

$$d_1^2 g_{mo} = d_1^2 \frac{\partial I}{\partial V_g} = \frac{2}{3}\epsilon\sqrt{\frac{2e}{m}\left(V_g + \frac{V_p}{\mu}\right)}\left(\frac{d_1}{D}\right)^3$$

$$= 3.512(V_g + V_p/\mu)^{1/2}(d_1/D)^{3/2} \text{ micromhos.} \tag{32}$$

The resulting variation with distance along the cathode is shown in Fig. 5.

Defining the figure of merit M at a point x along the cathode as the ratio

between the transconductance $\partial I/\partial V_g$ and the sum of C_{gc} and C_{pc} at this point, we find from (30)

$$M = (4J/3)^{1/3}[d_1 + (d_1 + d_2)/\mu]^{-1/3}\mu/(\mu + 1) \tag{33}$$

where $J = eI/m\epsilon$, $e/m = 1.77 \times 10^{11}$ coulombs/kg. From (31), we find on the other hand

$$M = (4J/3\, d_1)^{1/3}\mu/(\mu + 1) \tag{34}$$

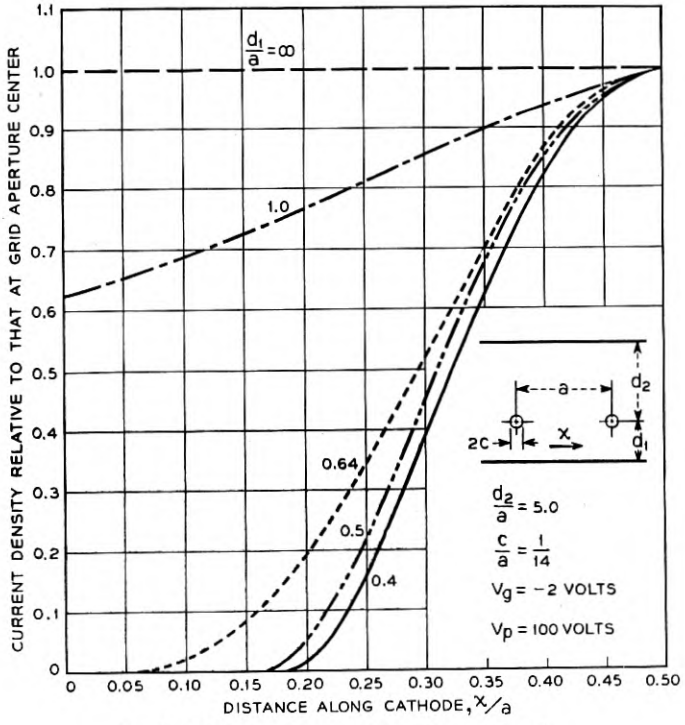


Fig. 4—Variation of current density in a triode.

Both formulas indicate that for a cathode capable of supplying a given current density the only means of improvement lies in decreasing the cathode-grid spacing. The improvement is extremely slow; doubling the figure of merit requires an eight-fold decrease in spacing.

We again emphasize that the calculated current densities and figures of merit are functions of x, the distance along the cathode. The total current between the two grid wires is found from (30) to be

$$I_T = 2K \int_{x_0}^{a/2} (V_g + V_p/\mu)^{3/2}\, dx/D^2 \tag{35}$$

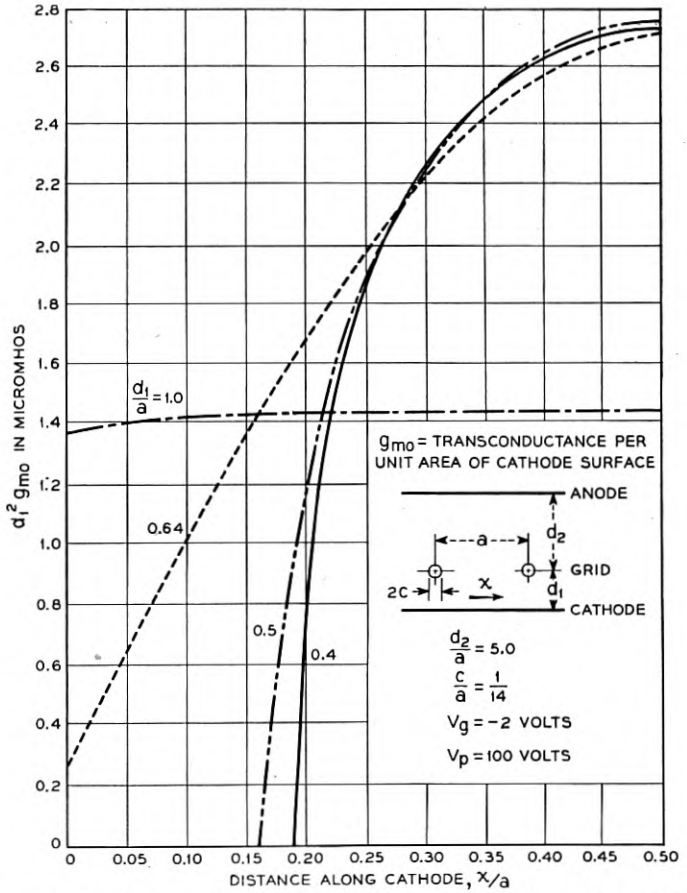


Fig. 5—Variation of transconductance along the cathode surface of a triode.

while, from (31),

$$I_T = \frac{2K}{{d_1}^2} \int_{x_0}^{a/2} [(\mu V_g + V_p)/(\mu + 1 + d_2/d_1)]^{3/2}\, dx \qquad (36)$$

where x_0 is given by

$$V_g + V_p/\mu(x_0) = 0 \qquad (37)$$

On the basis of several reasonable assumptions it may be shown that both (35) and (36) lead to an approximate 5/2 power law instead of 3/2 power law. Such a law has actually been observed in cases where shadow formation was suspected.

We wish to express our appreciation to Messrs. R. K. Potter, J. A. Morton,

and R. M. Ryder for their encouragement, and to Miss M. C. Packer for aid in the numerical computations.

References

1. E. D. McArthur and E. F. Peterson, "The Lighthouse Tube; A Pioneer Ultra-High-Frequency Development," *Proc. Nat. Electronic Conference*, Chicago, Oct. 1944, Vol. I, pp. 38–47.
2. J. H. Fremlin, R. N. Hall, and P. A. Shatford, "Triode Amplification Factors," *Electr. Comm.*, Vol. 23 (1946), pp. 426–435.
3. Hamilton, Knipp, and Kuper, "Klystrons and Microwave Triodes, Radiation Laboratory Series, New York, 1948, p. 153.
4. J. C. Maxwell, "A Treatise on Electricity and Magnetism," Vol. 1, pp. 310–316.
5. Riemann-Weber, "Differentialgleichungen der Physik," Vol. 2, p. 311.
6. L. Rosenhead and S. D. Daymond, "The Distribution of Potential in Some Thermionic Tubes," *Proc. Roy. Soc.*, Vol. 161 (1937), pp. 382–405.
7. M. J. O. Strutt, "Moderne Mehrgitter-Elektronenröhren," Berlin, 1940, S. 154.
8. R. C. Knight, *Proc. London Math. Soc.* (2) Vol. 39 (1935), pp. 272–281.
9. R. C. J. Howland and B. W. McMullen, "Potential Functions Related to Groups of Circular Cylinders," *Proc. Cambr. Phil. Soc.*, Vol. 32 (1936), pp. 402–415.
10. R. C. Knight and B. W. McMullen, "The Potential of a Screen of Circular Wires between two Conducting Planes," *Phil. Mag. Ser.* 7, Vol. 24, 1937, pp. 35–47.
11. Barkas, "Conjugate Potential Functions and the Problem of the Finite Grid," *Phys. Rev.*, Vol. 49 (1936), pp. 627–630.
12. J. H. Fremlin, "Calculation of Triode Constants," *Phil. Mag. Ser.* 7, Vol. 27 (1939), pp. 709–741; also *Electr. Comm.*, Vol. 18 (1939), pp. 33–49.
13. H. Herne, "Valve Amplification Factor," *Wireless Engineer*, Vol. 21 (1944), pp. 59–64.
14. Whittaker and Watson, "Modern Analysis," Third Edition, Cambridge (1940), Chapter XXI, p. 462.
15. R. W. King, "Thermionic Vacuum Tubes," *Bell System Technical Journal*, Vol. II (1923), pp. 31–100.
16. W. E. Benham, "A Contribution to Tube and Amplifier Theory," *Proc. I. R. E.*, Vol. 26 (1938), pp. 1093–1170.
17. W. R. Ferris, "Input Resistance of Vacuum Tubes at Ultra-High Frequencies," *Proc. I. R. E.*, Vol. 24 (1936), pp. 82–105.
18. D. O. North, "Analysis of the Effects of Space Charge on Grid Impedance," *Proc. I. R. E.*, Vol. 24 (1936), pp. 108–136.
19. F. B. Llewellyn and L. C. Peterson, "Vacuum Tube Networks," *Proc. I. R. E.*, Vol. 32 (1944), pp. 144–166.

Transconductance as a Criterion of Electron Tube Performance

By T. SLONCZEWSKI

QUANTITATIVE evaluation of electron tube performance has assumed added importance with the increasing extension of electronics into the fields of measurement and control. Simplification of the process of selection of suitable tube types and operating conditions from the general data available is of considerable value to all engineers concerned with electronics circuit design. The conventional procedure involves analysis of the plate current-grid voltage characteristics. The simpler method presented herein supplies the same information from an analysis of the transconductance-grid voltage characteristics. These are usually supplied by the manufacturer or can be obtained readily by measurement[1].

The method presented herein has been employed successfully for a number of years in the development of electronic measuring apparatus by a group of engineers who attended lectures on the subject given by the author. It applies chiefly to pentodes, where the internal plate impedance is high with respect to the load impedance. Its merit resides in the comparative brevity of the formulae, the ease of computation and the facility in obtaining the data from which the computations are made. It allows one to form a preliminary judgment of the performance of a tube from a brief glance at the characteristics furnished by the manufacturer better and faster than any other method known to the author. It should prove of value to the instructor teaching electron tube theory.

In the interest of simplicity some of the subscripts m, p, c and g appended usually to symbols for transconductance, plate current and grid voltages are deleted below. The scope of the discussion is so limited that no confusion may arise from this omission. The formulae are expressed in terms of amplitudes of voltage and current, capital letters being used for their symbols. All values are in peak volts. Levels are in decibels.

The g-e characteristic is introduced into the problem by starting with the general expression for the plate current

$$i = i_0 + \frac{\partial i}{\partial e} v + \frac{1}{2!} \frac{\partial^2 i}{\partial e^2} v^2 + \frac{1}{3!} \frac{\partial^3 i}{\partial e^3} v^3 + \cdots \qquad (1)$$

where v is the voltage measured from the bias point E_c, where the derivatives are taken, and utilizing the definition of the transconductance

[1] Radio Engineers' Handbook, F. E. Terman, McGraw-Hill, 1943, p. 961.

$$g = \frac{\partial i}{\partial e}. \tag{2}$$

Inserting (2) into (1) and calling G the value of g at E_c we obtain

$$i = \int_{-\infty}^{E_c} g de + Gv + \frac{1}{2}\frac{\partial g}{\partial e} v^2 + \frac{1}{6}\frac{\partial^2 g}{\partial e^2} v^3 + \frac{1}{24}\frac{\partial^3 g}{\partial e^3} v^4 \cdots \tag{3}$$

The first term of this expression is the space current of the tube at no load, that is when $v = 0$. On the g-e diagram, Fig. 1, it represents the area under the curve from the tube cut off C to the tube bias E_c. The second term represents the function of the tube as an amplifier. The third term represents the second-order modulation current. The latter is responsible for the objectionable generation of a second harmonic in an amplifier and the useful presence of the second harmonic in the frequency doubler, the direct current in a rectifier and the sidebands in a modulator.

The fourth and higher terms represent, in general, undesirable effects of modulation. They are usually smaller than the first two and, since their effects are additive, the first three terms of expression (3) may be studied profitably disregarding the others. If necessary, the effects of the higher-order terms may be added later.

The Idealized Parabolic Pentode

If, over a certain range of grid biases e_A to e_B, the effect of the fourth and higher terms of series (3) is negligible the g-e characteristic will be a straight line. Herein lies one of the advantages of the method, for a straight portion of a curve can be easily selected by inspection and checked with a straight edge. It is thus possible to select easily such a tube and operating point that third and higher-order modulation products are absent in the output. There is no such simple method of verifying whether a current characteristic is parabolic. That there are tubes having approximately straight portions of g-e characteristics can be verified by inspection of (Fig. 1) where the characteristic of the 6AG7 is given.

Since a portion of the g-e characteristic is a straight line, the third term coefficient $\frac{\partial g}{\partial e}$ may be replaced by the ratio $\frac{\Delta g}{\Delta e}$ where Δe is an arbitrary interval of grid voltage and Δg the corresponding change in g. In many of the following computations it will be advantageous to use for Δe the total excursion of the grid voltage.

On the basis of the simplifying assumption of a parabolic pentode it is possible to derive the simple formulae given below which cover the performance of the tube as an amplifier, rectifier and modulator.

The Parabolic Pentode Amplifier

Over the straight portion of the g-e curve the following relations hold for an input $v = P \cos pt$.

The fundamental current is $I_p = GP$ where P is the grid swing.

The second harmonic current is

$$I_{2p} = \frac{1}{4} \frac{\Delta g}{\Delta e} P^2$$

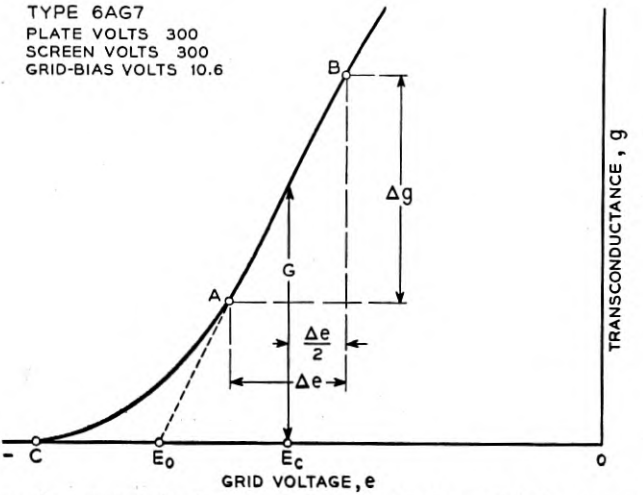


Fig. 1—g-e characteristic having principally second order modulation over part of the range.

To find the level H_2 of the second harmonic below the fundamental, prolong the straight portion of the characteristic down to the virtual cutoff E_0 (Fig. 1). Then

$$H_2 = 20 \log \frac{I_p}{I_{2p}} = 20 \log \frac{E_c - E_0}{P} + 12.$$

For the case when all of the parabolic characteristic is used

$$P = \frac{\Delta e}{2} \quad \text{and} \quad H_2 = 20 \log \frac{G}{\Delta g} + 18.$$

If it is desired to express H_2 in terms of the output current i_p and the space current i_0 the following approximate formula may be used:

$$H_2 = 20 \log \frac{i_0}{I_p} + 18.$$

This expression neglects the area under the characteristic on the left of the line AE_0, Fig. 1. The formula is useful in selecting a tube for closer consideration.

When a tube is used as a preamplifier in a wave analyzer an error of measurement may occur if two input frequencies intermodulate in the amplifier to produce a current of the same frequency as the one being measured. For instance, the fundamental $R \cos rt$ and the second harmonic $W \cos wt$ may intermodulate to form the third harmonic. If I_{r-w} is the disturbing current and I_p the wanted output, then

$$20 \log \frac{I_p}{I_{r-w}} = 20 \log \frac{E_c - E_0}{P} + 6 - 20 \log \frac{R}{P} - 20 \log \frac{W}{P}$$

The Rectifier

The portion of the plate current resulting from the rectification of a signal $P \cos pt$ is

$$I_{dc} = \frac{1}{4} \frac{\Delta g}{\Delta e} P^2.$$

If several frequencies were present, $P_1 \cos pt$, $P_2 \cos p_2 t$ and so on,

$$I_{dc} = \frac{1}{4} \frac{\Delta g}{\Delta e} (P_1^2 + P_2^2 + \cdots)$$

Thus I_{dc} is proportional to the square of the root-mean-square voltage input. This property of the parabolic tube of measuring the root-mean-square voltage is often useful in the measurement field.

If Δe is the parabolic range of the tube and Δg the corresponding change in g, the largest possible rectified current obeying the root-mean-square law will obtain for an amplitude $P = \frac{\Delta e}{2}$. Then

$$I_{\max} = \tfrac{1}{16} \Delta g \Delta e.$$

The Frequency Doubler

The second harmonic is given, as before, by

$$I_2 = \frac{1}{4} \frac{\Delta g}{\Delta e} P^2.$$

The largest possible output current is

$$I_{\max} = \tfrac{1}{16} \Delta g \Delta e.$$

In general, the level of the undesirable fundamental will be higher than the harmonic by

$$H_2 = 20 \log \frac{I_p}{I_{2p}} = \log \frac{E_c - E_0}{P} + 12.$$

For the maximum current case this reduces to

$$H_2 = 20 \log \frac{g}{\Delta g} + 18.$$

The Modulator

When two inputs $P \cos pt$ and $Q \cos qt$ are applied to the grid, the output is

$$i = i_0 + \frac{1}{4}\frac{\Delta g}{\Delta e}(P^2 + Q^2) + \frac{1}{4}\frac{\Delta g}{\Delta e}(P^2 \cos 2pt + Q^2 \cos 2qt) + \frac{1}{2}\frac{\Delta g}{\Delta e} PQ \cos (p + q)t + \frac{1}{2}\frac{\Delta g}{\Delta e} PQ \cos (p - q)t$$

The last two terms represent the sidebands.

In the case of a detector the available supply of the carrier voltage Q is copious. Putting $Q = \frac{\Delta e}{2}$ the well known result is obtained

$$I_{p+q} = I_{p-q} = \tfrac{1}{4}\Delta g P.$$

The conversion transconductance is

$$G_c = \frac{\Delta g}{4}.$$

To formulate filtering requirements the signal and carrier leaks must be found.

The signal leak is

$$20 \log \frac{I_p}{I_{p \pm q}} = 20 \log \frac{G}{\Delta g} + 12.$$

The carrier leak is

$$20 \log \frac{I_q}{I_{p \pm q}} = 20 \log \frac{G}{\Delta e} + 20 \log \frac{\Delta e}{P} + 6.$$

In the design of a heterodyne oscillator a generous supply of both input voltages is easily available and maximum output current is desirable. This occurs when $P = Q = \frac{\Delta e}{4}$. Then

$$I_{p-q} = \tfrac{1}{32}\Delta g \Delta e.$$

Whether P equals Q or not, the unwanted products are

$$20 \log \frac{I_q}{I_{p-q}} = 20 \log \frac{E_c - E_0}{Q} + 6$$

$$20 \log \frac{I_p}{I_{p-q}} = 20 \log \frac{E_c - E_0}{P} + 6$$

$$20 \log \frac{I_{2p}}{I_{p-q}} = 20 \log \frac{P}{Q} - 6$$

$$20 \log \frac{I_{2p}}{I_{p-q}} = 20 \log \frac{Q}{P} - 6$$

$$20 \log \frac{I_{r+q}}{I_{p-q}} = 0.$$

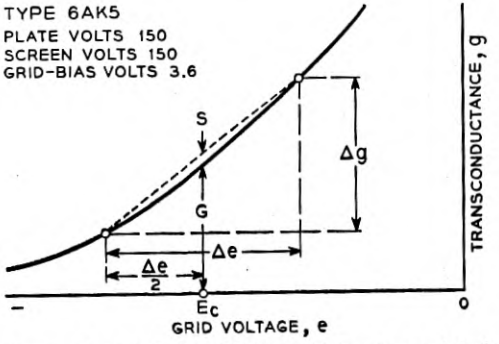


Fig. 2—g-e characteristic exhibiting third order modulation.

Third Order Modulation

When we do take into consideration the fourth term of equation (3), that is $\frac{1}{6}\frac{\partial^2 g}{\partial e^2} v^3$ the g-e characteristic will no longer be a straight line but a parabola. It turns out that all of the computations of the second-order effects as shown are so little affected that no corrections are necessary. The presence of third-order modulation caused by the term $\frac{1}{6}\frac{\partial^2 g}{\partial e^2} v^3$ will, however, add new types of modulation products to the output. These are usually objectionable.

A typical g-e characteristic with third-order modulation present is shown on Fig. 2. The curvature is usually concave upward. Instead of measuring the derivative $\frac{\partial^2 g}{\partial e^2}$ needed for the computations, it is more practical to scale

off the amount S by which the characteristic sags in the middle of the interval Δe (Fig. 2).

The plate current is then given by the expression

$$i = i_0 + Gv + \frac{1}{2}\frac{\Delta g}{\Delta e} v^2 + \frac{4S}{3(\Delta e)^2} v^3. \tag{4}$$

Single Frequency Input

When the signal input to an amplifier consists of a single frequency $v = P \cos pt$, the output current is given by

$$i = i_0 + \frac{\Delta g}{\Delta e} P^2 + \left[G + \frac{S}{(\Delta e)^2} P^2\right] P \cos pt + \frac{1}{4}\frac{\Delta g}{\Delta e} P^2 \cos 2pt + \frac{1}{3}\frac{S}{(\Delta e)^2} P^3 \cos 3pt.$$

The second-order effects consisting of the rectified current and second harmonic are seen to be unaffected by the presence of third-order modulation. However, the first-order effect, the fundamental output, ceases to be linear and a new product, the third harmonic, appears.

The change in fundamental output is expressible as a loading effect on transconductance. The effective transconductance of the tube is.

$$G_e = G + S\left(\frac{P}{\Delta e}\right)^2.$$

Expressed in db's the non-linearity effect is approximately

$$20 \log \frac{G_e}{G} = 8.6 \frac{S}{G}\left(\frac{P}{\Delta e}\right)^2.$$

When P is large it is convenient to select $\Delta e = 2P$ and get

$$G_e = G + \frac{S}{4}, \qquad 20 \log \frac{G_e}{G} = 2.15 \frac{S}{G}.$$

S is positive when the g-e curve is concave upward.

The third harmonic content of the output is

$$H_3 = 20 \log \frac{I_p}{I_{3p}} = 20 \log \frac{G}{S} + 40 \log \frac{\Delta e}{P} + 10.$$

If we select $\Delta e = 2P$ the second term drops out

$$H_3 = 20 \log \frac{G}{S} + 22.$$

If the curve is concave upward the third harmonic increases the peak value

of the wave. If it is concave downward the peak value of the wave decreases.

In the case of a two-frequency input $v = P \cos pt + Q \cos qt$ the second-order products are again unaffected. The third-order products will be:

$$I_{3p} = \frac{1}{3} \frac{S}{(\Delta e)^2} P^3$$

$$I_{3q} = \frac{1}{3} \frac{S}{(\Delta e)^2} Q^3$$

$$I_{2p\pm q} = \frac{S}{(\Delta e)^2} P^2 Q$$

$$I_{p\pm 2q} = \frac{S}{(\Delta e)^2} PQ^2.$$

There are situations in the design of detectors where the $I_{q\pm 2p}$ current may be disturbing. For example, in a wave analyzer modulation stage when measuring the second harmonic $P_2 \cos (2p)t$ the desired second-order product is $I_{q-(2p)}$. This is, however, of the same frequency as the third-order product I_{q-2p} generated by the intermodulations of the strong fundamental $P_1 \cos pt$ and the carrier $Q \cos qt$. The level of the wanted product with respect to the unwanted one is given by

$$20 \log \frac{I_{p-(2p)}}{I_{q-2p}} = 20 \log \frac{\Delta g}{S} + 20 \log \frac{\Delta e}{P_1} + 20 \frac{P_2}{P_1} - 6.$$

If there are two interfering inputs $R \cos rt$ and $W \cos wt$ they may, together with the carrier $Q \cos qt$, form an objectionable product $i_{r\pm w\pm q}$ of the same frequency as the wanted product i_{p-q}. The level of this disturbing product with respect to the wanted product is then given by

$$20 \log \frac{I_{p-q}}{I_{q\pm r\pm w}} = 20 \log \frac{\Delta g}{S} + 20 \log \frac{\Delta e}{R} + 20 \log \frac{P}{W} - 12.$$

When two input frequencies are present in the input, the effective transconductance of the tube becomes

$$G_e = G + \frac{S}{(\Delta e)^2} (P^2 + 2Q^2).$$

The amplifier gain depends on the level of all of the components of the input.

Fourth Order Modulation

If the fourth term of equation (3) is absent, but the fifth term $\frac{1}{24} \frac{\partial^3 g}{\partial e^3} v^4$ is present, fourth-order modulation will occur. The g-e characteristic will be a cubic with an inflection point at E_c.

TABLE I

If modulation up to the fourth order is present and $v = P\cos pt + Q\cos qt$, the plate current is $i = \sum_{m=0}^{m=4}\sum_{n=0}^{n=4} a_{mn}\cos(mp \pm nqt)$. The values of a_{mn} are:

m	n				
	0	1	2	3	4
0	$\frac{1}{4}\frac{\Delta g}{\Delta e}(F^2+Q^2) + \frac{3D}{4(\Delta e)^3}[4Q^2P^2+Q^4+P^4]$	$\left[G+\frac{S}{(\Delta e)^2}(2P^2+Q^2)\right]Q$	$\left[\frac{1}{4}\frac{\Delta g}{\Delta e}+\frac{3D}{(\Delta e)^3}(P^2+Q^2)\right]Q^2$	$\frac{1}{3}\frac{S}{(\Delta e)^2}Q^3$	$\frac{1}{4}\frac{D}{(\Delta e)^3}Q^4$
1	$\left[G+\frac{S}{(\Delta e)^2}(P^2+2Q^2)\right]P$	$\left[\frac{1}{2}\frac{\Delta g}{\Delta e}+\frac{3D}{(\Delta e)^3}(Q^2+P^2)\right]PQ$	$\frac{S}{(\Delta e)^2}PQ^2$	$\frac{D}{(\Delta e)^3}PQ^3$	0
2	$\left[\frac{1}{4}\frac{\Delta g}{\Delta e}+\frac{3D}{(\Delta e)^3}(P^2+Q^2)\right]P^2$	$\frac{S}{(\Delta e)^2}P^2Q$	$\frac{3}{2}\frac{D}{(\Delta e)^3}P^2Q^2$	0	0
3	$\frac{1}{3}\frac{S}{(\Delta e)^2}P^3$	$\frac{D}{(\Delta e)^3}P^3Q$	0	0	0
	$\frac{1}{4}\frac{D}{(\Delta e)^3}P^4$	0	0	0	0

$$g = G + \frac{\partial g}{\partial e} v + \frac{1}{6} \frac{\partial^3 g}{\partial e^3} v^3$$

To compute fourth-order effects a tangent is drawn to the curve in the middle of the range Δe (Fig. 3). It will represent a parabolic pentode over the range Δe and Δg. The departure D of the actual curve from the parabolic pentode at the extremes of the range Δe is measured. The g-e curve becomes

$$g = G + \frac{\Delta g}{\Delta e} v + \frac{8DV^3}{(\Delta e)^3}.$$

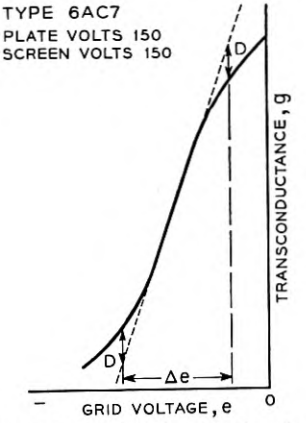


Fig. 3—g-e characteristic exhibiting fourth order modulation.

As a rule D is negative. It decreases the slope of the curve. The plate current is

$$i = i_0 + Gv + \frac{1}{2} \frac{\Delta g}{\Delta e} v^2 + \frac{2D}{(\Delta e)^3} v^4. \tag{5}$$

From this expression the fourth-order effects may be computed.

Single Frequency Input

Fourth-order modulation has no effect on the gain of the amplifier. It affects the second-order products, the rectified current and the second harmonic and adds the fourth harmonic.

In an amplifier it is unimportant to correct the value of H_2 for fourth-order modulation. The fourth harmonic is given by

$$H_4 = 20 \log \frac{G}{D} + 60 \log \frac{\Delta e}{P} + 12.$$

In a rectifier, the effect of fourth-order modulation is to destroy the square law of the rectifier. The error is

$$\frac{\Delta I_{dc}}{I_{dc}} = 3\frac{D}{\Delta g}\cdot\left(\frac{P}{\Delta e}\right)^2.$$

This correction is valid for the case of frequency doublers.

Two-Frequency Input

In a detector the fourth-order modulation produces an overloading effect; the modulator gain is then a function of the input

$$G = \frac{\Delta g}{4} + \tfrac{3}{8}D + \tfrac{3}{2}D\left(\frac{P}{\Delta e}\right)^2.$$

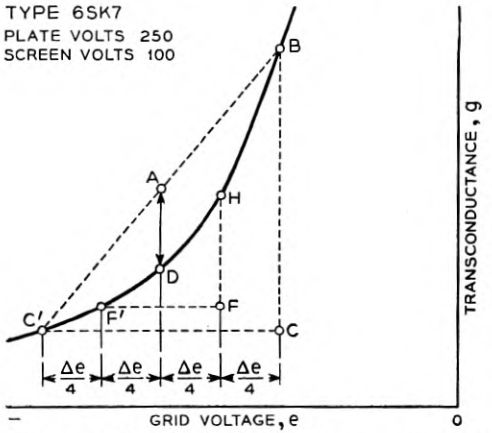


Fig. 4—g-e characteristic with second, third and fourth order modulation present. $D = \frac{2}{3}(\overline{CB} - 2\overline{FH})$; $\Delta g = \overline{CB} - 2D$; $S = \overline{AD}$

Neglecting $\frac{3}{8}D$ in comparison with $\frac{\Delta g}{4}$, the error in linearity is

$$\frac{\delta G_c}{G_c} = 6\frac{D}{\Delta g}\left(\frac{P}{\Delta e}\right)^2.$$

There will be also a cross-loading effect for an interfering signal $R\cos rt$.

$$\frac{\delta G_c}{G_c} = 2D\left(\frac{R}{\Delta e}\right)^2$$

In a heterodyne oscillator fourth-order modulation will produce a second harmonic at the output.

$$H_2 = 20\log\frac{\Delta g}{D} + 20\log\frac{\Delta e}{P} + 20\log\frac{\Delta e}{Q} - 10.$$

For the case of maximum output at $P = Q = \frac{\Delta e}{4}$

$$H_2 = 20 \log \frac{\Delta g}{D} + 14.$$

Accuracy of Computations

Before closing the subject, let us consider for a moment the accuracy involved. Very careful measurement of the characteristics of samples of some tubes, notably the 310A and 807 has shown that over a large range the fit with a parabola is excellent. On the other hand the electron tube bulletins give characteristics which are avowedly average. The drafting errors must be large and the temptation to use a straight edge instead of a french curve must be great. Probably no two observers will agree as to the extent of the straight part of a conductance curve. Even in the case of the 310A and 807 tubes mentioned above, manufacturing variations may affect the transconductance curve from tube to tube.

With this in mind we may conclude that the results obtained, that is, the values of the current obtained by the methods evolved above, must be approximate in character. The value of the analysis given lies in its simplicity rather than accuracy. Admitting that the situation is not satisfactory from the standpoint of reproducibility of results we must face the fact that there is no method nor the promise of a method for computing performance of a tube that would come within slide rule accuracy. We may console ourselves that in practice accuracy in estimating output levels and unwanted products is not really required. An error of 3 db or in the case of unwanted products an error of 6 db is of small consequence. This is about the variation to be expected to exist between two individual electron tubes and it must be included as a tolerance in determining performance requirements.

In connection with the third-order modulation the question arises whether the transconductance characteristic is really parabolic and not a curve of fourth or sixth degree and, if so, whether the equations derived above still apply.

While no accurate test can be applied conveniently we may compare a parabola with a quartic passing through the same three points. The parabola is characterized by a smooth curvature, while the quartic and the higher-degree curves have a flat middle portion with the ends turned up sharply. An inspection of the transconductance characteristics of tubes reveals that they are of the smoother curvature type—that is, that a square term is the chief contribution to the series representing the curve.

Should this not be the case and should we possibly mistake a quartic for a parabola, we still would obtain all of the phenomena caused by third-order

modulation; that is, we would get a third harmonic, a transconductance increment, a finite detector discrimination, and so on, only their values would be somewhat different. Moreover, a more elaborate analysis reveals that the computations as made above would be in error by relatively small amounts and, what is more important, they would always be on the safe side. The only important error would be the absence of the fifth harmonic, which cannot be produced by third-order modulation.

Experience with computations checked by measurement reveals that the equations apply in a great majority of practical cases. The computations may be used, therefore, as a guide in the selection of tubes, operating parameters, and in experiments, even though we must realize that they may not be fully justified theoretically and are not quite accurate.

Conclusion

The analysis of the transconductance characteristics of tubes could be pushed further to fifth, sixth and higher orders of modulation, but it is hardly worth it. The mathematical treatment becomes burdensome, the results are uninteresting and the applications are rare except in a qualitative way. Thus, having completed the discussion of the fourth-order modulation, we find an extension of the treatment to higher orders of modulation unprofitable.

The material presented in this article has consisted chiefly of a list of formulas which may be applied to practical computations of the transmission and quality of transmission of electron tubes. They will be found to be useful in design of electron tube circuits. The second objective of the analysis is to give the user a mental picture of the tube performance in terms of its conductance characteristic.

Thus in general the quality of transmission of an amplifier, its discrimination against interfering voltages, amplitude distortion, and second harmonic content are measured by the bias cut-off interval. The capacity of the tube to deliver large currents at small distortion is measured by the area under the characteristic. The gain is measured by the transconductance and, in some cases, by the steepness of the characteristic. Freedom from third- and fourth-harmonic distortion and input-output linearity are measured by the linearity of the characteristic.

In a modulator the current capacity of the tube is measured by the area between the characteristic and the lines defining the voltage and the transconductance range used. The conversion transconductance is measured by the transconductance range available. The linearity of the characteristic will measure the discrimination against third-order products. In the case of a rectifier the criteria will be the same as in the case of the modulator.

A tube with a small slope and large transconductance will deliver with

small distortion large currents whether used as an amplifier or modulator or rectifier. It will be a poor voltage amplifier, detector or voltmeter. A tube with a steep characteristic and large transconductance will be a good voltage amplifier, detector or voltmeter, but will carry little current.

Of course there are no rules without exception and special cases will be found in practice where the above recommendations may be violated. Electron tubes are used in many other ways than those discussed above and the present analysis covers only a small part of the field. The methods of computing modulation products shown above can be extended to triodes with little or no modification in cases of modulator or rectifier design where the external plate impedance is very low at the input frequencies.

Abstracts of Technical Articles by Bell System Authors

Administration of a Sampling Inspection Plan.[1] H. F. DODGE. Greatly expanded production during the war brought about extensive use of scientific sampling plans both within manufacturing plants and by procurement agencies. Perhaps most widely used were standard plans involving single sampling, double sampling, and multiple sampling for visual and gaging inspections on a go no-go basis. This paper discusses how a manufacturer may make use of these sampling plans in a manufacturing plant, how he goes about deciding on suitable levels of quality expressed in per cent defective, how he determnes whether sampling can be used advantageously in place of 100% inspection, and how he can choose and administer sampling plans that will limit the risks of sampling as well as provide the desired protection with a minimum amount of inspection. Particular attention is given to the operating characteristics and the mode of application of the standard AOQL (average outgoing quality limit) sampling plans published by Dodge and Romig.

The Bridge Erosion of Electrical Contacts. Part I.[2] J. J. LANDER and L. H. GERMER. Bridge erosion is the transfer of metal from one electrode to the others, which occurs when an electric current is broken in a low voltage circuit which is essentially purely resistive. It is associated with the bridge of molten metal formed between the electrodes as they are pulled apart, and more specifically with the ultimate boiling of some of the metal of this bridge before the contact is finally broken. This paper is concerned with fundamental studies of this molten bridge and with empirical measurements of the transfer of metal.

Electron Bombardment Conductivity in Diamond.[3] KENNETH G. MCKAY. A study has been made of electron bombardment conductivity in diamond using primary electrons of energies up to 14,000 ev. An alternating field method is used which reduces or eliminates the effects of internal space charge fields. Data on internal yields as a function of crystal field are given for both electron and positive hole carriers. Internal yields as high as 600 have been attained. The experimental curves are fitted to a theoretical curve for the space charge free crystal from which are derived reasonable values for the number of electrons produced in the conduction band per incident primary electron, the probable life time of the conduction

[1] *Industrial Quality Control*, November 1948.
[2] *Jour. Applied Physics*, October 1948.
[3] *Phys. Rev.*, December 1, 1948.

electrons and the crystal trap density. Experiments are described which lead to a hypothesis of space charge neutralization. A possible cause of the current fluctuations observed at high crystal fields is discussed.

The Philosophy of PCM.[4] B. M. OLIVER, J. R. PIERCE and C. E. SHANNON. Recent papers describe experiments in transmitting speech by PCM (pulse code modulation). This paper shows in a general way some of the advantages of PCM, and distinguishes between what can be achieved with PCM and with other broadband systems, such as large-index FM. The intent is to explain the various points simply, rather than to elaborate them in detail. The paper is for those who want to find out about PCM rather than for those who want to design a system. Many important factors will arise in the design of a system which are not considered in this paper.

Objectives for Sound Portrayal.[5] RALPH K. POTTER. Translation of sound into visible patterns is discussed in terms of broad objectives. It is suggested that no single design can be optimum and that perhaps the most useful standard of reference is the human ear. Special interests and complexity generally affect final design requirements.

A Waveguide Bridge for Measuring Gain at 4000 Mc.[6] A. L. SAMUEL and C. F. CRANDELL. A bridge has been constructed for measuring the gain and phase delay of amplifiers in the vicinity of 4000 Mc. The equipment is described, and the methods employed to reduce the possible errors are discussed. The general method may be adapted for use in any desired frequency range.

Video Distribution Facilities for Television Transmission.[7] ERNEST H. SCHREIBER. This paper describes the Bell System's plans for furnishing network and local video facilities. The Telephone Company is now using broad-band coaxial cable and microwave radio systems to provide regular message telephone service on a number of principal intercity routes throughout the nation. These facilities can be used to provide television transmission channels when properly equipped. Video service between Washington, D. C., New York, and Boston over these two types of facilities has been demonstrated. New facilities are rapidly being extended. Local video channels for pickup and metropolitan-area networks are provided by ordinary paper-insulated cable pairs, special shielded polyethylene-insulated pairs, by microwave radio systems, or by combinations of these systems. Amplifier and equalizing arrangements for providing wide-band transmission over these facilities are described. Present Bell System views of the availability of microwave and coaxial cable facilities on the principal routes,

[4] *Proc. I. R. E.*, November 1948.
[5] *Jour. Acous. Soc. Amer.*, January 1949.
[6] *Proc. I. R. E.—Waves and Electrons Section*—November 1948.
[7] *S. M. P. E. Journal*, December 1948.

types of circuits, bandwidths, bridging and terminating arrangements, and general information concerning the provision of television circuits are covered.

Communication in the Presence of Noise.[8] CLAUDE E. SHANNON. A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and compression of bandwidth and the threshold effect. Formulas are found for the maximum rate of transmission of binary digits over a system when the signal is perturbed by various types of noise. Some of the properties of "ideal" systems which transmit at this maximum rate are discussed. The equivalent number of binary digits per second for certain information sources is calculated.

Earth Conduction Effects in Transmission Systems[9] ERLING D. SUNDE. Earth conduction problems are encountered in both communication and power system engineering in connection with investigations of earth resistivity, grounding, corrosion of buried metallic structures, power system impedances and fault currents, inductive interference, lightning disturbances, and in connection with ground-wave radiation fields. Mr. Sunde deals comprehensively with the theory underlying various earth conduction effects and with its engineering applications, and brings together in unified manner many topics that hitherto have received only separate discussion in the literature. The author's purpose throughout has been to tie his discussion to practical considerations and problems.

Beginning with a review of the theory underlying various earth conduction effects and with its engineering applications, the book contains the following chapter headings: basic electromagnetic concepts and equations, earth resistivity testing and analysis, resistance of grounding arrangements, mutual impedance of insulated earth-return conductors, propagation characteristics of earth-return conductors, d-c earth conduction and corrosion protection, power system earth conduction and inductive interference, surge characteristics of earth-return conductors, lightning protection of cable and transmission lines.

A carefully compiled bibliography is included.

[8] *Proc. I. R. E.*, January 1949.

[9] Published by *D. VanNostrand Company, Inc.*, New York, London and Toronto, January, 1949. 373 pages. $6.00.

Contributors to This Issue

DIETRICH A. ALSBERG, Technical College of Stuttgart, 1938; Case School of Applied Science, postgraduate in electrical engineering, 1939–40. From 1940–43 Mr. Alsberg was engaged as development engineer by several organizations. From 1943–45 he served in the U. S. Army Ordnance Department at Aberdeen Proving Ground and in the European Theater. In 1945 he joined the Bell Telephone Laboratories where he is concerned with phase and transmission measurement problems.

JOHN BARDEEN, University of Wisconsin, B.S. in E.E., 1929, M.S., 1930; Gulf Research and Development Corporation, 1930–33; Princeton University, 1933–35, Ph.D. in Math. Phys., 1936; Junior Fellow, Society of Fellows, Harvard University, 1935–38; Assistant Professor of Physics, University of Minnesota, 1938–41; Prin. Phys., Naval Ordnance Laboratory, 1941–45. Bell Telephone Laboratories, 1945–. Dr. Bardeen is engaged in theoretical problems related to semiconductors.

W. R. BENNETT, B.S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Bennett has been active in the design and testing of multichannel communication systems, particularly with regard to modulation processes and the effects of nonlinear distortion. He is now engaged in research on various transmission problems.

WALTER H. BRATTAIN, B.S., Whitman College, 1924; M.A., University of Oregon, 1926; Ph.D., University of Minnesota, 1929; Major Phys., Bureau of Standards, 1928–29. Bell Telephone Laboratories, 1929–42. Columbia University, N.D.R.C., 1942–44. Bell Telephone Laboratories, 1944–. Dr. Brattain is engaged in the study of semiconductors.

C. H. DAGNALL, S.B., Massachusetts Institute of Technology, 1918; S.B., Harvard University, 1918; M.S., Cornell University, 1922. Signal Corps, U.S.A., 1918; General Electric Company, 1919; Instructor in Electrical Engineering, Cornell University, 1919–25. Bell Telephone Laboratories, 1925–. Mr. Dagnall has been chiefly concerned with the design of transmission networks.

F. S. FARKAS, E. E., 1929, Polytechnic Institute of Brooklyn. Engineer-

ing Department, Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Mr. Farkas was engaged in the development of transmission networks and is now concerned with the development of radio and network switching.

F. J. HALLENBECK, E.E., 1936, Polytechnic Institute of Brooklyn. Engineering Department, Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–. Mr. Hallenbeck has been concerned chiefly with the development of transmission networks for carrier systems.

R. A. LECONTE, E.E., Electrotechnical Institute, Grenoble University, France, 1908; French Army Corps of Engineers, 1915–20. Mr. Leconte came originally to the United States in 1917 with a French military mission and came back in 1919 to the French purchasing organization in this country. He joined the Engineering Department, Western Electric Company, in 1922; Bell Telephone Laboratories, 1925–. He has been concerned with voice frequency repeater and carrier terminal developments.

DANIEL LEED, B.S., College of the City of New York, 1941; Kollsman Instrument Company, 1941–43. Federal Telephone and Radio Corporation, 1943–44. Corps of Engineers, Los Alamos Laboratories of the Manhattan District, 1944–46. Bell Telephone Laboratories, 1946–. Mr. Leed is engaged in circuit development for phase and transmission measurement systems, particularly in the field of automatic frequency control.

D. B. PENICK, University of Texas, B.S. in Electrical Engineering, 1923, B.A., 1924; Columbia University, M.A., 1927. Engineering Department, Western Electric Company, 1924–25. Bell Telephone Laboratories, 1925–. Mr. Penick has been engaged in the development of carrier telephone systems.

LISS C. PETERSON, Chalmers Technical University, Gothenburg, 1921; Technical Universities of Charlottenburg and Dresden, 1921–23. American Telephone and Telegraph Company, 1925–30; Bell Telephone Laboratories, 1930–. Mr. Peterson has recently been concerned with the theory of hearing.

P. W. ROUNDS, A.B., Harvard University, 1929. Bell Telephone Laboratories, 1929–. Mr. Rounds has been engaged in the design of transmission networks.

C. W. SCHRAMM, B.S. in Electrical Engineering, Armour Institute

(now Illinois Institute) of Technology, 1927. Illinois Bell Telephone Company, 1927–29. Bell Telephone Laboratories, 1929–. Mr. Schramm has been concerned with the development of carrier telephone systems for both message and program use. During the war his attention was directed to the design of radar test equipment.

T. SLONCZEWSKI, B.S. in Electrical Engineering, Cooper Union Institute of Technology, 1926. Bell Telephone Laboratories, 1926–. Mr. Slonczewski has been engaged in the development of electrical measuring apparatus.

F. E. STEHLIK, B.E.E., 1933, M.E.E., 1935, Polytechnic Institute of Brooklyn. Bell Telephone Laboratories, 1936–. Mr. Stehlik was engaged in the design of crystal filters and is now concerned with the development of high-frequency networks.

E. D. SUNDE, B.S., Haugesund, Norway, 1922; E.E., Darmstadt, Germany, 1926. American Telephone and Telegraph Company, 1927–33; Bell Telephone Laboratories, 1933–. Mr. Sunde has been engaged in studies of interference in telephone circuits from power lines and railway electrification and is now concerned with studies of protection of the telephone plant against lightning damage.

H. M. TRUEBLOOD, B.S., Earlham College, 1902; B.S., Haverford College, 1903; Mass. Inst. Technology, 1908–09; Ph.D. (physics), Harvard University, 1913; Field Officer, U. S. Coast and Geodetic Survey, 1903–08; Instructor and Assistant Professor in Electrical Engineering, University of Pennsylvania, 1914–17; U. S. Naval Experimental Station, New London, Connecticut, 1917–19. American Telephone and Telegraph Company, Department of Development and Research, 1919–34; Bell Telephone Laboratories, 1934–. At present, Assistant Director of Transmission Engineering. Most of Dr. Trueblood's work has been on interference with communication systems from natural and other sources, with work on radar and radar testing equipment during World War II.

ANTHONY J. WIER, L.L.B., New Jersey Law School, 1935. New York Telephone Company, Plant Maintenance; and Western Electric Company, Installation and Equipment Engineering; 1914–28. Bell Telephone Laboratories, 1928–. Mr. Wier has been engaged in development work on toll telephone and telegraph equipment since 1928.